ACADEMIC PRESS

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT•

Journal of Computer and System Sciences 67 (2003) 341–364

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

http://www.elsevier.com/locate/jcss

The price of anarchy is independent of the network topology☆

Tim Roughgarden[1]

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

Received 14 June 2002; revised 14 November 2002

Abstract

edge of the network ⇒ road
edge congestion ⇒ queue length

We study the degradation in network performance caused by the selfish behavior of noncooperative network users. We consider a model of selfish routing in which the latency experienced by network traffic on an edge of the network is a function of the edge congestion, and network users are assumed to selfishly route traffic on minimum-latency paths. The quality of a routing of traffic is measured by the sum of travel times, also called the *total latency*. The outcome of selfish routing—a *Nash equilibrium*—does not in general minimize the total latency; hence, selfish behavior carries the cost of decreased network performance. We quantify this degradation in network performance via the *price of anarchy*, the worst-possible ratio between the total latency of a Nash equilibrium and of an optimal routing of the traffic. In this paper, we show that the price of anarchy is determined only by the simplest of networks. Specifically, we prove that under weak hypotheses on the class of allowable edge latency functions, the worst-case ratio between the total latency of a Nash equilibrium and of a minimum-latency routing for any multicommodity flow network is achieved by a single-commodity instance in a network of parallel links. In the special case where the class of allowable latency functions includes all of the constant functions, we prove that a network with only two parallel links suffices to achieve the worst-possible ratio. Our guarantee that simple networks always furnish worst-possible examples provides a powerful method for computing the price of anarchy with respect to an arbitrary class of latency functions. We apply this method to function classes that have been well studied in the literature, including degree-bounded polynomials and queuing delay functions. These are the first tight analyses of the price of anarchy for significant classes of latency functions outside the class of linear functions.

☆A preliminary version of this paper appeared in the Proceedings of the 34th Annual ACM Symposium on the Theory of Computing, May 2002.

E-mail address: timr@cs.cornell.edu.

[1] Supported by an NSF Graduate Fellowship, a Cornell University Fellowship, and ONR Grant N00014-98-1-0589. Research performed in part while visiting IBM Almaden.

0022-0000/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-0000(03)00044-8

1. Introduction

1.1. *Selfish routing and the price of anarchy*

We study the degradation in network performance caused by the selfish behavior of noncooperative network users. More precisely, we consider a directed network in which each edge possesses a latency function describing the common latency incurred by all traffic on the edge as a function of the edge congestion. Given a rate of traffic between each pair of nodes in the network, we aspire toward an assignment of traffic to paths minimizing the sum of all travel times (the *total latency*) of the network users. Unfortunately, in many settings network users are free to route their traffic in a selfish manner, without regard to the total latency.

Our goal is to quantify the cost of selfish routing in a network. We assume that an unregulated network user chooses the minimum-latency path from its source to its destination, given the link congestion caused by the rest of the network users, and expect the routes chosen by users to form a *Nash equilibrium* in the sense of classical game theory [17]. We further assume that each network user controls a negligible fraction of the overall traffic—for example, each user could represent a car in a highway system or a packet in a communication network. Feasible assignments of traffic to paths in the network can then be modeled in a continuous manner via *network flow*, with the amount of flow between a pair of nodes in the network equal to the rate of traffic between the two nodes. A Nash equilibrium then corresponds to a flow in which all flow paths between a given source and destination have minimum latency; in a flow without this property, some traffic can improve its travel time by switching from a longer path to a shorter one.

Traffic flows at Nash equilibrium do not in general minimize the total latency incurred by network users; this is a special case of the more general phenomenon that a Nash equilibrium in a noncooperative game need not optimize social welfare [6,20]. We can therefore study the cost of routing selfishly with the following question: for an arbitrary multicommodity flow network with congestion-dependent edge latencies, what is the worst-case ratio between the total latency of a flow at Nash equilibrium and that of a flow with minimum-possible total latency?

Roughgarden and Tardos [26] showed that this ratio, dubbed *the price of anarchy* by Papadimitriou [18], can be arbitrarily large unless additional structure is imposed on the classes of allowable edge latency functions and/or allowable network topologies. This observation motivates the central questions of our paper:

(1) Are there nontrivial classes of network topologies for which the price of anarchy is smaller than for arbitrary networks?
(2) Which classes of edge latency functions admit a finite price of anarchy? What is the price of anarchy in these cases?

1.2. *Our results*

We show that the underlying network topology plays no role in the determination of the price of anarchy, in the following sense. We show that under weak hypotheses on the class of allowable queue service rate: rate at which queue is shortening

latency functions,[2] the worst-case ratio between the total latency of a flow at Nash equilibrium and that of a minimum-latency flow in any multicommodity flow network is achieved by a single-commodity instance in a network of parallel links. Thus, for any fixed class of latency functions, no nontrivial restriction on the class of allowable network topologies (or on the number of commodities) will improve the price of anarchy. In the special case of a class of latency functions that includes all of the constant functions (functions ℓ of the form $\ell(x) = c$ for $c > 0$), we prove that a network with only two parallel links suffices to achieve the worst-possible ratio.

Our guarantee that simple networks always furnish worst-possible examples also provides a powerful method for computing the price of anarchy with respect to an arbitrary class of latency functions. For example, we show that the price of anarchy for networks with latency functions that are polynomials with nonnegative coefficients and degree at most p is achieved by a two-node, two-link network with latency functions $\ell(x) = 1$ and $\ell(x) = x^p$; a calculation then shows that the price of anarchy for such networks is precisely $[1 - p \cdot (p+1)^{-(p+1)/p}]^{-1}$, which is asymptotically $\Theta(\frac{p}{\ln p})$ as $p \to \infty$. We also consider latency functions of the form $\ell(x) = (u - x)^{-1}$ that arise as the delay functions of M/M/1 queues (where $u > 0$ should be interpreted as the edge capacity or the queue service rate) [8]; these latency functions have been extensively studied in the networking community [2,10,11,13,16]. As noted by Friedman [7], the price of anarchy can be finite in this setting only if we constrain the sum of all traffic rates to be at most a constant $R_{\max}$ and the minimum allowable edge capacity to be a constant $u_{\min} > R_{\max}$; in this case, we prove that the price of anarchy is precisely $[1 + \sqrt{u_{\min}/(u_{\min} - R_{\max})}]/2$. A summary of our work computing the price of anarchy for common classes of latency functions is given in Table 1. These results are the first tight analyses of the price of anarchy for significant classes of latency functions outside the class of linear functions, for which the worst-case ratio was shown to be 4/3 by Roughgarden and Tardos [26].

1.3. Related work

The idea of quantifying the inefficiency of selfish solutions using the framework of approximation is due to Koutsoupias and Papadimitriou [12], who proved upper and lower bounds on the worst-possible objective function value of a Nash equilibrium relative to that of an optimal solution in a simple load-balancing game. The price of anarchy in this model is also known as the *coordination ratio*. These bounds on the coordination ratio were subsequently improved by Mavronicolas and Spirakis [15], and Czumaj and Vöcking [4] provided a sharp analysis. Czumaj et al. [3] recently defined and studied generalizations of this load-balancing model.

The traffic routing model studied in this paper dates back to the 1950's [1,31] and has been studied extensively ever since (see [24] for further historical references). The price of anarchy in this model was first investigated by Roughgarden and Tardos [26], who proved that the price of anarchy in networks with linear edge latency functions is precisely 4/3. Roughgarden and Tardos

[2] For example, it suffices for the class to satisfy a mild and standard convexity assumption, to be closed under multiplication by positive scalars, and to possess some latency function that is positive when evaluated with zero congestion.

Table 1
The price of anarchy for common classes of edge latency functions

Description	Typical Representative	Price of Anarchy
Linear	$ax + b$	$\frac{4}{3} \approx 1.333$
Quadratic	$ax^2 + bx + c$	$\frac{3\sqrt{3}}{3\sqrt{3} - 2} \approx 1.626$
Cubic	$ax^3 + bx^2 + cx + d$	$\frac{4\sqrt[3]{4}}{4\sqrt[3]{4} - 3} \approx 1.896$
Polynomials of degree $\leqslant p$	$\sum_{i=0}^{p} a_i x^i$	$\Theta\left(\frac{p}{\log p}\right)$
M/M/1 delay functions	$(u - x)^{-1}$	$\frac{1}{2}\left(1 + \sqrt{\frac{u_{\min}}{u_{\min} - R_{\max}}}\right)$
M/G/1 delay functions	$\frac{1}{u} + \frac{x(1 + \sigma^2 u^2)}{2u(u - x)}$	See subsection subsection 5.2

The price of anarchy for linear latency functions was first determined in [26]; all other results are new to this paper. Polynomial coefficients are assumed nonnegative. The parameters u and σ are the expectation and standard deviation of the associated queue service rate distribution, see Section 5 for details. $R_{\max}$ denotes the maximum allowable amount of network traffic, and $u_{\min}$ denotes the minimum allowable edge service rate (or capacity).

[26] also showed that, assuming only that latency functions are continuous and nondecreasing, the total latency incurred by a flow at Nash equilibrium is at most that of an optimal flow forced to route twice as much traffic between each source-destination pair. The model and results of [26] have recently been extended in several different directions [7,25,28,30] that we will not detail here.

We know of no previous work proving that worst-case consequences of selfish behavior can always occur in simple games. Indeed, our guarantee that networks of parallel links suffice to capture worst-case examples is provably false in many variants of the selfish routing model studied in this paper. For example, Roughgarden and Tardos [24,26] proved that if network users carry a nonnegligible fraction of the overall traffic, or if they are allowed to route traffic on paths that are only approximately minimum-latency, then the worst-case cost of routing selfishly can be larger in general networks than in networks of parallel links. When network users carry a nonnegligible amount of traffic, even the basic issues of existence and uniqueness of Nash equilibria are more troublesome in general networks than in networks of parallel links [14,16]. Similarly, Roughgarden showed that the equilibria in a Stackelberg game related to the traffic model studied here are provably worse in general graphs than in networks of parallel links [23,24], that Braess's Paradox—the counterintuitive phenomenon that removing arcs from a network can *improve* the cost of selfish routing—grows increasingly severe as the underlying network becomes more complex [22], and that worst-case examples for selfish routing with respect to the *maximum* (rather than the total) latency grow in severity with the network size [21].

1.4. Organization

After reviewing some technical preliminaries in Section 2, in Section 3 we give an upper bound on the price of anarchy with respect to an arbitrary class of allowable latency functions. In Section 4 we give matching lower bounds using simple networks, thereby showing that the price of anarchy is essentially independent of the class of allowable network topologies. In Section 5 we leverage the fact that worst-case examples for selfish routing occur in simple networks to compute the price of anarchy for several well-studied classes of latency functions. We conclude in Section 6.

2. Preliminaries

2.1. The model

We consider a directed network $G = (V, E)$ with vertex set V, edge set E, and k source-destination vertex pairs $\{s_1, t_1\}, \ldots, \{s_k, t_k\}$. We allow parallel edges but have no use for self-loops. We denote the set of (simple) s_i-t_i paths by $\mathscr{P}_i$, and define $\mathscr{P} = \cup_i \mathscr{P}_i$. To avoid trivialities, we assume that $\mathscr{P}_i \neq \emptyset$ for each i. A *flow* is a function $f : \mathscr{P} \to \mathscr{R}^+$; for a fixed flow f we define $f_e = \sum_{P:e \in P} f_P$. We sometimes refer to a source–destination pair $\{s_i, t_i\}$ and the s_i–t_i paths of $\mathscr{P}_i$ as *commodity i*.

We associate a finite and positive *rate* r_i with each pair $\{s_i, t_i\}$, the amount of flow with source s_i and destination t_i; a flow f is said to be *feasible* if for all i, $\sum_{P \in \mathscr{P}_i} f_P = r_i$. Finally, each edge $e \in E$ possesses a congestion-dependent *latency* that we denote by $\ell_e(\cdot)$. For each edge $e \in E$, we assume that the latency function ℓ_e is nonnegative, differentiable, and nondecreasing. Unless otherwise noted, we will assume that latency functions are defined on all of $[0, \infty)$. The latency of a path P with respect to a flow f is defined as the sum of the latencies of the edges in the path, denoted by $\ell_P(f) = \sum_{e \in P} \ell_e(f_e)$. We will call the triple (G, r, ℓ) an *instance*.

We define the *cost* $C(f)$ of a flow f in G as the total latency incurred by f, so $C(f) = \sum_{P \in \mathscr{P}} \ell_P(f) f_P$. By summing over the edges in a path P and reversing the order of summation, we may also write $C(f) = \sum_{e \in E} \ell_e(f_e) f_e$. With respect to an instance (G, r, ℓ), a feasible flow minimizing $C(f)$ is said to be *optimal* or *minimum-latency*.

2.2. Flows at Nash equilibrium

Following Roughgarden and Tardos [26], we formalize our notion of a "selfishly defined traffic flow" in the next definition. Intuitively, we expect each unit of such a flow (no matter how small) to travel along the minimum-latency path available, where latency is measured with respect to the rest of the flow; otherwise, this unit of flow would reroute itself on a path with smaller latency.

Definition 2.1. A flow f feasible for instance (G, r, ℓ) is at *Nash equilibrium* (or is a *Nash flow*) if for all $i \in \{1, \ldots, k\}$, $P_1, P_2 \in \mathscr{P}_i$ with $f_{P_1} > 0$, and $\delta \in (0, f_{P_1}]$, we have $\ell_{P_1}(f) \leqslant \ell_{P_2}(\tilde{f})$, where

$$\tilde{f}_P = \begin{cases} f_P - \delta & \text{if } P = P_1, \\ f_P + \delta & \text{if } P = P_2, \\ f_P & \text{if } P \notin \{P_1, P_2\}. \end{cases}$$

Letting δ tend to 0, continuity and monotonicity of the edge latency functions give the following useful characterization of a flow at Nash equilibrium.

Proposition 2.2. *A flow f feasible for (G, r, ℓ) is at Nash equilibrium if and only if for every $i \in \{1, \ldots, k\}$ and $P_1, P_2 \in \mathcal{P}_i$ with $f_{P_1} > 0$, $\ell_{P_1}(f) \leqslant \ell_{P_2}(f)$.*

Briefly, Proposition 2.2 states that, in a flow at Nash equilibrium, all flow travels on minimum-latency paths. In particular, if f is at Nash equilibrium then all s_i–t_i flow paths (s_i–t_i paths to which f assigns a positive amount of flow) have equal latency, say $L_i(f)$. We can therefore express the cost $C(f)$ of a flow f at Nash equilibrium in a particularly nice form.

Proposition 2.3. *If f is a flow at Nash equilibrium for (G, r, ℓ), then*

$$C(f) = \sum_{i=1}^{k} L_i(f) r_i.$$

$L_i(f) = \ell_P(f)$ (for $P \in \mathcal{P}_i$)

It is also reassuring to note that flows at Nash equilibrium always exist and are essentially unique.

Proposition 2.4 (Beckmann et al. [1], Dafermos and Sparrow [5], Roughgarden and Tardos [26]). *Let (G, r, ℓ) be an instance with continuous, nondecreasing latency functions.*

(a) *(G, r, ℓ) admits a flow at Nash equilibrium.*
(b) *If f and f' are flows at Nash equilibrium for (G, r, ℓ), then $C(f) = C(f')$.*

2.3. *Characterizing optimal flows via marginal cost functions*

We have given in Proposition 2.2 a convenient characterization of flows at Nash equilibrium. Assuming mild extra conditions on our latency functions, there is an analogous characterization of optimal flows.

Definition 2.5. A latency function ℓ is *standard* if $x \cdot \ell(x)$ is convex on $[0, \infty)$. or $\ell^*(x)$ is increasing

Most but not all latency functions of interest are standard. All convex latency functions are standard, as are some well-behaved nonconvex functions such as $\log(1 + x)$. Nondecreasing and differentiable approximations of step functions are the most notable examples of nonstandard latency functions.

To state the characterization of optimal flows, we require one further definition.

Definition 2.6. If ℓ is a standard latency function, then the corresponding *marginal cost function* ℓ^* is defined by

$$\ell^* = \frac{d}{dx}(x \cdot \ell(x)).$$

By Definitions 2.5 and 2.6, the marginal cost function corresponding to a standard latency function is a nondecreasing function. Basic calculus (see e.g. [27, pp.108–109]) implies that such functions are also continuous.

We will typically denote an optimal flow by f^*. We denote the marginal cost function of an edge by ℓ^* because it is, in a sense, an "optimal latency function". Our final preliminary result makes this precise, and asserts that optimal flows arise as Nash flows with respect to the latency functions ℓ^*.

Proposition 2.7 (Beckmann et al. [1], Dafermos and Sparrow [5], Roughgarden and Tardos [26]). *Let (G, r, ℓ) be an instance with standard latency functions and corresponding marginal cost functions ℓ^*. Then, a flow f^* feasible for (G, r, ℓ) is optimal if and only if it is at Nash equilibrium for (G, r, ℓ^*).*

2.4. *Pigou's example and the inefficiency of Nash flows*

We next illustrate the definitions and propositions of this section with a simple but important example, essentially due to Pigou [19]. Consider a network with two nodes s and t, two parallel edges with latency functions $\ell(x) = 1$ and $\ell(x) = x$, and a traffic rate of 1 (see Fig. 1(a)). These two latency functions are standard in the sense of Definition 2.5. Routing all flow on the bottom link equalizes the latencies of the two available s–t paths at 1, and thus by Proposition 2.2 provides a flow f at Nash equilibrium. By Proposition 2.3 or by inspection, the cost $C(f)$ of f is 1.

Next, the marginal cost functions of the network are $\ell^*(x) = 1$ and $\ell^*(x) = 2x$, as shown in Fig. 1(b). Routing half of the traffic on each link equalizes the marginal costs of the two s–t paths at 1, and so by Proposition 2.7 furnishes a minimum-latency flow f^*. The cost of f^* is $C(f^*) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 1 = \frac{3}{4}$. Pigou's example thus demonstrates our assertion in Section 1 that Nash flows fail to optimize the total latency.

A qualitative explanation for the inefficiency of the Nash flow in Pigou's example is easily found. The lower edge of the network is capable of providing some traffic with a quick alternative to the upper edge, *provided it is used in moderation*. The optimal flow quite sensibly realizes this and divides traffic equally among the two routes, so that half of the traffic enjoys a relatively uncongested ride on the bottom edge. The selfish users in the Nash flow, by contrast, are unable to resist traveling along the bottom route until the congestion is so great as to render the edge useless.

The inefficiency of selfish routing can be more severe. Consider the minor modification of Pigou's example obtained by replacing the latency function $\ell(x) = x$ by the nonlinear one

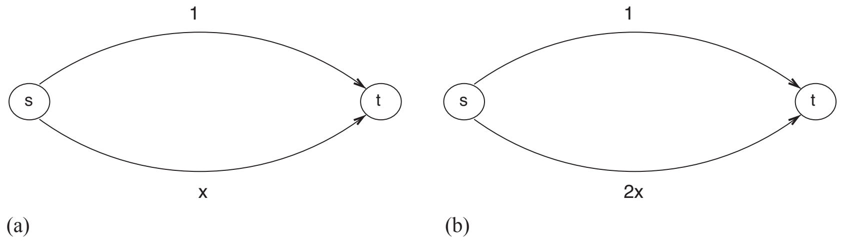

Fig. 1. Pigou's example. (a) Latency functions; (b) marginal cost functions.

$\ell(x) = x^p$ for $p \geqslant 2$. With a traffic rate of 1, the Nash flow f is the same as in Pigou's example; all flow is routed on the bottom link and the total latency is 1 (for any choice of p). On the other hand, the discrepancy between the latency functions (1 and x^p) and the marginal cost functions (1 and $(p+1)x^p$) is larger; now, the flow f^* that routes $(p+1)^{-1/p}$ units on the lower link and the remainder on the upper link equalizes the marginal costs of the two links at 1 and is thus optimal. The cost $C(f^*)$ of f^* is $1 - p \cdot (p+1)^{-(p+1)/p}$, which tends to 0 as $p \to \infty$. Thus, if arbitrarily steep latency functions are allowed, a flow at Nash equilibrium can be arbitrarily more costly than an optimal flow.

We have given a family of examples in which the price of selfishness, quantified by the ratio of the costs of the Nash and optimal flows, grows with the degree of nonlinearity of the network latency functions. From a *qualitative* perspective, however, all of these examples are identical: the Nash flow is inefficient because selfish users cannot resist overcongesting an edge that is beneficial when used in moderation. Put differently, the singular obstruction preventing Nash flows from optimizing the total latency in these examples is the inability (or unwillingness) of selfish users to discern which of two competing routes is superior from a global perspective.

In general multicommodity flow networks, we might expect additional problems to arise with unregulated traffic. For example, it is plausible that centralized control could be used to prevent different commodities from interfering with one another and thereby radically improve over the network performance achieved by selfish routing. Such problems would in turn complicate any explanation for the worst-case losses due to selfish routing. A central theme of this paper is that additional obstructions to selfish users optimizing the total latency *do not arise* in multicommodity flow networks, and that the worst-case inefficiency due to selfish routing can always be explained with the simplest of examples.

3. Upper bounding the price of anarchy

The goal of this section is to provide an upper bound on the worst-case ratio between the cost of a Nash flow and of an optimal flow, given a fixed but arbitrary class of allowable latency functions. We saw in Section 2.4 that this worst-case ratio depends crucially on how "nice" the class of allowable latency functions is, and one may therefore ask whether any meaningful upper bound is possible for an arbitrary class of latency functions. The answer is affirmative, provided that the upper bound is a *function of the class of allowable latency functions.*

To state the main result of this section precisely, denote by $\rho(G, r, \ell) \geqslant 1$ the ratio between the cost of a Nash and of an optimal flow for instance (G, r, ℓ); this ratio is well defined by Proposition 2.4. We will associate a real number $\alpha(\mathscr{L}) \geqslant 1$ to each class $\mathscr{L}$ of allowable edge latency functions that quantifies the "steepness" of the latency functions in $\mathscr{L}$, and will then prove that for any instance (G, r, ℓ) with latency functions in the class $\mathscr{L}$, $\rho(G, r, \ell) \leqslant \alpha(\mathscr{L})$. In Section 4 we will provide a matching lower bound by exhibiting, for any class $\mathscr{L}$, instances with latency functions in $\mathscr{L}$ and ρ-value arbitrarily close to $\alpha(\mathscr{L})$.

3.1. *Quantifying steepness with the anarchy value*

Our first task is to find a definition for a real number $\alpha(\mathscr{L})$ that captures how "nice" a class $\mathscr{L}$ of allowable latency functions is. Intuitively, we will define $\alpha(\mathscr{L})$ to be the worst-case ratio between the cost of a Nash and of an optimal flow in a Pigou-like example (see Section 2.4) with latency functions in $\mathscr{L}$. As a first step toward making this precise, we will consider a motivating example. It will be convenient to apply Proposition 2.7 to compute the optimal flow in this example; for this reason and others that will become clear later in this section, we will henceforth only consider networks with standard latency functions (see Definition 2.5).

Definition 3.1. A class $\mathscr{L}$ of latency functions is *standard* if it contains a nonzero function and each function $\ell \in \mathscr{L}$ is standard.

We now introduce the motivating example. Suppose we are given a standard class $\mathscr{L}$ of allowable latency functions, and wish to construct an example in which the Nash flow incurs much more latency than the optimal flow. A natural idea is to mimic the bad examples of Section 2.4 as best we can, given that $\mathscr{L}$ is the class of latency functions that we are allowed to work with. For simplicity, assume that $\mathscr{L}$ contains all of the constant functions and let ℓ_1 denote the function $\ell_1(x) = 1$. Then, we can consider the usual two-node, two-link network, assign the first link the latency function ℓ_1 and the second link the "steepest" latency function that we can find. More formally, suppose $\ell_2 \in \mathscr{L}$ is assigned to the second link where ℓ_2 satisfies $\ell_2(0) < 1$ and $\ell_2(x) > 1$ for x sufficiently large. Choosing $r > 0$ to satisfy $\ell_2(r) = 1$, we find that a Nash flow with traffic rate r routes all of its flow on the second edge for a total latency of r. Recalling the definition of a marginal cost function ℓ^* (Definition 2.6) and letting $\lambda \in [0,1]$ satisfy $\ell_2^*(\lambda r) = 1$, we find that an optimal flow routes λr units of flow on the second link and $(1-\lambda)r$ units of flow on the first link, for a total latency of $\lambda r \ell_2(\lambda r) + (1-\lambda)r$. Letting $\mu \in [0,1]$ denote $\ell_2(\lambda r)/\ell_2(r) = \ell_2(\lambda r)$, the ratio between the total latency of the Nash flow and of the optimal flow is $[\lambda\mu + (1-\lambda)]^{-1}$. Since this argument can be used with ℓ_1 replaced by any constant function, we arrive at the following expression for the worst-possible ρ-value arising in two-node, two-link networks with one link endowed with a constant function and the other endowed with the latency function ℓ.

Definition 3.2. Let ℓ be a nonzero standard latency function. The *anarchy value* $\alpha(\ell)$ of ℓ is

$$\alpha(\ell) = \sup_{r > 0 : \ell(r) > 0} [\lambda\mu + (1-\lambda)]^{-1}$$

where $\lambda \in [0,1]$ solves $\ell^*(\lambda r) = \ell(r)$ and $\mu = \ell(\lambda r)/\ell(r) \in [0,1]$.

The scalar $\lambda \in [0,1]$ exists because ℓ^* is continuous—recall the comments following Definition 2.6—and $\ell^*(0) = \ell(0) \leqslant \ell(r) \leqslant \ell^*(r)$. In most cases of interest λ will be uniquely determined by ℓ and r; otherwise, the assumption that ℓ is standard ensures that ℓ^* is nondecreasing and hence the anarchy value is well defined (i.e., that $[\lambda\mu + (1-\lambda)]^{-1}$ is independent of the choice of λ satisfying $\ell^*(\lambda r) = \ell(r)$).

The definition of the anarchy value of a latency function ℓ looks rather opaque from a mathematical perspective, but we emphasize that it is nothing more than the worst-case inefficiency of selfish routing in Pigou-like examples that make use of the latency function ℓ.

Since we are interested only in the most ill-behaved latency functions of a class, the next definition should be unsurprising.

Definition 3.3. The *anarchy value* $\alpha(\mathscr{L})$ of a standard class $\mathscr{L}$ of latency functions is

$$\alpha(\mathscr{L}) = \sup_{0 \neq \ell \in \mathscr{L}} \alpha(\ell).$$

Remark 3.4. (a) The anarchy value of a class lies in $[1, \infty]$ and need not be finite.

(b) The anarchy value may appear a fearsome expression to compute analytically, but we will see in Section 5 that it can typically be worked out in cases of practical interest.

We have already argued informally that if $\mathscr{L}$ is a standard class of latency functions containing the constant functions, then there are instances on a network with two nodes and two links and latency functions in $\mathscr{L}$ with ratio ρ arbitrarily close to the anarchy value $\alpha(\mathscr{L})$. On the other hand, there is no reason a priori to expect the anarchy value to have any connection to the ρ-value of instances defined on more general networks, or even to those defined in two-node networks with more than two parallel links). The central technical result of this paper is that, assuming only that the class $\mathscr{L}$ of allowable latency functions is standard, the anarchy value $\alpha(\mathscr{L})$ upper bounds the ratio $\rho(G, r, \ell)$ for *any* instance (G, r, ℓ) with latency functions in $\mathscr{L}$ (with an arbitrary network topology and an arbitrary number of commodities).

3.2. Proof approach

We next discuss our proof approach. At the highest level, the proof of the main theorem of this section is inspired by that of a theorem of Roughgarden and Tardos [26], which states: in an arbitrary network with linear latency functions (latency functions of the form $\ell(x) = ax + b$ for $a, b \geqslant 0$), the cost of a Nash flow is at most 4/3 times that of an optimal flow. The proof of this theorem has three steps, as follows. First, the characterizations of Nash and optimal flows (Propositions 2.2 and 2.7) are used to show that if f is a flow at Nash equilibrium for an instance (G, r, ℓ) with linear latency functions, then the scaled-down flow $f/2$ is optimal for the instance $(G, r/2, \ell)$. Second, the cost of $f/2$ is lower bounded in terms of the cost of f; this is not difficult since the scaled-down flow $f/2$ is a "significant fraction" of f. Finally, the cost of augmenting the flow $f/2$ to a flow optimal for (G, r, ℓ) is bounded below relative to the cost of f. This is the most difficult part of the proof; roughly, the argument leverages the connection between Nash and optimal flows given in Proposition 2.7 to show that the marginal cost of routing new flow with respect to $f/2$ is high, and thus augmenting the flow $f/2$ to a flow at the full set of traffic rates r is costly.

A direct attempt at adapting the three-step approach of [26] to networks with more general latency functions fails immediately. In networks with nonlinear latency functions (even networks with quadratic latency functions), there is no constant c for which a scaled-down version f/c of a Nash flow f is optimal for the reduced traffic rates r/c. Thus, it is not at all clear how to exploit our characterizations of Nash and optimal flows to relate their respective costs. To circumvent this problem, we view the proof approach of [26] in the following more general way: chop up an optimal flow into two "pieces" (in [26], $f/2$ and an augmentation from $f/2$ to a flow feasible for rates r) such that each piece can be lower-bounded in terms of the cost of a Nash flow. Guided by a desire to define the second piece of the optimal flow as an augmentation of the first and to lower bound its cost by means of marginal cost functions (as in [26]), we will define the first piece in a way that ensures that any augmentation with respect to it has large marginal cost. Unfortunately, this requires scaling down a Nash flow f by *different factors* on different edges, thereby producing an object which is *not* a flow—it is a more general object that need not obey conservation constraints, that we call a *pseudoflow*. This does not significantly complicate the lower bound for the cost of the scaled-down pseudoflow (it is a "significant fraction" of the Nash flow, as in [26]). However, a more careful analysis is now required to lower bound the cost of an augmentation from the scaled-down pseudoflow to a flow feasible for the original instance, as we are augmenting with respect to an object more complicated than simply a flow at reduced traffic rates.

3.3. *Proof of the upper bound*

We now turn toward making these ideas precise. We first define what we mean by a "scaled-down pseudoflow". The idea is to scale down the amount of Nash flow on a single edge until the value of the marginal cost function equals the original latency incurred by the Nash flow on that edge (this original latency is then our definition of "large marginal cost"). Formally, if f is a flow at Nash equilibrium, our scaled-down pseudoflow will be defined by $\{\lambda_e f_e\}_{e \in E}$ where λ_e satisfies $\ell_e^*(\lambda_e f_e) = \ell_e(f_e)$ (as in Definition 3.2). As discussed following Definition 3.2, these scaling factors always exist but need not be unique; our analysis must work with an arbitrary choice of scaling factors.

The next lemma formalizes the notion of "breaking up the optimal flow into two pieces". Again, the idea is to express the cost of the optimal flow as one term that is a scaled-down version of a Nash flow, and a second term that corresponds to an augmentation with respect to large marginal costs.

Lemma 3.5. *Let f^* and f be optimal and Nash flows, respectively, for instance (G, r, ℓ) with standard latency functions. For an edge e, let $\lambda_e \in [0, 1]$ solve $\ell_e^*(\lambda_e f_e) = \ell_e(f_e)$. Then,*

$$C(f^*) \geqslant \sum_e \left[\ell_e(\lambda_e f_e)\lambda_e f_e + (f_e^* - \lambda_e f_e)\ell_e(f_e)\right].$$

Proof. Since each edge latency function ℓ_e is standard, each marginal cost function ℓ_e^* is nondecreasing. For an edge e, we may thus write

$$\begin{aligned} \ell_e(f_e^*)f_e^* &= \ell_e(\lambda_e f_e)\lambda_e f_e + \int_{\lambda_e f_e}^{f_e^*} \ell_e^*(x)\,dx \\ &\geqslant \ell_e(\lambda_e f_e)\lambda_e f_e + (f_e^* - \lambda_e f_e)\ell_e^*(\lambda_e f_e) \\ &= \ell_e(\lambda_e f_e)\lambda_e f_e + (f_e^* - \lambda_e f_e)\ell_e(f_e) \end{aligned}$$

with the final equality following from the definition of λ_e. Summing over all edges proves the lemma. □

Neither the statement nor the proof of Lemma 3.5 assumes that the expression $f_e^* - \lambda_e f_e$ is nonnegative for all edges e; put differently, the augmentation from the pseudoflow defined by $\{\lambda_e f_e\}_{e \in E}$ to a flow f^* optimal for the original instance may increase *or decrease* the amount of flow on an edge.

To lower bound the right-hand side of Lemma 3.5, we require two more easy lemmas. The next lemma simply rephrases Definitions 3.2 and 3.3.

Lemma 3.6. *Let $\mathscr{L}$ be a standard class of latency functions with anarchy value $\alpha(\mathscr{L})$. For $\ell \in \mathscr{L}$ and $f > 0$, let $\lambda \in [0,1]$ solve $\ell^*(\lambda f) = \ell(f)$ and put $\mu = \ell(\lambda f)/\ell(f)$ (if $\ell(f) = 0$, put $\mu = 1$). Then $\lambda\mu + (1-\lambda) \geqslant \frac{1}{\alpha(\mathscr{L})}$.*

Our final lemma states that if f is a Nash flow for (G, r, ℓ), then f is a min-cost flow (in the classical sense of network flow theory [29]) with respect to the cost vector $\ell_e(f_e)_{e \in E}$. It is an easy consequence of the fact that, in a Nash flow f, all flow travels along minimum-latency paths (with latency measured with respect to f).

Lemma 3.7. *Let f be at Nash equilibrium and f^* feasible for instance (G, r, ℓ). Then,*

2.

$$\sum_e \ell_e(f_e) f_e \leqslant \sum_e \ell_e(f_e) f_e^*.$$

Proof. Let $L_i(f)$ denote the common latency of every s_i–t_i flow path of f, so that

$$\sum_{i=1}^{k} L_i(f) r_i = C(f) = \sum_e \ell_e(f_e) f_e$$

by Proposition 2.3. Since f is at Nash equilibrium, by Proposition 2.2 we have $\ell_P(f) \geqslant L_i(f)$ for every s_i–t_i path P. It follows that

$$\sum_e \ell_e(f_e) f_e^* = \sum_{i=1}^{k} \sum_{P \in \mathscr{P}_i} \ell_P(f) f_P^* \geqslant \sum_{i=1}^{k} L_i(f) r_i,$$

which proves the lemma. □

With all of the preliminaries now in place, we state and prove the main result of this section: the anarchy value of a standard class $\mathscr{L}$ of latency functions upper bounds the ratio ρ for any instance with latency functions in $\mathscr{L}$.

Theorem 3.8. *Let $\mathscr{L}$ be a standard class of latency functions with anarchy value $\alpha(\mathscr{L})$. Let (G, r, ℓ) denote an instance with latency functions drawn from $\mathscr{L}$. Then $\rho(G, r, \ell) \leqslant \alpha(\mathscr{L})$.*

Proof. Let f^* and f be optimal and Nash flows, respectively, for an instance (G, r, ℓ) with latency functions in the standard class $\mathscr{L}$. We begin by applying Lemma 3.5 to rewrite the cost $C(f^*)$ of the optimal flow in a form that is easier to relate to the cost $C(f)$ of the Nash flow:

$$\begin{aligned} C(f^*) &\geqslant \sum_e [\ell_e(\lambda_e f_e)\lambda_e f_e + (f_e^* - \lambda_e f_e)\ell_e(f_e)] \\ &= \sum_e [\mu_e \lambda_e f_e + (1 - \lambda_e)f_e + (f_e^* - f_e)]\ell_e(f_e) \\ &= \sum_e [\mu_e \lambda_e + (1 - \lambda_e)]\ell_e(f_e)f_e + \sum_e [f_e^* - f_e]\ell_e(f_e); \end{aligned}$$

following Definition 3.2, each scalar $\lambda_e \in [0, 1]$ is chosen (arbitrarily) to satisfy $\ell_e^*(\lambda_e f_e) = \ell_e(f_e)$, and $\mu_e = \ell_e(\lambda_e f_e)/\ell_e(f_e)$ (if $\ell_e(f_e) = 0$, put $\mu_e = 1$). In the second and third lines, we have rewritten the expression inherited from Lemma 3.5 so that the first sum enjoys a close connection with the anarchy value $\alpha(\mathscr{L})$, our desired upper bound for $\rho(G, r, \ell)$; the second sum can be regarded as an "error term". This error term is nonnegative by Lemma 3.7, so the inequality

$$C(f^*) \geqslant \sum_e [\mu_e \lambda_e + (1 - \lambda_e)]\ell_e(f_e)f_e$$

is valid. By Lemma 3.6, $\mu_e \lambda_e + (1 - \lambda_e) \geqslant 1/\alpha(\mathscr{L})$ for each edge e; thus, the quantities $[\mu_e \lambda_e + (1 - \lambda_e)]\ell_e(f_e)f_e$ and $\ell_e(f_e)f_e$ differ by at most an $\alpha(\mathscr{L})$ factor for each edge e. Summing over all edges, we find that the costs of f^* and f also differ by at most an $\alpha(\mathscr{L})$ factor:

$$C(f^*) \geqslant \sum_e \frac{\ell_e(f_e)f_e}{\alpha(\mathscr{L})} = \frac{C(f)}{\alpha(\mathscr{L})}.$$

The theorem is proved. □

4. Matching lower bounds in simple networks

With Theorem 3.8 in hand, it is now a relatively easy matter to prove the main results of the paper. In Section 4.1 we prove that, for a standard class of latency functions that contains the constant functions, the worst possible value of $\rho(G, r, \ell)$ for a multicommodity instance (G, r, ℓ) is realized (up to an arbitrarily small additive factor) by a single-commodity instance on a two-node, two-link network. In Section 4.2, we prove that under significantly weaker conditions on the class of allowable latency functions, the worst-case ρ-value is achieved (again, up to an arbitrarily small factor) by a single-commodity instance on a network of parallel links. We show in Section 4.3 that worst-case examples are simple for still broader classes of latency functions, including the queuing delay functions mentioned in Section 1.

4.1. Lower bounds in two-link networks

We begin by formalizing an argument of the previous section; the following lemma is essentially a restatement of Definitions 3.2 and 3.3.

Lemma 4.1. *Let G_2 denote the graph with one source vertex, one sink vertex, and two edges directed from source to sink. Let $\mathscr{L}$ denote a standard class of latency functions containing the constant functions, with anarchy value $\alpha(\mathscr{L})$. If $\mathscr{I}_2$ denotes the set of all single-commodity instances with underlying network G_2 and latency functions in $\mathscr{L}$, then*

$$\sup_{(G_2,r,\ell)\in\mathscr{I}_2} \rho(G_2,r,\ell) \geqslant \alpha(\mathscr{L}).$$

Combining Theorem 3.8 and Lemma 4.1, we find that the price of anarchy with respect to a class of latency functions containing the constant functions is essentially independent of the class of allowable network topologies, with the two-node, two-link networks of Fig. 2 always furnishing worst-case examples.

Theorem 4.2. *Let G_2 denote the graph with one source vertex, one sink vertex, and two edges directed from source to sink. Let $\mathscr{L}$ be a standard class of latency functions containing the constant functions. If $\mathscr{I}$ denotes the set of all instances with latency functions in $\mathscr{L}$ and $\mathscr{I}_2 \subseteq \mathscr{I}$ the single-commodity instances with underlying network G_2, then*

$$\sup_{(G_2,r,\ell)\in\mathscr{I}_2} \rho(G_2,r,\ell) = \alpha(\mathscr{L}) = \sup_{(G,r,\ell)\in\mathscr{I}} \rho(G,r,\ell).$$

4.2. Lower bounds in networks of parallel links

We now relax the assumption that the class of allowable latency functions contains all of the constant functions, and assume instead a much weaker condition that we call *diversity*.

Definition 4.3. A class $\mathscr{L}$ of latency functions is *diverse* if for each positive scalar $c>0$ there is a latency function $\ell \in \mathscr{L}$ satisfying $\ell(0) = c$.

For any class of latency functions that is closed under multiplication by positive scalars,[3] diversity merely asserts that some latency function is positive when evaluated at 0.

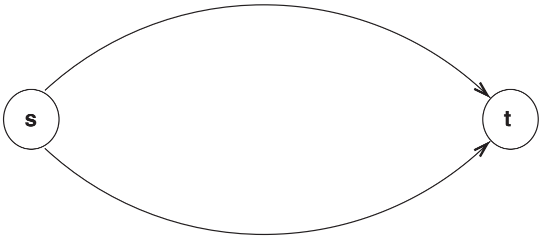

Fig. 2. Worst-case networks for standard classes of latency functions that include the constant functions.

We next show that networks of parallel links always provide worst-case examples of the inefficiency of selfish routing with respect to a standard diverse class of allowable latency functions.

Lemma 4.4. *Let G_m denote the graph with one source vertex, one sink vertex, and m edges directed from source to sink. Let $\mathscr{L}$ be a standard and diverse class of latency functions with anarchy value $\alpha(\mathscr{L})$. If $\mathscr{I}_m$ denotes the set of all single-commodity instances with underlying network G_m and latency functions in $\mathscr{L}$, then*

$$\sup_{(G,r,\ell)\in\cup_m \mathscr{I}_m} \rho(G,r,\ell) \geqslant \alpha(\mathscr{L}).$$

Proof. We will assume that $\alpha(\mathscr{L})$ is finite, and will leave the straightforward modifications necessary for the $\alpha(\mathscr{L}) = +\infty$ case to the interested reader.

Roughly speaking, the proof idea is to reduce the lemma to the previous case of a class of latency functions containing the constant functions. We compensate for the absence of a constant function $\ell(x) = c$ (which need not lie in $\mathscr{L}$) by "simulating" it with many parallel links endowed with latency functions satisfying $\ell(0) = c$.

For any $\varepsilon > 0$, choose a nonzero latency function $\ell_2 \in \mathscr{L}$, a positive number $r > 0$ with $\ell_2(r) > 0$, and a scalar $\lambda \in [0,1]$ satisfying $\ell_2^*(\lambda r) = \ell_2(r)$ so that $[\lambda\mu + (1-\lambda)]^{-1} \geqslant \alpha(\mathscr{L}) - \varepsilon/2$, where $\mu = \ell_2(\lambda r)/\ell_2(r)$. By the definition of the anarchy value, these parameter choices correspond to an instance in a two-node, two-link network with traffic rate r, latency functions $\ell_1(x) = \ell_2(r)$ and ℓ_2, and ρ-value at least $\alpha(\mathscr{L}) - \varepsilon/2$. This bad instance is not immediately useful to us because the constant latency function $\ell_1(x) = \ell_2(r)$ need not lie in $\mathscr{L}$. We will transform this bad instance into one with equally large ρ-value and latency functions in $\mathscr{L}$ by simulating the troublesome constant function ℓ_1 with parallel edges, all possessing a latency function $\ell \in \mathscr{L}$ satisfying $\ell(0) = \ell_2(r)$.

The class $\mathscr{L}$ is diverse, so there is a function $\ell \in \mathscr{L}$ with the property that $\ell(0) = \ell_2(r)$. Let m be so large that $\ell(\frac{(1-\lambda)r}{m-1}) \leqslant \ell_2(r) + \delta$ where δ is a sufficiently small positive number (depending on ε) to be chosen later; existence of the integer m follows from continuity of ℓ at 0. Define an instance on the network G_m of m parallel links with traffic rate r, latency function ℓ_2 on the last link, and latency function ℓ on the first $m-1$ links. The total latency incurred by the Nash flow is $\ell_2(r)r$ (all flow is routed on the last link). By our choice of m, the flow routing λr units of flow on the last link and $(1-\lambda)r/(m-1)$ units of flow on each of the first $m-1$ links has cost at most $\ell_2(r)r[\lambda\mu + (1-\lambda) + \frac{1-\lambda}{\ell_2(r)}\delta]$; choosing δ sufficiently small, we obtain an instance with ρ-value at least $\alpha(\mathscr{L}) - \varepsilon$. Since $\varepsilon > 0$ was arbitrary, the lemma follows. □

Theorem 3.8 and Lemma 4.4 together imply that, assuming only that the class of allowable latency functions is standard and diverse, the worst-case inefficiency of Nash flows occurs in networks of parallel links (see Fig. 3).

[3] Since a scalar multiplication of the latency functions can be effected merely by changing the units in which we measure latency, we expect many classes of interest to satisfy this property.

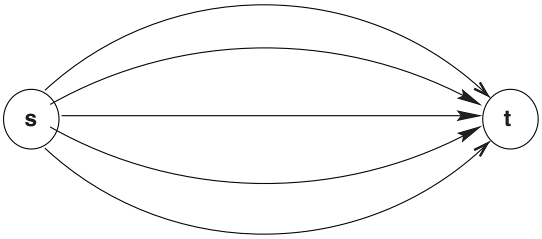

Fig. 3. Worst-case networks for standard diverse classes of latency functions. The number of parallel links can be arbitrarily large.

Theorem 4.5. *Let G_m denote the graph with one source vertex, one sink vertex, and m edges directed from source to sink. Let $\mathcal{L}$ be a standard and diverse class of latency functions. If $\mathcal{I}$ denotes the set of all instances with latency functions in $\mathcal{L}$ and $\mathcal{I}_m \subseteq \mathcal{I}$ the single-commodity instances with underlying network G_m, then*

$$\sup_{(G,r,\ell) \in \cup_m \mathcal{I}_m} \rho(G,r,\ell) = \alpha(\mathcal{L}) = \sup_{(G,r,\ell) \in \mathcal{I}} \rho(G,r,\ell).$$

Remark 4.6. The conclusion of the theorem is false with $\cup_m \mathcal{I}_m$ replaced by $\mathcal{I}_2$ (for a counterexample, take $\mathcal{L} = \{\ell(x) = a + x\text{: } a \geqslant 0\}$). The conclusion of the theorem is also false when the class of allowable latency functions need not be diverse (for a counterexample, take $\mathcal{L} = \{\ell(x) = 1 + x\}$).

4.3. Lower bounds in networks of disjoint paths

In this subsection we give our third and final result stating that worst-case examples for selfish routing are always simple. We will replace the hypothesis of diversity (Definition 4.3) with the still weaker condition that some available latency function is positive when evaluated with zero congestion. Our motivation is not generalization for its own sake; as we will see in Section 5, classes of latency functions common to networking applications need not be diverse. In addition to characterizing worst-possible network topologies for such function classes, the results of this subsection are essential for computing the price of anarchy with respect to these classes (a task we undertake in Section 5).

As in the previous subsection, we begin with a definition.

Definition 4.7. A class $\mathcal{L}$ of latency functions is *homogeneous* if $\ell(0) = 0$ for all $\ell \in \mathcal{L}$ and *inhomogeneous* otherwise.

It is clear that a diverse class of latency functions is inhomogeneous, but that the converse need not hold.

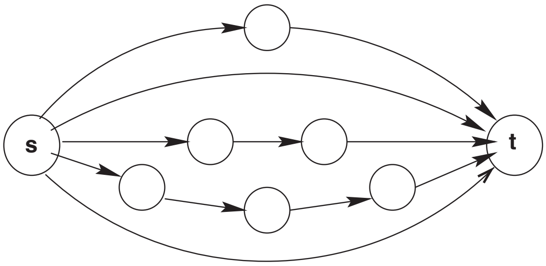

Fig. 4. Worst-case networks for standard inhomogeneous classes of latency functions. The number of paths and the number of edges in each path can be arbitrarily large.

Call a network a *union of paths* if it can be obtained from a network of parallel links by repeated edge subdivisions (see Fig. 4). We will show that unions of paths provide worst-case examples for the inefficiency of Nash flows with respect to a standard inhomogeneous class of latency functions.

Lemma 4.8. *Let $\mathscr{L}$ be a standard and inhomogeneous class of latency functions with anarchy value $\alpha(\mathscr{L})$. If $\mathscr{I}_u$ denotes the set of all single-commodity instances with underlying network a union of paths and latency functions in $\mathscr{L}$, then*

$$\sup_{(G,r,\ell)\in\mathscr{I}_u} \rho(G,r,\ell)\geqslant\alpha(\mathscr{L}).$$

Proof. Let $\mathscr{L}$ be a standard inhomogeneous class of allowable latency functions. We again assume for simplicity that $\alpha(\mathscr{L})$ is finite. As in the proof of Lemma 4.4, the idea is to work with a richer class of latency functions and then argue that any latency function in the class can be simulated with a collection of edges all possessing latency functions in $\mathscr{L}$.

Let $\overline{\mathscr{L}}$ denote the closure of $\mathscr{L}$ under multiplication by positive scalars, so that $\overline{\mathscr{L}} = \{\beta\ell\colon \ell\in\mathscr{L}, \beta>0\}$. Since $\mathscr{L}$ is standard and inhomogeneous, $\overline{\mathscr{L}}$ is standard and diverse. Moreover, since ℓ and $\beta\ell$ have equal anarchy value for any standard latency function ℓ and any $\beta>0$ (see Definition 3.2), $\alpha(\overline{\mathscr{L}}) = \alpha(\mathscr{L})$. For arbitrary $\varepsilon>0$, Lemma 4.4 then assures us of a single-commodity instance $(\bar{G},\bar{r},\bar{\ell})$ on a network $\bar{G}$ of parallel links with latency functions in $\overline{\mathscr{L}}$ and ρ-value at least $\alpha(\mathscr{L})-\varepsilon/2$. We next transform this instance into one on a union of paths with latency functions in $\mathscr{L}$ and ρ-value at least $\alpha(\mathscr{L})-\varepsilon$.

For each edge e of $\bar{G}$, write $\bar{\ell}_e = \beta_e\ell_e$ for $\beta_e>0$ and $\ell_e\in\mathscr{L}$. The ratio ρ is a continuous function of each scalar β_e (holding the network $\bar{G}$ and the traffic rate $\bar{r}$ fixed), so we may replace each β_e by a sufficiently close positive rational number γ_e to obtain a new instance with ρ-value at least $\alpha(\mathscr{L})-\varepsilon$. Clearing denominators, we may assume that each scalar γ_e is a positive integer (multiplying all latency functions of an instance by a common positive number does not affect its ρ-value).

The rest of the proof consists of observing that integral multiples of latency functions can be "simulated" with a path of edges, all possessing the original latency function. More precisely, define G by replacing each edge e of $\bar{G}$ by a directed path of γ_e new edges, each endowed with latency function ℓ_e. Since $\bar{G}$ is a network of parallel links, G is a union of paths. It is straightforward to check that the natural bijective correspondence between flows feasible for $(\bar{G}, \bar{r}, \gamma\ell)$ and flows feasible for $(G, \bar{r}, \ell)$ preserves both equilibria and total latency; therefore, $\rho(G, \bar{r}, \ell) \geqslant \alpha(\mathscr{L}) - \varepsilon$. Since $\varepsilon > 0$ is arbitrary, the lemma is proved. □

By Theorem 3.8 and Lemma 4.8, worst-case examples for an inhomogeneous standard class of allowable latency functions occur in networks that are unions of paths.

Theorem 4.9. *Let $\mathscr{L}$ be a standard and inhomogeneous class of latency functions, $\mathscr{I}$ the set of instances with latency functions in $\mathscr{L}$, and $\mathscr{I}_u \subseteq \mathscr{I}$ the single-commodity instances with underlying network a union of paths. Then*

$$\sup_{(G,r,\ell) \in \mathscr{I}_u} \rho(G, r, \ell) = \alpha(\mathscr{L}) = \sup_{(G,r,\ell) \in \mathscr{I}} \rho(G, r, \ell).$$

Remark 4.10. The conclusion of Theorem 4.9 fails if the set $\mathscr{I}_u$ of instances with underlying network a union of paths is replaced by the smaller set of instances defined on networks of parallel links. It also fails if the hypothesis of inhomogeneity is omitted, as a result of Roughgarden and Tardos [26, Corollary 4.2] implies that any instance (G, r, ℓ) with latency functions in the homogeneous class $\mathscr{L} = \{ax\colon a > 0\}$ satisfies $\rho(G, r, \ell) = 1 < 4/3 = \alpha(\mathscr{L})$. We do not know if the assumption that the function class is standard can be omitted. We leave open the problems of characterizing worst-case examples and computing the price of anarchy for classes of latency functions that fail to satisfy these two hypotheses.

5. Computing the price of anarchy

The past two sections have been devoted to finding simple examples that exhibit worst-possible losses due to selfish routing. One consequence of this work, Theorem 4.9, is that the price of anarchy (the largest ratio between the costs of the Nash and optimal flows) with respect to any standard inhomogeneous class $\mathscr{L}$ of allowable latency functions is nothing more than the anarchy value $\alpha(\mathscr{L})$ of Definition 3.3. This provides a general reduction from a combinatorial problem (finding a worst-case example among all possible multicommodity flow instances) to a much simpler analytical one (finding the "nastiest" latency function in a given class), which in turn permits the computation of the price of anarchy for many different function classes. In this section, we give three illustrative examples of such computations, and determine the price of anarchy for degree-bounded polynomials, delay functions of M/M/1 queues, and delay functions of M/G/1 queues. It will be obvious that other function classes can be treated in a similar way.

5.1. The price of anarchy for polynomial latency functions

For a positive integer p, let $\mathscr{L}_p$ denote the set of polynomials with nonnegative coefficients and degree at most p. As the first showcase for our machinery, we next compute the price of anarchy with respect to latency functions $\mathscr{L}_p$.

Proposition 5.1. *If $\mathscr{I}_p$ is the set of instances with latency functions in $\mathscr{L}_p$, then*

$$\sup_{(G,r,\ell)\in\mathscr{I}_p} \rho(G,r,\ell) = [1 - p\cdot(p+1)^{-(p+1)/p}]^{-1} = \Theta\left(\frac{p}{\ln p}\right).$$

Proof. Since $\mathscr{L}_p$ is standard and contains the constant functions, Theorem 4.2 implies that the price of anarchy is simply the anarchy value of $\mathscr{L}_p$. We claim that it suffices to compute the anarchy value of the smaller function class consisting of functions of $\mathscr{L}_p$ comprising only one term, namely $\widetilde{\mathscr{L}}_p \equiv \{ax^i\colon\ a\geqslant 0,\ \ i\in\{0,1,2,\dots,p\}\}$. This claim is valid because an instance (G,r,ℓ) with latency functions in $\mathscr{L}_p$ can be transformed into an equivalent instance with latency functions in $\widetilde{\mathscr{L}}_p$ by replacing an edge e of G with latency function $\ell_e(x) = \sum_{i=0}^{p} a_i x^i$ by a directed path of $p+1$ edges, with edge $i+1$ of the path possessing latency function $\tilde{\ell}_{e,i}(x) = a_i x^i$.[4]

We next compute the anarchy value $\alpha(\ell)$ of an arbitrary nonzero function $\ell(x) = ax^i$ of $\widetilde{\mathscr{L}}_p$ (recall Definition 3.2). If $i=0$ then $\alpha(\ell)=1$; otherwise, ℓ^* is strictly increasing and the scalar λ is uniquely determined by the choice of r. In this case, for $r>0$ we have $\lambda = (i+1)^{-1/i}$, hence $\mu = \lambda^i = (i+1)^{-1}$, hence $[\lambda\mu + (1-\lambda)]^{-1} = [(i+1)^{-(i+1)/i} + (1-(i+1)^{-1/i})]^{-1} = [1 - i\cdot(i+1)^{-(i+1)/i}]^{-1}$. Since this expression is independent of $r>0$, we obtain $\alpha(\ell) = [1 - i\cdot(i+1)^{-(i+1)/i}]^{-1}$. This expression is independent of a and is increasing in i on $[0,p]$ (as shown by a simple derivative test), so the functions of $\widetilde{\mathscr{L}}_p$ with largest anarchy value are those of the form ax^p for $a>0$; hence,

$$\sup_{(G,r,\ell)\in\mathscr{I}_p} \rho(G,r,\ell) = \alpha(\widetilde{\mathscr{L}}_p) = [1 - p\cdot(p+1)^{-(p+1)/p}]^{-1}. \qquad \square$$

Remark 5.2. A sharp lower bound on the left-hand side of Proposition 5.1 is provided by the bad examples of Section 2.4; the content of the proposition is that no worse example is possible, even in arbitrary multicommodity flow networks.

5.2. The price of anarchy for delay functions of M/M/1 queues

The latency function $\ell(x) = (u-x)^{-1}$ for $x<u$ arises as the (expected) delay function of an M/M/1 queue[5] with service rate (or capacity) u [8]. For this reason, such latency functions have been extensively studied in the networking literature [2,10,11,13,16]. These functions do not directly fit into our framework, since they are defined only on the set $[0,u)$, rather than on all of $[0,\infty)$. Nevertheless, only minor generalizations of our results are needed to compute the price of anarchy in this setting.

[4] This maneuver illustrates a general principle: if $\mathscr{L}$ is the cone generated by a (possibly infinite) class of latency functions S (i.e., $\mathscr{L}$ is the set of finite nonnegative linear combinations of functions in S), then the price of anarchy with respect to S and the price of anarchy with respect to $\mathscr{L}$ are equal.

[5] By M/M/1, we mean a single queue with Poisson arrivals and exponentially distributed service times [8].

We will fix two parameters, the largest allowable sum of all traffic rates $R_{\max}$ and the smallest allowable edge capacity $u_{\min}$. We will make the strong assumption that $R_{\max} < u_{\min}$; we will see that in the absence of this or similar assumptions, the price of anarchy is $+\infty$. Under this assumption, the restricted domains of the latency functions pose no difficulty; every feasible flow routes at most $R_{\max}$ units of flow on every edge and hence has a well-defined cost.

Let $\mathscr{L}$ denote the set of latency functions $\{\ell(x) = (u - x)^{-1}\colon u \geqslant u_{\min}\}$ and, for this example, redefine the anarchy value $\alpha(\ell)$ of a latency function ℓ to be $\alpha(\ell) = \sup_{r:0<r\leqslant R_{\max}} [\lambda\mu + (1-\lambda)]^{-1}$, where λ is the unique scalar satisfying $\ell^*(\lambda r) = \ell(r)$ and $\mu = \ell(\lambda r)/\ell(r)$. The key difference between this definition and the original definition of anarchy value (Definition 3.2) is that the range of traffic rates we consider is restricted to lie in $(0, R_{\max}]$ rather than $(0, \infty)$; this ensures that the equations defining λ and μ make sense.

Next, it is straightforward to check that Theorem 4.9 remains valid with our new definition of anarchy value, provided we only care about the worst-possible value of ρ achieved by instances whose sum of all traffic rates is at most $R_{\max}$. Since the class $\mathscr{L}$ satisfies both hypotheses of Theorem 4.9, computing the price of anarchy for instances with latency functions in $\mathscr{L}$ and sum of all traffic rates at most $R_{\max}$ reduces to computing the anarchy value of $\mathscr{L}$. Straightforward calculations, which can be found in [24], then yield the following proposition.

Proposition 5.3. *If $\mathscr{I}$ is the set of instances with latency functions in $\mathscr{L}$ and sum of all traffic rates at most $R_{\max} < u_{\min}$, then*

$$\sup_{(G,r,\ell)\in\mathscr{I}} \rho(G,r,\ell) = \frac{1}{2}\left(1 + \sqrt{\frac{u_{\min}}{u_{\min} - R_{\max}}}\right).$$

Remark 5.4. The class $\mathscr{L} = \{\ell(x) = (u-x)^{-1}\colon u \geqslant u_{\min}\}$ is *not* diverse, since $\ell(0) \leqslant \frac{1}{u_{\min}}$ for all $\ell \in \mathscr{L}$. Therefore, Theorems 4.2 and 4.5 are not sufficient to compute the price of anarchy in this application.

As foreshadowed above, the anarchy value of $\mathscr{L}$ and hence the worst possible value of ρ goes to $+\infty$ as $R_{\max} \to u_{\min}$. This fact was previously noted by Friedman [7]. We can therefore conclude that selfish routing is, in the worst case, quite costly in networks with M/M/1 delay functions.

On the bright side, Proposition 5.3 makes precise the intuition that selfish routing should not be costly in a lightly loaded network. It is also not hard to extend Proposition 5.3 to other classes of networks in which our draconian assumption that $R_{\max} < u_{\min}$ fails; roughly speaking, the price of anarchy will remain finite provided Nash flows are guaranteed to leave some bounded fraction of capacity unused on all edges of the network.

Finally, we mention a theorem of Roughgarden and Tardos [26] that suggests a simple design strategy for combating the inefficiency of selfish routing in congested networks with M/M/1 delay functions: doubling the capacity of all edges in the network reduces the total latency of selfish routing at least as much as routing traffic optimally.

5.3. The price of anarchy for delay functions of M/G/1 queues

As a final example, we extend the preceding analysis to queues that need not have exponentially distributed service times—that is, to M/G/1 delay functions (we retain our assumptions of a single queue and Poisson arrivals). Our solution will not be as clean as in the M/M/1 case, but will demonstrate that our techniques for computing the price of anarchy remain useful even for relatively complex classes of allowable latency functions.

Recall that if a queue service distribution (specifying the number of customers served in a time step) has finite expectation μ and finite standard deviation σ, then the expected waiting time with Poisson arrivals with rate $\lambda < \mu$ is

$$\frac{1}{\mu} + \frac{\lambda(1+\sigma^2\mu^2)}{2\mu(\mu-\lambda)};$$

see [8] or [9] for a derivation. To rephrase this formula in our usual notation, we view the parameter μ as the edge capacity u and the Poisson rate λ as the amount x of traffic assigned to the edge. As in the M/M/1 case, to achieve an interesting result we will need to assume a minimum allowable capacity $u_{\min}$ and a maximum allowable sum of all traffic rates $R_{\max} < u_{\min}$.

The anarchy value of such a function can be computed by the same method as for the M/M/1 case. After some (tedious) calculations and an application of Theorem 4.9, we obtain the following proposition.

Proposition 5.5. *Let $\mathscr{L}$ be a nonempty collection of $M/G/1$ delay functions with expected service rate at least $u_{\min}$. Then, the price of anarchy for instances with latency functions in $\mathscr{L}$ and sum of all traffic rates at most $R_{\max} < u_{\min}$ is precisely*

$$\sup_{\ell \in \mathscr{L}} \left(1 + \sqrt{\frac{u_\ell}{u_\ell - R_{\max}}}\right) \frac{2u_\ell + R_{\max}(\sigma_\ell^2 u_\ell^2 - 1)}{4u_\ell + (u_\ell + R_{\max} - \sqrt{u_\ell(u_\ell - R_{\max})})(\sigma_\ell^2 u_\ell^2 - 1)},$$

where u_ℓ and σ_ℓ denote the expectation and standard deviation of the service rate distribution associated with ℓ.

In the absence of additional assumptions on the class $\mathscr{L}$, we cannot simplify the expression of Proposition 5.5 further; this reflects the relative complexity of M/G/1 delay functions, which are specified by two independent parameters u_ℓ and σ_ℓ (unlike the simpler M/M/1 case). On the other hand, reducing the computation of the price of anarchy to computing the expression of Proposition 5.5 is both nontrivial and useful. When the class $\mathscr{L}$ possesses structure beyond merely being some collection of M/G/1 delay functions, the expression of Proposition 5.5 may become simple and transparent (as in the special case of M/M/1 delay functions, where $\sigma_\ell u_\ell = 1$ for all ℓ). Even for classes for which no analytical simplification is possible, Proposition 5.5 should permit the approximate (if not exact) computation of the price of anarchy with respect to $\mathscr{L}$ by straightforward numerical methods. In the simplest case where $\mathscr{L}$ is finite and not astronomically large—and we suspect almost all classes of M/G/1 delay functions can be closely approximated by such an $\mathscr{L}$—the price of anarchy can be computed simply by enumeration.

Table 2
Summary of main results, as stated in Theorems 4.2, 4.5, and 4.9

Class of allowable latency functions	Worst-case examples	Price of anarchy
Standard, includes constant functions	two-node, two-link networks	$\alpha(\mathscr{L})$
Standard, diverse	networks of parallel links	$\alpha(\mathscr{L})$
Standard, inhomogeneous	unions of paths	$\alpha(\mathscr{L})$

The expression $\alpha(\mathscr{L})$ denotes the anarchy value of a class $\mathscr{L}$ of latency functions, see Definition 3.3. A class $\mathscr{L}$ is standard if $x \cdot \ell(x)$ is convex for all $\ell \in \mathscr{L}$, is diverse if $\{\ell(0)\colon \ell \in \mathscr{L}\} \supseteq (0, \infty)$, and is inhomogeneous if $\{\ell(0)\colon \ell \in \mathscr{L}\} \neq \{0\}$.

We emphasize that without the assurance provided by our previous work that simple network topologies always provide worst-case examples, such an enumerative approach to computing the price of anarchy would be unthinkable.

6. Conclusion

We have studied the worst-possible degradation in network performance due to selfish routing, as quantified by the *price of anarchy*, the ratio of the total latency incurred by a flow at Nash equilibrium and by a minimum-latency flow. The price of anarchy is unbounded unless the class of allowable edge latency functions is restricted, a fact that motivated our investigation of the price of anarchy with respect to a fixed but arbitrary class of latency functions. Our work has two related aspects: we have shown that worst-case examples for selfish routing always occur in simple network topologies, and we have used this fact to give a general technique for computing the price of anarchy with respect to an arbitrary class of latency functions. While the precise description of worst-case examples depends on the hypotheses placed on the class $\mathscr{L}$ of allowable latency functions (see Table 2 for a summary), the price of anarchy is always, under the very weak assumptions that $\mathscr{L}$ is standard and inhomogeneous, the anarchy value $\alpha(\mathscr{L})$ of Definition 3.3.

In addition to characterizing worst-case examples and giving methods to compute the price of anarchy, our work demonstrates that the worst-possible consequences of selfish routing always have a simple, transparent explanation: flows at Nash equilibrium are inefficient because selfish users cannot resist overcongesting a route that is beneficial when used in moderation. We saw in Section 2.4 that this phenomenon occurs even in two-node, two-link networks, and is an obvious obstruction to selfish users achieving global optimality; our results imply that additional complications do not arise in arbitrary multicommodity flows networks.[6]

As discussed in Section 1, our central theorem stating that worst-case examples always occur in the simplest of networks is provably false in several related models of selfish routing, including models with other objective functions and models in which network users carry more than a negligible amount of traffic. It is therefore natural to ask: what are the properties of a selfish

[6] While worst-case examples occur in networks somewhat more complicated than those with two nodes and two links when edges cannot possess constant latency functions, the qualitative explanation for the inefficiency of selfish routing in such examples remains the same.

routing model that permit such a ''succinct explanation'' of the worst-case consequences of selfish behavior? Are there other natural classes of noncooperative games in which simple games always provide worst-case examples? We expect future work on these questions to guide us toward a deeper understanding of selfish behavior in many game-theoretic contexts.

Acknowledgments

We thank Amir Ronen for substantially simplifying our original proof of Theorem 3.8, Éva Tardos for helpful discussions and for suggestions on an earlier draft of this paper, and the anonymous referees for their comments.

References

[1] M. Beckmann, C.B. McGuire, C.B. Winsten, Studies in the Economics of Transportation, Yale University Press, 1956.

[2] D.P. Bertsekas, R. Gallager, Data Networks, 2nd Edition, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[3] A. Czumaj, P. Krysta, B. Vöcking, Selfish traffic allocation for server farms, in: Proceedings of the 34th Annual ACM Symposium on the Theory of Computing, Montreal, Canada, 2002, pp. 287–296.

[4] A. Czumaj, B. Vöcking, Tight bounds for worst-case equilibria, in: Proceedings of the 13th Annual Symposium on Discrete Algorithms, San Francisco, CA, 2002, pp. 413–420.

[5] S.C. Dafermos, F.T. Sparrow, The traffic assignment problem for a general network, J. Res. Nat. Bur. Standards Ser. B 73B (2) (1969) 91–118.

[6] P. Dubey, Inefficiency of Nash equilibria, Math. Oper. Res. 11 (1) (1986) 1–8.

[7] E.J. Friedman, A generic analysis of selfish routing, Working paper, Cornell University, 2001.

[8] D. Gross, C.M. Harris, Queuing Theory, 3rd Edition, Wiley, New York, 1998.

[9] S. Karlin, H.M. Taylor, A Second Course in Stochastic Processes, Academic Press, New York, 1981.

[10] Y.A. Korilis, A.A. Lazar, A. Orda, Capacity allocation under noncooperative routing, IEEE Trans. Automat. Control 42 (3) (1997) 309–325.

[11] Y.A. Korilis, A.A. Lazar, A. Orda, Avoiding the Braess paradox in noncooperative networks, J. Appl. Probab. 36 (1) (1999) 211–222.

[12] E. Koutsoupias, C. Papadimitriou, Worst-case equilibria, in: Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science, Trier, Germany, 1999, pp. 404–413.

[13] A.A. Lazar, A. Orda, D.E. Pendarakis, Virtual path bandwidth allocation in multiuser networks, IEEE/ACM Trans. Networking 5 (1997) 861–871.

[14] L. Libman, A. Orda, Atomic resource sharing in noncooperative networks, Telecomm. Systems 17 (4) (2001) 385–409.

[15] M. Mavronicolas, P. Spirakis, The price of selfish routing, in: Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing, Hersonissos, Crete, Greece, 2001, pp. 510–519.

[16] A. Orda, R. Rom, N. Shimkin, Competitive routing in multi-user communication networks, IEEE/ACM Trans. Networking 1 (1993) 510–521.

[17] G. Owen, Game Theory, 3rd Edition, Academic Press, New York, 1995.

[18] C. Papadimitriou, Algorithms, games, and the Internet, in: Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing, Hersonissos, Crete, Greece, 2001, pp. 749–753.

[19] A.C. Pigou, The Economics of Welfare, Macmillan, New York, 1920.

[20] A. Rapoport, A.M. Chammah, Prisoner's Dilemma, University of Michigan Press, Michigan, 1965.

[21] T. Roughgarden, The price of anarchy for the maximum latency of selfish routing, unpublished.

[22] T. Roughgarden, Designing networks for selfish users is hard, in: Proceedings of the 42nd Annual Symposium on Foundations of Computer Science, Las Vegas, NV, 2001, pp. 472–481.

[23] T. Roughgarden, Stackelberg scheduling strategies, in: Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing, Hersonissos, Crete, Greece, 2001, pp. 104–113.

[24] T. Roughgarden, Selfish Routing, Ph.D. Thesis, Cornell University, 2002.

[25] T. Roughgarden, É. Tardos, Bounding the inefficiency of equilibria in nonatomic congestion games, Technical Report TR2002-1866, Cornell University, 2002.

[26] T. Roughgarden, É. Tardos, How bad is selfish routing?, J. ACM 49 (2) (2002) 236–259.

[27] W. Rudin, Principles of Mathematical Analysis, 3rd Edition, McGraw-Hill, New York, 1976.

[28] A.S. Schulz, N. Stier Moses, Performance of user equilibria in traffic networks, in: Proceedings of the 14th Annual Symposium on Discrete Algorithms, Baltimore, MD, 2003, pp. 86–87.

[29] R.E. Tarjan, Data Structures and Network Algorithms, SIAM, Philadelphia, PA, 1983.

[30] A. Vetta, Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions, in: Proceedings of the 43rd Annual Symposium on Foundations of Computer Science, Vancouver, Canada, 2002, pp. 416–425.

[31] J.G. Wardrop, Some theoretical aspects of road traffic research, in: Proceedings of the Institute of Civil Engineers, London, Part II, Vol. 1, 1952, pp. 325–378.