How Bad is Selfish Routing?

Tim Roughgarden[*] Éva Tardos[†]

Department of Computer Science
Cornell University
Ithaca, NY 14853

June 3, 2003

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, including the Internet and other large-scale communication networks, 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.

[*]Research done while visiting the Computer Science Division, UC Berkeley, Berkeley, CA 94720. Supported by ONR grant N00014-98-1-0589. Email: `timr@cs.cornell.edu`.

[†]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 in order 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, link traversal time increases as the link becomes more congested.

In practice, it is often difficult or even impossible to impose optimal or near-optimal routing strategies on the traffic in a network, and thus network users are free to act according to their own interests, without regard to overall network performance. For example, existing Internet protocols place little restriction on how network traffic is routed, allowing network users to make decisions in a selfish or even malicious manner [**?**]. 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 *non-cooperative game* and expect the routes chosen by users to form a *Nash equilibrium* in the sense of classical game theory [**?**]. 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. 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" [**?**, **?**]. We are then interested in comparing the total latency of a Nash equilibrium with that of the optimal assignment of traffic to paths.

This line of research was recently initiated by Koutsoupias and Papadimitriou [15], who both considered network routing as a non-cooperative game (although in a different model than ours, and only for two-node networks) and proposed the worst-case ratio of the social welfare (suitably defined) achieved by a socially optimal set of strategies and by a Nash equilibrium as a measure of the performance degradation caused by a lack of regulation. As articulated in [15], this question studies the cost of the lack of *coordination* inherent in a non-cooperative 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 and the network a highway system, or agents might represent individual packets in a high-bandwidth communication network. 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 non-cooperative game then corresponds to a flow where all flow paths between a given source and destination have the same latency (if a flow does not have this property, some agent can improve its travel time by switching from the longer flow path to the shorter one). Beckman et al. [2] 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 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 [8, 22]), we show that a flow at Nash equilibrium has total latency at most $\frac{4}{3}$ that of the optimal flow. We also 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 Nash equilibrium flow with that of an optimal flow forced to route twice as much traffic between each pair of nodes. Our main result in this setting is that for any network with continuous nondecreasing latency functions, a flow at Nash equilibrium has total latency no more than that of an optimal flow forced to route twice as much traffic. We again give an example showing that our analysis is tight.

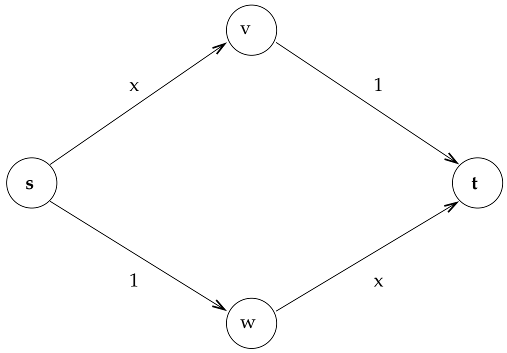
In the appendix, we examine two unrealistic assumptions made in the basic model: first, the assumption that agents can evaluate the latency of a path with infinite precision, and second, that there are an infinite number of agents each controlling a negligible fraction of the overall traffic. We define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

Related Work

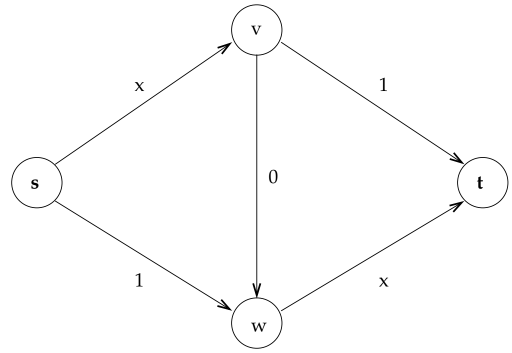
Unregulated traffic has been modeled as network flow with all flow paths between a given source-destination pair having equal latency since the 1950's [2, 23]. Beckman et al. [2] observed that such an equilibrium flow is an optimal solution to a related convex program (see also Section 2), while Dafermos and Sparrow [6] were perhaps the first authors interested in computing the equilibrium efficiently. Many subsequent papers gave increasingly efficient methods for computing equilibria (see [**?**] for a survey), while others have extended these results to more sophisticated models (see for example [1, 5, 10]).

In the past several decades, much of the work on this model has been inspired by a "paradox" first discovered by Braess [3] and later reported by Murchland [17]. 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 flow 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. As all agents are routed on a path of latency $\frac{3}{2}$, the flow has a total latency of $\frac{3}{2}$. 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 $s \to v \to w \to t$. The total latency of this flow is 2, as is the latency experienced by each individual agent. Thus, the intuitively helpful (or at least innocuous) action of adding a new zero-latency link may negatively impact *all* of the agents!

After Braess's Paradox was discovered (together with evidence of similarly counterintuitive and counterproductive traffic behavior following the construction of new roads in congested cities [**?**, 17]), researchers investigated the sensitivity of traffic equilibria to the properties of the underlying network [9], classified network topologies in which the addition of new links could degrade network performance [8, 22], discovered new types of "paradoxes" [21], and showed that Braess's Paradox has an analogue in queuing theory [4]. More recently, several papers [13, 14, 18] have investigated



(a) Before



(b) After

Figure 1: Braess's Paradox

a related model in which each agent controls a strictly positive amount of flow (and there are only finitely many agents); classes of network topologies and latency functions guaranteeing existence and uniqueness of Nash equilibria are studied by Orda et al. [18], while Korlis et al. [13, 14] study strategies for adding new edges and/or capacity to a network that guarantee an improvement in network performance.

Finally, the recent paper of Koutsoupias and Papadimitriou [15] is quite similar in spirit to ours, although their model is fairly different. In [15], a finite number of users share a series of parallel links, each of whom 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. We note that a Nash equilibrium in this model consists of a set of *mixed strategies* (i.e., agents select a distribution on paths) while in our model agents are confined to *pure strategies* (i.e., agents choose a single path); however, there is no essential distinction between pure and mixed strategies under the assumption that each agent controls a negligible amount of traffic. Different Nash equilibria may have different values in this model, so the *worst-case* Nash equilibrium is compared to a globally optimal choice of distributions. Koutsoupias and Papadimitriou obtain tight results in two-node, two-link networks and partial results for two-node networks with three or more parallel links.

Organization

The paper is structured as follows. 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 general edge latency functions. In Section 4 we prove a stronger and technically more involved result for networks with linear edge latency functions. The appendix considers several extensions to the basic model.

2 Preliminaries

2.1 The Model and a Non-Linear Programming Formulation

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, $\sum_{P\in\mathcal{P}_i} f_P = r_i$. Each edge $e \in E$ is given a load-dependent *latency* that we denote by $\ell_e(\cdot)$. The latency of a path P with respect to a flow f is then the sum of lengths of the edges on the path, denoted by $\ell_P(f) = \sum_{e\in P} \ell_e(f_e)$. For each $e \in E$, we assume that ℓ_e is non-negative, continuous, and nondecreasing. We define the *cost* $C(f)$ of a flow f in G as the total latency incurred by f, i.e.,

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

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

The problem of finding the minimum-latency feasible flow in a network is thus a special case of the following non-linear program

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

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). It is convenient at this point to assume that c_e is differentiable for each e, although the following results hold also under the weaker assumption of continuity. The fundamental result of this type is the Karush-Kuhn-Tucker Theorem (see, e.g., [19]). Roughly speaking, the version of the theorem most useful for our purposes states that a feasible solution y of a non-linear program is a local minimum if and only if the *gradient* of the objective function at y is non-negative in any locally feasible direction of movement. In our setting, a basic locally feasible move consists of removing flow from one s_i-t_i flow path and rerouting it on another s_i-t_i path; for a solution to be locally optimal such a move cannot decrease the objective function, implying that the gradient along a s_i-t_i flow path must be at most the gradient along any other s_i-t_i path. Letting $c'_P(f) = \sum_{e\in P} c'_e(f_e)$ and translating the Karush-Kuhn-Tucker Theorem into our setting then yields the desired characterization:

Lemma 1 *A flow f is a local optimum for (NLP) if and only if for every $i \in \{1, \ldots, k\}$ and $P_1, P_2 \in \mathcal{P}_i$ with $f_{P_1} > 0$, $c'_{P_1}(f) \leq c'_{P_2}(f)$.*

Moreover, if the objective function of (NLP) is convex, then Lemma 1 gives necessary and sufficient conditions for a flow to be *globally optimal*, as local and global minima of a convex function on a convex set coincide (see for example [19, Thm 2.3.4]).

2.2 Flows at Nash Equilibrium

We also wish to consider flows that behave "greedily" or "selfishly" in a certain sense, without regard to their cost. Intuitively, we expect each unit of 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 *A flow f in G is at* Nash equilibrium *if for all $i \in \{1, \dots, k\}$, $P_1, P_2 \in \mathcal{P}_i$, and $\delta \in [0, f_{P_1}]$, $\ell_{P_1}(f) \leq \ell_{P_2}(\tilde{f})$, 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 [10] or Wardrop's Principle [21, 22] in the literature, due to an influential paper of Wardrop [23].

Lemma 3 *A flow f is at Nash equilibrium if and only if for every $i \in \{1, \dots, k\}$ and $P_1, P_2 \in \mathcal{P}_i$ with $f_{P_1} > 0$, $\ell_{P_1}(f) \leq \ell_{P_2}(f)$.*

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 $L_i(f)$. Thus, we can express the cost $C(f)$ of a flow f at Nash equilibrium in a particularly nice form.

Lemma 4 *If f is a feasible flow at Nash equilibrium, then*

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

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) [**?**]. However, since in our model each agent carries a negligible fraction of the overall traffic, these two definitions are essentially equivalent.

The striking similarity between the characterizations of flows locally optimal for (NLP) and flows at Nash equilibrium was noticed early on by Beckman et al. [2], and provides an interpretation of an optimal flow as a flow at Nash equilibrium *with respect to a different set of edge latency functions*. More concretely, consider a minimum-latency flow f^* for (NLP). The flow f^* satisfies the conditions of Lemma 1, and so by Lemma 3 can be regarded as a flow at Nash equilibrium with respect to latency functions ℓ^*, where ℓ^*_e is the derivative of the function $\ell_e(x) \cdot x$ (we are assuming differentiability of ℓ_e for convenience only). Noting that $\ell^*_e(f_e) = \ell_e(f_e) + \ell'_e(f_e) \cdot f_e$, we can interpret ℓ_e as a function that quantifies only per-unit latency, while ℓ^*_e includes a second term accounting for the degradation in the total latency of the system. In a sense, the only difference

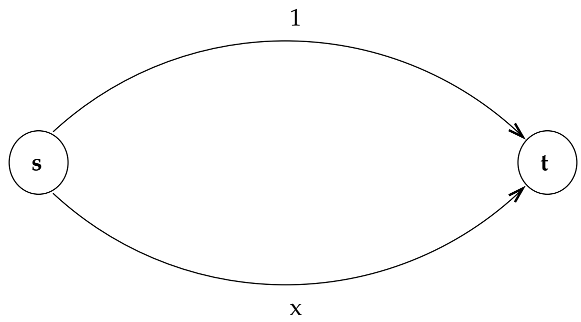


Figure 2: A Simple Bad Example

between an optimal flow and a flow at Nash equilibrium is that the former evaluates the cost of edge use (via ℓ^*) in a way that accounts for the latency experienced by all flow using the edge, while the latter "selfishly" evaluates edge latency by the per-unit rate ℓ.

Beckman et al. [2] also exploited this similarity between the two characterizations (in particular, that a flow at Nash equilibrium can be regarded as the optimal solution of a convex program of the form (NLP)) to prove the existence and essential uniqueness of Nash equilibria.

Lemma 5 ([2]) *A network G with continuous, nondecreasing latency functions admits a feasible flow at Nash equilibrium. Moreover, if f, f' are flows at Nash equilibrium, then $C(f) = C(f')$.*

3 A Bicriteria Result for Arbitrary Edge Costs

We have already seen (Figure 1(b)) that a flow at Nash equilibrium and a minimum-latency flow may have different costs. We next 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.

For a network G with rate vector r and edge latency functions 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 5.

We begin with some simple negative results. Recall in the canonical example demonstrating Braess's Paradox (Figure 1) a flow at Nash Equilibrium has total latency 2 while the optimal flow has total latency $\frac{3}{2}$; thus, in the above notation, $\rho = \frac{4}{3}$ in this particular instance. In fact, it is easy to construct an even simpler example (still with linear latency functions) with ratio $\frac{4}{3}$. In the network shown in Figure 2, with a single source-destination pair and rate 1, the flow at Nash equilibrium puts the entire unit of flow on the 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.)

Next we show that the ratio can be much worse when non-linear latency functions are allowed. For a positive integer k, we extend the example of Figure 2 by giving the lower link a latency function of x^k (other input data 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 $(k+1)^{-1/k}$ units to the lower link and the remainder to the upper link. This solution has a total latency of $1 - k \cdot (k+1)^{-(k+1)/k}$, which tends to 0 as $k \to \infty$. Thus, assuming only continuity and monotonicity of the edge latency functions, ρ cannot be bounded above.

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 $k \to \infty$. In particular, for any k 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 surprising generalization of this result to *any* network with continuous, nondecreasing edge latencies.

Theorem 6 *If f is feasible for (G, r, ℓ) and at Nash equilibrium, then for any f^* feasible for $(G, 2r, \ell)$, $C(f) \leq C(f^*)$.*

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 4). 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:

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

First we compare the cost of the flow f^* under the new latency functions $\bar{\ell}$ to its original cost $C(f^*)$. For any e, $\bar{\ell}_e(x) - \ell(x)$ 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 $x \geq 0$. Thus,

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

that is, evaluating the cost of f^* with latency functions $\bar{\ell}$ (rather than ℓ) changes 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 $P \in \mathcal{P}_i$. Since $\bar{\ell}_e$ is nondecreasing for each e, it follows that for any flow f' in G, $\bar{\ell}_P(f') \geq L_i(f)$ for any $P \in \mathcal{P}_i$. Thus, the cost of f^* with respect to $\bar{\ell}$ given by $\sum_P \bar{\ell}_P(f^*) f^*_P$ is bounded below by $\sum_i \sum_P L_i(f) f^*_P = \sum_i 2L_i(f) r_i = 2C(f)$.

Combining these results we obtain the theorem:

$$\begin{aligned} C(f^*) &\geq \sum_P \bar{\ell}_P(f^*) f^*_P - C(f) \\ &\geq 2C(f) - C(f) = C(f). \end{aligned}$$

■

Referring back to the bad example at the beginning of the section, we can see that Theorem 6 is essentially tight. More precisely, for any $\epsilon > 0$ one can take k sufficiently large to obtain an instance where the optimal flow feasible for rate $2 - \epsilon$ has cost strictly less than 1 (the cost of the flow at Nash equilibrium for the original rates) and the optimal flow feasible for rate 2 has cost at most $1 + \epsilon$.

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

Perhaps the simplest model of interest is that of a network with latency functions linear in the amount of flow on an edge. In this section, we consider the case where for each edge $e \in E$, $\ell_e(f_e) = a_e f_e + b_e$ for some $a_e, b_e \geq 0$. This is the setting in which Braess's paradox was originally discovered [3, 17], and several subsequent papers focused entirely on this model [8, 22].

We have already seen (Figures 1 and 2) 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 9) is a matching upper bound for networks with linear latency functions. Our proof techniques build on those of the previous section, 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 of a flow f is given by $\sum_e a_e f_e^2 + b_e f_e$; since $a_e \geq 0$ for all e, (NLP) is a convex program and thus Lemma 1 characterizes its optimal solutions. Also, in the notation of Section 2, if $\ell_e(f_e) = a_e f_e + b_e$, then $\ell_e^*(f_e) = 2a_e f_e + b_e$. For convenience, we summarize this discussion together with specialized versions of Lemmas 1 and 3 in the following lemma.

Lemma 7 *Suppose G is a directed network with k source-sink pairs and with edge latency functions $\ell_e = a_e f_e + b_e$ for each $e \in E$. Then,*

(a) a flow f is at Nash equilibrium in G if and only if for each i and $P, P' \in \mathcal{P}_i$ with $f_P > 0$,

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

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

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

The outline of the proof of the main theorem is as follows. It will be useful to think of the optimal flow for instance (G, r, ℓ) as the sum of two flows, each with rate vector $r/2$. We imagine that the two flows are sent in succession, and lower bound the cost of each separately. We will prove that the first flow has cost at least $\frac{1}{4}C(f)$ and the second at least $\frac{1}{2}C(f)$, where f is some flow at Nash equilibrium in G. The key observation (for both lower bounds) is that if f is a flow at Nash equilibrium feasible for rate vector r, then (by Lemma 7) $f/2$ is an *optimal* flow feasible for rate vector $r/2$; thus, the optimal flow can be viewed as the sum of $f/2$ and of a second flow corresponding to an augmentation of the flow $f/2$ to a flow feasible for rate vector r. The first lower bound follows easily from this observation (see the proof of Theorem 9 for details), but the second (for the cost of the augmentation, given that the first 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 gradient with respect to the current optimal flow.

Lemma 8 *Suppose (G, r, ℓ) has linear latency functions and f^* is an optimal flow in G. Let $L_i^*(f^*)$ be such that every s_i-t_i flow path P of f^* satisfies $\ell_P^*(f^*) = L_i^*(f^*)$. 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: $L_i^*(f^*)$ is well-defined for each i by Lemma 7(b). We can regard ℓ_e^* as the marginal cost of increasing the amount of flow on e (recall ℓ_e^* is derived from the gradient of the objective function of (NLP)), and thus $L_i^*(f^*)$ as the marginal cost of increasing the amount of flow between s_i and t_i. Thus, provided L_i^* is nondecreasing in r_i, the amount of flow sent between s_i and t_i, any flow of α units from s_i to t_i must have cost at least α times $L_i^*(f^*)$.

More formally, fix $\delta > 0$ and suppose f is feasible for $(G, (1+\delta)r, \ell)$. For any $e \in E$, by applying the Mean Value Theorem to the function $\ell_e(x) \cdot x$ we are assured of an x between f_e^* and f_e with $\ell_e^*(x)(f_e - f_e^*) = \ell_e(f_e)f_e - \ell_e(f_e^*)f_e^*$. Since ℓ_e^* is nondecreasing in its argument, we obtain

$$\ell_e^*(f_e^*)(f_e - f_e^*) \leq \ell_e(f_e)f_e - \ell_e(f_e^*)f_e^*.$$

Note that this bound holds both for $f_e^* \leq f_e$ and $f_e^* \geq f_e$. Thus,

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

■

We remark that Lemma 8 and its proof remain valid in much more general settings, for example when all the edge latency functions are convex. We are now prepared to prove the main theorem.

Theorem 9 *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 4). Since f satisfies the conditions of Lemma 7(a), it follows that $f/2$ satisfies the conditions of Lemma 7(b). Hence, as previously noted, $f/2$ is in fact an optimal solution to the instance $(G, r/2, \ell)$. Moreover, in the notation of Lemma 8, $L_i^*(f/2) = L_i(f)$ for each i. We also have

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

Taking $\delta = 1$ in Lemma 8, we find that the cost of any flow f' feasible for (G, r, ℓ) satisfies

$$C(f') \geq C(f/2) + \sum_{i=1}^{k} L_i^*(f/2) \frac{r_i}{2}$$

$$\begin{aligned} &\geq \frac{1}{4}C(f) + \frac{1}{2}\sum_{i=1}^{k} L_i(f) r_i \\ &= \frac{1}{4}C(f) + \frac{1}{2}C(f) = \frac{3}{4}C(f). \end{aligned}$$

■

We note that the analysis of this section can be extended to prove that in any network G with rate vector r where for some k, $\ell_e = a_e x^k + b_e$ for each e, $(G, r, \ell) \leq (1 - k \cdot (k+1)^{-(k+1)/k})^{-1}$. The example at the beginning of Section 3 shows that this result is tight.

Acknowledgements

We would like to 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 numerous insightful conversations about possible extensions of our results, and for his interest in our work.

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

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

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

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

[6] 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.

[7] J. Feigenbaum, C. Papadimitriou, and S. Shenker. Sharing the cost of multicast transmissions. In *Proceedings of the 32nd Annual ACM Symposium on the Theory of Computing*, 2000. To appear.

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

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

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

[11] Y. A. Korlis and A. A. Lazar. On the existence of equilibria in noncooperative optimal flow control. *Journal of the ACM*, 42(3):584–613, 1995.

[12] Y. A. Korlis, A. A. Lazar, and A. Orda. Achieving network optima using Stackelberg routing strategies. *IEEE/ACM Transactions on Networking*, 5(1):161–173, 1997.

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

[14] Y. A. Korlis, A. A. Lazar, and A. Orda. Avoiding the Braess paradox in noncooperative networks. *Journal of Applied Probability*, 42(3):309–325, 1999.

[15] 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.

[16] 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.

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

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

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

[20] S. J. Shenker. Making greed work in networks: A game-theoretic analysis of switch service disciplines. *IEEE/ACM Transactions on Networking*, 3(6):819–831, 1995.

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

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

[23] 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.

[24] Z. Zhang and C. Douligeris. Convergence of synchronous and asynchronous greedy algorithms in a multiclass telecommunications environment. *IEEE/ACM Transactions on Communcation*, 40:1277–1281, 1992.

A Extensions

In this section we extend the basic model in several ways, as the model of flow considered thus far suffers from several drawbacks. First, in practice agents cannot evaluate path latency exactly, only approximately. Subsection A.1 extends the notion of a flow at Nash equilibrium and Theorem 6 to this setting. Second, our basic model represents a scenario with infinitely many agents each controlling an infinitesimal amount of flow, while in practice we expect to encounter a finite number of agents, each controlling a strictly positive amount of flow. In subsection A.2 we prove an analogue of Theorem 6 for the case of finitely many agents, provided each agent can route its flow fractionally over any number of paths. In subsection A.3 we show that such an assumption is essentially necessary, in that no bicriteria result of the type of Theorem 6 holds when there are only finitely many agents, each of whom must route its flow on a single path.

A.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 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 an obvious modification of Definition 2. The analogue of Lemma 3 is then:

Lemma 10 *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 $f_{P_1} > 0$, $\ell_{P_1}(f) \leq (1+\epsilon)\ell_{P_2}(f)$.*

With minor modifications to the proof of Theorems 6, we can then prove the following result.

Theorem 11 *If f is feasible for (G, r, ℓ) and at ϵ-approximate Nash equilibrium, then for any f^* feasible for $(G, 2r, \ell)$, $C(f) \leq (1+\epsilon)C(f^*)$.*

A.2 Finitely Many Agents: Splittable Flow

Our basic model makes the unrealistic assumption that flow is comprised of infinitely many independent agents. 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, $f^{(i)} : \mathcal{P}_i \to \mathcal{R}^+$ for 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 $j \neq i$. It follows from results of Rosen [?] that such a flow exists (and in fact is essentially unique) under mild convexity assumptions, for example when all latency functions are convex.

Our main result for this model is an analogue of Theorem 6, proved using similar ideas.

Theorem 12 *If f is feasible for the finite splittable instance (G, r, ℓ) and at Nash equilibrium, then for any f^* feasible for the finite splittable instance $(G, 2r, \ell)$, $C(f) \leq C(f^*)$.*

Theorem 6 can thus 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.

A.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 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*

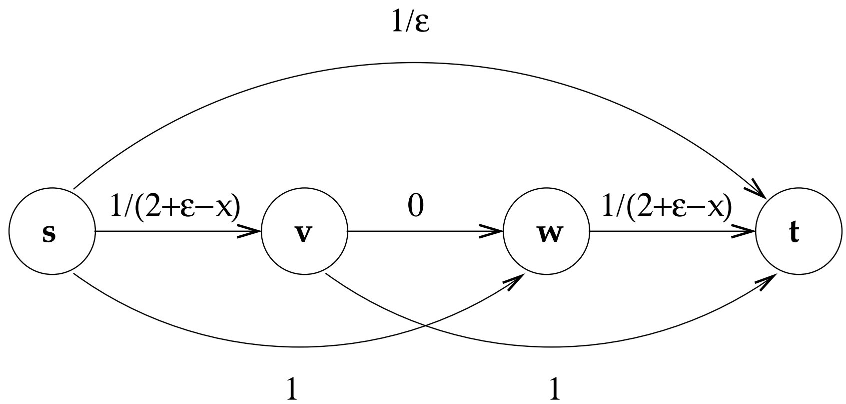


Figure 3: A Bad Example for Unsplittable Flow

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 3, 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 $s \to v \to t$ and the other $s \to w \to t$; the cost of this solution is less than 4 (for any $\epsilon > 0$). On the other hand, a solution with one agent choosing path $s \to v \to w \to t$ and the other routing on the $s \to t$ link is a flow at Nash equilibrium with cost greater than $\frac{1}{\epsilon}$; by choosing ϵ arbitrarily small this cost is arbitrary 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, this example easily extends to show that bicriteria statements analogous to the theorems of Sections 3 and A.2 are false when we require agents to route flow unsplittably. In particular, for any 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 $\frac{1}{2+\epsilon-x}$ and 0, a direct s-t link with constant latency function $\frac{1}{\epsilon}$, and arcs from s to v_{2i} and from v_{2i-1} to t with constant latency functions $\ell(x) = 1$ (e.g., $q = 1$ in Figure 3). 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 $s \to v_1 \to t$, the last $s \to v_{2q} \to t$, and otherwise the ith agent uses path $s \to v_{2i-2} \to v_{2i-1} \to t$). 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 as functions of the form $\ell(x) = 1/(u-x)$ have been used in several different models considered in the literature [13, 14, 16, 18] with the intention of modeling a link with capacity u. However, in networks where the largest possible change in edge latency resulting from a single agent rerouting its flow is bounded above, we can apply the results of Section A.1 to derive the following.

Theorem 13 *Suppose f is at Nash equilibrium in the finite unsplittable instance (G, ℓ, r), and for some $\alpha \geq 1$, $\ell_e(x + r_i) \leq \alpha \cdot \ell_e(x)$ for all $i \in \{1, \ldots, k\}, e \in E, x \in [0, \sum_{j \neq i} r_j]$. Then for any f^* feasible for $(G, 2r, \ell)$, $C(f) \leq \alpha \cdot C(f^*)$.*

For example, in an instance with linear latency functions (say $\ell_e(f_e) = a_e f_e + b_e$) with $b_e > 0$ for all edges e, we may apply Theorem 13 with $\alpha = 1 + \max_i r_i \cdot \max_e a_e / b_e$.