Performance Evaluation 68 (2011) 986–1001

Contents lists available at SciVerse ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Competition yields efficiency in load balancing games☆

Jonatha Anselmi [a,*], Urtzi Ayesta [a,b], Adam Wierman [c]

[a] BCAM – Basque Center for Applied Mathematics, Derio, 48160, Spain
[b] IKERBASQUE, Basque Foundation for Science, Bilbao, 48170, Spain
[c] Computer and Mathematical Sciences, Caltech, Pasadena, CA 91125, USA

ARTICLE INFO

Article history:
Available online 2 August 2011

Keywords:
Queueing games
Oligopolistic price competition
Parallel providers
Price of anarchy

ABSTRACT

We study a nonatomic congestion game with N parallel links, with each link under the control of a profit maximizing provider. Within this 'load balancing game', each provider has the freedom to set a price, or toll, for access to the link and seeks to maximize its own profit. Given prices, a Wardrop equilibrium among users is assumed, under which users all choose paths of minimal and identical effective cost. Within this model we have *oligopolistic price competition* which, in equilibrium, gives rise to situations where neither providers nor users have incentives to adjust their prices or routes, respectively. In this context, we provide new results about the existence and efficiency of oligopolistic equilibria. Our main theorem shows that, when the number of providers is small, oligopolistic equilibria can be extremely inefficient; however as the number of providers N grows, the oligopolistic equilibria become increasingly efficient (at a rate of $1/N$) and, as $N \to \infty$, the oligopolistic equilibrium matches the socially optimal allocation.

1. Introduction

As researchers in networking and telecommunications have become increasingly interested in economics and game theory, one of the topics that has spurred significant research is that of congestion games (a.k.a. routing games). Applications of congestion games are numerous, including settings such as communication networks and transportation networks, and resultantly a large literature has grown studying the existence and efficiency of equilibria in a variety of congestion games. See [1] for a recent survey.

In the classic formulation of a congestion game there are noncooperative agents sending traffic through a network and the agent strategies consist of possible routes through the network. In the 'nonatomic' version of these games, there are a continuous number of players, each of which has a negligible effect on the others. The utility of routes to agents corresponds to the 'latency' or 'congestion' experienced on the path. Most typically, the Wardrop equilibria, under which every agent sends traffic along its smallest latency path(s), are considered as the solution concept.

There is a large literature studying nonatomic congestion games and, at this point, properties related to the existence and inefficiency (a.k.a. price of anarchy) of the Wardrop equilibria are well understood. Interestingly, in many cases, there is little efficiency loss between the social optimal, which minimizes the total latency across all flows, and the Wardrop equilibrium. For example, for the case of affine latency functions the price of anarchy is bounded by 4/3 [2], i.e., the equilibria routes have total latency bounded by 4/3 that of the minimal latency. However, under more realistic latency functions there can

☆ Research partially supported by grant MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain), grant PI2010-2 (Department of Education and Research, Basque Government), and NSF CNS-0846025.

* Corresponding author.
E-mail addresses: anselmi@bcamath.org, jonatha.anselmi@gmail.com (J. Anselmi).

0166-5316/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2011.07.005

be significant efficiency loss. For example, if one considers latency functions defined by queueing models (as is appropriate for the Internet or transportation applications), the price of anarchy becomes unbounded: In the case of a 'load balancing game' with N parallel $M/M/1$ queues the price of anarchy is N [3] and when more general queueing models are considered the price of anarchy can be unbounded even in the case of two parallel queues [4]. Therefore, significant research has gone into understanding how to reduce the price of anarchy: Classic results in this direction include (i) using a Pigouvian tax (toll) [1], (ii) increasing the capacity of each link in the network [2], and (iii) controlling centrally a small fraction of the traffic [5,6].

In this paper, we show that *competition among profit-maximizing providers (corresponding to the links in the network) can induce efficient routing*. More specifically, the model we study in this paper consists of a nonatomic congestion game with N parallel links, with each link under the control of a profit maximizing provider. Within this 'load balancing game', each provider has the freedom to set a price, or toll, for access to the link and seeks to maximize its own profit (i.e., the product of price and traffic per time). Given prices, a Wardrop equilibria among users is assumed, under which users all choose paths of minimal cost, which is defined as latency plus price. See Section 2 for a formal description of the model. Within this model, we have *oligopolistic price competition* among the providers and we seek to understand the existence and efficiency of oligopolistic equilibria, which are situations where neither providers nor users have incentives to adjust their prices or routes, respectively.

Notice that the model we consider is perhaps more representative of many settings than classic congestion games. In particular, it provides a simple framework in which to study the impact of price and competition among providers, and as such has application to (i) competition among cloud service providers [1, Chapter 22], (ii) transportation networks where toll roads are not centrally controlled [7], (iii) competition among ISPs in communication networks [8], and (iv) freight transportation [9]. More generally, the model captures the impact of competition and price in congested markets, see [10] for a discussion of such markets.

In fact, the model we consider (or slight variants of it) has been studied in a number of recent papers [11,12,7,13]. Interestingly, the model has proven to be highly non-trivial analytically, and little is known about the existence and uniqueness of oligopolistic equilibria [7,10]. This is because the provider optimization problems are not convex and, thus, classical existence proofs do not apply. Despite limited understanding of existence, some progress has been made in the analysis of this model. In particular, [10] has studied the model from the perspective of the *social surplus* (or difference between the users' willingness to pay and delay costs), which is different from the price of anarchy, and shown that the ratio of the optimal attained social surplus to the one attained at the oligopolistic equilibrium (if it exists) is bounded from below by $5/6$. In parallel–serial networks, this ratio is lower bounded by $1/2$ if the latency functions are zero when evaluated at zero, otherwise it can be arbitrarily large [11]. If providers can compete over both prices and capacities, this ratio can be again arbitrarily large [14].

In this paper, we make two main contributions to the study of oligopolistic equilibria in congestion games.

First, we provide new results about existence and non-existence of oligopolistic equilibria (Section 3). We show existence in the case of homogeneous latency functions, and we show non-existence in the case where the system is 'under-provisioned' in the sense that the 'capacity' of the $N-1$ slowest servers is not enough to handle the full traffic. We also give numerical evidence that the best-response algorithm converges to a fixed-point, implying the existence of an oligopolistic equilibrium in an empirical sense.

Second, we provide a new understanding of the efficiency of oligopolistic equilibria (Section 4). Specifically, we show that if there are only two providers the oligopolistic equilibria can be arbitrarily worse than the optimal routing in terms of total latency (i.e., the price of anarchy can be unbounded). However, as the number of providers N grows, the price of anarchy drops at a rate of $1/N$ and, as $N \to \infty$, the price of anarchy converges to 1. As we scale $N \to \infty$, we allow the incoming traffic to scale proportionally with N, thus maintaining the same 'load' for the system while N grows. This highlights that the improvement in the price of anarchy is not due to a shrinking of the congestion in the system (as would happen if the traffic grew sublinearly with N), but is in fact due to the increasing competition among the providers. So, broadly speaking, this result can be seen as providing an example of the maxim that "competition yields efficiency", with the caveat (from our non-existence result) that the system "must not be under-provisioned".

For both existence and efficiency, we specialize our results using queueing-based latency functions. These are highly relevant for communication and transportation applications. Further, they highlight the power of competition among providers because the inefficiency that is evident for these latency functions in classic congestion games disappears with even a small amount of competition among providers. However, the queueing examples also highlight the complexity of understanding existence of oligopolistic equilibria.

The remainder of this paper is organized as follows. In Section 2, we describe the price competition game under investigation introducing the social optimum, Wardrop and oligopolistic equilibria. Section 3 presents our results on the existence of oligopolistic equilibria and Section 4 analyzes their efficiency from a performance standpoint. Finally, Section 5 draws the conclusions of our work and outlines future research.

2. Model and notation

As mentioned in the introduction, the model considered in this paper falls into the category of oligopolistic pricing games [1,12,10,7].

More specifically, we consider a network with $N > 1$ providers working in parallel. Each provider is assumed to own a single network resource and there is an infinite inelastic stream of infinitely-many arriving users that follows some stochastic process with intensity λ. The choice of inelastic traffic is motivated by the fact that in several cases, e.g., cloud-computing, the prices set by providers are small enough so that every user will choose to join the system; see, e.g., [15–17]. Upon joining the network, each user selects exactly one provider from which to receive service. Each user, thus, carries an infinitesimally small amount of traffic.

Let x_i denote the mean amount of traffic per unit of time to provider i in some stationary regime, which we define in the following, and denote the traffic vector by $\mathbf{x} = (x_1, \ldots, x_N)$, with $\sum_{i=1}^{N} x_i = \lambda$. In the remainder of the paper, indices i and j implicitly range from 1 to N if not otherwise specified.

Whenever a user selects provider i, it must pay an amount $p_i \geq 0$ to i. Let $\mathbf{p} = (p_1, \ldots, p_N)$ denote the vector of prices and define $\mathbf{p}_{-j} = (p_i)_{\forall i \neq j}$. We measure the *profit* per time unit of provider i by $p_i x_i$.

Our focus throughout is on the mean time it takes for a user to be served at i. We denote *latency of* i (a.k.a., the response time, sojourn time, or flow time of i) by $\ell_i(x_i)$. The form of the latency functions can differ greatly depending on the application considered. However, in transportation and communication networks, latency functions coming from queueing models are most typically used for modeling purposes. The form of the resulting function still depends on several factors, such as the scheduling discipline implemented by each provider and the details of the arrival process. One common model, which we use throughout the paper, assumes (i) a Poisson arrival process, (ii) Processor-Sharing scheduling[1] [18], (iii) i.i.d. service times, (iv) no limit to the number of users that each provider can handle simultaneously, and (v) the selection of a provider follows an i.i.d. probability law (as in, e.g., [3]). In this setting, for all i,

$$\ell_i(x_i) = \begin{cases} (\mu_i - x_i)^{-1} & \text{if } 0 \leq x_i < \mu_i \\ +\infty & \text{otherwise.} \end{cases} \tag{1}$$

Note that the parameter μ_i is interpreted as the mean service rate of provider i. The latency function in (1) is very popular in communication network performance modeling and corresponds to the mean response time of an $M/GI/1/$Processor-Sharing queue and an $M/M/1/$First-come-first-served queue [19].

Though we use the latency function (1) for illustrative purposes, the main results of the paper hold more generally. Specifically, in the remainder of the paper we will refer to the following assumptions for the latency functions.

Assumption 1. The functions $\ell_i(x_i),\ i = 1, \ldots, N$, are continuously differentiable and increasing over $[0, \mu_i)$, $\mu_i \leq \infty$.

Assumption 2. If there exists a vector of positive real numbers $[\mu_1, \ldots, \mu_N]$ such that $\lim_{x_i \to \mu_i} \ell_i(x_i) = +\infty$ for all i, then $\lambda < \sum_i \mu_i$.

Assumption 3. The functions $x_i \ell_i'(x_i)$ are non-decreasing for all i.

Clearly the latency function in (1) satisfies these assumptions, as do latency functions based on many other queueing models. Further, note that Assumption 3 does not necessarily require the convexity of ℓ_i, which is typically assumed (e.g., in [10,7]). Additionally, Assumption 2 guarantees the stability of the system in the optimal allocation $\mathbf{x}^{\text{Opt}}$ (see Definition 1).

Given the latency functions for each i, we define the *latency* of the network under the traffic allocation $\mathbf{x}$ as $\ell(\mathbf{x}) = \sum_i \frac{x_i}{\lambda} \ell_i(x_i)$. Note that the network latency can be interpreted as a direct measure of the overall quality of service at allocation $\mathbf{x}$.

Given the setup described to this point, we can now define the optimal traffic allocation, which serves as a benchmark for the performance of the oligopolistic equilibria we define later.

Definition 1. A traffic allocation $\mathbf{x}^{\text{Opt}}$ is said *socially optimal*, or a social optimum, if it minimizes the latency. That is, if

$$\mathbf{x}^{\text{Opt}} = \underset{\mathbf{x}: \sum_i x_i = \lambda, x_i \geq 0, \forall i}{\operatorname{argmin}} \ \ell(\mathbf{x}). \tag{2}$$

In the remainder of this section we introduce the Wardrop and oligopolistic equilibria in the context of the model described above.

2.1. Wardrop equilibrium

Since we assume that the incoming traffic is the aggregate flow of infinitely many selfish users that carry an infinitesimal amount of traffic, for a given price vector $\mathbf{p}$ the stationary regime is a Wardrop equilibrium (WE). This is characterized by Wardrop's first and second principles [20] (see also [21]): The distribution of traffic among the providers is such that the sum of the response time and the price of each provider, i.e., the *effective cost* incurred by each user, is minimum and equal

[1] Under Processor-Sharing scheduling a resource is shared evenly among all users present, i.e., if there are m users present they each receive $1/m$th of the resource.

at each provider. Wardrop's principles are used extensively in modeling the traffic distribution of communication networks and transportation networks and has received significant attention in the algorithmic game theory community, e.g., see the recent survey in [1].

In our context, a Wardrop equilibrium is defined as follows.

Definition 2. For a given price vector $\mathbf{p}$, a vector $\mathbf{x}^{\mathrm{WE}} \in \mathbb{R}^N$ is a *Wardrop equilibrium* if

$$\begin{aligned} &\ell_i(x_i^{\mathrm{WE}}) + p_i = \min_j \left\{\ell_j(x_j^{\mathrm{WE}}) + p_j\right\}, \quad \forall i : x_i^{\mathrm{WE}} > 0 \\ &\sum_i x_i^{\mathrm{WE}} = \lambda \\ &x_i^{\mathrm{WE}} \geq 0, \quad \forall i. \end{aligned} \tag{3}$$

Throughout, we denote by $W(\mathbf{p})$ the Wardrop equilibrium achieved with price vector $\mathbf{p}$. Note that our setting is essentially equivalent to nonatomic load balancing games, which have been studied in a number of recent papers [1]. In our context, it is immediate to conclude the existence and uniqueness of a Wardrop equilibrium, see [22].

Proposition 1. *For a fixed price vector* $\mathbf{p}$*, there exists exactly one Wardrop equilibrium.*

Though the proposition is well-known within our assumptions, it is useful to consider the proof. In particular, it can be shown that the set of Wardrop equilibria under $\mathbf{p}$ is given by the set of minimizers of

$$\sum_i \int_0^{x_i} \ell_i(z)\mathrm{d}z + p_i x_i \tag{4}$$

subject to $\sum_i x_i = \lambda$, $x_i \geq 0$, $\forall i$. Since (4) is a strictly convex function (ℓ_i is increasing by Assumption 1) defined over linear constraints, a unique Wardrop equilibrium exists.

2.2. Oligopolistic equilibrium

The key interaction that we wish to model is that of the competition among profit maximizing providers. Given a price vector $\mathbf{p}$, a provider may wish change its price to increase its individual profit in the Wardrop equilibrium that will result. Thus, profit maximizing providers compete with each other playing the so-called *price competition game*, and an oligopolistic equilibrium among the providers can arise.

An oligopolistic equilibrium represents the stationary situation where each provider has no incentives in unilaterally changing its price because otherwise it would decrease its profit. Specifically, we define an oligopolistic equilibrium in our setting as follows.

Definition 3. A vector $\mathbf{p}^{\mathrm{OE}}$ is an *oligopolistic equilibrium* if

$$p_i^{\mathrm{OE}}[W(\mathbf{p}^{\mathrm{OE}})]_i = \max_{p_i \geq 0} p_i[W(p_i, \mathbf{p}_{-i}^{\mathrm{OE}})]_i, \quad \forall i. \tag{5}$$

Provided that it exists, an oligopolistic equilibrium provides a prediction of the point where the strategic actions of both users and providers should converge. Throughout, we use $\mathbf{p}^{\mathrm{OE}}$ to be an oligopolistic equilibrium and $\mathbf{x}^{\mathrm{OE}} = W(\mathbf{p}^{\mathrm{OE}})$ to be the corresponding traffic allocation.

The question of existence of an oligopolistic equilibrium is not as simple as that of existence of a Wardrop equilibrium. However, the following proposition provides necessary conditions for a price vector to be an oligopolistic equilibrium. Further issues related to existence are our focus in Section 3.

Proposition 2. *Let* $\mathbf{p}^{\mathrm{OE}}$ *be an oligopolistic equilibrium and* $\mathbf{x}^{\mathrm{OE}} = W(\mathbf{p}^{\mathrm{OE}})$*. The following conditions must hold*

$$\begin{cases} p_i^{\mathrm{OE}} = x_i^{\mathrm{OE}} \dfrac{\partial \ell_i(x_i)}{\partial x_i}\Big|_{x_i = x_i^{\mathrm{OE}}} + \dfrac{x_i^{\mathrm{OE}}}{\sum\limits_{\substack{j \neq i: \\ A_j = 0}} \left(\dfrac{\partial \ell_j(x_j)}{\partial x_j}\Big|_{x_j = x_j^{\mathrm{OE}}}\right)^{-1}} & \forall i : A_i = 0, \\ p_i^{\mathrm{OE}} = 0, \quad \forall i : A_i > 0 \\ \ell_i(x_i^{\mathrm{OE}}) + p_i^{\mathrm{OE}} = B + A_i, \quad \forall i \\ A_i x_i = 0, \quad A_i \geq 0, \ \forall i \\ \sum_i x_i^{\mathrm{OE}} = \lambda \end{cases} \tag{6}$$

for some $B, A_i, \forall i$*, if at least two providers are used. Otherwise,* $\mathbf{p}^{\mathrm{OE}} = (\min_{i \neq k} \ell_i(0) - \ell_k(\lambda))\mathbf{e}_k$*, where* $k = \arg\min_j \ell_j(\lambda)$ *and* $\mathbf{e}_k$ *is the unit vector in direction* k*.*

Proof. The proof consists of merging together the KKT conditions for each of the optimization problems in (5). Each optimization problem i is analyzed assuming that the price vector $\mathbf{p}_{-i}^{\mathrm{OE}}$ is a constant. The details of this calculation are shown in the Appendix. Note that it follows from (6) that the variable A_i is zero if and only if provider i is used in an oligopolistic equilibrium. Further, B is interpreted as the effective cost (delay plus price) incurred by users. □

3. Equilibria existence and uniqueness

The primary questions to address about oligopolistic equilibria are those of existence and uniqueness. As mentioned above, understanding whether oligopolistic equilibria exist in our setting is non-trivial because of the non-convex structure of the equilibrium point (5). Further, it is easy to see that there are many natural cases where oligopolistic equilibria do not exist, which means that one cannot hope for as strong an existence result as holds for Wardrop equilibria.

The prior literature has begun to study the question of existence; however existence has only been proven formally in the simple case of linear functions of the type $\ell_i(x_i) = a_i x_i$ [10, Proposition 7], which allows the use of Kakutani's theorem. In this section we provide two new results regarding the existence and nonexistence of the oligopolistic equilibria.

One common property of latency functions used in communication and transportation networks is that providers have some 'capacity' on the traffic that can be handled above which the latency becomes infinite. For example, in the case of the queueing-based latency functions in (1), when $x_i \geq \mu_i$. It turns out that this capacity constraint on the providers can lead to settings where oligopolistic equilibria do not exist.

In particular, if the system is too heavily loaded, i.e., all providers need to be used to keep the congestion cost finite, no oligopolistic equilibrium exists.

Proposition 3. *Let Assumptions 1 and 2 hold. Furthermore, let* (i) $\lim_{x_i \to \mu_i} \ell_i(x_i) = \infty$, *for all* $i > 1$, (ii) $\mu_1 \geq \mu_i$, *for all* $i > 1$, (iii) $\lambda \geq \sum_{i>1} \mu_i$. *Then, there exists no oligopolistic equilibrium.*

Proof. The best response (price) of provider 1 with respect to any price set by the other providers is unbounded, because it will get at least $\lambda - \sum_{i>1} \mu_i$ traffic (this happens because $\lim_{x_i \to \mu_i} \ell_i(x_i) = \infty,\ i > 1$). Therefore, for any choice of a finite price by that provider, it has the incentive to increase it, which implies that there can be no oligopolistic equilibrium. □

An interpretation of Proposition 3 is that, if the system is under-provisioned, then it is possible for the provider with the highest capacity to exploit this fact and obtain arbitrarily large profits. As mentioned, this fact is particularly relevant in the case of queueing-based latency functions.

On the other hand, let us consider the simple case of $N = 2$ and with latency function (1) and $\lambda < \mu_2$. Thus, Proposition 3 does not apply. In this scenario, for any price set by provider i, provider $j \neq i$ cannot choose a price unboundedly large because provider i could handle all the traffic with an effective cost less than the price of j, which would cause the profit of j to become zero. Therefore, the prices must be bounded in this case. Continuing with this example, if the number of providers increases, then at a given point we must have $\lambda < \sum_{i>1} \mu_i$ (because $\mu_i = O(1)$), and the non-existence argument in Proposition 3 does not apply. Therefore, we may expect that an oligopolistic equilibrium eventually exists as the number of providers grows while keeping λ constant. Note, however, that this argument is not sufficient to establish the existence of an oligopolistic equilibrium, which is challenging even in this simple scenario.

Though it is difficult to provide a general existence result, we can characterize existence in the special case of homogeneous (symmetric) latency functions. Interestingly, even in the case of homogeneous latency functions, existence of an equilibrium is not guaranteed, the following condition on the latency functions is necessary.

Assumption 4. The function $\ell(x)$ satisfies the following inequality for all $x \in \left[0, \frac{\lambda}{N}\right)$

$$\frac{\ell\left(\frac{\lambda - x}{N-1}\right) - \ell(x)}{\frac{\lambda - x}{N-1} - x} < \frac{\lambda/N}{x} \ell'(\lambda/N). \tag{7}$$

A possible interpretation for previous assumption is that the latency function do not to increase "too fast" for $x > \lambda/N$. Also, assume that the arrival rate scales with N, i.e., $\lambda \leftarrow \lambda N$ (this scaling will be used and discussed in detail in the next section). Then, as $N \to \infty$, condition (7) becomes: For all $x \in [0, \lambda)$,

$$\frac{\ell(\lambda) - \ell(x)}{\lambda - x} < \frac{\lambda}{x} \ell'(\lambda). \tag{8}$$

Therefore, we have the following observation.

Observation 1. *Let Assumption 1 hold. If the arrival rate scales linearly with N and* $\ell(x)$ *is convex, which implies* $\frac{\ell(\lambda) - \ell(x)}{\lambda - x} < \ell'(\lambda),\ \forall x \in [0, \lambda)$, *then for N large enough Assumption 4 holds true.*

Proposition 4. *Let Assumptions 1 and 2 hold. Furthermore, let* (i) *all providers be homogeneous, i.e.,* $\ell_i(x) = \ell(x),\ \forall i$, (ii) $\ell(x)$ *be convex,* (iii) $\lim_{x\to\mu} \ell(x) = \infty$. *Then, prices*

$$p_i^{\mathrm{OE}} = p^{\mathrm{OE}} = \frac{\lambda}{N-1}\ell'(\lambda/N) \tag{9}$$

form the unique oligopolistic equilibrium of the price competition game if and only if $\lambda < (N-1)\mu$ *and Assumption 4 holds true.*

Proof. We prove existence using Definition 2. That is, assuming that all providers set price (9), we analyze the best response of each provider to understand under which conditions has the incentive of changing its price.

The best response of provider 1 (w.l.o.g) given that all the other providers have fixed price p^{OE} as in (9) is given by the optimizer(s) of

$$\begin{array}{ll} \max\limits_{p_1,x_1,x} & p_1 x_1 \\ \text{s.t. :} & \ell(x_1) + p_1 = \ell(x) + p^{\mathrm{OE}} \\ & x_1 + (N-1)x = \lambda \\ & x_1, x \geq 0 \end{array} \tag{10}$$

where we have used that the traffic $[W(p, \mathbf{p}_{-1}^{\mathrm{OE}})]_i$ are all equal (to x) for all providers $i > 1$ (implied by Definition 2). From the Lagrangian of previous optimization problem, we find the following KKT conditions

$$\begin{array}{l} p_1 = x_1\ell'(x_1) + \dfrac{x_1}{N-1}\ell'\left(\dfrac{\lambda - x_1}{N-1}\right) + Z \\ Z \geq 0 \\ (x_1 - \lambda)Z = 0 \\ \ell(x_1) + p_1 = \ell\left(\dfrac{\lambda - x_1}{N-1}\right) + p^{\mathrm{OE}} \\ 0 \leq x_1 \leq \lambda \end{array} \tag{11}$$

where Z is a Lagrange multiplier.

Note that, at the point $Z = 0, x_1 = \lambda/N, p_1 = p^{\mathrm{OE}}$, the equation $\ell(x_1) + x_1\ell'(x_1) + \frac{x_1}{N-1}\ell'\left(\frac{\lambda - x_1}{N-1}\right) + Z = \ell\left(\frac{\lambda - x_1}{N-1}\right) + p^{\mathrm{OE}}$ (from (11)) is satisfied. For $\lambda > x_1 \geq \lambda/N$ the left (right) hand term of previous equation is strictly increasing (decreasing) and therefore no other $x_1 \in [\lambda/N, \lambda)$ can solve it (here, we have used the fact that $\ell(\cdot)$ is convex and increasing). If $x_1 = \lambda$, we must have $\ell(\lambda) + \lambda\ell'(\lambda) + \frac{\lambda}{N-1}\ell'(0) + Z = \ell(0) + p^{\mathrm{OE}} = \ell(0) + \frac{\lambda}{N-1}\ell'(\lambda/N)$, with $C \geq 0$, which is again not possible. For $x_1 < \lambda/N$, there can be other solutions for conditions (11), but we must have that all of them do not yield a better profit for provider 1. This is true if $x\left(\ell\left(\frac{\lambda - x}{N-1}\right) - \ell(x) + p^{\mathrm{OE}}\right) < \frac{\lambda}{N}p^{\mathrm{OE}}$, for $x \in \left[0, \frac{\lambda}{N}\right)$. After substitution of Eq. (9) and some rearrangements, the inequality holds if and only if Assumption 4 is satisfied.

Therefore, given that there are no other better stationary points than the one specified above, the only other way provider 1 can increase its profit is to set p_1 arbitrarily large. This is because the optimization problem (10) is not convex, and thus the KKT conditions (11) are not sufficient for a point to be a global optimum. We have two cases:

(i) if $\lambda < (N-1)\mu$, then $[W(\infty, \mathbf{p}_{-1}^{\mathrm{OE}})]_1 = 0$ and $[W(\infty, \mathbf{p}_{-1}^{\mathrm{OE}})]_i = \lambda/(N-1), i > 1$, and no profit can be made by provider 1.

(ii) if $\lambda \geq (N-1)\mu$, then $[W(\infty, \mathbf{p}_{-1}^{\mathrm{OE}})]_1 \geq \lambda - (N-1)\mu$ and $[W(\infty, \mathbf{p}_{-1}^{\mathrm{OE}})]_i \leq \mu, i > 1$, and an infinite profit can be made by provider 1 because $\ell(\mu) = \infty$ (some users choose provider 1 to avoid the infinite congestion cost at other providers).

Therefore, we conclude that the best response of each provider is p^{OE} if $\lambda < (N-1)\mu$ and Assumption 4 holds true.

For the *only if* part, we observe that if (9) are the unique equilibrium prices, then the best response of provider 1 can be neither infinite (which implies $\lambda < (N-1)\mu$) nor at some other solution $\bar{x} \neq \frac{\lambda}{N}$ of (11). The latter implies that the profit gained getting $\bar{x}$ traffic is smaller than $\frac{\lambda}{N}p^{\mathrm{OE}}$. In particular, we must have $\bar{x}\left(\ell\left(\frac{\lambda - \bar{x}}{N-1}\right) - \ell(\bar{x}) + p^{\mathrm{OE}}\right) < \frac{\lambda}{N}p^{\mathrm{OE}}$, for all $\bar{x} \in \left[0, \frac{\lambda}{N}\right)$, which implies Assumption 4 after algebraic rearrangements. □

It is striking that even in the case of homogeneous latency functions, the existence of an equilibrium is not guaranteed. However, Assumption 4 is satisfied by most practical latency functions, and so we can expect equilibria to exist in many settings. For example, Fig. 1 plots the function $\frac{\lambda/N}{x}\ell'\left(\frac{\lambda}{N}\right) - \frac{\ell\left(\frac{\lambda - x}{N-1}\right) - \ell(x)}{\frac{\lambda - x}{N-1} - x}$ (see Assumption 4) when $\ell(x)$ are popular queueing latency functions. In particular, we consider the case where $\ell(x)$ is as in (1) and where $\ell(x) = \log\frac{\mu}{\mu - x}$, which corresponds to the case where the service discipline is Shortest-Remaining-Processing-Time (SRPT) [23,24]. In the figure, we have set $\lambda = 8, \mu = 10$ and $N = 2$. Both curves capture the qualitative behavior of $\frac{\lambda/N}{x}\ell'\left(\frac{\lambda}{N}\right) - \frac{\ell\left(\frac{\lambda - x}{N-1}\right) - \ell(x)}{\frac{\lambda - x}{N-1} - x}$ and reveal that the assumptions of Proposition 4 are satisfied.

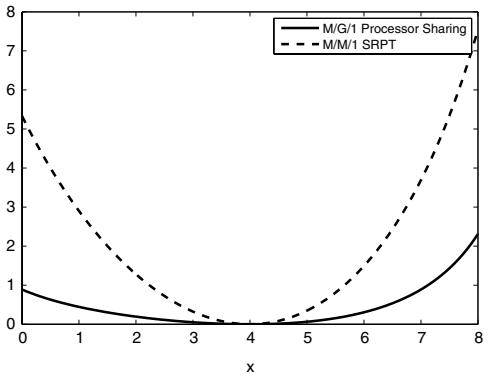

Fig. 1. Plot of the function $\frac{\lambda/N}{x}\ell'\left(\frac{\lambda}{N}\right) - \frac{\ell\left(\frac{\lambda-x}{N-1}\right)-\ell(x)}{\frac{\lambda-x}{N-1}-x}$ for all $x \in [0, \lambda]$ (see Assumption 4) for two popular cases of latency functions from queueing theory. Both functions are positive in $[0, \lambda/2)$, meaning that Assumption 4 holds true.

Also, Observation 1 says that if N is large enough and the arrival rate scales linearly with N, then Assumption 4 holds true and existence is shown.

A final remark about Proposition 4 is that from (9) the equilibrium profit can be easily calculated, and interestingly it can be seen to be increasing with λ.

To conclude the section, we present a result that will be of use later in the paper. In particular, even when the assumptions of Proposition 4 are not met, and so equilibrium existence is not characterized, if an equilibrium exists it must be symmetric.

Proposition 5. *Let Assumptions 1 and 3 hold. Furthermore, let all providers be homogeneous, i.e., $\ell_i(x) = \ell(x)$, $\forall i$. If an oligopolistic equilibrium exists, then it is unique and such that all equilibrium prices are equal.*

Proof. Let us define $\ell'(x_{i_1}^{\mathrm{OE}}) \stackrel{\mathrm{def}}{=} \left.\frac{\partial \ell(x_{i_1})}{\partial x_{i_1}}\right|_{x_{i_1}=x_{i_1}^{\mathrm{OE}}}$ for simplicity.

Assume that $\mathbf{p}^{\mathrm{OE}}$ is an equilibrium such that $0 = p_{i_1}^{\mathrm{OE}} < p_{i_2}^{\mathrm{OE}}$, for some i_1 and i_2. Then, we must have $x_{i_1} = 0$ (otherwise i_1 would set some positive price), which implies $\ell(0) > \ell(x_{i_2}) + p_{i_2}^{\mathrm{OE}}$. This is a contradiction because $\ell(\cdot)$ is increasing by Assumption 1, thus $0 = p_{i_1}^{\mathrm{OE}} < p_{i_2}^{\mathrm{OE}}$ cannot be an oligopolistic equilibrium.

Assume that $\mathbf{p}^{\mathrm{OE}}$ is an equilibrium such that $0 < p_{i_1}^{\mathrm{OE}} < p_{i_2}^{\mathrm{OE}}$, for some i_1 and i_2. Then, we must have

$$\ell(x_{i_1}^{\mathrm{OE}}) + p_{i_1}^{\mathrm{OE}} = \ell(x_{i_2}^{\mathrm{OE}}) + p_{i_2}^{\mathrm{OE}}, \tag{12}$$

where $\mathbf{x}^{\mathrm{OE}} = W(\mathbf{p}^{\mathrm{OE}})$, which implies that $x_{i_1}^{\mathrm{OE}} > x_{i_2}^{\mathrm{OE}}$ because $\ell(\cdot)$ is increasing by Assumption 1.

By substituting the expression of the prices (6) in (12), after little algebra we obtain

$$\ell(x_{i_1}^{\mathrm{OE}}) + x_{i_1}^{\mathrm{OE}}\ell'(x_{i_1}^{\mathrm{OE}}) + \frac{x_{i_1}^{\mathrm{OE}}x_{i_2}^{\mathrm{OE}}\ell'(x_{i_2}^{\mathrm{OE}})}{\sum_{j\neq i_1,i_2} x_{i_2}^{\mathrm{OE}}\frac{\ell'(x_{i_2}^{\mathrm{OE}})}{\ell'(x_j^{\mathrm{OE}})} + x_{i_2}^{\mathrm{OE}}} = \ell(x_{i_2}^{\mathrm{OE}}) + x_{i_2}^{\mathrm{OE}}\ell'(x_{i_2}^{\mathrm{OE}}) + \frac{x_{i_1}^{\mathrm{OE}}x_{i_2}^{\mathrm{OE}}\ell'(x_{i_1}^{\mathrm{OE}})}{\sum_{j\neq i_1,i_2} x_{i_1}^{\mathrm{OE}}\frac{\ell'(x_{i_1}^{\mathrm{OE}})}{\ell'(x_j^{\mathrm{OE}})} + x_{i_1}^{\mathrm{OE}}}. \tag{13}$$

Now, since $x\ell'(x)$ is increasing and $x_{i_1}^{\mathrm{OE}} > x_{i_2}^{\mathrm{OE}}$, we must have

$$\begin{aligned}
&\frac{x_{i_1}^{\mathrm{OE}}\ell'(x_{i_1}^{\mathrm{OE}})\sum_{j\neq i_1,i_2} x_{i_2}^{\mathrm{OE}}\frac{\ell'(x_{i_2}^{\mathrm{OE}})}{\ell'(x_j^{\mathrm{OE}})}}{\sum_{j\neq i_1,i_2} x_{i_2}^{\mathrm{OE}}\ell'(x_{i_2}^{\mathrm{OE}})/\ell'(x_j^{\mathrm{OE}}) + x_{i_2}^{\mathrm{OE}}} + \frac{x_{i_1}^{\mathrm{OE}}x_{i_2}^{\mathrm{OE}}(\ell'(x_{i_1}^{\mathrm{OE}}) + \ell'(x_{i_2}^{\mathrm{OE}}))}{\sum_{j\neq i_1,i_2} x_{i_2}^{\mathrm{OE}}\ell'(x_{i_2}^{\mathrm{OE}})/\ell'(x_j^{\mathrm{OE}}) + x_{i_2}^{\mathrm{OE}}} \\
&> \frac{x_{i_2}^{\mathrm{OE}}\ell'(x_{i_2}^{\mathrm{OE}})\sum_{j\neq i_1,i_2} x_{i_1}^{\mathrm{OE}}\frac{\ell'(x_{i_1}^{\mathrm{OE}})}{\ell'(x_j^{\mathrm{OE}})}}{\sum_{j\neq i_1,i_2} x_{i_1}^{\mathrm{OE}}\ell'(x_{i_1}^{\mathrm{OE}})/\ell'(x_j^{\mathrm{OE}}) + x_{i_1}^{\mathrm{OE}}} + \frac{x_{i_1}^{\mathrm{OE}}x_{i_2}^{\mathrm{OE}}(\ell'(x_{i_1}^{\mathrm{OE}}) + \ell'(x_{i_2}^{\mathrm{OE}}))}{\sum_{j\neq i_1,i_2} x_{i_1}^{\mathrm{OE}}\ell'(x_{i_1}^{\mathrm{OE}})/\ell'(x_j^{\mathrm{OE}}) + x_{i_1}^{\mathrm{OE}}}.
\end{aligned} \tag{14}$$

On the other hand, even $\ell(x_{i_1}^{\mathrm{OE}}) > \ell(x_{i_2}^{\mathrm{OE}})$ because $\ell(\cdot)$ is increasing. This contradiction implies the uniqueness and symmetry of an equilibrium. □

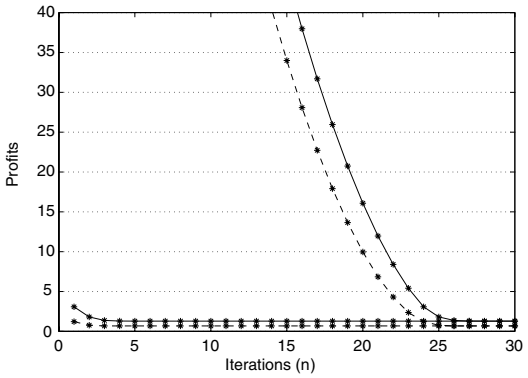

Fig. 2. Providers' profits starting from initial price vectors $\mathbf{p}(0) = (1, 1)$ (the two lines on the bottom) and $\mathbf{p}(0) = (100, 100)$ (the two lines at the top). The continuous (dashed) lines refer to the profit of provider 1 (2).

Table 1
Providers' profits by varying λ with $\mu_1 = 6$, $\mu_2 = 4$.

λ	$p_1(\infty)[W(\mathbf{p}(\infty))]_1$	$p_2(\infty)[W(\mathbf{p}(\infty))]_2$
1	0.071	0.003
2	0.23	0.06
3	0.60	0.25
3.9	1.28	0.67
≥ 4	∞	∞

3.1. Processor sharing queues

Since the known results about the existence of oligopolistic equilibria are limited, in this section we consider the specific example of latency function (1) and use numeric experiments to illustrate the existence of oligopolistic equilibrium in non-homogeneous scenarios.

In order to find oligopolistic equilibria, we apply best response (BR) dynamics, which are known to converge to Nash equilibria in many settings, see [25] for example. Under best response dynamics the system starts with some (non-equilibrium) price vector $\mathbf{p}(0)$. At time $n \geq 0$, provider 1 observes the resulting Wardrop equilibrium $W(\mathbf{p}(n))$ and chooses the price $p_1(n+1)$ that maximizes its profit conditioned on the prices of the others providers, i.e.,

$$p_1(n+1) = \arg\max_{p \geq 0} p[W(p, \mathbf{p}_{-1}(n))]_1. \tag{15}$$

In turn, provider 2 then observes the Wardrop equilibrium resulting from $[p_1(n+1), \mathbf{p}_{-1}(n)]$ and chooses its best price $p_2(n+1)$. In accordance with (5), the sequence of prices $p_i(n)$ represent the *best response* of i.

In Fig. 2, we illustrate the application of best response dynamics in the case of $N = 2$ and latency functions (1). Specifically, Fig. 2 shows the sequence of profits $[p_i(n) \cdot [W(\mathbf{p}(n))]_i]$ when $\mu_1 = 6$, $\mu_2 = 4$, $\lambda = 3$ and different initial vectors $\mathbf{p}(0)$. Notice that the initial price vectors seem to not affect the asymptotic profits. In fact, this is no accident as we can prove that in this setting an oligopolistic equilibrium is unique if it exists.

Specifically, the (necessary) conditions in (6) allow us to immediately derive conditions for the number of possible oligopolistic equilibria, and we have the following proposition.

Proposition 6. *Assume $N = 2$, latency functions (1) with $\mu_1 > \mu_2$, and $\lambda < \mu_2$. The number of oligopolistic equilibria is less than or equal to the number of roots in the interval $(\lambda - \mu_2, \mu_1)$ of the cubic*

$$(\mu_1 - \lambda + x_1)(\mu_2 - \lambda + x_1)^2 = (\mu_2 - x_1)(\mu_1 - x_1)^2. \tag{16}$$

Proof. The result follows after a little algebra and the substitution of (1) in (6). □

From this proposition, we can see that if μ_1 is sufficiently large, it is possible to show that the discriminant of the above cubic is negative (the dominant term is $-O(\mu_1^6)$). Therefore, only one real root exists, meaning that at most one oligopolistic equilibrium can arise in the price competition game.

The results of other experiments with best response dynamics are summarized in Table 1, which shows the equilibrium prices of all servers as λ is varied. Again, these prices do not seem to be dependent of the initial conditions. As one might expect, the profits increase with the load. Further, when $\lambda \geq 4$, this system becomes under-provisioned and Proposition 3 applies.

4. Equilibria efficiency

We now move from the existence of oligopolistic equilibria to the efficiency of oligopolistic equilibria. Before presenting the results it is worth placing them in the context of results about the efficiency of Wardrop equilibria for nonatomic load balancing games. In particular, there are some settings where Wardrop equilibria are quite efficient, for example in the case of affine $\ell_i(\cdot)$ where the price of anarchy is $4/3$ independent of N. However, more commonly the price of anarchy can be large. For example, when polynomial ℓ_i are considered the price of anarchy approximately grows with the degree of the polynomial [2]. Further, when queueing latency functions are considered the price of anarchy can be unbounded. In the case of $M/GI/1$/Processor-Sharing queues the price of anarchy is N [3] and if other scheduling policies such as Shortest-Remaining-Processing-Time first are considered the price of anarchy can be unbounded already in the case of $N = 2$ [26].

One motivation for studying oligopolistic pricing is to understand how competition among profit maximizing providers affects this inefficiency. Thus, we focus again on the price of anarchy (PoA), defined formally as follows in our setting:

$$\mathrm{PoA}(N; \ell_1, \ldots, \ell_N) \stackrel{\text{def}}{=} \sup_{\mathbf{p}^{\mathrm{OE}}} \frac{\ell(\mathbf{x}^{\mathrm{OE}})}{\ell(\mathbf{x}^{\mathrm{Opt}})} \geq 1, \tag{17}$$

where $\mathbf{p}^{\mathrm{OE}}$ is an oligopolistic equilibrium.

This is a natural measure of efficiency to consider for applications in communication and transportation networks where latency provides a measure of the user experience (or quality of service). For example, in the context of a cloud-computing 'infrastructure-as-a-service' provider, user experience is of key importance and any increased latency compared to the optimal allocation indicates suboptimal resource allocation.

Though we focus on the price of anarchy as defined above, other complementary efficiency measures have been looked at previously in the literature. For example, the oligopolistic pricing game under investigation has been studied in [10] with a focus on an efficiency measure that is defined as the ratio between the social surplus at the worst oligopolistic equilibrium and at the optimal allocation. Under this measure, it is shown that the efficiency is remarkably high, bounded from below by $5/6$.[2] Thus, the oligopolistic equilibrium is quite efficient under that measure. However, those results do not imply any bound on the price of anarchy defined above.

The main conclusion of the results in this section is that, though the price of anarchy can be arbitrarily large if N is small (and so there is little competition), it converges to 1 with rate $1/N$ as the number of providers N increases to infinity. This holds even under queueing latency functions, where the contrast is stark. If N is large the Wardrop equilibrium can be extremely inefficient, while the oligopolistic equilibrium is extremely efficient. This highlights the benefit of increasing competition among providers to the underlying architecture that hosts the providers. In the context of cloud-computing, an infrastructure provider such as [15–17] should therefore encourage competition in order to optimally exploit its available resources.

4.1. Inefficiency under limited competition

The fact that Wardrop equilibria can be extremely inefficient already provides intuition for the fact that the oligopolistic equilibria can be inefficient for a small number of providers. However, the situation is even worse than one might expect. We show in the following that the price of anarchy of oligopolistic equilibria is unbounded even in the case when affine latency functions are considered, which we noted earlier have a $4/3$ price of anarchy in the case of the Wardrop equilibrium. Thus, limited competition can induce inefficiency even in situations that are, in some sense, 'inherently' efficient.

Proposition 7. $\sup_{N;\ell_1,\ldots,\ell_N} \mathrm{PoA}(N; \ell_1, \ldots, \ell_N) = \infty$.

Proof. To prove the result it suffices to give a family of examples for which the price of anarchy are unbounded. For this, we consider $N = 2$ and $\ell_1(x_1) = a_1 x_1$, $\ell_2(x_2) = x_2$. From [10] we know that with linear latency functions, an oligopolistic equilibrium exists and that in equilibrium both providers will be used.

We first calculate the optimal allocation (see Definition 1). Solving

$$\begin{aligned} \mathbf{x}^{\mathrm{Opt}} = \; & \arg\min \frac{x_1}{\lambda} a_1 x_1 + \frac{x_2}{\lambda} x_2 \\ & \text{s.t.} : \; x_1 + x_2 = \lambda, \; x_1, x_2 \geq 0, \end{aligned}$$

we obtain

$$x_1^{\mathrm{Opt}} = \frac{\lambda}{a_1 + 1}. \tag{18}$$

[2] We observe that our notion of socially optimal allocation slightly differs from the one in [10], where the authors use the additional parameter R called *reservation utility*. However, if R is sufficiently large, then both notions of social optimum are equivalent.

We next calculate the oligopolistic equilibrium. Since in equilibrium both providers will be used, from (6) we get

$$\begin{cases} 2a_1x_1 + x_1 = 2x_2 + a_1x_2 \\ \sum_i x_i = \lambda, \quad x_i \geq 0 \ \forall i. \end{cases} \tag{19}$$

i.e.,

$$x_1^{\mathrm{OE}} = \frac{\lambda}{3}\frac{a_1+2}{a_1+1}. \tag{20}$$

To show that the price of anarchy can be unbounded, we need to bound the following quantity

$$\frac{\ell(\mathbf{x}^{\mathrm{OE}})}{\ell(\mathbf{x}^{\mathrm{Opt}})} = \frac{a_1\frac{(\lambda(a_1+2))^2}{9(a_1+1)^2} + \left(\lambda - \frac{\lambda(a_1+2)}{3(a_1+1)}\right)^2 + \left(\lambda - \frac{\lambda(a_1+2)}{3(a_1+1)}\right)}{a_1\frac{\lambda^2}{(a_1+1)^2} + \left(\lambda - \frac{\lambda}{a_1+1}\right)^2 + \left(\lambda - \frac{\lambda}{a_1+1}\right)}. \tag{21}$$

Notice that if a_1 is large in the above, then

$$\frac{\ell(\mathbf{x}^{\mathrm{OE}})}{\ell(\mathbf{x}^{\mathrm{Opt}})} \sim \frac{a_1\frac{\lambda^2}{9} + \left(\lambda - \frac{\lambda}{3}\right)^2}{\lambda^2} \sim a_1/9, \tag{22}$$

which implies that the price of anarchy can be made arbitrarily large as desired. □

4.2. Efficiency under increased competition

Having shown that inefficiency can occur when the number of providers is small, we now turn to the case where there is significant competition, i.e., the case of large N. Our goal in this section is to highlight that as $N \to \infty$ the increased competition among providers yields efficient traffic allocations. However, to show such a result we need to first define a scaling of our system that allows us to take $N \to \infty$.

The scaling we define in order to consider 'large' systems is the following. We consider there are $C \geq 1$ provider 'types' or 'classes', where all providers belonging to a given class have the same latency function. In other words, providers of the same class are statistically equivalent. We then define β_c as the proportion of class-c providers, and consider a limit where N grows but $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_C)$ remains fixed.

This does not completely define the scaling however, it is additionally important that the arrival rate grows as the size of the system N grows, otherwise the system becomes trivial. In particular, if the arrival rate is not scaled, then as N grows the congestion of the system disappears. Thus, we seek to scale the arrival rate so that the congestion of the system remains 'constant' through the scaling. To achieve this we choose the arrival rate when there are N servers as λN. This scaling is natural because it keeps the 'load' constant, i.e., ratio of the arrival rate to the total service capacity, of the system the same as N is scaled. Additionally, this scaling models the fact that the increasing number of providers is likely a response to a growing market, i.e., arrival rate. In the queueing theory literature, this scaling of the arrival rate with the system capacity is quite popular, e.g., [27,28]. Other scalings can be also considered [29] and generalizing the results in this section to such scalings is an interesting topic for future work. However, with respect to *heavy-traffic* scalings where the arrival rate approaches the overall network capacity, Proposition 3 already says that an oligopolistic equilibrium cannot exist.

In the context of the scaling described above, we can now state our main result.

Theorem 1. *Let Assumptions 1–3 hold. As $N \to \infty$, there exists at most one oligopolistic equilibrium and, for any C and $\boldsymbol{\beta}$,*

$$\lim_{N\to\infty,\boldsymbol{\beta}} \mathrm{PoA}(N; \ell_1, \ldots, \ell_C) = 1. \tag{23}$$

Further this limit is approached at a rate of $1/N$.

Proof. Let c range from 1 to C, the number of classes, and let x_c^{OE} be the oligopolistic-equilibrium traffic that a provider of class c gets, i.e., $x_c^{\mathrm{OE}} = x_i^{\mathrm{OE}}$ for each provider i of class c. Let also $\mathcal{S}_N$ be the set of all oligopolistic-equilibrium traffic allocations assuming that the total number of providers is N. It is clear that $\mathcal{S}_N \subseteq \mathcal{T}_N \stackrel{\mathrm{def}}{=} \{\mathbf{x} : \mathbf{x}\text{is a solution of (6)}\}$. In the following, we prove the theorem by showing that $|\mathcal{T}_N| \to 1$ as $N \to \infty$ (keeping fixed $\boldsymbol{\beta}$) and that the limiting allocation is $\mathbf{x}^{\mathrm{Opt}}$.

We have that, as $N \to \infty$:

(i) within the constraints in (6), $\left.\frac{\partial \ell_i(x_i)}{\partial x_i}\right|_{x_i = x_i^{\mathrm{OE}}}$ is strictly positive by Assumption 1 and it can be bounded from above by means of Assumption 2 (in the conditions of Assumption 2, for any fixed arrival and service rates, one has $\left.\frac{\partial \ell_i(x_i)}{\partial x_i}\right|_{x_i = x_i^{\mathrm{OE}}} < \left.\frac{\partial \ell_i(x_i)}{\partial x_i}\right|_{x_i = \mu_i - \epsilon}$, for ϵ small enough),

(ii) providers of the same class must have the same equilibrium price and traffic allocation by symmetry (this follows by using the same by-contradiction argument used in the proof of Proposition 5), and
(iii) at least one class of providers is used.

Thus, for all c we have

$$\frac{x_c^{\mathrm{OE}}}{\sum\limits_{i\neq j:A_i=0}\left(\frac{\partial \ell_i(x_i)}{\partial x_i}\Big|_{x_i=x_i^{\mathrm{OE}}}\right)^{-1}} = \frac{x_c^{\mathrm{OE}}}{\sum\limits_{\substack{c'=1:\\ A_{c'}=0}}^{C}(N-1_{\{c'=c\}})\beta_{c'}\left(\frac{\partial \ell_{c'}(x_{c'})}{\partial x_{c'}}\Big|_{x_{c'}=x_{c'}^{\mathrm{OE}}}\right)^{-1}} \to 0, \tag{24}$$

where x_c^{OE} is the traffic to one provider of class c.

Therefore, in the limit,

$$\begin{cases} p_c^{\mathrm{OE}} = x_c^{\mathrm{OE}}\dfrac{\partial \ell_i(x_c)}{\partial x_c}\Big|_{x_c=x_c^{\mathrm{OE}}}, & \forall c: A_c=0 \\ p_c^{\mathrm{OE}} = 0, \quad \forall i: A_c>0 \\ \ell_c(x_c^{\mathrm{OE}}) + p_c^{\mathrm{OE}} = B^{\mathrm{OE}} + A_c, & \forall c \\ A_c x_c = 0, \quad A_c \geq 0, \ \forall c \\ \sum\limits_c \beta_c x_c^{\mathrm{OE}} = \lambda. \end{cases} \tag{25}$$

In the same limiting regime, the optimality conditions of (2) become

$$\begin{aligned} &\ell_c(x_c) + x_c\frac{\partial \ell_c(x_c)}{\partial x_c} = \lambda(W+Y_c), \quad \forall c \\ &Y_c x_c = 0, \quad Y_c \geq 0, \ \forall c \\ &\sum_c \beta_c x_c = \lambda, \quad x_c \geq 0, \ \forall c, \end{aligned} \tag{26}$$

where W and Y_c are Lagrange multipliers ($Y_c = 0$ if and only if c is used), and x_c is the traffic to one provider of class c.

Given that the optimization problem (2) is strictly convex (by Assumption 1), the KKT conditions (26) are both necessary and sufficient for the existence of a unique solution.

This means that the conditions (25) must also have a unique solution coinciding with the solution of (26). In fact, a solution for (25) exists if and only if $A_c = \lambda Y_c$ and $B^{\mathrm{OE}} = \lambda W$ because both (26) and (25) have the same structure.

Provided that an oligopolistic equilibrium exists, this immediately implies $\lim_{N\to\infty,\boldsymbol{\beta}} \mathrm{PoA}(N;\ell_1,\ldots,\ell_N) = 1$. To complete the proof, we note that, as N grows, the allocation at any oligopolistic equilibrium (provided that it exists) converges to its asymptotic value with rate $1/N$. This is evident given (24). □

Although, it is not unexpected to see that the price of anarchy for oligopolistic equilibria converges to one given the maxim that "competition yields efficiency", it is perhaps surprising how quickly this convergence happens (at a rate of $1/N$). This highlights that, not only are oligopolistic equilibria efficient when N is large, but they become efficient quickly as N grows. Another observation that follows from the proof of Theorem 1 is that the prices charged by the providers in the oligopolistic equilibrium converge to the Pigouvian taxes [1], which are known to induce optimal behavior. Further, the convergence to these prices happens at a rate of $1/N$, so even for small N the prices are close to the socially optimal ones. We explore both of these observations further in the context of queueing-based latency functions in Section 4.3.

To conclude this section, let us briefly remark about one simple, but important, extension of Theorem 1. Specifically, note that the optimal allocation we have used as a benchmark to this point, though it is commonly used as a benchmark for load balancing and congestion games, is not truly the optimal allocation. That is, it assumes that requests are allocated to servers probabilistically in a i.i.d. manner, i.e., that the probability a job is forwarded to provider i is $x_i^{\mathrm{Opt}}/\lambda$. In practice, for example in cloud computing systems, upon the arrival of a user request it is possible to exploit information about where recent requests were sent in order to improve the overall latency. For example, if $\ell_i = \ell_j$ for all i, j, then a central broker implementing a Round-Robin allocation[3] outperforms any probabilistic routing policy, see [30].

Thus, to understand the true inefficiency among all possible routing policies (rather than among just probabilistic routing policies), the task is more complex. To illustrate this, note that the arrival process of users to providers changes considerably and this affects the latency functions in a non-trivial manner. However, recent results have provided a bound on the difference between latencies of the probabilistic and non-probabilistic optimal routing schemes, see [31]. In particular, if we let $\tilde{\mathbf{x}}^{\mathrm{Opt}}$ represent the non-probabilistic optimal traffic allocation and $\tilde{\ell}_i$ represent the induced latency functions, then [31] proves that $\tilde{\ell}(\tilde{\mathbf{x}}^{\mathrm{Opt}}) \geq \ell(\mathbf{x}^{\mathrm{Opt}})/2$.

[3] Under Round-Robin, request i is sent to provider i mod N.

Applying this result to the setting of the current paper then simply doubles the price of anarchy proven in Theorem 1. Thus, as $N \to \infty$, the oligopolistic equilibrium provides latency within a factor of 2 of that provided by the optimal, non-probabilistic allocation. This is important when contrasting the distributed setting modeled by an oligopolistic equilibrium (which cannot induce non-probabilistic routing) with a centralized approach where non-probabilistic routing is standard.

4.3. Processor sharing queues

To obtain more insight into the efficiency of oligopolistic equilibria we now focus on one specific class of latency functions where explicit results can be obtained—the $M/GI/1$/Processor-Sharing queue (1). Further, this class represents a particularly important model for communication systems.

The social optimal allocation in this setting is well understood; see, for example, [32,3,4]. Assuming for simplicity that $\infty > \mu_1 > \mu_2 > \cdots > \mu_N > 0$, we have

$$x_j^{\text{Opt}} = \mu_j - \sqrt{\mu_j}\frac{\sum_{i=1}^{\bar{i}} \mu_i - \lambda}{\sum_{i=1}^{\bar{i}} \sqrt{\mu_i}}, \quad \forall j \le \bar{i}, \tag{27}$$

where $\mu_{N+1} = 0$,

$$\bar{i} = \min\left\{ i \ge 1 : \sqrt{\mu_{i+1}} \le \left(\sum_{j=1}^{i} \mu_j - \lambda\right) \Big/ \sum_{j=1}^{i} \sqrt{\mu_j} \right\}$$

and $x_j^{\text{Opt}} = 0$, for all $j > \bar{i}$.

The above allocation yields a socially-optimal cost of

$$\ell(\mathbf{x}^{\text{Opt}}) = \frac{1}{\lambda}\frac{\left(\sum_{i=1}^{\bar{i}} \sqrt{\mu_i}\right)^2}{\sum_{i=1}^{\bar{i}} \mu_i - \lambda} - \frac{\bar{i}}{\lambda}. \tag{28}$$

The Wardrop equilibrium is also well understood for this setting, [32,3,4]. Additionally, the price of anarchy for the Wardrop equilibrium has been bounded by N, which has been shown to be tight [3].

We now move to characterizing the oligopolistic equilibrium. Within this setting, we have already seen in the numerical experiments of Section 3 that oligopolistic equilibria exist if the system is not 'over-provisioned'. Further, we have seen that there is often a unique oligopolistic equilibrium. We now derive explicit bounds on the price of anarchy, traffic allocation, and prices that emerge in an oligopolistic equilibrium.

Theorem 2. *For each i, let $\ell_i(x_i)$ as in (1). If $\mathbf{p}^{\text{OE}}$ denotes the price vector at some OE that uses k providers, then the following inequalities hold true for each $i \le k$:*

$$\frac{x_i^{\text{Opt}}}{\mu_i - x_i^{\text{Opt}}} \le p_i^{\text{OE}} \le \frac{x_i^{\text{Opt}}}{\mu_i - x_i^{\text{Opt}}} + \frac{\mu_i}{\sum_{j \le k: j \ne i} \mu_j} B \tag{29}$$

$$x_i^{\text{OE}} \le \mu_i - \sqrt{\frac{\mu_i}{B}} \tag{30}$$

$$\frac{\mu_i}{(\mu_i - x_i^{\text{Opt}})^2} \le B \le \frac{\mu_i}{(\mu_i - x_i^{\text{Opt}})^2}\left(\max_{i' \le k} 1 - \frac{\mu_{i'}}{\sum_{j \le k: j \ne i'} \mu_j}\right)^{-1}. \tag{31}$$

Proof. Without loss of generality, assume that the number of providers used is N (if only k providers are used, the following sums and indices should be up to k). Using Proposition 2, in an oligopolistic equilibrium $\mathbf{p}^{\text{OE}}$ the following conditions must hold

$$\begin{cases} \dfrac{x_i^{\text{OE}}}{(\mu_i - x_i^{\text{OE}})^2} \le p_i^{\text{OE}} \le \dfrac{x_i^{\text{OE}}}{(\mu_i - x_i^{\text{OE}})^2} + \dfrac{x_i^{\text{OE}}}{\sum_{j \ne i}(\mu_j - x_j^{\text{OE}})^2}, & \forall i \\ (\mu_i - x_i^{\text{OE}})^{-1} + p_i^{\text{OE}} = B, & \forall i \\ \sum_i x_i^{\text{OE}} = \lambda. \end{cases} \tag{32}$$

Using the first inequality, we have the conditions

$$\begin{cases} (\mu_i - x_i^{\mathrm{OE}})^{-1} + \dfrac{x_i^{\mathrm{OE}}}{(\mu_i - x_i^{\mathrm{OE}})^2} \le B, \quad \forall i \\ \displaystyle\sum_i x_i^{\mathrm{OE}} = \lambda \end{cases} \tag{33}$$

which coincide with the KKT conditions of (2) provided that $\le$ is replaced by $=$. This means that $\mathbf{x}^{\mathrm{Opt}}$ satisfies (33) when

$$B \ge B^{\mathrm{Opt}} = (\mu_i - x_i^{\mathrm{Opt}})^{-1} + \frac{x_i^{\mathrm{Opt}}}{(\mu_i - x_i^{\mathrm{Opt}})^2}. \tag{34}$$

Using the inequality in (33) and the second equality of (32), we find

$$\mu_i (B - p_i^{\mathrm{OE}})^2 - B \le 0. \tag{35}$$

Since $B - p_i^{\mathrm{OE}} > 0$, we get

$$\sqrt{\frac{B}{\mu_i}} \ge B - p_i^{\mathrm{OE}} = (\mu_i - x_i^{\mathrm{OE}})^{-1}. \tag{36}$$

We use (36) to bound the second term in the second inequality of (32) obtaining

$$p_i^{\mathrm{OE}} \le \frac{x_i^{\mathrm{OE}}}{(\mu_i - x_i^{\mathrm{OE}})^2} + \frac{\mu_i}{\sum_{j \ne i} \mu_j} B. \tag{37}$$

Substituting (37) in (32), we find

$$\begin{cases} (\mu_i - x_i^{\mathrm{OE}})^{-1} + \dfrac{x_i^{\mathrm{OE}}}{(\mu_i - x_i^{\mathrm{OE}})^2} \ge a_{\max} B, \quad \forall i \\ \displaystyle\sum_i x_i^{\mathrm{OE}} = \lambda \end{cases} \tag{38}$$

where $a_{\max} = \max_i \left\{ 1 - \frac{\mu_i}{\sum_{j \ne i} \mu_j} \right\}$. Conditions (38) are again the KKT conditions of (2) provided that $\ge$ is replaced by $=$. This means that $\mathbf{x}^{\mathrm{Opt}}$ satisfies (38) when

$$a_{\max} B \le B^{\mathrm{Opt}}. \tag{39}$$

Inequalities (29) and (31) follow by the first inequality in (32) and (34), (37), (39). □

Inequalities (29) and (31) provide bounds on the equilibrium price, traffic allocation, and profit of each provider, and these bounds are asymptotically exact since, as N grows, $\mu_i / \sum_{j \ne i} \mu_j$ approaches zero. This is in agreement with Theorem 1.

Corollary 1. *For each i, let $\ell_i(x_i)$ as in (1). Assume that one oligopolistic equilibrium using the same providers of the socially-optimal allocation exists. If $k > 1$ denote the number of providers used, then*

$$\mathrm{PoA}(N; \ell_1, \ldots, \ell_N) \le \left(\max_{i \le k} 1 - \frac{\mu_i}{\sum_{j \le k; j \ne i} \mu_j} \right)^{-1/2}. \tag{40}$$

Proof. From (36) and (39), we find

$$x_i^{\mathrm{OE}} \le \mu_i - \sqrt{\frac{\mu_i}{B}} \le \mu_i - \sqrt{\frac{\mu_i}{B^{\mathrm{Opt}}} a_{\max}}. \tag{41}$$

Since $(\mu_i - x_i)^{-1}$ is increasing in x_i, substituting (41) in $\ell(\mathbf{x}^{\mathrm{OE}})$

$$\ell(\mathbf{x}^{\mathrm{OE}}) \le \frac{1}{\lambda} \sum_i \sqrt{\frac{\mu_i B^{\mathrm{Opt}}}{a_{\max}}} - \frac{N}{\lambda}, \tag{42}$$

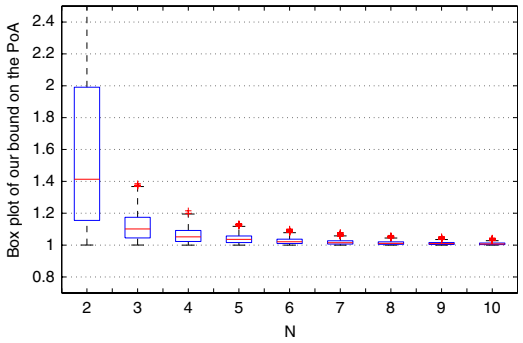

Fig. 3. Boxplot of the bound on the PoA as given in (40) under a varying number of providers. The service rates are chosen uniformly at random from [1, 10 000].

and using (28) we finally get

$$\frac{\ell(\mathbf{x}^{\mathrm{OE}})}{\ell(\mathbf{x}^{\mathrm{Opt}})} \leq \frac{\frac{1}{\lambda}\sum_i \sqrt{\frac{\mu_i B^{\mathrm{Opt}}}{a_{\max}}}}{\frac{1}{\lambda}\sum_i \sqrt{\mu_i B^{\mathrm{Opt}}}} = \frac{1}{\sqrt{a_{\max}}}. \quad \square \tag{43}$$

To provide some more insight into Corollary 1, we provide the results of some numerical experiments in Fig. 3. Specifically, Fig. 3 varies the network size N and plots our bound on the price of anarchy (40) for a random instance of a load balancing game where the service rates have been randomly generated from set [1, 10 000] according to a uniform distribution. Fig. 3 shows a boxplot for each N, which refers to 10,000 experiments. The point of this figure is to illustrate the speed with which the price of anarchy converges to 1. In the figure, we observe that the average performance loss at the worst oligopolistic equilibrium is remarkably small even when N is very small.

5. Concluding remarks

In this paper we have studied the existence and efficiency of oligopolistic equilibria in load balancing games. This model is meant to provide insight into the impact of competition and price in congested markets, e.g., transportation networks and communication networks. One particularly interesting setting that the model studied here provides insight into is that of the cloud computing 'infrastructure-as-a-service' model, where a infrastructure provider such as [15–17] rents out resources to content providers, who then serve traffic requests independently, and are often in competition with each other.

The main insights that stem from our results on existence and efficiency of oligopolistic equilibria are that (i) competition can be disastrous if the system is under-provisioned, i.e., an oligopolistic equilibrium will not exist and the provider prices will grow unboundedly, and (ii) if the system is properly provisioned and an oligopolistic equilibrium exists, then competition yields efficiency, i.e., as the $N \to \infty$, the price of anarchy converges to 1 at a rate of $1/N$.

One important technical contribution of this work is that we define and analyze a scaling of the load balancing game which allows a non-trivial limit to emerge as $N \to \infty$, and thus allows the proof of result (ii) above. This scaling is quite natural and allows the traffic to grow linearly with N. In the queueing theory literature, this is a popular scaling for studying multi-server systems [27,28,7]. An interesting direction for extending this work is to study other scalings of the system in order to understand how robust the conclusion that 'competition yields efficiency' is for load balancing games.

More broadly, there are many open questions that remain about oligopolistic equilibria among competing providers. In particular, the question of existence remains very open—even in simple settings existence has not been characterized. Additionally, there are many interesting variations of the model that are worthy of study, including considering bounded prices for the providers and inelastic traffic.

Acknowledgments

The authors are very grateful to Olivier Brun, Balakrishna Prabhu and the anonymous reviewers for their insightful comments that significantly increased the quality of this paper.

Appendix

Proof of Proposition 2. As done for the proof of Proposition 9 in [10], the proof derives the KKT conditions of optimization problem (5). Point $W(\mathbf{p}^{\mathrm{OE}})$ is uniquely given by the minimizers of (4) whose KKT conditions are

$$\begin{aligned}
&\ell_i(x_i) + p_i^{\mathrm{OE}} - A_i = B, \quad \forall i \\
&\sum_i x_i = \lambda, \quad x_i \geq 0, \ \forall i \\
&A_i x_i = 0, \quad A_i \geq 0, \ \forall i
\end{aligned} \tag{44}$$

where A_i and B are Lagrange multipliers. Assume that an oligopolistic equilibrium exists and denote it by $\mathbf{p}^{\mathrm{OE}}$. Using the optimality conditions (44), for provider j

$$\begin{aligned}
p_j^{\mathrm{OE}} x_j^{\mathrm{OE}} = \max \quad & p_j x_j \\
\text{s.t.}: \quad & \ell_j(x_j) + p_j - A_j = B, \\
& \ell_i(x_i) + p_i^{\mathrm{OE}} - A_i = B, \quad \forall i \neq j \\
& \sum_i x_i = \lambda \\
& A_i x_i = 0, \quad A_i \geq 0 \, \forall i
\end{aligned} \tag{45}$$

where each provider $i \neq j$ has fixed equilibria price p_i^{OE} (a constant here). The constraint $x_i \geq 0, \ \forall i$, does not appear because it is redundant (for a fixed price vector there exists a unique Wardrop equilibria). Without loss of generality, assume that the first $\tilde{N}$ providers are used. Since $p_i^{\mathrm{OE}} > 0$ if and only if $i \leq \tilde{N}$ if and only if $A_i = 0$ if and only if $x_i > 0$, (45) can be rewritten as (for $j \leq \tilde{N}$)

$$\begin{aligned}
p_j^{\mathrm{OE}} x_j^{\mathrm{OE}} = \max \quad & p_j x_j \\
\text{s.t.}: \quad & \ell_j(x_j) + p_j = B, \\
& \ell_i(x_i) + p_i^{\mathrm{OE}} = B, \quad \forall i \neq j : i \leq \tilde{N} \\
& \sum_i x_i = \lambda.
\end{aligned} \tag{46}$$

Clearly, if $j > \tilde{N}$, then $p_j^{\mathrm{OE}} x_j^{\mathrm{OE}} = 0$. With respect to multipliers V_i and W, the KKT conditions of (46) become

$$\begin{cases}
-x_j + V_j = 0 \\
-p_j + V_j \dfrac{\partial \ell_j(x_j)}{\partial x_j} + W = 0 \\
V_i \dfrac{\partial \ell_i(x_i)}{\partial x_i} + W = 0, \quad \forall i \neq j : i \leq \tilde{N} \\
-\displaystyle\sum_{i=1}^{\tilde{N}} V_i = 0 \\
\ell_j(x_j) + p_j = B \\
\ell_i(x_i) + p_i^{\mathrm{OE}} = B, \quad \forall i \neq j : i \leq \tilde{N} \\
\displaystyle\sum_i x_i = \lambda,
\end{cases} \tag{47}$$

and after some algebra we obtain

$$\begin{cases}
p_j = x_j \dfrac{\partial \ell_j(x_j)}{\partial x_j} + \dfrac{x_j}{\displaystyle\sum_{i=1:i\neq j}^{\tilde{N}} \left(\frac{\partial \ell_i(x_i)}{\partial x_i} \right)^{-1}} \\
\ell_j(x_j) + p_j = B \\
\ell_i(x_i) + p_i^{\mathrm{OE}} = B, \quad \forall i \neq j : i \leq \tilde{N} \\
\displaystyle\sum_i x_i = \lambda.
\end{cases}$$

If $\mathbf{p}^{\mathrm{OE}}$ is such that $p_i^{\mathrm{OE}} = 0$ for all $i \neq k$, then the best-response of k is to set the price $(\min_{i\neq k} \ell_i(0) - \ell_k(\lambda))\mathbf{e}_k$ (over this threshold at least one other provider would have the incentive to deviate its price). Provider k must be such that $k = \arg\min_j \ell_j(\lambda)$. In fact, if provider k' is such that $\ell_{k'}(\lambda) \leq \ell_k(\lambda)$, then it would get a non-null proportion of the traffic in the corresponding Wardrop equilibrium, against the assumption. □

References

[1] N. Nisan, T. Roughgarden, E. Tardos, V.V. Vazirani, Algorithmic Game Theory, Cambridge University Press, New York, NY, USA, 2007.
[2] T. Roughgarden, E. Tardos, How bad is selfish routing? J. ACM 49 (2002) 236–259.
[3] M. Haviv, T. Roughgarden, The price of anarchy in an exponential multi-server, Oper. Res. Lett. 35 (4) (2007) 421–426.

[4] E. Altman, U. Ayesta, B. Prabhu, Load balancing in processor sharing systems, Telecommun. Syst. 47 (1–2) (2011) 35–48.
[5] Y. Korilis, A. Lazar, A. Orda, Achieving network optima using stackelberg routing strategies, IEEE Trans. Netw. 5 (1) (1997) 161–173.
[6] T. Roughgarden, Stackelberg scheduling strategies, SIAM J. Comput. 33 (2) (2004) 332–350.
[7] E. Engel, R. Fischer, A. Galetovic, Toll competition among congested roads, NBER Technical Working Papers 0239, National Bureau of Economic Research, Inc., May 1999.
[8] T. Wu, D. Starobinski, A comparative analysis of server selection in content replication networks, IEEE/ACM Trans. Netw. 16 (6) (2008) 1461–1474.
[9] R. Cominetti, J.R. Correa, N.E. Stier-Moses, The impact of oligopolistic competition in networks, Oper. Res. 57 (2009) 1421–1437.
[10] D. Acemoglu, A. Ozdaglar, Competition and efficiency in congested markets, Math. Oper. Res. 32 (1) (2007) 1–31.
[11] D. Acemoglu, A. Ozdaglar, Competition in parallel-serial networks, Games Econom. Behav. 25 (2007) 1180–1192.
[12] A. Hayrapetyan, E. Tardos, T. Wexler, A network pricing game for selfish traffic, in: Proceedings of the Twenty-Fourth Annual ACM Symposium on Principles of Distributed Computing, PODC'05, ACM, New York, NY, USA, 2005, pp. 284–291.
[13] P. Dube, R. Jain, Bertrand games between multi-class queues, in: IEEE CDC, 2009, pp. 8588–8593.
[14] D. Acemoglu, K. Bimpikis, A. Ozdaglar, Price and capacity competition, Games Econom. Behav. 66 (1) (2009) 1–26.
[15] http://aws.amazon.com/ec2/.
[16] http://www.gogrid.com.
[17] http://www.slicehost.com.
[18] L. Kleinrock, Queueing Systems, vol. 2, John Wiley and Sons, 1976.
[19] U.N. Bhat, An Introduction to Queueing Theory: Modeling and Analysis in Applications, Birkhäuser Verlag, 2008.
[20] J.G. Wardrop, Some theoretical aspects of road traffic research, Proc. Inst. Civil Eng. 1 (1952) 325–378.
[21] J.R. Correa, A.S. Schulz, N.E. Stier-Moses, Selfish routing in capacitated networks, Math. Oper. Res. 29 (4) (2004) 961–976.
[22] M.J. Beckmann, C.B. McGuire, C.B. Winsten, Studies in the Economics of Transportation, Yale University Press, New Haven, CT, 1956.
[23] M. Lin, A. Wierman, B. Zwart, Heavy-traffic analysis of mean response time under shortest remaining processing time, Perform. Eval., in press (doi:10.1016/j.peva.2011.06.001).
[24] A. Wierman, Scheduling for today's computer systems: bridging theory and practice, Ph.D. Thesis, Computer Science, Carnegie Mellon University, 2007.
[25] A. Orda, R. Rom, N. Shimkin, Competitive routing in multiuser communication networks, IEEE/ACM Trans. Netw. 1 (1993) 510–521.
[26] H. Chen, J. Marden, A. Wierman, The effect of local scheduling in load balancing designs, in: Proceedings of IEEE Infocom, 2009.
[27] W. Whitt, Stochastic Process Limits, Springer, New York, 2002.
[28] F. Kelly, Loss networks, Ann. Appl. Probab. 1 (1991) 319–378.
[29] H. Kushner, Heavy Traffic Analysis of Controlled Queueing and Communication Networks, Springer-Verlag, New York, 2001.
[30] J. Walrand, An Introduction to Queueing Networks, Prentice-Hall, 1988.
[31] J. Anselmi, B. Gaujal, The price of anarchy in parallel queues revisited, SIGMETRICS Perform. Eval. Rev. 38 (1) (2010) 353–354.
[32] C.H. Bell, S. Stidham, Individual versus social optimization in the allocation of customers to alternative servers, Manage. Sci. 29 (1983) 83–839.

Jonatha Anselmi received the M.Sc. and Ph.D. degrees in computer engineering from Politecnico di Milano (Italy) in 2005 and 2009, respectively.

In 2008, he worked at the Mathematical Sciences Department of IBM T.J. Watson, Yorktown Heights (US), as a summer student. After the Ph.D., he joined INRIA-Grenoble (France) as a postdoc working for the project MESCAL. Currently, he is a postdoc at the Basque Center for Applied Mathematics—BCAM, Bilbao (Spain).

His research interests focus on the performance analysis and optimization of distributed systems over the Internet.

Urtzi Ayesta is an IKERBASQUE researcher at BCAM, the Basque Center for Applied Mathematics, Bilbao, Spain. Previously he has been a CNRS researcher working at LAAS, Toulouse, France and an ERCIM postdoc fellow at CWI, Amsterdam, the Netherlands. He received the Ph.D. degree in Computer Science from Université de Nice-Sophia Antipolis (France). His Ph.D. research work was carried out at the research laboratories of INRIA Sophia-Antipolis and France Telecom R&D. Urtzi Ayesta holds an M.Sc. degree in Electrical Engineering from Columbia University (US) and a Diplome in Telecomunication Engineering from Nafarroako Unibertsitate Publikoa-Universidad Publica de Navarra (Spain). His main research interests include scheduling theory, queueing theory, stochastic processes, game theory and their application to the performance evaluation, conception and dimensioning of telecommunication networks and distributed systems.

Adam Wierman is an Assistant Professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology, where he is a member of the Rigorous Systems Research Group (RSRG). He received his Ph.D., M.Sc. and B.Sc. in Computer Science from Carnegie Mellon University in 2007, 2004, and 2001, respectively. His research interests center around resource allocation and scheduling decisions in computer systems and services. More specifically, his work focuses both on developing analytic techniques in stochastic modeling, queueing theory, scheduling theory, and game theory, and applying these techniques to application domains such as energy-efficient computing, data centers, social networks, and the electricity grid.

He received the ACM SIGMETRICS Rising Star award in 2011, and has also received best paper awards at ACM SIGMETRICS, IFIP Performance, IEEE INFOCOM, and ACM GREENMETRICS. He was named a Seibel Scholar, received an Okawa Foundation grant, and received an NSF CAREER grant. Additionally, his dissertation received the CMU School of Computer Science Distinguished Dissertation Award and was given an honorable mention for the INFORMS Doctoral Dissertation Award for Operations Research in Telecommunications. He has also received multiple teaching awards, including the Associated Students of the California Institute of Technology (ASCIT) Teaching Award.