How Bad is Selfish Routing?*

Tim Roughgarden† Éva Tardos‡

December 5, 2001

Abstract

We consider the problem of routing traffic to optimize the performance of a congested network. We are given a network, a rate of traffic between each pair of nodes, and a latency function for each edge specifying the time needed to traverse the edge given its congestion; the objective is to route traffic such that the sum of all travel times—the total latency—is minimized.

In many settings, it may be expensive or impossible to regulate network traffic so as to implement an optimal assignment of routes. In the absence of regulation by some central authority, we assume that each network user routes its traffic on the minimum-latency path available to it, given the network congestion caused by the other users. In general such a "selfishly motivated" assignment of traffic to paths will not minimize the total latency; hence, this lack of regulation carries the cost of decreased network performance.

In this paper we quantify the degradation in network performance due to unregulated traffic. We prove that if the latency of each edge is a linear function of its congestion, then the total latency of the routes chosen by selfish network users is at most $\frac{4}{3}$ times the minimum possible total latency (subject to the condition that all traffic must be routed). We also consider the more general setting in which edge latency functions are assumed only to be continuous and nondecreasing in the edge congestion. Here, the total latency of the routes chosen by unregulated selfish network users may be arbitrarily larger than the minimum possible total latency; however, we prove that it is no more than the total latency incurred by optimally routing *twice* as much traffic.

*A preliminary version of this paper appeared in the Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science, November 2000.

†Department of Computer Science, Cornell University, Ithaca NY 14853. Research done in part while visiting the Computer Science Division, UC Berkeley, Berkeley, CA 94720. Supported by ONR grant N00014-98-1-0589. Email: `timr@cs.cornell.edu`.

‡Department of Computer Science, Cornell University, Ithaca NY 14853. Research done in part while a visiting Miller Professor at the Computer Science Division, UC Berkeley, Berkeley, CA 94720. Partially supported by a Guggenheim Fellowship, NSF grant CCR-9700163 and ONR grant N00014-98-1-0589. Email: `eva@cs.cornell.edu`.

1 Introduction

A fundamental problem arising in the management of large-scale traffic and communication networks is that of routing traffic to optimize network performance. One problem of this type is the following: given the rate of traffic between each pair of nodes in a network, find an assignment of traffic to paths so that the sum of all travel times (the *total latency*) is minimized. A difficult aspect of this problem is that the amount of time needed to traverse a single link of a network is typically *load-dependent*, that is, the common latency suffered by all traffic on the link increases as the link becomes more congested.

It is often difficult or even impossible to impose optimal or near-optimal routing strategies on the traffic in a network; in these settings, network users are free to act according to their own interests, without regard to overall network performance. The central question of this paper is *how much does network performance suffer from this lack of regulation?*

As a first step toward formalizing this question mathematically, we assume that, in the absence of network regulation, users act in a purely selfish (but not malicious) manner. Under this assumption, we can view network users as independent agents participating in a *noncooperative game* and expect the routes chosen by users to form a *Nash equilibrium* in the sense of classical game theory [36]. In other words, we assume that each agent uses the minimum-latency path from its source to its destination, given the link congestion caused by the rest of the network users.[1] It is well known that Nash equilibria do not in general optimize social welfare; perhaps the most famous example is that of "The Prisoner's Dilemma" [14, 36]. We are then interested in comparing the total latency of a Nash equilibrium with that of the optimal assignment of traffic to paths.

Our approach to quantifying the inefficiency inherent in a selfishly defined solution (dubbed the "price of anarchy" by Papadimitriou [37]) is inspired by recent work of Koutsoupias and Papadimitriou [28]. In [28], network routing was modeled as a noncooperative game (though with a different model than ours, and only for two-node networks) and the worst-case ratio of the social welfare achieved by a Nash equilibrium and by a socially optimal set of strategies was proposed as a measure of the performance degradation caused by a lack of regulation. As articulated in [28], we study the cost of the lack of *coordination* inherent in a noncooperative game, as opposed to the cost of a lack of *unbounded computing power* (studied via approximation algorithms) or the cost of a lack of *complete information* (studied via on-line algorithms).

For most of the paper we assume that each agent controls a negligible fraction of the overall traffic. For example, each agent could represent a car in a highway system, or a packet in a communication network; an equilibrium then represents a steady-state in the system (perhaps best achieved in a road network by daily commuters during rush hour and in a communication network by persistent or long-running applications). Under this assumption, a feasible assignment of traffic to paths in the network can be modeled as *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 in the aforementioned noncooperative game

[1] Friedman [19] has pointed out that this equilibrium arises not only in settings where network users can select their own paths, but also in network protocols that determine routes for users via a shortest path computation (provided link latency is used as the distance metric).

then corresponds to a flow where all flow paths between a given source and destination have equal (and smallest possible) latency—if a flow does not have this property, some agent can improve its travel time by switching from a longer flow path to a shorter one.

Beckman et al. [3] showed that if the latency of each network link is a continuous nondecreasing function of the flow on the link, then a flow corresponding to a Nash equilibrium always exists and moreover all such flows have the same total latency. Thus, we can study the cost of routing selfishly via the following question: among all networks with continuous, nondecreasing link latency functions, what is the worst-case ratio between the total latency of a flow at Nash equilibrium and that of an optimal flow (i.e., a flow minimizing the total latency)?

Our Results

In networks in which the latency of each edge is a linear function of the edge congestion (a model that has been the focus of several previous papers [7, 18, 51]), we show that a flow at Nash equilibrium has total latency at most $\frac{4}{3}$ times that of the optimal flow. We give examples showing that this result is tight.

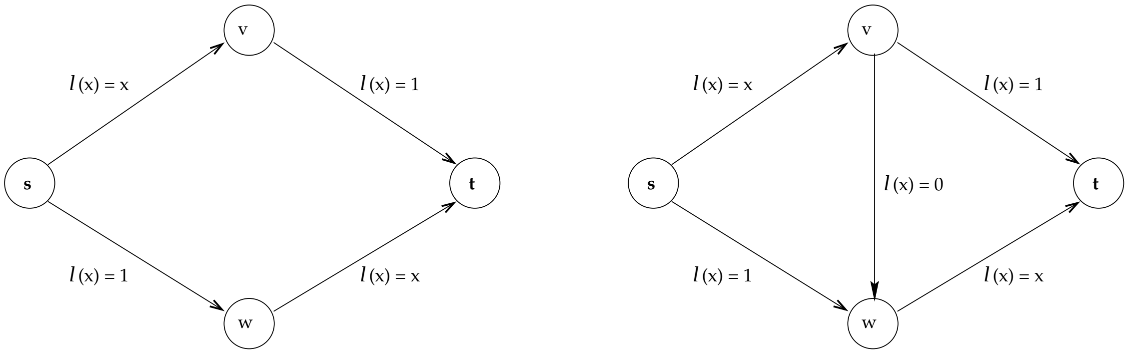
We also consider the model in which link latency functions are assumed only to be continuous and nondecreasing. We first show that the ratio between the total latency of a flow at Nash equilibrium and that of an optimal flow may be unbounded in this model. We then work toward *bicriteria* results; in particular, we compare the total latency of a flow at Nash equilibrium with that of an optimal flow that routes *additional traffic* between each pair of nodes.[2] Our main result in this setting is that for any network with continuous nondecreasing latency functions, 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. We again give an example showing that our analysis is tight.

Finally, we examine two unrealistic assumptions made in the basic model: first, the assumption that agents can evaluate the latency of a path with arbitrary precision, and second, that there is an infinite number of agents each controlling a negligible fraction of the overall traffic. We define extensions to the basic model and use them to analyze the sensitivity of our results to these assumptions.

Related Work

Unregulated traffic has been modeled as network flow with all flow paths between a given source-destination pair having equal latency since the 1950's [3, 52] (see also Knight [24]). Beckman et al. [3], observing that such an equilibrium flow is an optimal solution to a related convex program (see also Section 2), gave existence and uniqueness results for traffic equilibria. Dafermos and Sparrow [13] were perhaps the first authors interested in computing the equilibrium efficiently, and many subsequent papers gave increasingly efficient methods for computing equilibria (see [16, 17] for a survey). Since these early works, many properties

[2]This approach is thus in the spirit of the analyses of online algorithms via *resource augmentation* given by Kalyanasundaram and Pruhs [23] and Phillips et al. [39].



(a) Before

(b) After

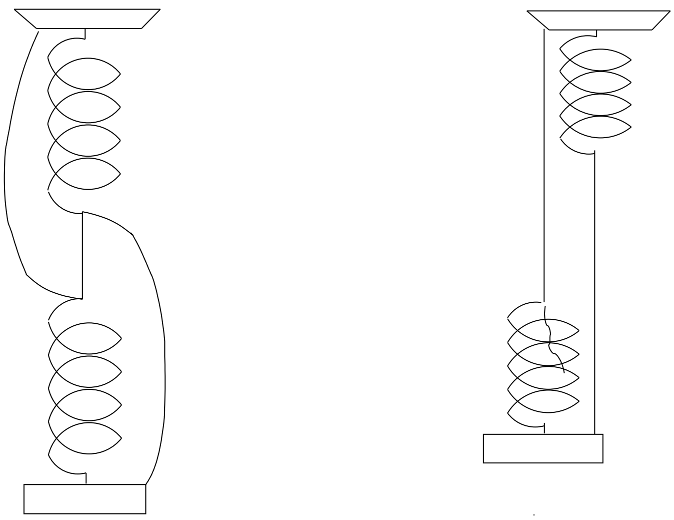
Figure 1: Braess's Paradox. The addition of an intuitively helpful link can negatively impact all of the users of a congested network.

of and extensions to this traffic model have been studied (see for example [1, 11, 16, 21, 32, 33, 34, 35, 41, 43, 47, 49]).

In the past several decades, much of the work on this traffic model has been inspired by a "paradox" first discovered by Braess [6] and later reported by Murchland [31] (see also [2] for a non-technical account). The essence of Braess's Paradox is captured by the example shown in Figure 1, where the edges are labeled with their latency functions (each a function of the link congestion x). Suppose one unit of traffic needs to be routed from s to t in the first network of Figure 1. In the unique flow at Nash equilibrium, which coincides with the optimal flow, half of the traffic takes the upper path and the other half travels along the lower path, and thus all agents are routed on a path of latency $\frac{3}{2}$. Next suppose a fifth edge of latency 0 (independent of the congestion) is added to the network, with the result shown in Figure 1(b). The optimal flow is unaffected by this augmentation (there is no way to use the new link to decrease the total latency) while in the new (unique) flow at Nash equilibrium, all traffic follows path $s \to v \to w \to t$; here, the latency experienced by each individual agent is 2. Thus, the intuitively helpful (or at least innocuous) action of adding a new zero-latency link may negatively impact *all* of the agents!

Motivated by the discovery of Braess's Paradox and evidence of similarly counterintuitive and counterproductive traffic behavior following the construction of new roads in congested cities [25, 31], researchers attempted to classify networks in which the addition of a single link could degrade network performance [18, 51], discovered new types of "paradoxes" [12, 15, 20, 48, 50], and proved that detecting Braess's Paradox (even in its worst-possible manifestation) is algorithmically difficult [42]. In a related model with finitely many agents, each controlling a strictly positive amount of flow, Korilis et al. [26, 27] studied strategies for adding new edges and/or capacity to a network that guarantee an improvement in network performance.

In contrast to this previous work, we are interested in quantifying the difference in social welfare between equilibrium and optimal traffic flows. To the best of our knowledge, the only previous work with this goal is the paper of Koutsoupias and Papadimitriou [28]; however,



(a) Before (b) After

Figure 2: Strings and springs. Severing a taut string results in the rise of a heavy weight.

the model of this paper is quite different from ours. In [28], a finite number of users share a collection of parallel links, and each user chooses a distribution on the set of links (specifying the probability that the agent will route all of its flow on a given link). Each agent wishes to minimize the expected congestion it will experience, while the global objective is to minimize the expected load on the most congested edge. Different Nash equilibria may have different objective function values in the model of [28], so the *worst-case* Nash equilibrium is compared to a globally optimal choice of distributions. Koutsoupias and Papadimitriou obtain a tight analysis of this worst-case ratio in two-node, two-link networks and partial results for two-node networks with three or more parallel links; tight results have recently been obtained for parallel networks with any number of links by Mavronicolas and Spirakis [30] (for a special case) and by Czumaj and Vöcking [10] (for the general case).

Equilibria in Other Settings

Braess's Paradox is not particular to traffic in networks; perhaps the most compelling analogue occurs in a mechanical network of strings and springs, constructed by Cohen and Horowitz [7] and shown in Figure 2. In this device, one end of a spring is attached to a fixed support, and the other end to a string. A second identical spring is hung from the free end of the string and carries a heavy weight. Finally, strings are connected (with some slack) from the support to the upper end of the second spring and from the lower end of the first spring to the weight. Assuming that the springs are ideally elastic, the stretched length of a spring is a linear function of the force applied to it. We may thus view the network of strings and springs as a traffic network, where forces correspond to flows and physical distance corresponds to latency. With a suitable choice of string and spring lengths and spring constants, the equilibrium position of this mechanical network is described by Figure 2(a). Contrary to intuition, severing the taut string causes the weight to rise, as shown in Figure 2(b); this corresponds to deleting the zero-latency arc of Figure 1(b), thereby obtaining the network of Figure 1(a) with its improved Nash equilibrium.

Our result for traffic equilibria in networks with linear latency functions provides a quantitative limit on the extent to which this phenomenon can occur. In particular, we show that our result implies that for any system of strings and springs carrying a single weight, the distance between the support and the weight after severing an arbitrary collection of strings and springs is at least $\frac{3}{4}$ times the original support-weight distance.

Further examples of analogous phenomena have been exhibited in two-terminal electrical networks [7] (where our results give analogous bounds on the largest possible increase in conductivity obtainable by removing conducting links) and queuing networks [8].

Organization

In Section 2 we give a formal definition of our network model and of flows at Nash equilibrium, and state several lemmas needed for our main results. In Section 3 we prove our main bicriteria result for networks with arbitrary edge latency functions. In Section 4 we prove a stronger and technically more involved result for networks with linear edge latency functions. Section 5 considers several extensions to the basic model, and Section 6 concludes with a discussion of recent work.

2 Preliminaries

In this section we present the basic definitions and preliminary results needed in the rest of the paper. Subsections 2.1 and 2.2 describe the basic model and flows at Nash equilibrium, and are prerequisites for all that follows. Subsection 2.3 gives a characterization of minimum-latency flows that is crucial for our result for networks with linear latency functions (Section 4), but unnecessary for our bicriteria result for networks with arbitrary latency functions (Section 3). In Subsection 2.4 we prove the existence and essential uniqueness of flows at Nash equilibrium and in Subsection 2.5 we observe that this proof gives a good (but not optimal) upper bound on the cost of selfish routing for networks with sufficiently well-behaved edge latency functions. Except for the fact that Nash flows exist and are unique, no results or techniques from these two subsections will be needed in the rest of the paper.

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 denote the set of (simple) s_i-t_i paths by $\mathcal{P}_i$, and define $\mathcal{P} = \cup_i \mathcal{P}_i$. A *flow* is a function $f : \mathcal{P} \to \mathcal{R}^+$; for a fixed flow f we define $f_e = \sum_{P:e\in P} f_P$. 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, $\sum_{P\in\mathcal{P}_i} f_P = r_i$. Finally, each edge $e \in E$ is given a load-dependent *latency function* that we denote by $\ell_e(\cdot)$. For each $e \in E$, we assume that the latency function ℓ_e is nonnegative, differentiable[3], and nondecreasing. We will call the triple (G, r, ℓ) an *instance*. 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 define the *cost* $C(f)$ of a flow f in G as the total latency incurred by f, that is,

$$C(f) = \sum_{P \in \mathcal{P}} \ell_P(f) f_P.$$

By summing over the edges in a path P and reversing the order of summation, we may also write $C(f) = \sum_{e \in E} \ell_e(f_e) f_e$.

2.2 Flows at Nash Equilibrium

We wish to consider flows that represent an equilibrium among many noncooperative agents—i.e., flows that behave "greedily" or "selfishly", without regard to the overall cost. Intuitively, we expect each unit of such a flow (no matter how small) to travel along the minimum-latency path available to it, where latency is measured with respect to the rest of the flow; otherwise, this flow would reroute itself on a path with smaller latency. We formalize this idea in the next definition.

Definition 2.1 *A flow f feasible for instance (G, r, ℓ) is at* Nash equilibrium *if for all $i \in \{1, \ldots, k\}$, $P_1, P_2 \in \mathcal{P}_i$, and $\delta \in [0, f_{P_1}]$, we have $\ell_{P_1}(f) \le \ell_{P_2}(\tilde{f})$, 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, occasionally called a Wardrop equilibrium [22] or Wardrop's Principle [50, 51] in the literature, due to an influential paper of Wardrop [52].

Lemma 2.2 *A flow f feasible for instance (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 $f_{P_1} > 0$, $\ell_{P_1}(f) \le \ell_{P_2}(f)$.*

In particular, if f is at Nash equilibrium then all s_i-t_i flow paths (i.e., s_i-t_i paths to which f assigns a positive amount of flow) have equal latency, say $L_i(f)$. We can thus express the cost $C(f)$ of a flow f at Nash equilibrium in a particularly nice form.

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

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

[3] We make this assumption for simplicity only. All of our results hold, with minor modifications to the proofs, under the weaker assumption of continuity.

Remark. Our definition of a flow at Nash equilibrium corresponds to an equilibrium in which each agent chooses a single path of the network (a *pure strategy*), whereas in classical game theory a Nash equilibrium is typically defined via *mixed strategies* (in which an agent may choose a probability distribution over pure strategies) [36]. However, since in our model each agent carries a negligible fraction of the overall traffic, these two definitions are essentially equivalent (see [22] for a rigorous discussion).

2.3 Characterizing Optimal Flows via Convex Programming

We now investigate the properties of an optimal flow—that is, a flow that minimizes total latency. Recalling that the cost of a flow f may be expressed $C(f) = \sum_{e \in E} \ell_e(f_e) f_e$, observe that the problem of finding the minimum-latency feasible flow in a network is a special case of the following non-linear program

$$
(NLP) \qquad
\begin{array}{lll}
\text{Min} & \displaystyle\sum_{e \in E} c_e(f_e) & \\
\text{subject to:} & & \\
& \displaystyle\sum_{P \in \mathcal{P}_i} f_P = r_i & \forall i \in \{1, \ldots, k\} \\
& f_e = \displaystyle\sum_{P \in \mathcal{P} : e \in P} f_P & \forall e \in E \\
& f_P \geq 0 & \forall P \in \mathcal{P}
\end{array}
$$

where in our problem, $c_e(f_e) = \ell_e(f_e) f_e$.

For simplicity we have given a formulation with an exponential number of variables, but it is not difficult to give an equivalent compact formulation (with decision variables only on edges and explicit conservation constraints) that requires only polynomially many variables and constraints.

Next, we characterize the local optima of (NLP). Intuitively, we expect a flow to be locally optimal if and only if moving flow from one path to another can only increase the flow's cost. Put differently, we expect a flow to be locally optimal when the marginal benefit of decreasing flow along any s_i-t_i flow path is at most the marginal cost of increasing flow along any other s_i-t_i path. Since the local and global minima of a convex function on a convex set coincide (see, e.g., [38, Thm 2.3.4]), this condition should be necessary and sufficient for a flow to be *globally* optimal whenever the objective function of (NLP) is convex (as is the case when for each edge $e \in E$ we have $c_e(f_e) = \ell_e(f_e) f_e$ with a convex latency function ℓ_e).

We formalize this characterization of global optima in convex programs of the form (NLP) in the next lemma. Let c'_e denote the derivative $\frac{d}{dx} c_e(x)$ of c_e and define $c'_P(f)$ by $c'_P(f) = \sum_{e \in P} c'_e(f_e)$. We then have the following.[4]

Lemma 2.4 ([3, 13]) *A flow f is optimal for a convex program of the form (NLP) if and only if for every $i \in \{1, \ldots, k\}$ and $P_1, P_2 \in \mathcal{P}_i$ with $f_{P_1} > 0$, $c'_{P_1}(f) \leq c'_{P_2}(f)$.*

[4]For a formal derivation via the Karush-Kuhn-Tucker Theorem [38], see [3, 13].

Remark: To see that this characterization still makes sense under the weaker assumption of continuous (and not necessarily differentiable) latency functions, define $c_e^-(x), c_e^+(x)$ to be the left and right derivatives of c_e at x, respectively; since c_e is assumed to be convex, c_e^+ and c_e^- exist everywhere. Define $c_P^+(f)$ and $c_P^-(f)$ for a path P in the obvious way. Then, the proof of Lemma 2.4 can easily be extended to show that a flow f is optimal in this more general setting if and only if for any P_1, P_2 as above, $c_{P_1}^-(f) \leq c_{P_2}^+(f)$.

The striking similarity between the characterizations of optimal solutions to a convex program of the form (NLP) and of flows at Nash equilibrium was noticed early on by Beckman et al. [3], and provides an interpretation of an optimal flow as a flow at Nash equilibrium *with respect to a different set of edge latency functions.* To make this relationship precise, denote the marginal cost of increasing flow on edge e by $\ell_e^*(f_e) = (\ell_e(f_e)f_e)' = \ell_e(f_e) + \ell_e'(f_e)f_e$. Lemmas 2.2 and 2.4 then yield the following corollary.

Corollary 2.5 ([3, 13]) *Let (G, r, ℓ) be an instance in which $x \cdot \ell_e(x)$ is a convex function for each edge e, with marginal cost functions ℓ^* defined as above. Then a flow f feasible for (G, r, ℓ) is optimal if and only if it is at Nash equilibrium for the instance (G, r, ℓ^*).*

Remark: We will typically denote a minimum-latency flow for an instance by f^*. The marginal cost functions are denoted by ℓ^* as they are "optimal latency functions" in a sense made precise by Corollary 2.5: the optimal flow f^* arises as a flow at Nash equilibrium with respect to latency functions ℓ^*.

Observe that the function $\ell_e^*(x)$ describing the marginal cost of increasing flow on edge e has one term $\ell_e(x)$ capturing the per-unit latency incurred by the additional flow and a second term $x \cdot \ell_e'(x)$ accounting for the increased congestion experienced by the flow already using the edge. Essentially, the only difference between an optimal flow and a flow at Nash equilibrium is that the former accounts for this "conscientious" second term while the latter disregards it.

2.4 Existence of Flows at Nash Equilibrium

In this subsection, we exploit the similarity between the characterizations of Nash and of minimum-latency flows (Lemmas 2.2 and 2.4) to prove the existence and essential uniqueness of Nash equilibria. This result is originally due to Beckman et al. [3] and was later reproved by Dafermos and Sparrow [13]; we include a proof both for completeness and because the techniques will prove useful in the next subsection.

Lemma 2.6 ([3, 13]) *An instance (G, r, ℓ) with continuous, nondecreasing latency functions admits a feasible flow at Nash equilibrium. Moreover, if $f, \tilde{f}$ are flows at Nash equilibrium, then $C(f) = C(\tilde{f})$.*

Proof: Set $h_e(x) = \int_0^x \ell_e(t)dt$. Since each h_e is differentiable with nondecreasing derivative ℓ_e, each h_e is convex. Now consider the convex program

$$\text{Min} \quad \sum_{e \in E} h_e(f_e)$$

subject to:

$$(NLP2) \qquad \sum_{P\in\mathcal{P}_i} f_P = r_i \qquad \forall i \in \{1,\ldots,k\}$$

$$f_e = \sum_{P\in\mathcal{P}:e\in P} f_P \qquad \forall e \in E$$

$$f_P \geq 0 \qquad \forall P \in \mathcal{P}$$

and notice that the optimality conditions of Lemma 2.4 for $(NLP2)$ precisely match the characterization of flows at Nash equilibrium in Lemma 2.2. In other words, the optimal solutions for $(NLP2)$ are precisely the flows at Nash equilibrium for (G, r, ℓ). Existence of a Nash equilibrium then follows from the facts that $(NLP2)$ has a continuous objective function and a compact feasible region. Next, suppose $f, \tilde{f}$ are flows in G at Nash equilibrium (and hence global optima for $(NLP2)$). By convexity of the objective function of $(NLP2)$, whenever $f_e \neq \tilde{f}_e$ the function h_e must be linear between these two values (otherwise any convex combination of $f, \tilde{f}$ would be a feasible solution for $(NLP2)$ with smaller objective function value) and hence ℓ_e must be constant between these two values. This implies that $\ell_e(f_e) = \ell_e(\tilde{f}_e)$ for all $e \in E$, hence $L_i(f) = L_i(\tilde{f})$ for all i, and hence (by Lemma 2.3) $C(f) = C(\tilde{f})$. ■

2.5 A Good but Not Optimal Upper Bound

The proof of Lemma 2.6 provides a fairly general method for upper-bounding the ratio between the cost of a flow at Nash equilibrium and of a minimum-latency flow for instances with latency functions that are, in some sense, "not too steep". Before making this statement precise, we require some additional notation. For an instance (G, r, ℓ) admitting an optimal flow f^* and a flow at Nash equilibrium f, we denote the ratio $\frac{C(f)}{C(f^*)}$ by $\rho = \rho(G, r, \ell)$; note that ρ is well-defined by Lemma 2.6. We then have the following corollary.

Corollary 2.7 *Suppose the instance (G, r, ℓ) and the constant $\alpha \geq 1$ satisfy*

$$x \cdot \ell_e(x) \leq \alpha \cdot \int_0^x \ell_e(t)dt$$

for all edges e and all positive real numbers x. Then

$$\rho(G, r, \ell) \leq \alpha.$$

Proof: Roughly speaking, the corollary holds since a flow at Nash equilibrium for (G, r, ℓ) optimizes an objective function (the objective function of $(NLP2)$ in the proof of Lemma 2.6) that is at most a factor α away from the true objective function $C(\cdot)$. More formally, let f and f^* denote Nash and optimal flows for (G, r, ℓ), respectively; we can then derive

$$\begin{aligned} C(f) &= \sum_{e\in E} \ell_e(f_e) f_e \\ &\leq \alpha \sum_{e\in E} \int_0^{f_e} \ell_e(t)dt \end{aligned}$$

$$
\begin{aligned}
&\le \alpha \sum_{e \in E} \int_0^{f_e^*} \ell_e(t)dt \\
&\le \alpha \sum_{e \in E} \ell_e(f_e^*) f_e^* \\
&= \alpha \cdot C(f^*)
\end{aligned}
$$

where the first inequality follows from the hypothesis, the second inequality from the fact that the Nash flow f optimizes the objective function $\sum_e \int_0^x \ell_e(t)dt$ (see Lemma 2.6), and the third inequality from the assumption that every latency function ℓ_e is nondecreasing. ■

While the hypothesis of Corollary 2.7 is somewhat opaque, the corollary nevertheless gives a non-trivial upper bound on the cost of selfish routing for many instances, such as instances with latency functions that are polynomials with nonnegative coefficients.

Corollary 2.8 *Suppose every latency function ℓ_e of the instance (G, r, ℓ) has the form $\ell_e(x) = \sum_{i=0}^{p} a_{e,i} x^i$ for a positive integer p and nonnegative reals $a_{e,i}$. Then,*

$$\rho(G, r, \ell) \le p + 1.$$

In particular, Corollary 2.8 states that in an instance with linear latency functions (that is, every latency function ℓ_e has the form $\ell_e(x) = a_e x + b_e$ for $a_e, b_e \ge 0$), the cost of a flow at Nash equilibrium is at most twice that of a minimum-latency flow. This upper bound is non-trivial (to the best of our knowledge, it was not known prior to this work) but it is not sharp: in Section 4 we prove that if (G, r, ℓ) is an instance with linear latency functions, then $\rho(G, r, \ell) \le \frac{4}{3}$. The upper bound of Corollary 2.8 is not best possible for higher-degree polynomials, either; Roughgarden [44] has recently shown that if all latency functions of an instance (G, r, ℓ) are polynomials with degree at most p, then $\rho(G, r, \ell) = O(\frac{p}{\ln p})$ (see Section 6 for details). On the other hand, for each value of $\alpha \ge 1$, there is an instance (with rather contrived latency functions) for which the inequality of Corollary 2.7 holds with equality.

3 A Bicriteria Result for General Latency Functions

We have already seen (Figure 1(b)) that a flow at Nash equilibrium and a minimum-latency flow may have different costs. In the next two sections, we analyze the *ratio* of the cost of a flow at Nash equilibrium to that of the minimum-latency flow. In this section we work with general (continuous, nondecreasing) latency functions, while in Section 4 we will specialize to the case of linear latency functions.

We begin with some simple negative results. Recall from Subsection 2.5 that by $\rho(G, r, \ell)$ we mean the ratio between the cost of a Nash flow and of an optimal flow for instance (G, r, ℓ). For example, in the second network of Braess's Paradox (Figure 1(b)), a flow at Nash equilibrium has total latency 2 while the optimal flow has total latency $\frac{3}{2}$; thus, $\rho = \frac{4}{3}$ in this instance. In fact, it is easy to construct an even simpler example (still with linear latency functions) with ratio $\rho = \frac{4}{3}$. In the network shown in Figure 3, with a single source-destination pair and rate 1, the flow at Nash equilibrium puts the entire unit of flow on the

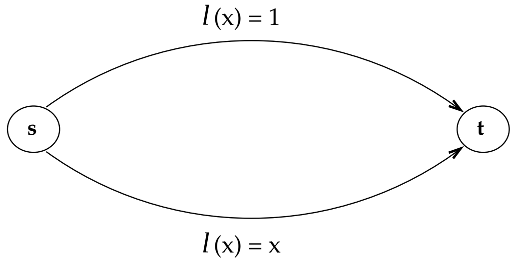


Figure 3: A Simple Bad Example

lower link (with a total latency of 1) while the minimum-latency flow spreads flow evenly across the two links, thereby incurring a cost of $\frac{3}{4}$. Thus, $\rho = \frac{4}{3}$ in this simple instance as well. (In the next section we will prove that this is the worst possible ratio for instances with linear latency functions.)

Unfortunately, the ratio can be much worse when non-linear latency functions are allowed. For a positive integer p, consider modifying the example of Figure 3 by giving the lower link a latency function of $\ell(x) = x^p$ (everything else remains unchanged). The flow at Nash equilibrium again places the entire unit on the lower link, incurring a cost of 1, while the optimal flow assigns $(p+1)^{-1/p}$ units to the lower link and the remainder to the upper link. This solution has a total latency of $1 - p \cdot (p+1)^{-(p+1)/p}$, which tends to 0 as $p \to \infty$. Thus, assuming only continuity and monotonicity of the edge latency functions, ρ cannot be bounded above (even as a function of the network size).

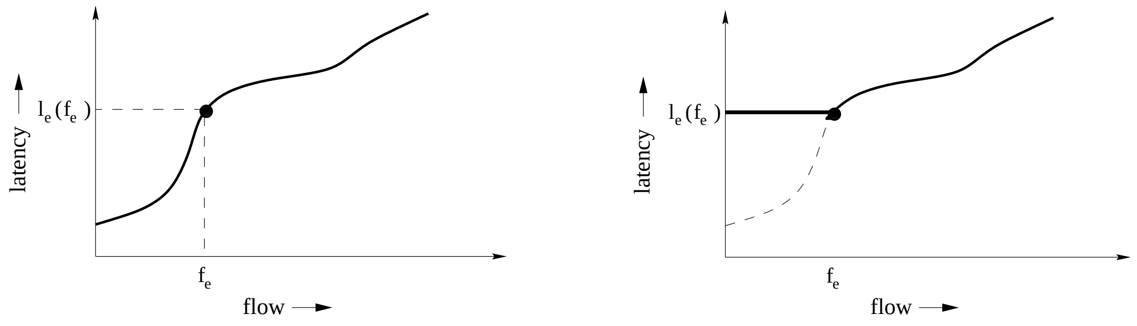
On the other hand, this example does not rule out interesting *bicriteria* results. Toward this end, we compare the cost of a flow at Nash equilibrium to an optimal flow feasible for *increased rates*. In the example above, an optimal flow feasible for rate $r \geq 1$ assigns the additional flow to the upper link, now incurring a cost that tends to $r - 1$ as $p \to \infty$. In particular, for any p an optimal flow feasible for twice the rate ($r = 2$) has total latency at least that of the flow at Nash equilibrium (feasible for the original rates). Our main result of this section is a proof of the generalization of this result to *any* network with continuous, nondecreasing edge latencies.

Theorem 3.1 *If f is a flow at Nash equilibrium for (G, r, ℓ) and f^* is feasible for $(G, 2r, \ell)$, then $C(f) \leq C(f^*)$.*

Proof: Suppose f, f^* satisfy the hypotheses of the theorem. For $i = 1, \ldots, k$, let $L_i(f)$ be the latency of an s_i-t_i flow path (of f), so that $C(f) = \sum_i L_i(f) r_i$ (see Lemma 2.3). We seek a set of latency functions $\bar{\ell}$ that on one hand approximates the original ones (in the sense that the cost of a flow with respect to latency functions $\bar{\ell}$ is close to its original cost) and, on the other hand, allows us to easily lower bound the cost (with respect to $\bar{\ell}$) of *any* feasible flow. With this goal in mind, we define new latency functions $\bar{\ell}$ as follows:

$$\bar{\ell}_e(x) = \begin{cases} \ell_e(f_e) & \text{if } x \leq f_e \\ \ell_e(x) & \text{if } x \geq f_e. \end{cases}$$

Figure 4 illustrates this construction.



(a) Graph of latency function ℓ_e and its value at flow value f_e

(b) Graph of latency function $\bar{\ell}_e$

Figure 4: Construction in the proof of Theorem 3.1 of modified latency function $\bar{\ell}_e$ given original latency function ℓ_e and Nash flow value f_e. Solid lines denote graphs of functions.

First we compare the cost of the flow f^* under the new latency functions $\bar{\ell}$ to its original cost $C(f^*)$. For any edge e, $\bar{\ell}_e(x) - \ell_e(x)$ is zero for $x \geq f_e$ and bounded above by $\ell_e(f_e)$ for $x < f_e$, so $x(\bar{\ell}_e(x) - \ell_e(x)) \leq \ell_e(f_e)f_e$ for all $x \geq 0$. Notice that the left-hand side (the discrepancy between $x\bar{\ell}_e(x)$ and $x\ell_e(x)$) is maximized when x is slightly smaller than f_e and when $\ell_e(x) = 0$; in this case, the value of the left-hand side is essentially the area of the rectangle enclosed by dashed lines in Figure 4(a). The difference between the new cost (with respect to $\bar{\ell}$) and the old cost (with respect to ℓ) can now be bounded as follows:

$$\begin{aligned}
\sum_e \bar{\ell}_e(f_e^*)f_e^* - C(f^*) &= \sum_{e\in E} f_e^*(\bar{\ell}_e(f_e^*) - \ell_e(f_e^*)) \\
&\leq \sum_{e\in E} \ell_e(f_e)f_e \\
&= C(f).
\end{aligned}$$

In other words, evaluating f^* with latency functions $\bar{\ell}$ (rather than ℓ) increases its cost by at most an additive $C(f)$ factor.

On the other hand, if f_0 denotes the zero flow in G, then by construction $\bar{\ell}_P(f_0) \geq L_i(f)$ for any path $P \in \mathcal{P}_i$. Since $\bar{\ell}_e$ is nondecreasing for each edge e, it follows that $\bar{\ell}_P(f^*) \geq L_i(f)$ for each path $P \in \mathcal{P}_i$. Thus, the cost of f^* with respect to $\bar{\ell}$ can be bounded below in the following manner:

$$\begin{aligned}
\sum_P \bar{\ell}_P(f^*)f_P^* &\geq \sum_i \sum_{P\in\mathcal{P}_i} L_i(f)f_P^* \\
&= \sum_i 2L_i(f)r_i = 2C(f).
\end{aligned}$$

Combining these two results we obtain the theorem:

$$C(f^*) \geq \sum_P \bar{\ell}_P(f^*)f_P^* - C(f)$$

$$\geq \; 2C(f) - C(f) = C(f).$$

■

The same proof also shows the following more general result.

Theorem 3.2 *If f is a flow at Nash equilibrium for (G, r, ℓ) and f^* is feasible for $(G, (1+\gamma)r, \ell)$, then $C(f) \leq \frac{1}{\gamma}C(f^*)$.*

Referring back to the bad example at the beginning of the section (the network of Figure 3 with latency functions $\ell(x) = 1$ and $\ell(x) = x^p$), we see that Theorem 3.2 is essentially tight for all values of γ. More precisely, by taking p sufficiently large we can obtain an instance admitting an optimal flow feasible for a traffic rate arbitrarily close to $(1+\gamma)$ with cost strictly less than γ (recall the cost of the flow at Nash equilibrium for the original rate $r = 1$ is 1) and an optimal flow feasible for rate $1+\gamma$ with cost arbitrarily close to γ.

4 Worst-Case Ratio of $\frac{4}{3}$ with Linear Latency Functions

In this section, we consider the scenario where the latency of each edge e is linear in the edge congestion—that is, where for each edge $e \in E$, $\ell_e(x) = a_e x + b_e$ for some $a_e, b_e \geq 0$. This is the setting in which Braess's paradox was originally discovered [6, 31], and several subsequent papers focused entirely on this model [18, 51]. In addition, linear latency functions are important for other applications: we will see later in this section that the mechanical networks of strings and springs mentioned in the Introduction can be modeled as traffic networks with linear latency functions, and Friedman [19] shows how linear latency functions naturally arise in a simple model of selfish users transferring files over a network employing a congestion control protocol (such as TCP).

We have already seen (Figures 1 and 3) two examples with linear latency functions for which ρ, the ratio of the cost of a flow at Nash equilibrium and the cost of an optimal flow, is $\frac{4}{3}$. Our main result for this section (Theorem 4.5) is a matching upper bound for networks with linear latency functions. Our proof techniques build on those of the previous two sections, the primary extension being a more refined approach to lower bounding the cost of an optimal flow.

The results of Section 2 have particularly simple and useful forms in the special case of linear latency functions. First, the total latency $C(f)$ of a flow f is given by $C(f) = \sum_e a_e f_e^2 + b_e f_e$; since $a_e \geq 0$ for all e, the non-linear program (NLP) of Subsection 2.3 is a convex (quadratic) program and thus Lemma 2.4 characterizes its optimal solutions. Also, in the notation of Subsection 2.3, if the latency function ℓ_e of edge e is $\ell_e(x) = a_e x + b_e$, then the marginal cost function ℓ_e^* of e is simply $\ell_e^*(x) = 2a_e x + b_e$. For convenience, we summarize this discussion together with specialized versions of Lemmas 2.2 and 2.4 in the following lemma.

Lemma 4.1 *Let (G, r, ℓ) be an instance with edge latency functions $\ell_e(x) = a_e x + b_e$ for each $e \in E$. Then,* *(a) a flow f is at Nash equilibrium in G if and only if for each source-sink pair i and $P, P' \in \mathcal{P}_i$ with $f_P > 0$,*

$$\sum_{e \in P} a_e f_e + b_e \le \sum_{e \in P'} a_e f_e + b_e$$

(b) a flow f^* is (globally) optimal in G if and only if for each source-sink pair i and $P, P' \in \mathcal{P}_i$ with $f_P^* > 0$,

$$\sum_{e \in P} 2a_e f_e^* + b_e \le \sum_{e \in P'} 2a_e f_e^* + b_e.$$

As an aside, we note that Lemma 4.1 immediately gives a simple proof of the following non-trivial result regarding networks in which the latency of each edge is proportional to its congestion; this result is implicit in the work of Dafermos and Sparrow [13], and other properties of this special case have been investigated in the context of electrical networks [5, 9].

Corollary 4.2 *Let G be a network in which each edge latency function ℓ_e is of the form $\ell_e(x) = a_e x$. Then for any rate vector r, a flow feasible for (G, r, ℓ) is optimal if and only if it is at Nash equilibrium.*

Proof: A feasible flow for such an instance satisfies the conditions of Lemma 4.1(a) if and only if it satisfies the conditions of Lemma 4.1(b). ■

A second corollary of Lemma 4.1 will play a crucial role in our proof of the main theorem of this section.

Lemma 4.3 *Suppose (G, r, ℓ) has linear latency functions and f is a flow at Nash equilibrium. Then,*

(a) the flow $f/2$ is optimal for $(G, r/2, \ell)$

(b) the marginal cost of increasing the flow on a path P with respect to $f/2$ equals the latency of P with respect to f.

Proof: For part (a), simply note that if f satisfies the conditions of Lemma 4.1(a) for (G, r, ℓ), then $f/2$ satisfies the conditions of Lemma 4.1(b) for $(G, r/2, \ell)$. For the second part, recall that if edge e has latency function $\ell_e(x) = a_e x + b_e$ then e has marginal cost function $\ell_e^*(x) = 2a_e x + b_e$. Thus, $\ell_e^*(f_e/2) = \ell_e(f_e)$ for each edge e and hence $\ell_P^*(f/2) = \ell_P(f)$ for each path P. ■

An outline of the proof of the main theorem is as follows. It will be useful to think about creating an optimal flow for the instance (G, r, ℓ) via a two-step process: in the first step, a flow optimal for the instance $(G, r/2, \ell)$ is sent through G, and in the second step this flow is augmented to one optimal for (G, r, ℓ) (note that this augmentation may increase *or decrease* the amount of flow on any given arc). We will show that the first flow has cost at least $\frac{1}{4}C(f)$ and that the augmentation has cost at least $\frac{1}{2}C(f)$, where f is some flow at Nash equilibrium.

We will see in the proof of Theorem 4.5 that the first lower bound follows easily from Lemma 4.3(a), but the second (for the cost of the augmentation, given that the first flow has already been routed) requires more work, and in particular the following lemma. Intuitively, the lemma simply claims that the per-unit cost of increasing the amount of flow through a network is at least the marginal cost of increasing flow on any path with respect to the current optimal flow.

Lemma 4.4 *Suppose (G, r, ℓ) is an instance with linear latency functions for which f^* is an optimal flow. Let $L_i^*(f^*)$ be the minimum marginal cost of increasing flow on an s_i-t_i path with respect to f^*. Then for any $\delta > 0$, a feasible flow for the problem instance $(G, (1+\delta)r, \ell)$ has cost at least*

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

Proof: First, note that if each L_i^* is nondecreasing in r_i, then routing δr_i additional units of flow from s_i to t_i would cost at least $\delta L_i^*(f^*) r_i$ and the lemma would then follow easily by summing over s_i-t_i pairs. Although it is intuitively plausible that marginal costs are increasing in the amount of flow (it is certainly true for each edge individually), the proof requires a little work.

Formally, fix $\delta > 0$ and suppose f is feasible for $(G, (1+\delta)r, \ell)$. In general f_e may be larger or smaller than f_e^*. For any edge $e \in E$, convexity of the function $x \cdot \ell_e(x) = a_e x^2 + b_e x$ implies that

$$\ell_e(f_e) f_e \geq \ell_e(f_e^*) f_e^* + (f_e - f_e^*) \ell_e^*(f_e^*).$$

In essence, this inequality states that estimating the cost of changing the flow value on edge e from f_e^* to f_e by $(f_e - f_e^*)\ell_e^*(f^*)$ (i.e., by the marginal cost of flow increase at f_e^* times the size of the perturbation) only underestimates the actual cost of an increase (when $f_e > f_e^*$) and overestimates the actual benefit of a decrease (when $f_e < f_e^*$). We may thus derive

$$\begin{aligned} C(f) &= \sum_{e \in E} \ell_e(f_e) f_e \\ &\geq \sum_{e \in E} \ell_e(f_e^*) f_e^* + \sum_{e \in E} (f_e - f_e^*) \ell_e^*(f_e^*) \\ &= C(f^*) + \sum_{i=1}^{k} \sum_{P \in \mathcal{P}_i} \ell_P^*(f^*)(f_P - f_P^*). \end{aligned}$$

Since we have $L_i^*(f^*) \leq \ell_P^*(f^*)$ for each i and each $P \in \mathcal{P}_i$ and equality holds unless $f_P^* = 0$ (see Lemma 4.1(b)), we obtain

$$\begin{aligned} C(f) &\geq C(f^*) + \sum_{i=1}^{k} L_i^*(f^*) \sum_{P \in \mathcal{P}_i} (f_P - f_P^*) \\ &= C(f^*) + \delta \sum_{i=1}^{k} L_i^*(f^*) r_i, \end{aligned}$$

completing the proof. ■

We remark that Lemma 4.4 and its proof remain valid in much more general settings; all that is required is convexity of the function $x \cdot \ell_e(x)$ for each edge e (which holds when, for example, each edge latency function ℓ_e is convex).

We are now prepared to prove the main theorem.

Theorem 4.5 *If (G, r, ℓ) has linear latency functions, then $\rho(G, r, \ell) \leq \frac{4}{3}$.*

Proof: Let f be a flow in G at Nash equilibrium. Let $L_i(f)$ be the latency of an s_i-t_i flow path, so that $C(f) = \sum_i L_i(f) r_i$ (see Lemma 2.3). By Lemma 4.3(a), $f/2$ is an optimal solution to the instance $(G, r/2, \ell)$. Moreover, by Lemma 4.3(b), $L_i^*(f/2) = L_i(f)$ for each i (in words, marginal costs with respect to $f/2$ and latencies with respect to f coincide); this establishes the necessary connection between the cost of augmenting $f/2$ to a flow feasible for (G, r, ℓ) and the cost of a flow at Nash equilibrium.

Taking $\delta = 1$ in Lemma 4.4, we find that the cost of any flow f^* feasible for (G, r, ℓ) satisfies

$$\begin{aligned} C(f^*) &\geq C(f/2) + \sum_{i=1}^{k} L_i^*(f/2) \frac{r_i}{2} \\ &= C(f/2) + \frac{1}{2} \sum_{i=1}^{k} L_i(f) r_i \\ &= C(f/2) + \frac{1}{2} C(f). \end{aligned}$$

Finally, it's easy to lower bound the cost of $f/2$:

$$\begin{aligned} C(f/2) &= \sum_e \frac{1}{4} a_e f_e^2 + \frac{1}{2} b_e f_e \\ &\geq \frac{1}{4} \sum_e a_e f_e^2 + b_e f_e \\ &= \frac{1}{4} C(f) \end{aligned}$$

and thus $C(f^*) \geq \frac{3}{4} C(f)$. ■

We note that the analysis of this section can easily be extended to prove that in any instance (G, r, ℓ) where for some p, $\ell_e(x) = a_e x^p + b_e$ (with $a_e, b_e \geq 0$) for each edge e, $\rho(G, r, \ell) \leq (1 - p \cdot (p+1)^{-(p+1)/p})^{-1} = \Theta(\frac{p}{\ln p})$. The example at the beginning of Section 3 shows that this result is tight. Roughgarden [44] has recently shown (via a different analysis) that this upper bound holds more generally for instances with polynomial latency functions with nonnegative coefficients and any number of terms with degree at most p.

Consequences for Strings and Springs

We now return to the mechanical networks of strings and springs discussed in the Introduction and Figure 2. Viewing the support as a source and the suspended weight as a sink, with each string and spring as an arc, the equilibrium position of the mechanical device can be modeled as a Nash equilibrium in a traffic network G, with the support-weight distance corresponding to the common latency of any source-sink flow path. Strings (as perfectly inelastic objects) are modeled as links with constant latency functions while (perfectly elastic) springs correspond to links with latency functions that include a term of the form ax. Severing a string or spring corresponds to deleting an edge from a traffic network; thus any realizable equilibrium of the mechanical network (after possibly destroying some of its constituent parts) corresponds to a Nash equilibrium in a subgraph of the corresponding traffic network G.

Although Theorem 4.5 is concerned with the total latency of flows (a concept with no natural analogue in our mechanical networks), we can use the result in the following way. By Theorem 4.5, every traffic flow in G (and in particular every flow at Nash equilibrium in a subgraph of G) has total latency at least $\frac{3}{4}$ times that of a Nash flow f in G. By Lemma 2.3, it follows that if the common latency of every flow path of f is L and $\hat{f}$ is a flow at Nash equilibrium in a subgraph of G, then the common latency of every flow path of $\hat{f}$ is at least $\frac{3}{4}L$. Reinterpreting this result for networks of strings and springs, we obtain the following corollary of Theorem 4.5.

Corollary 4.6 *In any network of strings and springs carrying a single weight with support-weight distance D, the support-weight distance after severing an arbitrary collection of strings and springs is at least $\frac{3}{4}D$.*

5 Extensions

The basic traffic model of this paper suffers from several deficiencies; in this section we attempt to rectify some of them by extending the basic model in several different ways. First, agents can often only evaluate path latency approximately, rather than exactly. Subsection 5.1 extends the notion of a flow at Nash equilibrium and Theorem 3.1 to this setting. Second, our basic model represents a scenario with infinitely many agents each controlling an infinitesimal amount of flow, while we typically expect to encounter a finite number of agents, each controlling a strictly positive amount of flow. In Subsection 5.2 we prove an analogue of Theorem 3.1 for the case of finitely many agents, provided each agent can route its flow fractionally over any number of paths. In Subsection 5.3 we show that such an assumption is essentially necessary, in that no bicriteria result analogous to Theorem 3.1 holds when there are only finitely many agents, each of whom must route its flow on a single path; however, a version of Theorem 3.1 does hold if agents do not control too much flow and the edge latency functions are not too steep. Finally, we remark that Theorems 3.1 and 4.5 can be extended to a broader class of games (including games without the structure provided by a network); this direction of research is pursued in a companion paper [46].

5.1 Flows at Approximate Nash Equilibrium

It is unreasonable to expect agents to be able to evaluate the latency of different paths with arbitrary precision. We next investigate the sensitivity of our results to this assumption. We suppose that an agent can only distinguish between paths that differ significantly in their latency (say by more than a $(1+\epsilon)$ factor for some $\epsilon > 0$). Our definition of a flow at *ϵ-approximate Nash equilibrium* is then an obvious modification of Definition 2.1:

Definition 5.1 *A flow f feasible for instance (G, r, ℓ) is at ϵ-*approximate Nash equilibrium *if for all $i \in \{1, \ldots, k\}$, $P_1, P_2 \in \mathcal{P}_i$, and $\delta \in [0, f_{P_1}]$, we have $\ell_{P_1}(f) \leq (1+\epsilon)\ell_{P_2}(\tilde{f})$, where*

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

The analogue of Lemma 2.2 is then:

Lemma 5.2 *A flow f is at ϵ-approximate Nash equilibrium if and only if for every $i \in \{1, \ldots, k\}$ and $P_1, P_2 \in \mathcal{P}_i$ with $f_{P_1} > 0$, $\ell_{P_1}(f) \leq (1+\epsilon)\ell_{P_2}(f)$.*

The next theorem provides an analogue of Theorem 3.1 for flows at ϵ-approximate Nash equilibrium.

Theorem 5.3 *If f is at ϵ-approximate Nash equilibrium with $\epsilon < 1$ for (G, r, ℓ) and f^* is feasible for $(G, 2r, \ell)$, then $C(f) \leq \frac{1+\epsilon}{1-\epsilon} C(f^*)$.*

The proof closely follows the proof of Theorem 3.1 and is omitted. A simple example on a network similar to that of Braess's Paradox (Figure 1(b)) shows that the $\frac{1+\epsilon}{1-\epsilon}$ factor in Theorem 5.3 cannot be improved.

5.2 Finitely Many Agents: Splittable Flow

Our basic model makes the often unrealistic assumption that there are an infinite number of noncooperative agents, each controlling a negligible fraction of the overall traffic. In this subsection we extend the basic model to the case of finitely many agents, each of whom controls a strictly positive amount of flow. In this subsection we allow an agent to split flow along any number of paths; the next subsection investigates the case where each agent must route all of its flow on a single path.

We are given a network G with continuous nondecreasing latency functions ℓ as before, and in addition k *agents*. We assume that agent i intends to send r_i units of flow from source s_i to destination t_i. Distinct agents may have identical source-destination pairs. We continue to denote an instance by (G, r, ℓ), and we call the instance *finite splittable*. A *flow f* now consists of k functions, with one function $f^{(i)} : \mathcal{P}_i \to \mathcal{R}^+$ for each agent i. For a flow f, we denote by $C_i(f)$ the total latency experienced by agent i; thus, $C_i(f) = \sum_{P \in \mathcal{P}_i} \ell_P(f) f_P^{(i)}$. As usual, a flow is *at Nash equilibrium* if no agent can decrease the latency it experiences by rerouting its flow. In this setting, a flow f is at Nash equilibrium if and only if for each i, $f^{(i)}$ minimizes $C_i(f)$ given $f^{(j)}$ for $j \neq i$. We will focus on the case where for each edge e, $x \cdot \ell_e(x)$ is a convex function; under this assumption, results of Rosen [40] imply that a flow at Nash equilibrium must exist and will be essentially unique.

Our main result for this model is an analogue of Theorem 3.1.

Theorem 5.4 *If f is at Nash equilibrium for the finite splittable instance (G, r, ℓ) with $x \cdot \ell_e(x)$ convex for each e, and f^* is feasible for the finite splittable instance $(G, 2r, \ell)$, then $C(f) \leq C(f^*)$.*

Proof: Fix f, f^* and define latency functions $\bar{\ell}$ as in the proof of Theorem 3.1. As in that proof, evaluating f^* with latency functions $\bar{\ell}$ (rather than ℓ) increases its cost by at most an additive $C(f)$ factor.

We claim that f is optimal for the instance $(G, r, \bar{\ell})$. We proceed by contradiction, showing that if f is not optimal for $(G, r, \bar{\ell})$ then f fails to be at Nash equilibrium for (G, r, ℓ). Suppose f is not optimal; since the instance $(G, r, \bar{\ell})$ defines a convex optimization problem of the form (NLP) (see Subsection 2.3), by Lemma 2.4 there are two paths P_1, P_2, an agent i such that $P_1, P_2 \in \mathcal{P}_i$ with $f^{(i)}_{P_1} > 0$, and a sufficiently small $\delta \in (0, f^{(i)}_{P_1}]$ such that moving δ units of flow from P_1 to P_2 yields a new flow with cost (with respect to $\bar{\ell}$) strictly less than that of f. Our goal is to show that the same local move will be beneficial for agent i in the instance (G, r, ℓ). We may assume that P_1, P_2 are disjoint (otherwise, the following argument may be applied to the symmetric difference of P_1 and P_2). The benefit (with respect to $\bar{\ell}$) of removing δ units of flow from path P_1 is then $\delta \cdot \bar{\ell}_{P_1}(f) = \delta \cdot \ell_{P_1}(f)$ (since $\bar{\ell}_e(x) = \ell_e(f_e)$ when $x \leq f_e$) while the cost (with respect to $\bar{\ell}$) of adding δ units of flow to P_2 is $\sum_{e \in P_2} [\ell_e(f_e + \delta)(f_e + \delta) - \ell_e(f_e) f_e]$; we are assuming that the former exceeds the latter. On the other hand, agent i is capable of making an identical local change to $f^{(i)}$ in the instance (G, r, ℓ), and doing so provides a benefit to agent i of at least $\delta \cdot \ell_{P_1}(f)$ with respect to ℓ (since latency functions are nondecreasing) and a cost (with respect to ℓ) of

$$\sum_{e \in P_2} [\ell_e(f_e + \delta)(f^{(i)}_e + \delta) - \ell_e(f_e) f^{(i)}_e]$$

which is at most

$$\sum_{e \in P_2} [\ell_e(f_e + \delta)(f_e + \delta) - \ell_e(f_e) f_e]$$

since ℓ_e is nondecreasing and $f^{(i)}_e \leq f_e$ for each edge e. Thus, moving δ units of flow from path P_1 to path P_2 yields a better outcome for agent i in the instance (G, r, ℓ), so f fails to be at Nash equilibrium for (G, r, ℓ).

We have determined that any flow feasible for $(G, r, \bar{\ell})$ must have cost at least $C(f)$. Since every latency function is nondecreasing, it follows that any flow feasible for $(G, 2r, \bar{\ell})$ (and in particular f^*) must have cost at least $2C(f)$ (such a flow may be expressed as the sum of two flows feasible for $(G, r, \bar{\ell})$, and the cost of their sum is at least the sum of their individual costs). Since the cost of f^* with respect to $\bar{\ell}$ exceeds its cost with respect to ℓ by at most $C(f)$, the theorem follows. ■

Theorem 3.1 can be regarded as the limiting case of the above theorem, as the number of agents tends to infinity and the amount of flow controlled by each agent tends to 0.

5.3 Finitely Many Agents: Unsplittable Flow

In this subsection we continue our investigation of selfish routing with finitely many agents, each controlling a non-negligible amount of flow. It is easy to imagine scenarios in which

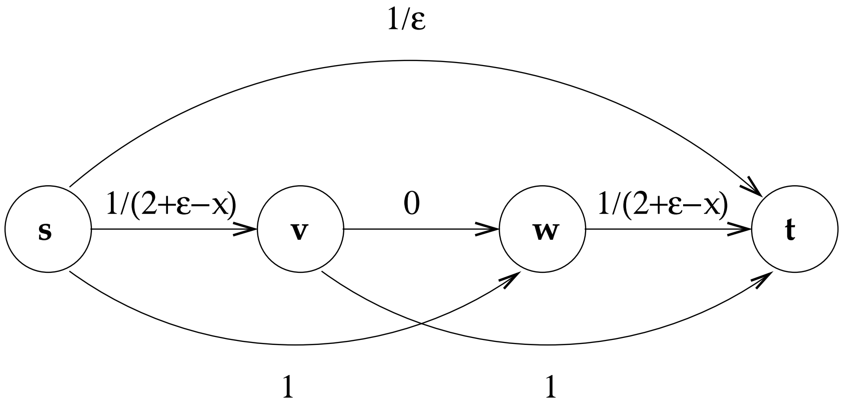


Figure 5: A Bad Example for Unsplittable Flow

agents cannot route flow on several different paths, but instead must select a single path for routing. Our previous results have made crucial use of the "infinitely divisible" nature of flow, and we next show that this assumption is essentially necessary.

Consider an instance (G, r, ℓ) as in the previous subsection (with k agents and the ith agent controlling r_i units of flow), but with the additional constraint that each agent selects a *single* path on which to route all of its flow. We call such an instance *finite unsplittable*. Adapting the definition of the previous subsection to this new setting, a flow f (now consisting only of k paths) is at Nash equilibrium if and only if for each i, agent i routes its flow on a path minimizing $\ell_P(f)$ (with P ranging over all paths in $\mathcal{P}_i$), given the paths chosen by the other $k-1$ agents.

We first consider a simple example showing that a flow at Nash equilibrium may have cost arbitrarily larger than that of an optimal flow. Consider the network given in Figure 5, and suppose there are two agents, each of whom has source s, destination t, and one unit of flow to send; $\epsilon > 0$ is arbitrary. In the optimal solution, one agent chooses path $s \rightarrow v \rightarrow t$ and the other $s \rightarrow w \rightarrow t$; the cost of this solution is less than 4 (for any $\epsilon > 0$). On the other hand, a solution with one agent choosing path $s \rightarrow v \rightarrow w \rightarrow t$ and the other routing on the $s \rightarrow t$ link is a flow at Nash equilibrium with cost greater than $\frac{1}{\epsilon}$; by choosing ϵ arbitrarily small this cost is arbitrarily large, and hence arbitrarily more costly than optimal.

In light of the example at the beginning of Section 3, such a result is hardly surprising; however, we can extend this example to show that bicriteria statements analogous to Theorems 3.1 and 5.4 are false when we require agents to route flow unsplittably. For a positive integer q, consider the network G_q consisting of $2q+2$ vertices arranged in a path $s, v_1, v_2, \ldots, v_{2q}, t$ with edges along the path alternately having latency functions $\ell(x) = \frac{1}{2+\epsilon-x}$ and $\ell(x) = 0$, a direct s-t link with constant latency function $\ell(x) = \frac{1}{\epsilon}$, and arcs from s to v_{2i} and from v_{2i-1} to t with constant latency functions $\ell(x) = 1$ (observe that this construction produces the network of Figure 5 when $q = 1$). As in the previous paragraph, there is a flow at Nash equilibrium with two agents, each controlling one unit of flow, with cost greater than $\frac{1}{\epsilon}$. On the other hand, it is possible for $q+1$ agents to each send one unit of flow through G_q at total cost at most $3q$ (the first agent uses path $s \rightarrow v_1 \rightarrow t$, the last $s \rightarrow v_{2q} \rightarrow t$, and otherwise the ith agent uses path $s \rightarrow v_{2i-2} \rightarrow v_{2i-1} \rightarrow t$). Letting ϵ tend to 0 for each fixed value of q, we see that an optimal flow can send *arbitrarily more flow* at *arbitrarily less cost* than a flow at Nash equilibrium.

In the above bad example, the network has latency functions with unbounded derivatives; in this situation, routing a strictly positive amount of additional flow on an edge may increase the latency of that edge by an arbitrarily large amount. This example is of particular interest because functions of the form $\ell(x) = 1/(u-x)$ arise as the delay functions of $M/M/1$ queues with capacity or service rate u (see, e.g., [4]) and are therefore common in the networking literature [4, 26, 27, 29, 35]. However, in networks where the largest possible change in edge latency resulting from a single agent rerouting its flow is not too large, we can apply the results of Subsection 5.1 to derive the following.

Theorem 5.5 *Suppose f is at Nash equilibrium in the finite unsplittable instance (G, r, ℓ), and for some $\alpha < 2$, we have $\ell_e(x + r_i) \leq \alpha \cdot \ell_e(x)$ for all agents $i \in \{1, \ldots, k\}$, edges $e \in E$, and $x \in [0, \sum_{j \neq i} r_j]$. Then for any flow f^* feasible for $(G, 2r, \ell)$, $C(f) \leq \frac{\alpha}{2-\alpha} \cdot C(f^*)$.*

Proof: We may interpret f and f^* as (fractional) flows feasible for instances (G, r', ℓ) and $(G, 2r', \ell)$ of the original type (that is, instances in the sense of Sections 2–4), where r'_i is the total amount of flow controlled by agents with source s_i and destination t_i in the original instance. The hypotheses ensure that f is at $(\alpha - 1)$-approximate Nash equilibrium for (G, r', ℓ), so the result follows from Theorem 5.3. ■

For example, in an instance with linear latency functions (say $\ell_e(x) = a_e x + b_e$) with $b_e > 0$ for all edges e, we may apply Theorem 5.5 with $\alpha = 1 + \max_i r_i \cdot \max_e a_e/b_e$.

6 Recent Work

Since the publication of a preliminary version of this paper [45], several new results have been discovered. Closest in spirit to this work is the recent paper of Roughgarden [44], who proves that for almost any class of allowable latency functions, single-commodity instances on two-node networks of parallel links furnish the worst possible examples for the cost of selfish routing. The matching lower and upper bounds for linear latency functions given in Figure 3 and Theorem 4.5 of this paper are thus a special case of the more general statement in [44]. This result permits the computation of the worst-case ratio between the cost of a Nash flow and of an optimal flow with respect to an arbitrary class of allowable latency functions; for example, it is shown in [44] that the cost of selfish routing with polynomial latency functions (for any fixed degree bound) is maximized by the two-node, two-link examples of Section 3.

Also, Friedman [19] has recently shown that in any network, for "most" traffic rate vectors the cost of selfish routing is much smaller than the worst-case value. To state his result more precisely, fix a network G with latency functions ℓ, and let $N(r)$ be the cost of a Nash flow for instance (G, r, ℓ). Friedman uses the ratio $\Gamma(r) = N(r)/N(r/2)$ as a sensitivity measure of the problem instance (G, r, ℓ). Applying Theorem 3.1 to $(G, r/2, \ell)$ shows that the ratio $\rho(G, r, \ell)$ between the cost of the Nash and optimal flows for (G, r, ℓ) is bounded above by $\Gamma(r)$. Friedman [19] shows that for "most" traffic rate vectors in $[r/2, r]$, the cost of selfish routing is only $O(\log \Gamma(r))$.

Finally, we mention two recent efforts to control the inefficiency inherent in selfishly-defined equilibria. In the first, Roughgarden [43] shows that if a fraction of the network traffic is centrally controlled and is carefully routed by a network manager, then the corresponding induced equilibrium (with selfish users taking into account the congestion caused by the centrally routed traffic) is less inefficient then a flow at Nash equilibrium. In particular, with no assumptions on the edge latency functions (a setting in which Nash flows may be arbitrarily more costly than optimal flows, recall Section 3), a network manager controlling a constant fraction of the network traffic can induce an equilibrium that is only a constant-factor more costly than an optimal assignment of all of the traffic. Unfortunately, only the restricted setting of networks of parallel links (or, equivalently, of machine scheduling) is considered in [43].

In a different direction, Roughgarden [42] studies the problem of designing networks that admit efficient Nash flows. In particular, the following simple network design problem is considered in [42]: given an instance (G, r, ℓ) with a single source-sink pair, find the subnetwork H of G minimizing the latency experienced by all traffic in a Nash flow of (H, r, ℓ). Braess's Paradox shows that in general a proper subgraph of G will be the optimal solution to such a network design problem. Sadly, the results of [42] are negative: the obvious heuristic that always returns the entire network G (in essence, ignoring the possibility of Braess's Paradox) is a $\frac{4}{3}$-approximation algorithm[5] for networks with linear latency functions and a $\lfloor \frac{n}{2} \rfloor$-approximation algorithm for networks with arbitrary latency functions (where n is the number of nodes of G), and no better approximation is possible in polynomial time unless $P = NP$.

Acknowledgements

We thank Leonard Schulman, Christos Papadimitriou, and Satish Rao for a discussion about Braess's paradox, and for suggesting the comparison of the social value of a Nash equilibrium to the optimal social value. This discussion was the start of our research. We are also grateful to Scott Shenker for his interest in our results and for many helpful conversations about directions for future work, to Leslie Ann Goldberg for bringing [2] and [7] to our attention, and to an anonymous referee for several useful suggestions.

References

[1] H. Z. Aashtiani and T. L. Magnanti. Equilibria on a congested transportation network. *SIAM Journal on Algebraic and Discrete Methods*, 2(3):213–226, 1981.

[2] T. Bass. Road to ruin. *Discover*, 13:56–61, 1992.

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

[5] A *c-approximation algorithm* for a minimization problem runs in polynomial time and returns a solution no more than c times as costly as an optimal solution.

[4] D. P. Bertsekas and R. Gallager. *Data Networks*. Prentice-Hall, 1992. Second Edition.

[5] R. Bott and R. J. Duffin. On the algebra of networks. *Transactions of the AMS*, 74:99–109, 1953.

[6] D. Braess. Uber ein paradoxon der verkehrsplanung. *Unternehmensforschung*, 12:258–268, 1968.

[7] J. E. Cohen and P. Horowitz. Paradoxical behavior of mechanical and electrical networks. *Nature*, 352:699–701, 1991.

[8] J. E. Cohen and F. P. Kelly. A paradox of congestion in a queuing network. *Journal of Applied Probability*, 27:730–734, 1990.

[9] R. M. Cohn. The resistance of an electrical network. *Proceedings of the AMS*, 1:316–324, 1950.

[10] A. Czumaj and B. Vöcking. Tight bounds for worst-case equilibria. In *Proceedings of the 13th Annual Symposium on Discrete Algorithms*, pages 413–420, 2002.

[11] S. Dafermos. Traffic equilibrium and variational inequalities. *Transportation Science*, 14(1):42–54, 1980.

[12] S. C. Dafermos and A. Nagurney. On some traffic equilibrium theory paradoxes. *Transportation Research, Series B*, 18B:101–110, 1984.

[13] S. C. Dafermos and F. T. Sparrow. The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Series B*, 73B(2):91–118, 1969.

[14] P. Dubey. Inefficiency of Nash equilibria. *Mathematics of Operations Research*, 11(1):1–8, 1986.

[15] C. Fisk. More paradoxes in the equilibrium assignment problem. *Transportation Research*, 13B:305–309, 1979.

[16] M. Florian. Nonlinear cost network models in transportation analysis. *Mathematical Programming Study*, 26:167–196, 1986.

[17] M. Florian and D. Hearn. Network equilibrium models and algorithms. In M. O. Ball, T. Magnanti, C. Monma, and G. Nemhauser, editors, *Network Routing*, chapter 6, pages 485–550. Elsevier Science, 1995.

[18] M. Frank. The Braess Paradox. *Mathematical Programming*, 20:283–302, 1981.

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

[20] J. N. Hagstrom and R. A. Abrams. Characterizing Braess's paradox for traffic networks. In *IEEE Conference on Intelligent Transportation Systems*, pages 837–842, 2001.

[21] M. A. Hall. Properties of the equilibrium state in transportation networks. *Transportation Science*, 12(3):208–216, 1978.

[22] A. Haurie and P. Marcotte. On the relationship between Nash-Cournot and Wardrop equilibria. *Networks*, 15:295–308, 1985.

[23] B. Kalyanasundaram and K. Pruhs. Speed is as powerful as clairvoyance. *Journal of the ACM*, 47(4):617–643, 2000. Preliminary version in *FOCS '95*.

[24] F. H. Knight. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics*, 38:582–606, 1924.

[25] W. Knödel. *Graphentheoretische Methoden und ihre Anwendungen*. Springer-Verlag, 1969.

[26] Y. A. Korilis, A. A. Lazar, and A. Orda. Capacity allocation under noncooperative routing. *IEEE Transactions on Automatic Control*, 42(3):309–325, 1997.

[27] Y. A. Korilis, A. A. Lazar, and A. Orda. Avoiding the Braess paradox in noncooperative networks. *Journal of Applied Probability*, 36(1):211–222, 1999.

[28] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413, 1999.

[29] A. A. Lazar, A. Orda, and D. E. Pendarakis. Virtual path bandwidth allocation in multiuser networks. *IEEE/ACM Transactions on Networking*, 5:861–871, 1997.

[30] M. Mavronicolas and P. Spirakis. The price of selfish routing. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, pages 510–519, 2001.

[31] J. D. Murchland. Braess's paradox of traffic flow. *Transportation Research*, 4:391–394, 1970.

[32] A. Nagurney. *Sustainable Transportation Networks*. Edward Elgar, 2000.

[33] Y. Nesterov. Stable flows in transportation networks. CORE Discussion Paper 9907, 1999.

[34] Y. Nesterov and A. De Palma. Stable dynamics in transportation systems. CORE Discussion Paper 00/27, 2000.

[35] A. Orda, R. Rom, and N. Shimkin. Competitive routing in multi-user communication networks. *IEEE/ACM Transactions on Networking*, 1:510–521, 1993.

[36] G. Owen. *Game Theory*. Academic Press, 1995. Third Edition.

[37] C. Papadimitriou. Algorithms, games, and the Internet. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, pages 749–753, 2001.

[38] A. L. Peressini, F. E. Sullivan, and J. J. Uhl. *The Mathematics of Nonlinear Programming*. Springer-Verlag, 1988.

[39] C. A. Phillips, C. Stein, E. Torng, and J. Wein. Optimal time-critical scheduling via resource augmentation. *Algorithmica*, 32(2):163–200, 2002. Preliminary version in *STOC '97*.

[40] J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965.

[41] R. W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

[42] T. Roughgarden. Designing networks for selfish users is hard. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, pages 472–481, 2001.

[43] T. Roughgarden. Stackelberg scheduling strategies. In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, pages 104–113, 2001.

[44] T. Roughgarden. The price of anarchy is independent of the network topology. Submitted for publication, 2002.

[45] T. Roughgarden and É. Tardos. How bad is selfish routing? In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 93–102, 2000.

[46] T. Roughgarden and É. Tardos. Bounding the inefficiency of Nash equilibria. In preparation, 2002.

[47] Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, 1985.

[48] M. J. Smith. In a road network, increasing delay locally can reduce delay globally. *Transportation Research*, 12:419–422, 1978.

[49] M. J. Smith. The existence, uniqueness and stability of traffic equilibria. *Transportation Research*, 13B:295–304, 1979.

[50] R. Steinberg and R. E. Stone. The prevalence of paradoxes in transportation equilibrium problems. *Transportation Science*, 22(4):231–241, 1988.

[51] R. Steinberg and W. I. Zangwill. The prevalence of Braess' paradox. *Transportation Science*, 17(3):301–318, 1983.

[52] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Pt. II*, volume 1, pages 325–378, 1952.