How Bad is Selfish Routing?

Tim Roughgarden* Éva Tardos†

December 22, 2000

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.

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, the common latency suffered by all traffic on the link 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

*Department of Computer Science, Cornell University, Ithaca NY 14853. 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.

†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.

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 [5]. 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 [28]. 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" [12, 28]. We are then interested in comparing the total latency of a Nash equilibrium with that of the optimal assignment of traffic to paths.

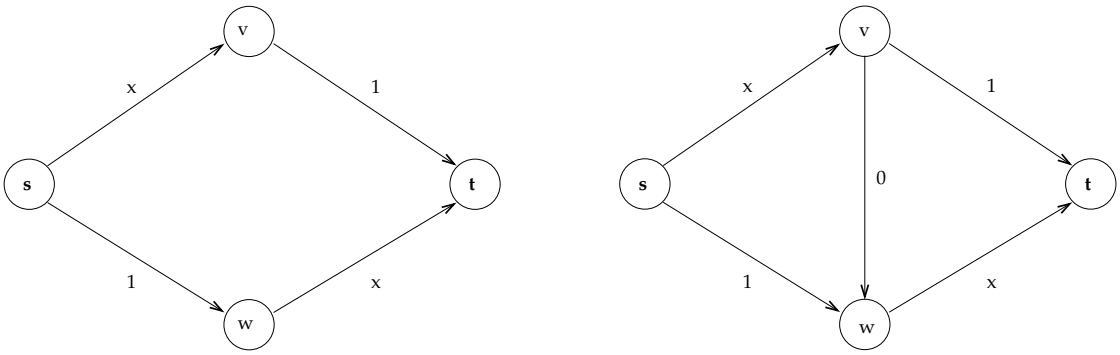
This line of research was recently initiated by Koutsoupias and Papadimitriou [22], who both considered network routing as a non-cooperative game (although in a different model than ours, and only for two-node networks) and proposed the worst-case ratio of the social welfare (suitably defined) achieved by a Nash equilibrium and by a socially optimal set of strategies as a measure of the performance degradation caused by a lack of regulation. As articulated in [22], 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; 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 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 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, 15, 36]), we show that a flow at Nash equilibrium has total latency at most $\frac{4}{3}$ 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



(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.

bicriteria results; in particular, we compare the total latency of a Nash equilibrium flow with that of an optimal flow that routes *additional 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.

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 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 [3, 37] (see also Knight [18]). 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 [11] were perhaps the first authors interested in computing the equilibrium efficiently, and many subsequent papers gave increasingly efficient methods for computing equilibria (see [14] for a survey); others have extended these results to more sophisticated models (see for example [1, 10, 14, 17, 25, 26, 31, 33, 34]).

In the past several decades, much of the work on this model has been inspired by a "paradox" first discovered by Braess [6] and later reported by Murchland [24] (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!

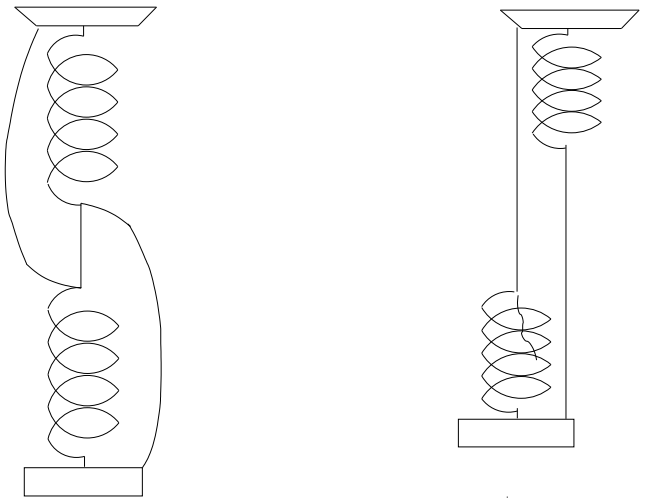
After Braess's Paradox was discovered (together with evidence of similarly counterintuitive and counterproductive traffic behavior following the construction of new roads in congested cities [19, 24]), researchers investigated the sensitivity of traffic equilibria to the properties of the underlying network [16], classified network topologies in which the addition of new links could degrade network performance [15, 36], and discovered new types of "paradoxes" [13, 35]. More recently, several papers [20, 21, 27] have investigated 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. [27], while Korlis et al. [20, 21] study strategies for adding new edges and/or capacity to a network that guarantee an improvement in network performance. To the best of our knowledge, however, no previous work has attempted to quantify the difference in social welfare between equilibrium and optimal traffic flows.

Finally, the recent paper of Koutsoupias and Papadimitriou [22] is quite similar in spirit to ours, although their model is fairly different. In [22], 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. 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 use *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 the model of [22], 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.

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 string 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.

On the other hand, 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



(a) Before (b) After

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

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

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. Section 5 considers several extensions to the basic model.

2 Preliminaries

2.1 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 ℓ_e is nonnegative, differentiable[1], and nondecreasing. We will call the triple (G, r, ℓ) an *instance*.

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

The latency of a path P with respect to a flow f is defined as the sum of 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, 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 $C(f) = \sum_{e \in E} \ell_e(f_e) f_e$.

2.2 Flows at Nash Equilibrium

We wish to consider flows that represent an equilibrium among many non-cooperative agents—i.e., flows that behave "greedily" or "selfishly", without regard to their 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 in G 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 $\ell_{P_1}(f) \leq \ell_{P_2}(\tilde{f})$, 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 [17] or Wardrop's Principle [35, 36] in the literature, due to an influential paper of Wardrop [37].

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

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 2.3 *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) [28]. However, since in our model each agent carries a negligible fraction of the overall traffic, these two definitions are essentially equivalent (see [17] for a rigorous argument).

2.3 A Characterization of Optimal Flows via Non-Linear Programming

We now investigate the properties of an optimal flow—i.e., 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

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

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 P is at most the marginal cost of increasing flow along any other s_i-t_i path. Finally, since the local and global minima of a convex function on a convex set coincide (see, e.g., [29, 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 (e.g., when each edge latency function is convex).

The following lemma formalizes the preceding discussion. Letting $c'_P(f) = \sum_{e \in P} c'_e(f_e)$, and applying the Karush-Kuhn-Tucker Theorem (see, e.g., [29]) to a convex program of the form (NLP), we obtain the following characterization of optimal flows (for a careful derivation see [3, 11]).

Lemma 2.4 *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 $f_{P_1} > 0$, $c'_{P_1}(f) \leq c'_{P_2}(f)$.*

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*. More concretely, consider a minimum-latency flow f^* for a convex program of the form (NLP). The flow f^* satisfies the conditions of Lemma 2.4, and so by Lemma 2.2 can be regarded as a flow at Nash equilibrium with respect to latency functions c'.

Now consider the special case of (NLP) where $c_e(f_e)$ has the form $\ell_e(f_e)f_e$ for each edge e. Denoting the marginal cost of increasing flow on edge e by $\ell_e^*(f_e) = (\ell_e(f_e)f_e)' = \ell_e(f_e) + \ell_e'(f_e)f_e$, we see that any flow f^* at Nash equilibrium with respect to latency functions ℓ^* is optimal with respect to the original latency functions ℓ. Interpreting ℓ_e^* as a function with one term capturing per-unit latency and a second term accounting for the degradation in total latency due to increased congestion, we may say that the only essential difference between an optimal flow and a flow at Nash equilibrium is that the former conscientiously 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. [3] 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. We include a proof for completeness.

Lemma 2.5 ([3]) *A network G 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

$$
\begin{aligned}
&\text{Min} && \sum_{e \in E} h_e(f_e) && \\
&\text{subject to:} && && \\
(NLP2) \qquad & && \sum_{P \in \mathcal{P}_i} f_P = r_i && \forall i \in \{1, \ldots, k\} \\
& && f_e = \sum_{P \in \mathcal{P}: e \in P} f_P && \forall e \in E \\
& && f_P \geq 0 && \forall P \in \mathcal{P}
\end{aligned}
$$

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) $C(f) = C(\tilde{f})$. ■

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.

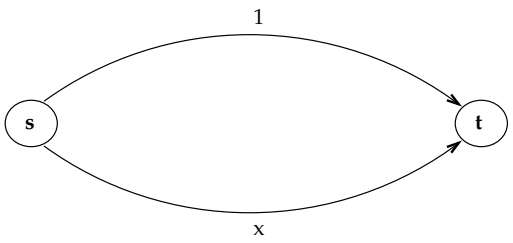


Figure 3: A Simple Bad Example

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 2.5.

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 3, 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.)

Unfortunately, the ratio can be much worse when non-linear latency functions are allowed. For a positive integer k, we extend the example of Figure 3 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 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:

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

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, the difference between the new cost (w.r.t. $\bar{\ell}$) and the old cost (w.r.t. ℓ) can 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) 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 $\bar{\ell}_P(f^*) \geq L_i(f)$ for any $P \in \mathcal{P}_i$. Thus, the cost of f^* with respect to $\bar{\ell}$ 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:

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

■

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, we can see that Theorem 3.2 is essentially tight for all values of γ. More precisely, for any $\epsilon > 0$ one can take k sufficiently large to obtain an instance where the optimal flow feasible for rate $(1+\gamma) - \epsilon$ has cost strictly less than γ (recall the cost of the flow at Nash equilibrium for the original rates is 1) and the optimal flow feasible for rate $1 + \gamma$ has cost at most $\gamma + \epsilon$.

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 [6, 24], and several subsequent papers focused entirely on this model [15, 36]. In addition, we will see that the mechanical networks of strings and springs mentioned in the Introduction can be modeled as traffic networks with linear latency functions.

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 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 (quadratic) program and thus Lemma 2.4 characterizes its optimal solutions. Also, in the notation of subsection 2.3, if $\ell_e(f_e) = a_e f_e + b_e$, then $\ell_e^*(f_e)$, the marginal cost of increasing flow on e, is simply $2a_e f_e + b_e$. For convenience, we summarize this discussion together with specialized versions of Lemmas 2.2 and 2.4 in the following lemma.

Lemma 4.1 *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 $f_P^* > 0$,

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

As an aside, we note that Lemma 4.1 immediately gives a simple proof of the following nontrivial result regarding networks in which the latency of each edge is proportional to its congestion; other properties of this special case have been investigated in the context of electrical networks [4, 9].

Corollary 4.2 *Let G be a network in which each edge latency function is of the form $\ell_e(f_e) = a_e f_e$. 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) for each $i = 1, \ldots, k$, the common marginal cost of increasing flow on an s_i-t_i flow path of $f/2$ equals the common latency of an s_i-t_i flow path of 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 all s_i-t_i flow paths of f have equal latency (see Lemma 2.2) and that if $\ell_e(f_e) = a_e f_e + b_e$ then the marginal cost of increasing flow on edge e is given by $\ell_e^*(f_e) = 2a_e f_e + b_e$. Thus, $\ell_e^*(f_e/2) = \ell_e(f_e)$ for any edge e and hence $\ell_P^*(f/2) = \ell_P(f)$ for any 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 common marginal cost of increasing flow on an s_i-t_i flow path P of f^*, so that $\ell_P^*(f^*) = L_i^*(f^*)$ for each such P. 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: The marginal cost of increasing flow on each s_i-t_i flow path, $L_i^*(f^*)$, is well-defined for each i by Lemma 4.1(b). If we knew that each L_i^* was nondecreasing in r_i, then routing α additional units of flow from s_i to t_i would cost at least $\alpha \cdot L_i^*(f^*)$ 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 $e \in E$, convexity of the function $\ell_e(f_e) f_e = a_e f_e^2 + b_e f_e$ 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 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 $f_e < f_e^*$).

Thus,

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

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

■

We remark that Lemma 4.4 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 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 w.r.t. $f/2$ and latencies w.r.t. 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, f.

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

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

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 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, $\rho(G, r, \ell) \leq (1 - k \cdot (k+1)^{-(k+1)/k})^{-1}$. The example at the beginning of Section 3 shows that this result is tight.

Consequences for Strings and Springs

We now return to the mechanical networks of strings and springs mentioned 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 Nash equilibrium in a subgraph of G) has total latency at least $\frac{3}{4}$ times that of a Nash equilibrium f in G. By Lemma 2.3, it follows that if the common latency of every flow path of f is L and f^* is a flow at Nash equilibrium in a subgraph of G, then the common latency of every flow path of 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

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 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 in practice we 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. Our results can also 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 [32].

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

Following the method of proof for Theorem 3.1, we can show the following result.

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

Proof: Suppose f, f^* satisfy the hypotheses of the theorem. For $i = 1, \ldots, k$, let $L_i(f)$ be the minimum latency of any s_i-t_i path (w.r.t. f); since f is at ϵ-approximate Nash equilibrium, every s_i-t_i flow path has latency at most $(1+\epsilon)L_i(f)$ and hence $C(f) \leq (1+\epsilon)\sum_i L_i(f)r_i$.

As in the proof of Theorem 3.1, define a new set of latency functions $\bar{\ell}$ by

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

As before, the cost of a flow with respect to $\bar{\ell}$ exceeds its cost with respect to ℓ by at most an additive factor of $C(f)$. The rest of the proof of Theorem 3.1 can be imitated to give the bound $C(f) \leq \frac{1+\epsilon}{1-\epsilon}C(f^*)$; obtaining the better bound of the theorem statement requires the following slightly more involved argument.

First, noting that $\bar{\ell}_e(x) \geq \ell_e(f_e)$ for any e and $x \geq 0$, we may lower bound $C(f^*)$ as follows:

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

Summing over paths instead of edges we obtain

$$\begin{aligned} C(f^*) &\geq \sum_i \sum_{P \in \mathcal{P}_i} \ell_P(f)(f_P^* - f_P) \\ &\geq \sum_i L_i(f) \sum_{P \in \mathcal{P}_i} (f_P^* - f_P) \\ &= \sum_i L_i(f) r_i \\ &\geq \frac{C(f)}{1+\epsilon} \end{aligned}$$

as desired. ■

5.2 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 $j \neq i$. We will focus on the case where for each edge e, $x \cdot \ell_e(x)$ is a (weakly) convex function; under this assumption, results of Rosen [30] 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 proofs of Theorems 3.1 and 5.3. As in previous proofs, evaluating f^* with latency functions $\bar{\ell}$ (rather than ℓ) changes 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), there is an agent i, two paths $P_1, P_2 \in \mathcal{P}_i$, and a $\delta > 0$ such that moving δ units of flow from P_1 to P_2 yields a new flow with cost (w.r.t. $\bar{\ell}$) strictly less than that of f (cf., subsection 2.3). 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 (w.r.t. $\bar{\ell}$) 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 (w.r.t. $\bar{\ell}$) 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 $\delta \cdot \ell_{P_1}(f)$ w.r.t. ℓ (since latency functions are nondecreasing) and a cost (w.r.t. ℓ) of

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

since ℓ_e is nondecreasing and $f_e^{(i)} \leq f_e$ for each 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^* w.r.t. $\bar{\ell}$ exceeds its cost w.r.t. 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 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.

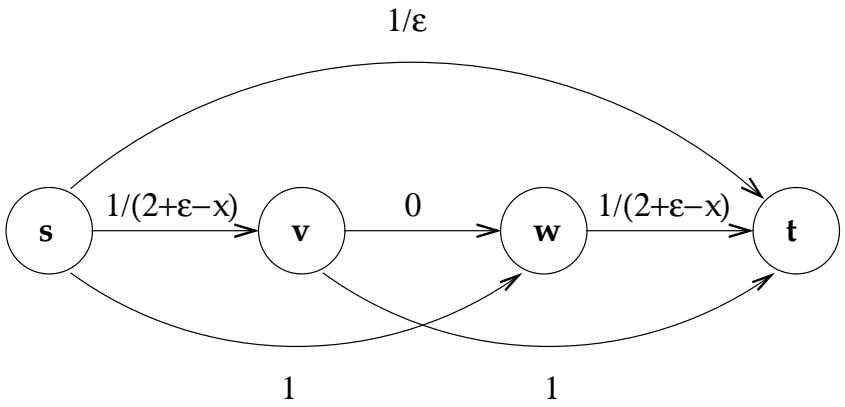


Figure 4: A Bad Example for Unsplittable Flow

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 4, 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 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, this example easily extends to show that bicriteria statements analogous to the Theorems 3.1 and 5.4 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 4). 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 [20, 21, 23, 27] 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 subsection 5.1 to derive the following.

Theorem 5.5 *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 $(G, 2r, \ell)$, $C(f) \leq \alpha \cdot C(f^*)$.*

Proof: We may interpret f and f^* as (fractional) flows feasible for instances (G, r', ℓ) and $(G, 2r', \ell)$ of the original type (i.e., 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 $\ell_e(f_e) = a_e f_e + b_e$) with $b_e > 0$ for all edges e, we may apply Theorem 5.5 with $\alpha = 1 + \max_i r_i \cdot \max_e a_e/b_e$.

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 numerous insightful conversations about possible extensions of our results, and for his interest in our work. Finally, we thank Leslie Ann Goldberg for bringing [2] and [7] to our attention.

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 Economices of Transportation.* Yale University Press, 1956.

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

[5] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobsob, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang. Recommendations on queue management and congestion avoidance in the Internet. Network Working Group Request for Comments 2309, April 1998.

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

[11] 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.

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

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

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

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

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

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

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

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

[32] T. Roughgarden and É. Tardos. The limits of inefficiency of Nash equilibria. In preparation.

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

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

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

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

[37] 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.