MIT Sloan School of Management

Working Paper 4447-03
November 2003

Computational Complexity, Fairness, and the Price of Anarchy of the Maximum Latency Problem

Jose R. Correa, Andreas S. Schulz, Nicolas E. Stier Moses

© 2003 by Jose R. Correa, Andreas S. Schulz, Nicolas E. Stier Moses. All rights reserved.
Short sections of text, not to exceed two paragraphs, may be quoted without
explicit permission, provided that full credit including © notice is given to the source.

This paper also can be downloaded without charge from the
Social Science Research Network Electronic Paper Collection:
http://ssrn.com/abstract=473342

COMPUTATIONAL COMPLEXITY, FAIRNESS, AND THE PRICE OF ANARCHY OF THE MAXIMUM LATENCY PROBLEM

JOSÉ R. CORREA, ANDREAS S. SCHULZ, AND NICOLÁS E. STIER MOSES

Operations Research Center
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
{jcorrea,schulz,nstier}@mit.edu

Abstract. We study the problem of minimizing the maximum latency of flows in networks with congestion. We show that this problem is NP-hard, even when all arc latency functions are linear and there is a single source and sink. Still, one can prove that an optimal flow and an equilibrium flow share a desirable property in this situation: all flow-carrying paths have the same length; i.e., these solutions are "fair," which is in general not true for the optimal flow in networks with nonlinear latency functions. In addition, the maximum latency of the Nash equilibrium, which can be computed efficiently, is within a constant factor of that of an optimal solution. That is, the so-called price of anarchy is bounded. In contrast, we present a family of instances that shows that the price of anarchy is unbounded for instances with multiple sources and a single sink, even in networks with linear latencies. Finally, we show that an s-t-flow that is optimal with respect to the average latency objective is near optimal for the maximum latency objective, and it is close to being fair. Conversely, the average latency of a flow minimizing the maximum latency is also within a constant factor of that of a flow minimizing the average latency.

1. Introduction

We study static network flow problems in which each arc possesses a latency function, which describes the common delay experienced by all flow on the arc as a function of the volume of the arc flow. Load-dependent arc costs have a variety of applications in situations in which one wants to model congestion effects, which are bound to appear, e.g., in communication networks, road traffic, or evacuation problems. In this context, a unit of flow frequently denotes a huge number of "users" (or "agents"), which might represent data packages in the Internet, drivers on a highway system, or individuals fleeing from a building. Depending on the concrete circumstances, the operators of these networks can pursue a variety of system objectives. For instance, they might elect to minimize the average latency, they might aim at minimizing the maximum latency, or they might try to ensure that users between the same origin-destination pair experience essentially the same latency. In fact, the ideal solution would be simultaneously optimal or near optimal with respect to all three objectives.

For linear latencies, we prove the existence of an s-t-flow that is at the same time optimal for two of the three objectives while its average latency is within a factor of 4/3 of that of an optimum. As attractive as this solution might be, we also show that it is NP-hard to compute. Moreover, there is a surprising difference between linear and nonlinear latency functions. Namely, this particular flow remains optimal with respect to the maximum latency and near optimal with respect to the average latency, but it does in general not guarantee that different users face the same latency. However,

Key words and phrases. System Optimum, User Equilibrium, Selfish Routing, Price of Anarchy, Approximation Algorithms, Multicriteria Optimization, Multicommodity Flows.

an optimal s-t-flow for the average latency objective can be computed in polynomial time, and we show that the latency of any one user is within a constant factor of that of any other user. In particular, the maximum latency is within the same constant factor of the maximum latency of an optimal solution to the latter objective. This constant factor only depends on the class of allowable latency functions. For instance, its value is 2 for the case of linear latencies.

Linear latencies are sufficient for certain congestion phenomena to occur. One interesting example is Braess' paradox (1968), which refers to the fact that the addition of an arc can actually increase the (average and maximum) latency in a network in which users act selfishly and independently. This user behavior is captured by the Nash equilibrium of the underlying game in which each user picks a minimal latency path, given the network congestion due to other users (Wardrop 1952). While the inefficiency of this so-called user equilibrium and hence the severity of Braess' paradox had previously been bounded in terms of the average latency, it turns out that it is also bounded with respect to the maximum latency. Indeed, the latencies encountered by different users between the same origin-destination pair are the same. The user equilibrium therefore represents another flow that can be computed in polynomial time and that is optimal or close to optimal for all the three objectives introduced earlier.

The Model. We consider a directed graph $G = (N, A)$ together with a set of source-sink pairs $K \subseteq N \times N$. For each terminal pair $k = (s_k, t_k) \in K$, let $\mathcal{P}_k$ be the set of directed (simple) paths in G from s_k to t_k, and let $d_k > 0$ be the demand rate associated with commodity k. Let $\mathcal{P} := \bigcup_{k\in K} \mathcal{P}_k$ be the set of all paths between terminal pairs, and let $d := \sum_{k\in K} d_k$ be the total demand. A feasible flow f assigns a nonnegative value f_P to every path $P \in \mathcal{P}$ such that $\sum_{P\in\mathcal{P}_k} f_P = d_k$ for all $k \in K$. In the context of single-source single-sink instances, we will drop the subindex k. Each arc a has a load-dependent latency $\ell_a(\cdot)$. We assume that the functions $\ell_a : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are nonnegative, nondecreasing, and differentiable. We define the latency of a path $P \in \mathcal{P}$ under a given flow f as $\ell_P(f) := \sum_{a\in P} \ell_a(\sum_{Q\in\mathcal{P}:Q\ni a} f_Q)$.

The maximum latency of a feasible flow f is $L(f) := \max\{\ell_P(f) : P \in \mathcal{P}, f_P > 0\}$. We call a feasible flow that minimizes the maximum latency a *min-max flow* and denote it by $\hat{f}$. The *maximum latency problem* consists of finding a min-max flow. The average latency of a feasible flow f is defined as $C(f) := \sum_{P\in\mathcal{P}} \ell_P(f) f_P/d$. We refer to the optimal solution with respect to this objective function as the *system optimum* and denote it by f^*. A feasible flow is at *Nash equilibrium* (or is a *user equilibrium*) if for every $k \in K$ and every two paths $P_1, P_2 \in \mathcal{P}_k$ with $f_{P_1} > 0$, $\ell_{P_1}(f) \leq \ell_{P_2}(f)$. In other words, all flow-carrying s_k-t_k-paths have equal (and actually minimal) latency. In particular, equilibrium flows are "fair," i.e., they have unfairness 1, if the *unfairness* of a feasible flow f is defined as $\max_{k\in K} \max\{\ell_{P_1}(f)/\ell_{P_2}(f) : P_1, P_2 \in \mathcal{P}_k, f_{P_1}, f_{P_2} > 0\}$.

Main Results. While a user equilibrium can be computed in polynomial time (Beckmann, McGuire, and Winsten 1956), and so can a system optimum if $x\,\ell_a(x)$ is convex for all arcs $a \in A$, we show in Section 2 that it is an NP-hard problem to compute a min-max flow. This result still holds if all latencies are linear and there is a single source-sink pair. Note that the flows that we are considering are *not* required to be integer, neither on paths nor on arcs.

As pointed out earlier, a Nash equilibrium has unfairness 1 by construction. In Section 3, we establish the somewhat surprising existence of a min-max flow that is fair too, when latencies are linear and there is a single source and a single sink. In addition, although it is well known that system optima are unfair, we provide a tight bound that quantifies the severity of this effect. This bound applies to general multicommodity flows and arbitrary latency functions.

Finally, in Section 4, we show that in the single-source single-sink case under arbitrary latency functions, there actually exist solutions that are simultaneously optimal or near optimal with respect to all three criteria (maximum latency, average latency, and unfairness). In fact, this property is shared by the min-max flow, the system optimum and the user equilibrium, albeit with different

	maximum latency		average latency		unfairness	
min-max flow	1		4/3	Thm. 11	1	Thm. 5
system optimum	2	Thm. 10	1		2	Thm. 6
Nash equilibrium	4/3	Thm. 8	4/3	Thm. 7	1	

TABLE 1. Summary of results for single-source single-sink networks with linear latency functions. The first entry in each cell represents a worst-case bound on the ratio of the value of the flow associated with the corresponding column to the value of an optimal flow for the objective function denoted by the corresponding row. The second entry refers to the theorem in this paper in which the respective result is proved. All bounds are tight, as examples provided after each theorem demonstrate. The bound of 4/3 on the ratio of the average latency of the user equilibrium to that of the system optimum was first proved by Roughgarden and Tardos (2002); we give a simpler proof in Theorem 7. Weitz (2001) observed first that this bound carries forward to the maximum latency objective for the case of only one source and sink; we present a generalization of this observation to multicommodity flows in Theorem 8.

bounds. Table 1 presents the bounds obtained for the three criteria in the single-source single-sink case with linear latencies. An important consequence of these results is that computing a user equilibrium or a system optimum constitutes a constant-factor approximation algorithm for the NP-hard maximum latency problem. On the other hand, already in networks with multiple sources and a single sink, the ratio of the maximum latency of a Nash equilibrium to that of the min-max flow is not bounded by a constant, even with linear latency functions.

Related Work. Most papers on evacuation problems consider constant travel times; we refer the reader to the surveys by Aronson (1989) and Powell, Jaillet, and Odoni (1995) for more details. One notable exception is the work by Köhler and Skutella (2002). They considered a dynamic quickest flow problem with load-dependent transit times, for which they established strong NP-hardness. They also provided an approximation algorithm by considering the average of a flow over time, which is a static flow. Köhler, Langkau, and Skutella (2002) proposed to use time-expanded networks to derive approximation algorithms for a similar problem.

The concept of the price of anarchy, which is the ratio of the performance of a Nash equilibrium to that of an optimal solution, was introduced by Koutsoupias and Papadimitriou (1999) in the context of a game motivated by telecommunication networks. This inspired considerable subsequent work, including Mavronicolas and Spirakis 2001; Koutsoupias, Mavronicolas, and Spirakis 2002; Czumaj and Vöcking 2002; Czumaj, Krysta, and Vöcking 2002. These papers study the maximum latency of transmissions in two-node networks consisting of multiple links connecting a single source with a single sink. Indeed, under certain assumptions, when users are selfish, the maximum latency is not too large compared to the best coordinated solution. Although these results are similar in nature to some of ours, their model is not comparable to ours because they work with a finite number of players and consider mixed strategies. In contrast, in our setting, every player just controls an infinitesimal amount of flow, making mixed strategies irrelevant. Moreover, we work with arbitrary networks. For more details on the various routing games, we refer the reader to the excellent survey by Czumaj (2004).

Roughgarden and Tardos (2002), Roughgarden (2003), Schulz and Stier Moses (2003), and Correa, Schulz, and Stier Moses (2003) studied the price of anarchy with respect to the average travel time in general networks and for different classes of latency functions. In particular, if $\mathcal{L}$ is the set of allowable latency functions, the ratio of the average travel time of a user equilibrium to that of a system optimum is bounded by $\alpha(\mathcal{L})$, where $\alpha(\mathcal{L})$ is a constant that only depends on $\mathcal{L}$. For example, in case $\mathcal{L}$ only contains concave functions, $\alpha(\mathcal{L}) = 4/3$. We will later make use of this result (Section 4).

For the maximum latency objective, Weitz (2001) was the first to observe that the price of anarchy is bounded in single-source single-sink networks. He also presented a family of examples that showed that Nash equilibria can be arbitrarily bad in multiple commodity networks. Roughgarden (2004) gave a tight bound for the single-source single-sink case that depends on the size of the network.

Game-theoretic concepts seem to offer an attractive way of computing approximate solutions to certain hard problems. Indeed, Anshelevich et al. (2003) approximated optimal solutions to a network design problem that is NP-hard with the help of Nash and approximate Nash equilibria. A related idea was used by Fotakis et al. (2002) and Feldmann et al. (2003) to show that although it is hard to find the best and worst equilibrium of the telecommunication game described before, there exists an approximation algorithm for computing a Nash equilibrium with minimal social cost. Correa et al. (2003) pursued the same idea by computing a provably good Nash equilibrium in a setting with multiple equilibria in which computing the best equilibrium is hard.

In the context of Section 3, we should point out that there exist multiple (nonequivalent) definitions of (un)fairness. The definition we use here comes from the competition between different agents in the routing game. Roughgarden (2002) defined unfairness as the ratio of the maximum latency of a system optimum to the latency of a user equilibrium; we later recover the bounds that he obtained. Jahn et al. (2002) considered the definition of unfairness presented here; they looked for flows that minimize the total travel time among those with bounded unfairness.

2. Computational Complexity

In our model, both the system optimum and the Nash equilibrium can be computed efficiently because they represent optimal solutions to certain convex programs. On the other hand, it follows from the work of Köhler and Skutella (2002) on the quickest s-t-flow problem with load-dependent transit times that the maximum latency problem considered here is NP-hard (though not necessarily in NP) when latencies include arbitrary nonlinear functions or when there are explicit arc capacities. Lemma 1 below implies that the general maximum latency problem is in NP, while Theorem 3 establishes its NP-hardness, even in the case of linear latencies and a single source and a single sink.

Note that the following result does not follow from ordinary flow decomposition as it is not clear how to convert a flow on arcs into a path flow such that the latency of the resulting paths remains bounded; in fact, it is a consequence of Theorem 3 that the latter problem is NP-hard, too.

Lemma 1. *Let f be a feasible flow for a multicommodity flow network with load-dependent arc latencies. Then there exists another feasible flow f' such that $L(f') \le L(f)$, and f' uses at most $|A|$ paths for each source-sink pair.*

Proof. Consider an arbitrary commodity $k \in K$. Let $P_1, \dots, P_r$ be s_k-t_k-paths such that $f_{P_i} > 0$ for $i = 1, \dots, r$, and $\sum_{i=1}^{r} f_{P_i} = d_k$. Slightly overloading notation, we let $P_1, \dots, P_r$ also denote the arc incidence vectors of these paths. Let's assume that $r > |A|$. (Otherwise we are done.) Hence, the vectors $P_1, \dots, P_r$ are linearly dependent and $\sum_{i=1}^{r} \lambda_i P_i = 0$ has a nonzero solution. Let's assume without loss of generality that $\lambda_r \neq 0$. We define a new flow f'' (not necessarily feasible) by setting $f''_{P_i} := f_{P_i} - \frac{\lambda_i}{\lambda_r} f_{P_r}$ for $i = 1, \dots, r$, and $f''_P := f_P$ for all other paths P. Notice that under f'', the flow on arcs does not change:

$$\sum_{i=1}^{r} P_i f''_{P_i} = \sum_{i=1}^{r-1} P_i f_{P_i} - \sum_{i=1}^{r-1} \frac{\lambda_i}{\lambda_r} P_i f_{P_r} = \sum_{i=1}^{r} P_i f_{P_i} \ .$$

Here, we used the linear dependency for the last equality. In particular, $L(f'') \le L(f)$. Let us consider a convex combination f' of f and f'' that is nonnegative and uses fewer paths than f. Note that such a flow always exists because $f''_{P_r} = 0$, and the flow on some other paths $P_1, \dots, P_{r-1}$

might be negative. Moreover, $L(f') \leq L(f)$, too. If f' still uses more than $|A|$ paths between s_k and t_k, we can iterate this process so long as necessary to prove the claim. □

Corollary 2. *The recognition version of the maximum latency problem is in NP.*

Proof. Lemma 1 shows the existence of a succinct certificate. Indeed, there is a min-max flow using no more than $|K| \cdot |A|$ paths. □

We are now ready to prove that the maximum latency problem is in fact NP-hard. We present a reduction from Partition:

> Given: A set of n positive integer numbers $q_1, \ldots, q_n$.
> Question: Is there a subset $I \subset \{1, \ldots, n\}$ such that $\sum_{i \in I} q_i = \sum_{i \notin I} q_i$?

Theorem 3. *The recognition version of the maximum latency problem is NP-complete, even when all latencies are linear functions and the network has a single source-sink pair.*

Proof. Given an instance of Partition, we define an instance of the maximum latency problem as follows. The network consists of nodes $0, 1, \ldots, n$ with 0 representing the source and n the sink. The demand is one. For $i = 1, \ldots, n$, the nodes $i-1$ and i are connected with two arcs, namely a_i with latency $\ell_{a_i}(x) = q_i\, x$ and $\tilde{a}_i$ with latency $\ell_{\tilde{a}_i}(x) = q_i$.

Let $L := \frac{3}{4} \sum_{i=1}^{n} q_i$. Notice that the system optimum f^* has cost equal to L and $f^*_a = 1/2$ for all $a \in A$. We claim that the given instance of Partition is a Yes-instance if and only if there is a solution to the maximum latency problem of maximum latency equal to L. Indeed, if there is a partition I, the flow that routes half a unit of flow along the 0-n-path composed of arcs a_i, $i \in I$, and $\tilde{a}_i$, $i \notin I$, and the other half along the complementary path has maximum latency L.

To prove the other direction, assume that we have a flow f of maximum latency equal to L. Therefore, $C(f) \leq L$ (there is unit demand), which implies that $C(f) = L$ (it cannot be better than the optimal solution). As the arc flows of a system optimum are unique, this implies that $f_a = 1/2$ for all $a \in A$. Take any path P such that $f_P > 0$ and partition its arcs such that I contains the indices of the arcs $a_i \in P$. Then, $\frac{3}{4} \sum_{i=1}^{n} q_i = L = \ell_P(f) = \sum_{i \in I} \frac{q_i}{2} + \sum_{i \notin I} q_i$, and subtracting the left-hand side from the right-hand side yields $\sum_{i \in I} \frac{q_i}{4} = \sum_{i \notin I} \frac{q_i}{4}$. □

Corollary 4. *Let f be a (path) flow in an s-t-network with linear latencies. Let $(f_a : a \in A)$ be the associated flow on arcs. Given just $(f_a : a \in A)$ and $L(f)$, it is NP-hard to compute a decomposition of this arc flow into a (path) flow f' such that $L(f') \leq L(f)$. In particular, it is NP-hard to recover a min-max flow even though its arc values are given.*

Note that Corollary 4 neither holds for the system optimum nor the user equilibrium. In both cases any flow derived from an ordinary flow decomposition is indeed an optimal flow respectively equilibrium flow.

Let us finally mention that Theorem 4.3 in Köhler and Skutella (2002) implies that the maximum latency problem is APX-hard when latencies can be arbitrary nonlinear functions or when there are explicit arc capacities.

3. Fairness

User equilibria are fair by definition. Indeed, all flow-carrying paths between the same source and sink have equal latency. The next result establishes the same property for min-max s-t-flows in the case of linear latencies. Namely, a fair min-max flow always exists. Therefore, the difference between a Nash equilibrium and a min-max flow is that the latter may leave paths unused that are shorter than the ones carrying flow, a situation that cannot happen in equilibrium. This result is not true for nonlinear latencies, as we shall see later.

Theorem 5. *Every instance of the single-source single-sink maximum latency problem with linear latency functions has an optimal solution that is fair.*

Proof. Consider an instance with demand d and latency functions $\ell_a(f_a) = q_a f_a + r_a$, for $a \in A$. Among all min-max flows, let $\hat{f}$ be the one that uses the smallest number of paths. Let $P_1, P_2, \ldots, P_k$ be these paths. Consider the following linear program:

$$\min \quad z \tag{1a}$$

$$\text{s.t.} \quad \sum_{a \in P_i} \Big(q_a \big(\sum_{P_h \ni a} f_{P_h} \big) + r_a \Big) \le z \qquad \text{for } i = 1, \ldots, k, \tag{1b}$$

$$\sum_{i=1}^{k} f_{P_i} = d \tag{1c}$$

$$f_{P_i} \ge 0 \qquad \text{for } i = 1, \ldots, k. \tag{1d}$$

Note that this linear program has $k + 1$ variables. Furthermore, by construction, it has a feasible solution with $z = L(\hat{f})$, and there is no solution with $z < L(\hat{f})$. Therefore, an optimal basic feasible solution gives a min-max flow that satisfies with equality k of the inequalities (1b) and (1d). As $f_{P_i} > 0$ for all i because of the minimality assumption, all inequalities (1b) have to be tight. □

A byproduct of this proof is that an arbitrary flow can be transformed into a fair one without increasing its maximum latency. In fact, just solve the corresponding linear program. An optimal basic feasible solution will either be fair or it will use fewer paths. In the latter case, eliminate all paths carrying zero flow and repeat until a fair solution is found.

Notice that the min-max flow may not be fair for nonlinear functions. Indeed, the instance displayed in Figure 1 features high unfairness with latencies that are polynomials of degree p.

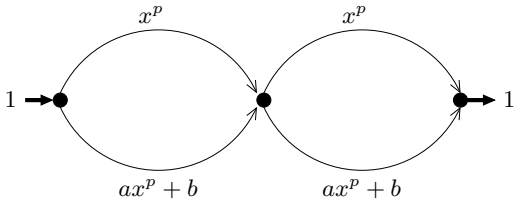

FIGURE 1. Instance with nonlinear latencies illustrating that fair min-max flows may not exist.

When $a = (1+\varepsilon)^{p-1}$ and $b = 2 - \left(\frac{1+\varepsilon}{2+\varepsilon}\right)^{p-1} - \delta$ for some $\varepsilon > 0$ and $\delta > 0$ such that $b > 1$, the min-max flow routes $\frac{1}{2+\varepsilon}$ units of flow along the "top-bottom" and "bottom-top" paths, respectively, and $\frac{\varepsilon}{2+\varepsilon}$ units of flow along the "top-top" path. It is not hard to see that this flow is optimal. Indeed, the "bottom-bottom" path is too long to carry any flow. Moreover, by symmetry, the "top-bottom" and "bottom-top" paths have to carry the same amount of flow. Therefore, the optimal solution can be computed by solving a one-dimensional minimization problem, whose only variable is the amount x of flow on the "top-top" path. The unique optimal solution to this problem is $x = \frac{\varepsilon}{2+\varepsilon}$.

Let us compute the unfairness of this solution. The "top-top" path has latency equal to $2\left(\frac{1+\varepsilon}{2+\varepsilon}\right)^p$, which tends to $\left(\frac{1}{2}\right)^{p-1}$ as $\varepsilon \to 0$. The latency of the other two paths used by the optimum is equal to $2 - \delta$. Therefore, the unfairness of this min-max flow is arbitrarily close to 2^p.

A typical argument against using the system optimum in the design of route-guidance devices for traffic assignment is that, in general, it assigns some drivers to unacceptably long paths in order

to use shorter paths for most other drivers; see, e.g., Beccaria and Bolelli (1992). The following theorem quantifies the severity of this effect by characterizing the unfairness of the system optimum. It turns out that there is a relation to earlier work by Roughgarden (2002), who compared the maximum latency of a system optimum in a single-sink single-source network to the latency of a user equilibrium. He showed that for a given class of latency functions $\mathcal{L}$, this ratio is bounded from above by $\gamma(\mathcal{L})$, which is defined to be the smallest value that satisfies $\ell_a^*(x) \le \gamma(\mathcal{L})\ell_a(x)$ for all $\ell \in \mathcal{L}$ and all $x \ge 0$. Here, $\ell_a^*(x) := \ell_a(x) + x\,\ell_a'(x)$ is the function that makes a system optimum for the original instance a user equilibrium of an instance in which the latencies are replaced by ℓ^* (Beckmann, McGuire, and Winsten 1956). For instance, $\gamma(\text{polynomials of degree } p) = p+1$. We prove that the unfairness of a system optimum is in fact bounded by the same constant, even for general instances with multiple commodities. The same result was independently obtained by Roughgarden (personal communication, October 2003).

Theorem 6. *Let f^* be a system optimum in a multicommodity flow network with arc latency functions drawn from a class $\mathcal{L}$. Then, the unfairness of f^* is bounded from above by $\gamma(\mathcal{L})$.*

Proof. We will prove the result for the single-source single-sink case. The extension to the general case is straightforward. As a system optimum is a user equilibrium with respect to latencies ℓ^*, there exists L^* such that $\ell_P^*(f^*) = L^*$ for all paths $P \in \mathcal{P}$ with $f_P^* > 0$. From the definitions of ℓ^* and $\gamma(\mathcal{L})$, we have that $\ell_a(x) \le \ell_a^*(x) \le \gamma(\mathcal{L})\ell_a(x)$ for all x. Let $P_1, P_2 \in \mathcal{P}$ be two arbitrary paths with $f_{P_1}^*, f_{P_2}^* > 0$. Hence, $\ell_{P_1}(f^*) \le L^*$ and $\ell_{P_2}(f^*) \ge L^*/\gamma(L)$. It follows that $\ell_{P_1}(f^*)/\ell_{P_2}(f^*) \le \gamma(L)$. □

Notice that Theorem 6 implies Roughgarden's earlier bound for the single-source single-sink case. Indeed, for a Nash equilibrium f, $\min\{\ell_P(f^*) : P \in \mathcal{P}, f_P^* > 0\} \le \min\{\ell_P(f) : P \in \mathcal{P}, f_P > 0\}$. Otherwise, $C(f^*) > C(f)$, which contradicts the optimality of f^*. In addition, the example shown in Figure 2 proves that the bound given in Theorem 6 is tight. Indeed, it is easy to see that the system optimum routes half of the demand along each arc, implying that the unfairness is $\ell^*(d/2)/\ell(d/2)$. Taking the supremum of that ratio over $d \ge 0$ and $\ell \in \mathcal{L}$, we get $\gamma(\mathcal{L})$.

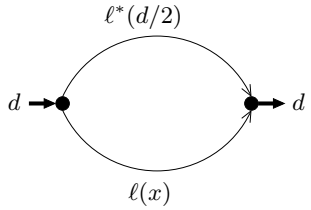

FIGURE 2. Instance showing that Theorem 6 is tight.

4. Price of Anarchy and Related Approximation Results

Nash equilibria in general and user equilibria in particular are known to be inefficient, as evidenced by Braess' paradox (1968). Koutsoupias and Papadimitriou (1999) suggested measuring this degradation in performance, which results from the lack of central coordination, by the worst-case ratio of the value of an equilibrium to that of an optimum. This ratio has become known as the "price of anarchy," a phrase coined by Papadimitriou (2001). It is quite appealing (especially for evacuation situations) that in the routing game considered here, the price of anarchy is small; i.e., the selfishness of users actually drives the solution close to optimality. Recall that the user equilibrium results from everybody choosing a shortest path under the prevailing congestion conditions. Since a user equilibrium can be computed in polynomial time, this also leads to an approximation algorithm for the maximum latency problem.

In order to derive a bound on the price of anarchy for the maximum latency objective, we use a corresponding bound for the average latency of Nash equilibria, which was first proved for linear latency functions by Roughgarden and Tardos (2002) and then extended to different classes of latency functions by Roughgarden (2003) and Correa, Schulz, and Stier Moses (2003). For the sake of completeness, let us include a simpler proof of Roughgarden and Tardos' result (see also Correa et al. 2003).

Theorem 7 (Roughgarden and Tardos 2002)**.** *Let f be a user equilibrium and let f^* be a system optimum in a multicommodity flow network with linear latency functions. Then $C(f) \le \frac{4}{3} C(f^*)$.*

Proof. Let $\ell_a(x) = q_a x + r_a$ with $q_a, r_a \ge 0$ for all $a \in A$. Then,

$$C(f) = \sum_{a\in A}(q_a f_a + r_a) f_a \le \sum_{a\in A}(q_a f_a + r_a) f_a^* \le \sum_{a\in A}(q_a f_a^* + r_a) f_a^* + \frac{1}{4}\sum_{a\in A} q_a f_a^2 \le C(f^*) + \frac{1}{4} C(f) \ .$$

The first inequality holds since the equilibrium flow f uses shortest paths with respect to the arc latencies caused by itself. The second inequality follows from $(f_a^* - f_a/2)^2 \ge 0$. □

In general,

$$C(f) \le \alpha(\mathcal{L}) C(f^*), \text{ where } \alpha(\mathcal{L}) := \Big(1 - \sup_{\ell\in\mathcal{L},\, 0\le x\le d} \Big\{ \frac{x\big(\ell(d) - \ell(x)\big)}{d\,\ell(d)} \Big\}\Big)^{-1} \ , \tag{2}$$

and the proof is similar to the one given above for Theorem 7; see Roughgarden (2003) and Correa et al. (2003) for details. For polynomials with nonnegative coefficients of degree 2, $\alpha(\mathcal{L})$ equals 1.626; for those with degree 3, $\alpha(\mathcal{L}) = 1.896$; in general, $\alpha(\mathcal{L}) = \Theta(p/\ln p)$ for polynomials of degree p.

It was first noted by Weitz (2001) that in networks with only one source and one sink, any upper bound on the price of anarchy for the average latency is an upper bound on the price of anarchy for the maximum latency. We include a multicommodity version of this result.

Theorem 8. *Consider a multicommodity flow network with latency functions in $\mathcal{L}$. Let f be a Nash equilibrium and $\hat{f}$ a min-max flow. For each commodity $k \in K$, $L_k(f) \le \frac{d}{d_k}\alpha(\mathcal{L}) L(\hat{f})$, where L_k is the maximum latency incurred by commodity k, d_k is its demand rate, and d is the total demand.*

Proof. Let f^* be the system optimum. Then,

$$d_k L_k(f) \le d\, C(f) \le d\, \alpha(\mathcal{L}) C(f^*) \le d\, \alpha(\mathcal{L}) C(\hat{f}) \le d\, \alpha(\mathcal{L}) L(\hat{f}) \ .$$

Here, the first inequality holds because f is a Nash equilibrium, the second inequality is exactly Equation (2), the third one comes from the optimality of f^*, and the last one just says that the average latency is less than the maximum latency. □

This implies that, for the single-source single-sink case, computing a Nash equilibrium is an $\alpha(\mathcal{L})$-approximation algorithm for the maximum latency problem. Notice that this guarantee is

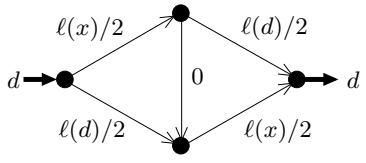

FIGURE 3. Instance showing that Theorem 8 is tight for single-commodity networks.

tight as shown by the example given in Figure 3, which goes back to Braess (1968). Indeed, the

latency of a Nash equilibrium is $\ell(d)$ while the maximum latency of a min-max flow, which coincides with the system optimum, is

$$\ell(d) - \max_{0 \le x \le d} \left\{ \frac{x}{d} \left(\ell(d) - \ell(x) \right) \right\} \quad .$$

Taking the supremum over $d \geq 0$ and $\ell \in \mathcal{L}$, the ratio of the latency of the Nash equilibrium to that of the min-max flow is arbitrarily close to $\alpha(\mathcal{L})$.

For instances with multiple sources and a single sink, the maximum latency of a user equilibrium is unbounded with respect to that of an optimal solution, even with linear latencies. In fact, we will show that the price of anarchy cannot be better than $\Omega(n)$, where n is the number of nodes in the network. Note that this also implies that the price of anarchy is unbounded in single-source single-sink networks with explicit arc capacities. Weitz (2001) showed that the price of anarchy is unbounded in the case of multiple commodities, and Roughgarden (2004) proved that it is bounded by $n - 1$ if there is a common source and sink.

Theorem 9. *The price of anarchy in a single-commodity network with multiple sources and a single sink is $\Omega(n)$, even if all latencies are linear functions.*

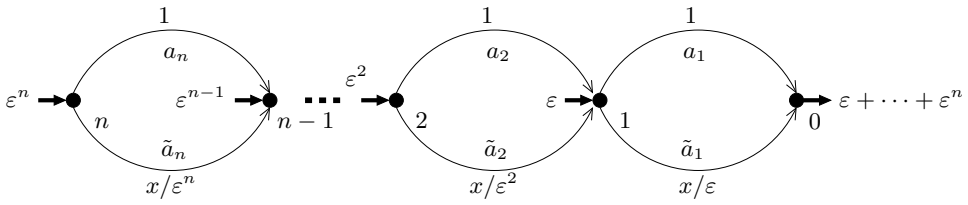

FIGURE 4. Instance showing that Nash equilibria can be arbitrarily bad when multiple sources are present.

Proof. Fix a constant $\varepsilon > 0$ and consider the instance presented in Figure 4. Nodes $n, n-1, \ldots, 1$ are the sources while node 0 is the sink. Nodes i and $i-1$ are connected with two arcs: a_i with constant latency equal to 1 and $\tilde{a}_i$ with latency equal to x/ε^i. Let the demand entering node $i > 0$ be ε^i. The user equilibrium of this instance routes the flow along paths of the form $\tilde{a}_i, a_{i-1}, \ldots, a_1$ and has maximum latency n. To show the claim, it suffices to exhibit a good solution. For instance, for origin i, let its demand flow along the path $a_i, \tilde{a}_{i-1}, \ldots, \tilde{a}_1$. Under this flow, the load of $\tilde{a}_i$ is equal to $\varepsilon^{i+1} + \cdots + \varepsilon^n$ and its traversal time is $(\varepsilon^{i+1} + \cdots + \varepsilon^n)/\varepsilon^i = \varepsilon^1 + \cdots + \varepsilon^{n-i}$. Hence, we can bound the maximum latency from above by $1 + \frac{n\varepsilon}{1-\varepsilon}$, which tends to 1 when $\varepsilon \to 0$. □

In the single-source single-sink case, not only Nash equilibria represent good approximations to the maximum latency problem; an immediate corollary of Theorem 6 is that system optima are also close to optimality with respect to the maximum latency objective.

Theorem 10. *For single-source single-sink instances with latency functions drawn from $\mathcal{L}$, computing a system optimum is a $\gamma(\mathcal{L})$-approximation algorithm for the maximum latency problem.*

Proof. Theorem 6 states that the length of a longest path used by the system optimum f^* is at most $\gamma(\mathcal{L})$ times the length of a shortest flow-carrying path. The latter value cannot be bigger than the maximum latency of a path used by the min-max flow because f^* is optimal for the average latency; the result follows. □

The bound given in Theorem 10 is best possible. To see this, consider the instance depicted in Figure 5. The min-max flow routes the entire demand along the lower arc, for a small enough $\varepsilon > 0$. On the other hand, the unique system optimum has to satisfy $\ell^*(f^*) = \ell^*(d) - \varepsilon$, where f^* is

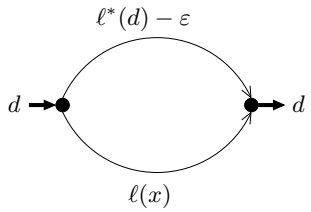

FIGURE 5. Instance showing that Theorem 10 is tight.

the flow along the lower arc. Therefore, the upper arc has positive flow and the maximum latency is $\ell^*(d) - \varepsilon$. The ratio between the maximum latencies of the two solutions is arbitrarily close to $\ell^*(d)/\ell(d)$. Taking the supremum over $d \geq 0$ and $\ell \in \mathcal{L}$ shows that the bound in Theorem 10 is tight.

To complete Table 1, let us prove that the average latency of the min-max flow is not too far from that of the system optimum.

Theorem 11. *Let $\hat{f}$ be a min-max flow and let f^* be a system optimum for an instance with a single source, a single sink and latencies drawn from $\mathcal{L}$. Then, $C(\hat{f}) \leq \alpha(\mathcal{L})C(f^*)$.*

Proof. Note that $C(\hat{f}) \leq L(\hat{f}) \leq L(f) = C(f) \leq \alpha(\mathcal{L})C(f^*)$, where f is the Nash equilibrium of the instance. □

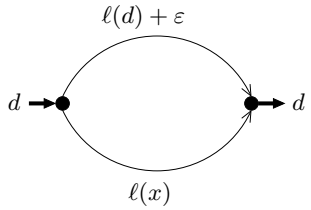

FIGURE 6. Instance showing that Theorem 11 is tight.

Again, the guarantee given in the previous theorem is tight. To show this, it is enough to note that the equilibrium flow and the min-max flow coincide in the example of Figure 6, and their average latency is $\ell(d)$. Moreover, the average latency of the system optimum is arbitrary close to

$$\ell(d) - \max_{0 \leq x \leq d} \left\{ \frac{x}{d} \left(\ell(d) - \ell(x) \right) \right\} \quad .$$

Taking the supremum of the ratio of these two values over $d \geq 0$ and $\ell \in \mathcal{L}$ completes the argument.

In Table 2, we summarize the findings for single-source single-sink networks with latencies drawn from a given class $\mathcal{L}$ of allowable latency functions.

	maximum latency	average latency	unfairness
min-max flow	1	$\alpha(\mathcal{L})$	?
system optimum	$\gamma(\mathcal{L})$	1	$\gamma(\mathcal{L})$
user equilibrium	$\alpha(\mathcal{L})$	$\alpha(\mathcal{L})$	1

TABLE 2. Overview of approximation guarantees for single-source single-sink networks when latencies belong to a given set $\mathcal{L}$. All bounds are tight. The "?" indicates that no upper bound is known; recall from the example depicted in Figure 1 that 2^p is a lower bound for polynomials of degree p.

References

Anshelevich, E., A. Desgupta, É. Tardos, and T. Wexler (2003). Near-optimal network design with selfish agents. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC)*, San Diego, CA, pp. 511–520. ACM Press, New York, NY.

Aronson, J. E. (1989). A survey of dynamic network flows. *Annals of Operations Research 20*, 1–66.

Beccaria, G. and A. Bolelli (1992). Modelling and assessment of dynamic route guidance: the MARGOT project. In *Proceedings of the IEEE Vehicle Navigation & Information Systems Conference*, Oslo, Norway, pp. 117–126.

Beckmann, M. J., C. B. McGuire, and C. B. Winsten (1956). *Studies in the economics of transportation*. Yale University Press, New Haven, CT.

Braess, D. (1968). Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung 12*, 258–268.

Correa, J. R., A. S. Schulz, and N. E. Stier Moses (2003). Selfish routing in capacitated networks. MIT, Sloan School of Management, Working Paper No. 4319-03.

Czumaj, A. (2004). Selfish routing on the Internet. In J. Leung (Ed.), *Handbook of scheduling: algorithms, models, and performance analysis*. CRC Press, Boca Raton, FL. To appear.

Czumaj, A., P. Krysta, and B. Vöcking (2002). Selfish traffic allocation for server farms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, Montreal, Canada, pp. 287–296. ACM Press, New York, NY.

Czumaj, A. and B. Vöcking (2002). Tight bounds for worst-case equilibria. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, CA, pp. 413–420. SIAM, Philadelphia, PA.

Feldmann, R., M. Gairing, T. Lücking, B. Monien, and M. Rode (2003). Nashification and the coordination ratio for a selfish routing game. In J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger (Eds.), *Proceedings of the 30th International Colloquium on Automata, Languages, and Programming (ICALP)*, Eindhoven, The Netherlands, Volume 2719 of *Lecture Notes in Computer Science*, pp. 514–526. Springer, Berlin.

Fotakis, D., S. Kontogiannis, E. Koutsoupias, M. Mavronicolas, and P. Spirakis (2002). The structure and complexity of Nash equilibria for a selfish routing game. In P. Widmayer, F. Triguero, R. Morales, M. Hennessy, S. Eidenbenz, and R. Conejo (Eds.), *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming (ICALP)*, Málaga, Spain, Volume 2380 of *Lecture Notes in Computer Science*, pp. 123–134. Springer, Berlin.

Jahn, O., R. H. Möhring, A. S. Schulz, and N. E. Stier Moses (2002). System-optimal routing of traffic flows with user constraints in networks with congestion. MIT, Sloan School of Management, Working Paper No. 4394-02.

Köhler, E., K. Langkau, and M. Skutella (2002). Time-expanded graphs with flow-dependent transit times. In R. H. Möhring and R. Raman (Eds.), *Proceedings of the 10th Annual European Symposium on Algorithms (ESA)*, Rome, Italy, Volume 2461 of *Lecture Notes in Computer Science*, pp. 599–611. Springer, Berlin.

Köhler, E. and M. Skutella (2002). Flows over time with load-dependent transit times. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, CA, pp. 174–183. SIAM, Philadelphia, PA.

Koutsoupias, E., M. Mavronicolas, and P. Spirakis (2002). Approximate equilibria and ball fusion. In *Proceedings of the 9th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, Andros, Greece. Carleton Scientific, Ottawa, Canada.

Koutsoupias, E. and C. H. Papadimitriou (1999). Worst-case equilibria. In C. Meinel and S. Tison (Eds.), *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, Trier, Germany, Volume 1563 of *Lecture Notes in Computer Science*, pp. 404–413. Springer, Berlin.

Mavronicolas, M. and P. Spirakis (2001). The price of selfish routing. In *Proceedings of the 33th Annual ACM Symposium on Theory of Computing (STOC)*, Hersonissos, Greece, pp. 510–519. ACM Press, New York, NY.

Papadimitriou, C. H. (2001). Algorithms, games, and the Internet. In *Proceedings of the 33th Annual ACM Symposium on Theory of Computing (STOC)*, Hersonissos, Greece, pp. 749–753. ACM Press, New York, NY.

Powell, W. B., P. Jaillet, and A. Odoni (1995). Stochastic and dynamic networks and routing. In M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser (Eds.), *Networks*, Volume 4 of *Handbook in Operations Research and Management Science*, pp. 141–295. Elsevier Science, Amsterdam.

Roughgarden, T. (2002). How unfair is optimal routing? In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, San Francisco, CA, pp. 203–204. SIAM, Philadelphia, PA.

Roughgarden, T. (2003). The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences 67*, 341–364.

Roughgarden, T. (2004). The maximum latency of selfish routing. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, New Orleans, LA. SIAM, Philadelphia, PA. To appear.

Roughgarden, T. and É. Tardos (2002). How bad is selfish routing? *Journal of the ACM 49*, 236–259.

Schulz, A. S. and N. E. Stier Moses (2003). On the performance of user equilibria in traffic networks. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Baltimore, MD, pp. 86–87. SIAM, Philadelphia, PA.

Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers 1*, Part II, 325–378.

Weitz, D. (2001). The price of anarchy. Unpublished manuscript.