MIT Sloan School of Management

Working Paper 4274-02
December 2002

On The Performance of User Equilibria in Traffic Networks

Andreas S. Schulz and Nicolás Stier Moses

© 2002 by Andreas S. Schulz and Nicolás 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_id=366583

ON THE PERFORMANCE OF USER EQUILIBRIA IN TRAFFIC NETWORKS

ANDREAS S. SCHULZ AND NICOLÁS STIER MOSES

Sloan School of Management and Operations Research Center
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
E-mail: {schulz,nstier}@mit.edu

Abstract. According to Wardrop's first principle, drivers in a traffic network choose their routes selfishly; that is, they travel on a shortest path under the prevailing traffic conditions between their respective origin and destination. This behavior is captured by the Nash equilibrium of the underlying non-cooperative game, commonly called user equilibrium in this context. Because Nash equilibria do usually not optimize any global criterion per se, there is no apparent reason why a user equilibrium should be close to a system optimum, which is a solution of minimal total (and, therefore, average) travel time. In this paper, we extend recent positive results on the efficiency of user equilibria in simple networks to models that are more realistic. First, we introduce and analyze user equilibria in capacitated networks. In particular, we show that the worst ratio of the total travel time of the best user equilibrium to the total travel time of the system optimum does not change if capacities are included in the model. Second, we propose to compare the efficiency of user equilibria to a more restricted version of system optimum. In fact, the ordinary system optimum typically treats some drivers unfairly in that it assigns them to considerably longer paths than others. For this reason, a system optimum is often considered inadequate for purposes of traffic planning. We analyze the performance guarantee of user equilibria when compared to constrained system optima, which are designed to be more fair, and establish improved bounds in this setting.

1. Introduction

While Wardrop (1952) had introduced the concept of Nash equilibrium to *describe* user behavior in traffic networks, traffic engineers have proposed to utilize user equilibria in route-guidance systems to *prescribe* user behavior. Yet, Nash equilibria in general and user equilibria in particular are known to be inefficient (Dubey 1986), and critics favored in principle the difficult-to-implement system optimum, which guarantees that the average travel time is minimal. Hence, the recent result that user equilibria are near optimal (Roughgarden and Tardos 2002) came as a welcome surprise, which may help to justify the use of user equilibria in retrospect.

More precisely, Roughgarden and Tardos showed that the sum of all travel times (also called total latency) of a user equilibrium in an uncapacitated multicommodity flow network is at most that of an optimal routing of twice as much traffic in the same network. Moreover, the total latency of selfish routing is at most 4/3 times that of the best coordinated routing, when link

Date: December 2002.

1991 *Mathematics Subject Classification.* Primary 90C35; 90B20, 90C25, 90C27, 90C90.

Key words and phrases. Route Guidance, Traffic Assignment, System Optimum, User Equilibrium, Nash Equilibrium, Performance Guarantee, Multicommodity Flow.

This work was supported by the High Performance Computation for Engineered Systems (HPCES) programme of the Singapore-MIT Alliance (SMA) and by a General Motors Innovation Grant awarded to the first author.

An extended abstract appears in the Proceedings of the 14th ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, January 12-14, 2003.

delays depend linearly on congestion. Furthermore, Roughgarden (2002) discovered that the worst-case inefficiency due to selfish routing is independent of the network topology, which permits to compute the degradation in network performance caused by selfish users for a wide array of latency functions.

In this paper, we extend the work of Roughgarden and Tardos (2002) and Roughgarden (2002) in two directions. First, we introduce and analyze user equilibria in *capacitated networks*. In contrast to networks without capacities (the framework of Roughgarden and Tardos' work), the set of user equilibria is no longer convex and an equilibrium can be arbitrarily worse than the system optimum, even if arc latency functions are linear. However, adding capacities does not change the worst ratio between the *best* user equilibrium and the system optimum, given a fixed but arbitrary class of allowable latency functions. In other words, while Roughgarden showed that the worst ratio of total latency in user equilibrium to that of system optimum does not depend on the topology of the network, we establish that this ratio is also independent of arc capacities, so long as one considers the best equilibrium.

Second, we compare the performance of user equilibria to *constrained system optima*. The latter concept was introduced by Jahn, Möhring, and Schulz (2000) and studied by Jahn, Möhring, Schulz, and Stier Moses (2002) in an attempt to create route guidance with efficient solutions that are at the same time attrative to users. In a constrained system optimum, drivers are only routed along paths that are not too long compared to a shortest path under an a-priori estimate of the travel time. Hence, the acceptance rate of corresponding route-guidance devices is supposedly higher than that of an ordinary system optimum. In particular, a constrained system optimum provides a more realistic lower bound for the performance of traffic than a system optimum; consequently, performance guarantees for user equilibria relating to constrained system optima may be more meaningful.

This paper is organized as follows: Section 2 introduces the details of the model together with the required notation. In Section 3, we discuss user equilibria in networks with arc capacities. Constrained system optima are the object of study in Section 4. We conclude with additional remarks in Section 5.

2. The Basic Model

We consider a directed network $G = (V, A)$, and a set of k origin-destination (OD) pairs (s_i, t_i) with $s_i, t_i \in V$, $i = 1, \dots, k$. A flow of rate d_i must be routed from s_i to t_i. In the context of traffic networks, these demands are assumed to be arbitrarily divisible; in fact, there are infinitely many individuals (i.e., car drivers) and the route decision of one single individual has only an infinitesimal impact on other users. Let P_i be the set of directed (simple) paths from s_i to t_i in G. Each arc $a \in A$ has a non-negative, non-decreasing and differentiable *latency function* $\ell_a(\cdot)$, which maps the flow on a to the time needed to traverse a. Note that the traversal time of an arc a does indeed only depend on the congestion (i.e., flow) in a. We assume that the functions ℓ_a belong to a given set $\mathcal{L}$ of latency functions, and that $\ell_a(f_a)f_a$ is convex in f_a, for all $a \in A$. A flow f is a function that assigns a non-negative value to every path $P \in \mathcal{P}$, where $\mathcal{P} := \cup P_i$. Given a path flow, the corresponding arc flow is easily calculated as $f_a = \sum_{P \ni a} f_P$, for each $a \in A$. A flow is feasible when the demand is met, that is, $\sum_{P \in P_i} f_P = d_i$, for $i = 1, \dots, k$. The travel time along a path P is $\ell_P(f) := \sum_{a \in P} \ell_a(f_a)$. Generally, the quality of different flows is compared by using the total travel time in the network, which is defined as $C(f) := \sum_{P \in \mathcal{P}} \ell_P(f) f_P = \sum_{a \in A} \ell_a(f_a) f_a$.

A feasible flow f^* that minimizes $C(f)$ is called a *system optimum*. It can be computed by solving a convex min-cost multicommodity flow problem. The corresponding first-order optimality conditions imply that a flow is optimal iff the objective function value cannot be improved by re-routing flow from a single path to another path (Beckman, McGuire, and Winsten 1956). Formally, the flow f is optimal if and only if for all OD-pairs i and all paths $P, Q \in P_i$ such that $f_P >$

$0: \ \ell_P^*(f) \leq \ell_Q^*(f)$, where $\ell_P^*(f) := \sum_{a \in P} \ell_a^*(f_a)$ is the gradient along path P. Here, $\ell_a^*(f_a) := \ell_a(f_a) + \ell_a'(f_a) f_a$.

A feasible flow f is a *user equilibrium* if for all OD-pairs i and all paths $P, Q \in P_i$ such that $f_P > 0: \ \ell_P(f) \leq \ell_Q(f)$ (Wardrop 1952). The common value of the travel time for OD-pair i is $L_i(f) := \min_{P \in P_i} \ \ell_P(f)$, $i = 1, \ldots, k$. Therefore, the cost of the equilibrium can be expressed as $C(f) = \sum_i L_i(f) d_i$. This characterization of the user equilibrium coincides with the optimality conditions of a related, suitably constructed convex min-cost multicommodity flow problem (Beckman et al. 1956). In particular, a user equilibrium always exists and can be computed efficiently using standard procedures. For an extensive discussion of solution techniques and related aspects, we refer the reader to Florian (1986) and Sheffi (1985).

At times, we will restrict latency functions to be linear; our motivation for doing so is twofold. On the one hand, only then we can compare our results with previous results; on the other hand, the models considered become more tractable. We represent linear latency functions as $\ell_a(f_a) = q_a f_a + r_a$, where q_a and r_a are non-negative constants. Even though the linear case might appear restrictive, it is enough for congestion phenomena to occur. One interesting example is the so-called Braess Paradox (Braess 1968), which demonstrates that by adding a link to a network, the performance (i.e., travel time) of each user can degrade, instead of the network becoming less congested.

3. The Capacitated Case

We extend the basic model to include arc capacities c_a. Given a corresponding instance, a system optimum solves the following convex min-cost multicommodity flow problem:

$$\begin{aligned}
\min \quad & \sum_{a \in A} \ell_a(f_a) f_a & & \\
\text{s.t.} \quad & \sum_{P \ni a} f_P = f_a & & \text{for all } a \in A, \\
& \sum_{P \in P_i} f_P = d_i & & \text{for all } i = 1, \ldots, k, \\
& f_a \leq c_a & & \text{for all } a \in A, \\
& f_P \geq 0 & & \text{for all } P \in \mathcal{P}.
\end{aligned}$$

Notice that the above-mentioned optimality conditions for the uncapacitated case do not work in the more general case with capacities. For the former, we are able to consider just two paths between the same OD-pair at a time. Here, with capacities, this is not true anymore. A flow f^* is a system optimum if and only if

$$\text{for all feasible directions } h: \ \sum_{a \in A} h_a \ell_a^*(f_a^*) \geq 0 \ .$$

Given the new setting with explicit arc capacities, we also need to define "user equilibrium." The natural extension is to assume that no user can switch to a shorter route *with residual capacity.*

Definition 3.1. A flow f is a *(capacitated) user equilibrium* if for all OD-pairs $i \in \{1, \ldots, k\}$ and paths $P, Q \in P_i$ such that $f_P > 0$ and $\min_{a \in Q} \{c_a - f_a\} > 0 : \ell_P(f) \leq \ell_Q(f)$.

In contrast to the case without arc capacities, flow-carrying paths between the same OD-pair now can have different latencies. If we define $\bar{L}_i(f) := \max\{\ell_P(f) : P \in P_i, f_P > 0\}$, a user equilibrium f satisfies the following conditions:

$$\text{If } \ell_P(f) > \bar{L}_i(f), \text{ then } f_P = 0; \text{ if } \ell_P(f) < \bar{L}_i(f), \text{ then } \min_{a \in P} \{c_a - f_a\} = 0.$$

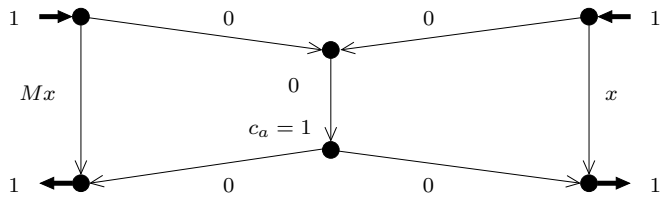

FIGURE 1. *Instance with multiple equilibria.*

In other words, we can partition the set of paths that serve demand i in three sets: short and up-to-capacity paths, paths that have residual capacity and a common length of $\bar{L}_i(f)$, and long paths without flow.

3.1. **Multiple User Equilibria.** In networks without capacities, the user equilibrium is essentially unique; in particular, different equilibria, if any, share the same total latency. An important effect of arc capacities is the existence of multiple equilibria, which is caused by the saturation of some arcs that afterwards restrict the route choice for the remaining users. Figure 1 provides an example with two commodities of unit demand each. The nodes on the left represent one OD-pair, while the nodes on the right form the other OD-pair. Arc labels indicate the corresponding latency functions; M is a fixed non-negative constant, and the arc in the center is the only arc with bounded capacity. Every user has two options: the route that goes through the center, which is always the shortest, and the alternative at the side. So long as there is remaining capacity on the common path, users will take that route. As soon as its capacity is exhausted, the flow becomes an equilibrium. Indeed, consider the two feasible flows f_1 and f_2 in which respectively the first and the second commodity saturates the arc in the middle. Both solutions are user equilibria, and so are their convex combinations. It is interesting to note (and unfortunate for optimization purposes) that the set of all user equilibria generally forms a non-convex region in the space of flows. In particular, computing the most efficient user equilibrium involves optimizing a convex function over a non-convex feasible region, in general. This explains why the "best" user equilibrium is difficult to characterize. We will therefore pick a particular user equilibrium that has a good characterization.

3.2. **Beckman User Equilibrium.** The natural way of extending the mathematical programming approach of Beckman et al. (1956) for computing user equilibria is the inclusion of capacities as additional constraints. To that effect, we define the *Beckman user equilibrium* to be the solution to the following problem:

$$\begin{aligned}
\min \quad & \sum_{a \in A} \int_0^{f_a} \ell_a(x)\, dx & \\
\text{s.t.} \quad & \sum_{P \ni a} f_P = f_a & \text{for all } a \in A, \\
& \sum_{P \in P_i} f_P = d_i & \text{for all } i = 1, \dots, k, \\
& f_a \le c_a & \text{for all } a \in A, \\
& f_P \ge 0 & \text{for all } P \in \mathcal{P}.
\end{aligned}$$

As this amounts to minimizing a convex function over a polytope, the arc variables f_a of an optimal solution f are unique if the objective function is strictly convex; however, path variables f_P do not need to be unique. Standard optimality conditions imply that a flow f is a Beckman user equilibrium

if and only if

$$\text{for all feasible directions } h: \ \sum_{a\in A} h_a \ell_a(f_a) \geq 0 \ . \tag{3.1}$$

This condition is crucial when proving results on the efficiency of user equilibria. First, however, let us show that a Beckman user equilibrium is indeed an equilibrium in the sense of Definition 3.1.

Lemma 3.2. *If f is a Beckman user equilibrium, then it is a user equilibrium.*

Proof. The flow f satisfies condition (3.1) by assumption. Therefore, the gradient along any feasible direction is non-negative. Suppose that f is not a user equilibrium in terms of Definition 3.1. Then, there are two paths P and P' between the same OD-pair with $f_P > 0$ and $\min_{a\in P'}\{c_a - f_a\} > 0$ such that $\ell_P(f) > \ell_{P'}(f)$. We will re-route flow from P to P' to arrive at a contradiction. Define a circulation h by setting $h_a := \varepsilon(\chi_a^{P'} - \chi_a^P)$, for all $a \in A$. Here, $\chi^Q \in \{0,1\}^A$ stands for the incidence vector of path Q, i.e., $\chi_a^Q = 1$ for all $a \in Q$, and 0 otherwise. The flow $f + h$ is feasible for sufficiently small $\varepsilon > 0$. Hence, h is a feasible direction at f. In particular, the gradient along h is non-negative. That is, $0 \leq \sum_{a\in A} h_a \ell_a(f_a) = \varepsilon(\ell_{P'}(f) - \ell_P(f))$. Thus, $\ell_P(f) \leq \ell_{P'}(f)$, which is a contradiction. □

Notice that a Beckman user equilibrium is not necessarily the most efficient equilibrium; it is just one that has a good characterization. It is this structure that helps us to carry forward some of the results known from networks without capacities.

3.3. **Best and Worst Equilibria.** Let us return to the example in Figure 1. The total latency of user equilibria f_1 and f_2 is 1 and M, respectively. The system optimum f^*, which also is the Beckman user equilibrium, routes $1/(M+1)$ units of flow through the left path and $M/(M+1)$ units of flow through the right path of each OD-pair. Its total travel time is $M/(M+1)$. Hence, the sum of all travel times of the worst equilibrium f_2 is not bounded by a constant times that of the system optimum, even when latencies are linear. In fact, if $M \to \infty$, then $C(f_2)/C(f^*) \to \infty$. Note that this finding is in clear contrast to the case without arc capacities, in which the total latency of every user equilibrium is within a factor of 4/3 of that of the system optimum, for linear arc delay functions (Roughgarden and Tardos 2002). In general, this factor is equal to $\alpha(\mathcal{L})$ if arc latency functions belong to the set $\mathcal{L}$ (Roughgarden 2002). At this, the so-called *anarchy value* $\alpha(\mathcal{L})$ of a set $\mathcal{L}$ of functions is defined as $\alpha(\mathcal{L}) = \sup_{0\neq\ell\in\mathcal{L}} \alpha(\ell)$, with the following meaning of $\alpha(\ell)$.

Definition 3.3 (Roughgarden 2002)**.** The *anarchy value* $\alpha(\ell)$ of a latency function ℓ is

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

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

Note that $\alpha(\ell) = 4/3$ if ℓ is linear. Despite the previous example, we will now extend these results to the *best* user equilibrium in a capacitated network—by actually looking at a Beckman user equilibrium. We start with a useful observation.

Lemma 3.4. *Let g be a feasible flow and f be a Beckman user equilibrium of a network with arc capacities. Then,*

$$\sum_{a\in A} (g_a - f_a)\ell_a(f_a) \geq 0 \ .$$

Proof. Note that $g - f$ is a feasible direction at f; the claim then follows from condition (3.1). □

We are now ready to prove the main result of this part of the paper. It basically extends Theorem 3.8 in Roughgarden (2002) to networks with arc capacities. Indeed, the proof follows the same outline, but makes use of Lemma 3.4.

Theorem 3.5. *Let f^* be a system-optimal flow and $\tilde{f}$ be the best user equilibrium for a capacitated network with latency functions drawn from $\mathcal{L}$. Then $C(\tilde{f}) \le \alpha(\mathcal{L})C(f^*)$.*

Proof. Let f be a Beckman user equilibrium and define the constants $\lambda_a \in [0,1]$ such that $\ell_a^*(\lambda_a f_a) = \ell_a(f_a)$ and $\mu_a = \ell_a(\lambda_a f_a)/\ell_a(f_a) \in [0,1]$. If we add and subtract $\ell_a(\lambda_a f_a)\lambda_a f_a$ from the cost of f^*, we can write:

$$\begin{aligned}
C(f^*) &= \sum_{a\in A} \left(\ell_a(\lambda_a f_a)\lambda_a f_a + \int_{\lambda_a f_a}^{f_a^*} \ell_a^*(x)\, dx \right) \\
&\ge \sum_{a\in A} \big(\ell_a(\lambda_a f_a)\lambda_a f_a + (f_a^* - \lambda_a f_a)\ell_a^*(\lambda_a f_a) \big) \\
&= \sum_{a\in A} \big(\ell_a(\lambda_a f_a)\lambda_a f_a + (f_a^* - \lambda_a f_a)\ell_a(f_a) \big) \\
&= \sum_{a\in A} \big(\lambda_a\mu_a + (1-\lambda_a) \big)\ell_a(f_a) f_a + \sum_{a\in A} (f_a^* - f_a)\ell_a(f_a) \\
&\ge \sum_{a\in A} \frac{\ell_a(f_a) f_a}{\alpha(\mathcal{L})} \\
&\ge C(\tilde{f})/\alpha(\mathcal{L}).
\end{aligned}$$

For the first inequality, we used that $\ell_a^*(\cdot)$ is non-decreasing, whereas for the second one, we bounded the first two terms with the anarchy value and the last one with Lemma 3.4. □

In particular, the ratio of the total travel time in the best user equilibrium to that of the system optimum in a capacitated network with linear latency functions is at most 4/3.

4. The Constrained System Optimum

In this section, we return to networks without explicit arc capacities in order to compare the total latency of the user equilibrium to that of a more restricted version of system optimum. It is widely accepted that the system optimum in a traffic network cannot be implemented in route-guidance systems; some users would be routed on considerably longer paths for the benefit of others and hence would typically not follow the route recommendations. Jahn et al. (2000, 2002) therefore proposed to use solutions of minimal total travel time in a system where all users are assigned to paths that are not too long. We believe that this concept also allows for a more realistic assessment of the quality of user equilibria.

To set the stage for our analysis, we need to introduce some additional notation. With each arc in the network, we now additionally associate a *normal travel time* $\hat{\ell}_a$. Its value represents an a-priori belief of users and hence does not depend upon the actual flow in the arc. We define $\hat{\ell}_P := \sum_{a\in P} \hat{\ell}_a$ as the normal travel time along path P; let $\hat{L}_i$ be the smallest normal travel time among all paths connecting OD-pair i, for $i = 1, \dots, k$. For $\varepsilon \ge 0$, we let $\mathcal{P}_\varepsilon$ be the subset of $\mathcal{P}$ that consists of those paths $P \in P_i$ that satisfy $\hat{\ell}_P \le (1+\varepsilon)\hat{L}_i$, $i = 1, \dots, k$.

For example, the normal travel time of an arc could be the geographic distance between its endpoints, the free-flow travel time (i.e., the travel time in uncongested state), or the travel time in user equilibrium. The only requisite is that it is fixed in advance. Figure 2 compares the set of paths in an ordinary network to the set of allowable paths when working with geographic distances as normal travel times. In this paper, we study two versions of normal travel times. First, in Sections 4.1–4.3, we will concentrate on the free-flow travel time $\hat{\ell}_a = \ell_a(0)$. Subsequently, in Section 4.4, we will work with $\hat{\ell}_a = \ell_a(f)$, where f is the user equilibrium for the instance considered.

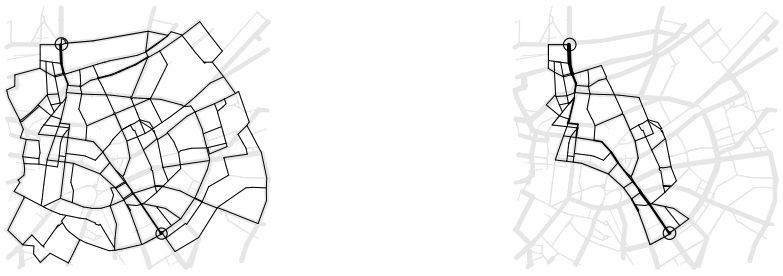

FIGURE 2. *The set of all paths connecting the highlighted OD-pair and the restricted set.*

An *ε-constrained system optimum* f^*_ε of a given instance is defined as an optimal solution to the following problem:

$$\begin{aligned} \min \quad & C(g) \\ \text{s.t.} \quad & g \text{ is a feasible flow} , \\ & g_P = 0 \quad \text{for all } P \notin \mathcal{P}_\varepsilon . \end{aligned} \tag{4.1}$$

Obviously, $C(f^*) \leq C(f^*_\varepsilon)$, where f^* is an ordinary system optimum. We intend to characterize the relationship between the cost of a user equilibrium f and that of a constrained system optimum f^*_ε, for each $\varepsilon \geq 0$. We therefore introduce the efficiency of a network G with demand vector d and latency vector ℓ as

$$\rho_\varepsilon(G, d, \ell) := \frac{C(f)}{C(f^*_\varepsilon)} \ .$$

We define the function $\gamma(\varepsilon)$ as the worst ratio that can possibly be achieved:

$$\gamma(\varepsilon) := \sup_{(G,d,\ell)} \rho_\varepsilon(G, d, \ell) \ .$$

Our goal is equivalent to characterizing the function $\gamma(\cdot)$ for the above-mentioned choices of normal travel times. It is immediately clear that $\gamma(\varepsilon)$ is non-decreasing because $C(f^*_\varepsilon)$ is a non-increasing function of ε. Furthermore, it also is easy to verify that $\gamma(\varepsilon) \geq 1$ for all $\varepsilon \geq 0$. In addition, for linear latencies and any $\varepsilon \geq 0$, we get $C(f) \leq \frac{4}{3} C(f^*) \leq \frac{4}{3} C(f^*_\varepsilon)$, using Theorem 3.5. This implies that in the linear case, $\gamma(\varepsilon) \leq \frac{4}{3}$, for all $\varepsilon \geq 0$. Moreover, for instances with $\hat{L}_i > 0$ for all OD-pairs $i = 1, \ldots, k$, $C(f^*_\varepsilon) = C(f^*)$ when ε is sufficiently large.

4.1. **Free-Flow Normal Travel Times.** For the following three sections, we assume that normal travel times are defined as travel times in the uncongested network; i.e., $\hat{\ell}_a = \ell_a(0)$ for all $a \in A$. We also write $L_i(0)$ for $\hat{L}_i$. Under this assumption, it turns out that user equilibria have improved performance guarantees when compared to constrained system optima instead of ordinary system optima. This improvement arises from the fact that constrained system optima do not perform as well as system optima. Indeed, for small values of ε they are worse than the user equilibrium itself, as we shall see. Jahn et al. (2002) conducted a detailed computational study of this fact, which supports the theory that we develop here.

We start our study of $\gamma(\cdot)$ by introducing a construction that permits to modify a given instance for some ε so as to obtain an instance for a different value of ε, but with similar efficiency. Let ε and δ be fixed positive constants; let $\mathcal{I}$ be an instance that satisfies $\rho_\varepsilon(G, d, \ell) \geq \gamma(\varepsilon) - \delta$. We create a new instance $\mathcal{I}'$ by adding a new vertex s'_i for every demand i. The nodes s'_i are only connected to s_i using arcs of constant latency M_i, specified below. The rest of $\mathcal{I}'$ is kept as in $\mathcal{I}$. Notice first that the natural extensions f' and $f^{*\prime}$ of a user equilibrium f and a system optimum f^* in $\mathcal{I}$ are a user equilibrium and a system optimum in $\mathcal{I}'$, respectively. The next lemma establishes a relation between constrained system optima.

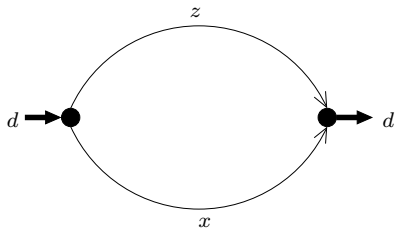

FIGURE 3. *Simple tight instance.*

Lemma 4.1. *Consider a fixed* $0 < \varepsilon' < \varepsilon$ *and set* $M_i := \frac{\varepsilon-\varepsilon'}{\varepsilon'} L_i(0)$. *If* f_ε^* *is an* ε*-constrained system optimum of* $\mathcal{I}$*, then the natural extension* $f_\varepsilon^{*\prime}$ *is an* ε'*-constrained system optimum of* $\mathcal{I}'$.

Proof. All paths in $\mathcal{I}$ that are used for OD-pair i have a normal length between $L_i(0)$ and $(1+\varepsilon)L_i(0)$. After adding M_i to each of them, their lengths are between $M_i + L_i(0)$ and $M_i + (1+\varepsilon)L_i(0) = (1+\varepsilon')(M_i + L_i(0))$. It follows that $f_\varepsilon^{*\prime}$ is an ε'-constrained system optimum. □

Observe that extending a flow g in $\mathcal{I}$ to a flow g' in $\mathcal{I}'$ changes its cost by a fixed amount M; that is, $C(g') = M + C(g)$, where $M = \sum_{i=1}^k M_i d_i$. Moreover, $M = \frac{\varepsilon-\varepsilon'}{\varepsilon'} \sum_i L_i(0) d_i \le \frac{\varepsilon-\varepsilon'}{\varepsilon'} C(g)$, because $\ell_P(g) \ge L_i(0)$ for any path $P \in P_i$.

Theorem 4.2. *The function* $\gamma(\varepsilon)/\varepsilon$ *is non-increasing.*

Proof. Setting $g = f_\varepsilon^*$ in the preceding paragraph, we obtain

$$\gamma(\varepsilon') \ge \frac{C(f')}{C(f_\varepsilon^{*\prime})} = \frac{M + C(f)}{M + C(f_\varepsilon^*)} \ge \frac{C(f)}{\frac{\varepsilon}{\varepsilon'} C(f_\varepsilon^*)} \ge \frac{\varepsilon'}{\varepsilon} \left(\gamma(\varepsilon) - \delta\right) \qquad \text{for all } \varepsilon' < \varepsilon \ .$$

As δ was arbitrary, $\gamma(\varepsilon') \ge \frac{\varepsilon'}{\varepsilon}\, \gamma(\varepsilon)$ for all $\varepsilon' < \varepsilon$. □

Corollary 4.3. *The function* $\gamma(\varepsilon)$ *is subadditive.*

4.2. **Tight and Worst Instances for the Linear Case.** In this section, we will characterize classes of instances that have bad performance for the case of linear latency functions; in other words, they dominate the shape of $\gamma(\cdot)$. We call an instance *tight* when $\rho_\varepsilon(G, d, \ell) = \frac{4}{3}$ and *worst* when $\rho_\varepsilon(G, d, \ell) = \gamma(\varepsilon)$.

For the (unconstrained) system optimum, there are simple examples of tight instances (Roughgarden and Tardos 2002). A particularly simple one is given by two parallel arcs joining two nodes with a demand of d units; see Figure 3. The latency of the top arc is the constant z, which satisfies $d \le z \le 2d$; the latency of the bottom arc is equal to its flow value. It is not hard to see that the user equilibrium f is $(0, d)$ and the system optimum f^* is $(d - z/2, z/2)$, where the pair of values represents the flow in the top arc and in the bottom arc, respectively. The objective function values are $C(f) = d^2$ and $C(f^*) = zd - z^2/4$. For a tight instance, we simply set $z = d$ to get $\rho = \frac{4}{3}$.

By a careful analysis of the proof of Theorem 3.8 in Roughgarden (2002), one can establish conditions that characterize tight instances in the unconstrained case. Assume, as before, that f is the user equilibrium and f^* the system optimum.

Observation 4.4. *A linear instance is tight if and only if the following three conditions are satisfied:*

(a) *for all* $a \in A$: $\quad f_a^* = f_a/2$ *or* $q_a = 0$,
(b) *for all* $P \in P_i$: $\quad f_P^* = 0$ *or* $\ell_P(f) = L_i(f)$,
(c) *for all* $a \in A$: $\quad r_a = 0$ *or* $f_a = 0$.

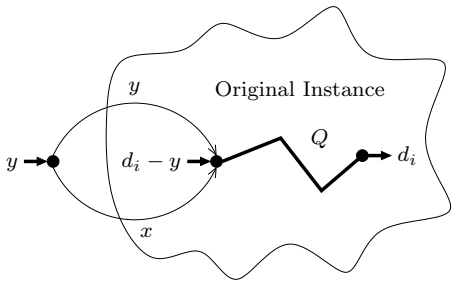

FIGURE 4. *Modified instance used to prove Theorem 4.7.*

We will use Observation 4.4 to show that there cannot exist tight instances for the constrained case (although for some ε, $\gamma(\varepsilon)$ could still be 4/3). But first we need an additional observation. We say that a flow is an *ε-constrained user equilibrium* if it is a user equilibrium in the network restricted to paths in $\mathcal{P}_\varepsilon$ only.

Lemma 4.5. *For instances with linear latencies, a* 0*-constrained system optimum is a* 0*-constrained user equilibrium, and vice versa.*

Proof. The two optimality conditions are equivalent because all paths used are shortest with respect to normal travel times. □

Theorem 4.6. *Let (G, d, ℓ) be an instance with linear latency functions, and let $0 \le \varepsilon < \infty$. Then, $\rho_\varepsilon(G, d, \ell) < \frac{4}{3}$.*

Proof. Suppose that $\gamma(\varepsilon) = 4/3 = \rho_\varepsilon(G, d, \ell)$ for some $\varepsilon > 0$. In that case, an ε-constrained system optimum f^*_ε is a system optimum, too, because the instance is also tight for the unconstrained case. Indeed,

$$\frac{4}{3} = \frac{C(f)}{C(f^*_\varepsilon)} \le \frac{C(f)}{C(f^*)} \le \frac{4}{3} \ .$$

Here, f denotes a user equilibrium and f^* an ordinary system optimum, as usual.

Observation 4.4(c) implies that $r_a = 0$ for all arcs a with $f_a > 0$. In other words, the latency of any flow-carrying arc a is of the form $q_a f_a$. Hence, $L_i(0)$ is zero for all OD-pairs i, which makes the actual value of ε irrelevant. In particular, f^*_ε is an ε'-constrained system optimum for all $0 \le \varepsilon' \le \infty$. Moreover, f is a 0-constrained user equilibrium, too. Lemma 4.5 therefore implies that f and f^*_ε have the same total latency, which is a contradiction. □

We turn our attention to characterizing the worst instances for a fixed ε. We say that a path $P \in P_i$ is *long* if $\ell_P(0) = (1+\varepsilon)L_i(0)$, and it is *short*, otherwise. We again use the idea of modifying a network to prove that if an instance admits flow that is routed along a short path, the instance can be changed into one that has a higher value of ρ.

Theorem 4.7. *Let $\varepsilon > 0$. Let f^*_ε be an ε-constrained system optimum for a given instance with linear latencies. If the instance is worst (but not tight), then f^*_ε uses long paths only.*

Proof. To simplify notation, let $\tilde{f} := f^*_\varepsilon$. Suppose that for some OD-pair i there is a short path $Q \in \mathcal{P}_\varepsilon$ such that $\tilde{f}_Q > 0$. We will construct a new instance that is worse.

Let $y := \min\{(1+\varepsilon)L_i(0) - \ell_Q(0), 2\tilde{f}_Q, d_i\}$. We modify the given instance by incorporating two new arcs as illustrated in Figure 4. The joint head of the new arcs is s_i, i.e., the origin of path Q. Moreover, we reassign y units of demand from s_i to the joint tail of the added arcs. The latencies of the new arcs are y and x, respectively.

Consider a path P in the original network that serves the same commodity as Q. After the modification, there are two possible extensions of P. The path P_1 (resp. P_2) starts with the upper (resp. lower) arc of the new subnetwork and then continues along P. The path P_2 belongs to $\mathcal{P}_\varepsilon$ because $\ell_{P_2}(0) = \ell_P(0)$. If $\ell_P(0) \le \ell_Q(0)$, $P_1 \in \mathcal{P}_\varepsilon$, too. The user equilibrium and the constrained system optimum of the new instance are simple extensions of f and $\tilde{f}$. Indeed, the vector f' with $f'_a = f_a$ for all $a \in A$ and with values 0 and y for the upper and lower new arc, respectively, is a user equilibrium for the new instance of total latency $C(f') = C(f) + y^2$. Similarly, $\tilde{f}$ can be extended by defining $\tilde{f}'$ as $\tilde{f}_a$ for $a \in A$ and $(y/2, y/2)$ in the new arcs. Its cost is $C(\tilde{f}') = C(\tilde{f}) + \frac{3}{4}y^2$. Therefore,

$$\rho' = \frac{C(f) + y^2}{C(\tilde{f}) + \frac{3}{4}y^2} \ ,$$

which is a convex combination of ρ and $\frac{4}{3}$. As the former is smaller than the latter, the new instance has a worse performance (i.e., $\rho' > \rho$). □

4.3. **Bounds for $\gamma(\varepsilon)$.** In this section, we present upper and lower bounds for the function $\gamma(\cdot)$. We start with an upper bound that, for small ε and linear latencies, improves on the bound presented in Theorem 3.5.

Theorem 4.8. *If we restrict the definition of $\gamma(\cdot)$ to instances with linear latency functions only, then $\gamma(\varepsilon) \le 1/(1-\varepsilon)$ for all $0 \le \varepsilon < 1$. In particular, $\gamma(0) = 1$ and $\gamma(\varepsilon) < 4/3$ for $\varepsilon < 1/4$.*

Proof. Consider a fixed $\varepsilon < 1$ and let $\tilde{f}$ be an ε-constrained system optimum. We define the function $h(t) := C(f + t(\tilde{f} - f))$. Due to the convexity of $C(\cdot)$, $h(1) \ge h(0) + h'(0)$. It is enough to prove that $h(0) + h'(0) \ge (1-\varepsilon)h(0)$ because then $C(\tilde{f}) = h(1) \ge (1-\varepsilon)h(0) = (1-\varepsilon)C(f)$, as required. Now,

$$\begin{aligned} h'(0) &= \sum_a \ell_a^*(f_a)(\tilde{f}_a - f_a) = \sum_a \left[2\ell_a(f_a) - \ell_a(0)\right](\tilde{f}_a - f_a) \\ &\ge 2\left[\sum_i L_i(f)d_i - \sum_i L_i(f)d_i\right] + \sum_i L_i(0)d_i - (1+\varepsilon)\sum_i L_i(0)d_i \\ &= -\varepsilon \sum_i L_i(0)d_i \ge -\varepsilon C(f) = -\varepsilon h(0). \end{aligned}$$

The first inequality comes from the fact that $\ell_P(f) = L_i(f)$ for every P such that $f_P > 0$, and $\ell_P(f) \ge L_i(f)$ in general. Similarly, $\ell_P(0) \le (1+\varepsilon)L_i(0)$ for every P such that $\tilde{f}_P > 0$, and $\ell_P(0) \ge L_i(0)$ in general. □

We now give lower bounds for $\gamma(\varepsilon)$ by providing corresponding instances. For the linear case, we use a collection of instances based on a modified Braess Paradox network. For the general case, we use polynomials of large degree. If we consider arbitrary latency functions, Lemma 4.10 rules out many of the results proved for the linear case.

Lemma 4.9. *Under the assumption of Theorem 4.8, $\gamma(\varepsilon) \ge 1 + 1/(3 + \frac{2}{\varepsilon})$.*

Proof. Consider the network depicted in Figure 5, which is similar to the original instance of the Braess Paradox, but with the latency of the middle arc set to the constant z instead of 0 and a demand of d units. Every combination of z and d gives a different instance; optimizing over z and d in order to obtain the best lower bound for $\gamma(\varepsilon)$ yields the claim. □

Figure 6 displays and summarizes the bounds for $\gamma(\varepsilon)$ in the linear case.

Lemma 4.10. *For the general case: $\gamma(\varepsilon) \ge \max\{1+\varepsilon, 2\}$.*

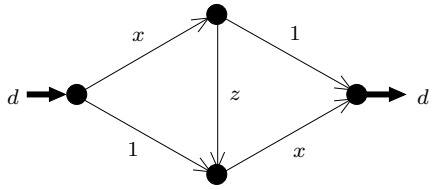

FIGURE 5. *Instance used in Lemma 4.9.*

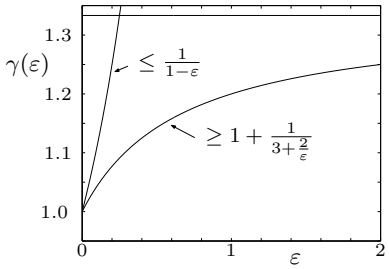

FIGURE 6. *Bounds on $\gamma(\varepsilon)$ in the case of linear latencies.*

Proof. We start with the first bound. Consider a network consisting of two parallel arcs and with unit demand. The latencies of the two arcs are $\ell_1(x) = 1$ and $\ell_2(x) = x^p + q$ for given p and $q < 1$. Denoting flows as usual, it is clear that $C(f) = 1$ for any choice of p and q because at equilibrium $L(f) = 1$. The condition for being able to use both arcs in a constrained system optimum is $1 \le (1+\varepsilon)q$; in that situation, system optimum and constrained system optimum are the same. Notice also that $C(f^*) \to q$ when $p \to \infty$. To get a lower bound for $\gamma(\varepsilon)$, we set $q = (1+\varepsilon)^{-1}$, which allows the ε-constrained system optimum to use both arcs. Thus, $C(f)/C(f^*) \to 1/q = 1+\varepsilon$ as $p \to \infty$. Hence, $\gamma(\varepsilon) \ge 1 + \varepsilon$.

For the second part, consider a similar network with demand of 2 units and latencies $\ell_1(x) = x$ and $\ell_2(x) = x^p$. In this case, all paths have zero latency with respect to normal travel times; therefore, all paths are allowed for any ε. The user equilibrium f is the flow $(1, 1)$ with cost 2 and f^*_ε $(= f^*)$ is the flow $(2 - z, z)$, where z is the root of the polynomial $\ell^*_2(x) - \ell^*_1(2 - x) = (p+1)x^p + 2x - 4$ that satisfies $0 \le z \le 2$. Its cost is $(2-z)^2 + z^{p+1}$ and its efficiency ρ_p is twice the reciprocal of the cost. Computing the limit, we obtain $\rho_p \to 2$ when $p \to \infty$, □

4.4. **User Equilibrium Normal Travel Times.** In this section, we assume that the normal travel times are set equal to the travel times in user equilibrium. Compared to the free-flow travel times studied above, this definition has the advantage of arising from a loaded network; in particular, it better captures the experience of the users. In addition, constrained system optima under this choice of normal travel times have very good quality. While Jahn et al. (2002) explored the practical aspects of this definition, we here give the first theoretical analysis.

One important consequence of this choice of normal travel times is that the user equilibrium f itself is a feasible solution to Problem (4.1). Moreover, it is not hard to see that $f^*_\varepsilon \to f^*$ when $\varepsilon \to \infty$. Therefore, we have

$$C(f^*) = C(f^*_\infty) \le C(f^*_\varepsilon) \le C(f^*_0) \le C(f) \text{ for all } \varepsilon \ge 0 \ . \tag{4.2}$$

As in the previous section, we obtain a lower bound for the function $\gamma(\cdot)$ by providing an appropriate instance. Take the tight example in Figure 3 (i.e., $z = d$). Both paths have latency d

in equilibrium. Hence, for every $\varepsilon \geq 0$, a constrained system optimum can use both paths; consequently, it also is a system optimum. As the example is tight within the class of instances having linear latencies, $\gamma(\varepsilon) = 4/3$ uniformly for all $\varepsilon \geq 0$. Moreover, this example can easily be generalized to almost all relevant classes of latency functions.

Lemma 4.11. *Consider networks with latency functions drawn from the class $\mathcal{L}$ and assume that the definition of $\gamma(\cdot)$ is restricted to such instances. Then, $\gamma(\varepsilon)$ is uniformly equal to $\alpha(\mathcal{L})$.*

The lemma implies that user equilibria do in general not display improved performance when compared to constrained system optima under user equilibrium normal travel times. Indeed, the worst-case guarantee is the same as with respect to system optima.

Given that the quality of constrained system optima is rather good, we may go in the opposite direction. That is, how well does a constrained system optimum approximate the true system optimum? Put differently, what are the properties of the function $\Gamma(\varepsilon)$ defined as the worst ratio that can possibly be attained; i.e., $\Gamma(\varepsilon) := \sup_{(G,d,\ell)} \{C(f^*_\varepsilon)/C(f^*)\}$. Apparently, $\Gamma(\varepsilon)$ is a non-increasing function of ε, and $\Gamma(\infty) = 1$.

Lets restrict our consideration to instances with linear latencies. In that case, (4.2) implies that $\Gamma(\varepsilon) \leq 4/3$ for all $\varepsilon \geq 0$. Consider once more the instance depicted in Figure 3, this time with $d = 1$ and $1 \leq z \leq 2$. The constrained system optimum is the system optimum if $z \leq 1 + \varepsilon$; it is the user equilibrium, otherwise. This gives a lower bound for $\Gamma(\varepsilon)$ equal to

$$\left[(\varepsilon + 1)\left(1 - \frac{\varepsilon + 1}{4}\right)\right]^{-1},$$

for $0 \leq \varepsilon \leq 1$. The bound forces $\Gamma(0)$ to be $\frac{4}{3}$. Finally, note that for $\varepsilon = 0$ in the linear case, we have two extreme examples: in the first, $C(f^*) = C(f^*_0) = \frac{3}{4}C(f)$, and $C(f) = C(f^*_0) \approx \frac{4}{3}C(f^*)$ in the second.

5. Conclusion

Lately, the proper use of route-guidance devices with the objective of improving the utilization of road networks by giving more information to drivers has been one of the most active research areas in traffic engineering. Indeed, the ultimate goal of *Intelligent Transportation Systems* would be making the actual traffic to be close to the system optimum. Yet, not all drivers would have the incentive to follow a corresponding route recommendation; actually, some would face rather long "detours." Therefore, most of the recent approaches in transportation science are content with computing a user equilibrium, which also is referred to as traffic assignment in this context. That is, users are guided onto paths they would — in theory — take anyway. Our results give an a posteriori justification for doing so; in fact, we have shown for a broader class of networks than considered before that the expense of working with user equilibria instead of system optima is limited.

The introduction of arc capacities gives rise to multiple equilibria. In particular, the price of anarchy jumps to infinity, even in the case of linear link delay functions. Nevertheless, it is reassuring and encouraging that the best user equilibrium is still close to the system optimum, despite of the presence of capacities.

User equilibria perform even better (in a relative sense) if the best state that can potentially be induced in the network is a constrained system optimum with respect to zero-flow normal travel times. The latter was recently proposed as another concept for use in route-guidance systems. An upper bound for $\gamma(\cdot)$ in the case of arbitrary latencies is still missing, although it could be uniformly infinity. When the constrained system optimum is computed under user equilibrium normal travel times, its quality improves and so does the "fairness" to users.

References

Beckman, 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.

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

Florian, M. (1986). Nonlinear cost network models in transportation analysis. *Mathematical programming 26*, 167–196.

Jahn, O., R. H. Möhring, and A. S. Schulz (2000). Optimal routing of traffic flows with length restrictions in networks with congestion. In *Operations Research Proceedings 1999*, pp. 437–442. Springer.

Jahn, O., R. H. Möhring, A. S. Schulz, and N. 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.

Murchland, J. (1970). Braess's paradox of traffic flow. *Transportation Research 4*, 391–394.

Roughgarden, T. (2002). The price of anarchy is independent of the network topology. *Manuscript*. To appear in Journal of Computer and System Sciences.

Roughgarden, T. and E. Tardos (2002). How bad is selfish routing? *Journal of the ACM 49*(2), 236–259.

Sheffi, Y. (1985). *Urban Transportation Networks*. Prentice-Hall, Englewood, NJ, 1985.

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