MATHEMATICS OF OPERATIONS RESEARCH
Vol. 29, No. 4, November 2004, pp. 961–976
ISSN 0364-765X | EISSN 1526-5471 | 04 | 2904 | 0961

DOI 10.1287/moor.1040.0098

Selfish Routing in Capacitated Networks

José R. Correa
Departamento de Ciencias de la Computacion, Universidad de Chile, Avenida Blanco Encalada 2120, Santiago, Chile C.P. 837-0459, jcorrea@dcc.uchile.cl

Andreas S. Schulz
Sloan School of Management, Massachusetts Institute of Technology, Office E53-36, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, schulz@mit.edu

Nicolás E. Stier-Moses
Columbia Business School, Uris Hall, Room 418, 3022 Broadway, New York, New York 10027-6902, nicolas.stier@columbia.edu

According to Wardrop's first principle, agents in a congested network choose their routes selfishly, a behavior that is captured by the Nash equilibrium of the underlying noncooperative game. A Nash equilibrium does not optimize any global criterion per se, and so there is no apparent reason why it should be close to a solution of minimal total travel time, i.e., the system optimum. In this paper, we offer positive results on the efficiency of Nash equilibria in traffic networks. In contrast to prior work, we present results for networks with capacities and for latency functions that are nonconvex, nondifferentiable, and even discontinuous.

The inclusion of upper bounds on arc flows has early been recognized as an important means to provide a more accurate description of traffic flows. In this more general model, multiple Nash equilibria may exist and an arbitrary equilibrium does not need to be nearly efficient. Nonetheless, our main result shows that the best equilibrium is as efficient as in the model without capacities. Moreover, this holds true for broader classes of travel cost functions than considered hitherto.

Key words: selfish routing; price of anarchy; traffic assignment; system optimum; Nash equilibrium; performance guarantee; multicommodity flow
MSC2000 subject classification: Primary 90B10, 90B20, 91A10; secondary: 90C25, 90C27, 90C35, 91A13, 91A43
OR/MS subject classification: Primary: Networks/graphs: multicommodity, theory; secondary: games: noncooperative; mathematics: combinatorics; transportation: models
History: Received June 23, 2003; revised February 10, 2004.

1. Introduction. It is a common behavioral assumption in the study of traffic networks modeling congestion effects and therefore featuring flow-dependent link travel times, that travelers choose routes that they perceive as being the shortest under the prevailing traffic conditions. In other words, travelers minimize their individual travel times (Kohl 1841). The situation resulting from these individual decisions is one where drivers cannot reduce their journey times by unilaterally choosing another route, which prompted Knight (1924) to call the resulting traffic pattern an equilibrium. Nowadays it is indeed known as the user equilibrium (Dafermos and Sparrow 1969), and it is effectively thought of as a steady state evolving after a transient phase where travelers successively adjust their route choices until a situation with stable route travel costs and route flows has been reached (Larsson and Patriksson 1999). In a seminal contribution, Wardrop (1952, p 345) stated two principles that formalize this notion of equilibrium and the alternative postulate of the minimization of the total travel costs. His first principle reads:

> The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Wardrop's first principle of route choice, which is identical to the notion postulated by Kohl (1841) and Knight (1924), became accepted as a sound and simple behavioral principle to describe the spreading of trips over alternate routes due to congested conditions (Florian 1999). The connection between a traffic pattern satisfying Wardrop's first principle and a Nash equilibrium of a network game among the trip-makers was first formulated by Charnes and Cooper (1961). Indeed, in real urban traffic systems, observed flows are likely to be

closer to a user than a system optimum (Downs 1962). The system optimum is characterized by Wardrop's second principle (p. 345):

> The average journey time is a minimum.

Not surprisingly, the total (or equivalently, average) travel time is generally not minimized by the user equilibrium, because users do not pay for their external costs (Dupuit 1849, Pigou 1920, Knight 1924). Hence, the recent result that user equilibria are *near* optimal (Roughgarden and Tardos 2002) came as a welcome surprise. In fact, they showed that the total travel time (also called total latency) of a user equilibrium in an *uncapacitated* multicommodity flow network (the framework of the work discussed above) 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 latencies depend linearly on congestion. Furthermore, Roughgarden (2003) proved that the worst-case inefficiency due to selfish routing is independent of the network topology. More specifically, any given family $\mathcal{L}$ of latency functions gives rise to a parameter $\alpha(\mathcal{L})$, which can be computed on a simple, single-commodity network, so that for any *uncapacitated* network with multiple commodities the total latency of the user equilibrium is at most $\alpha(\mathcal{L})$ times that of the system optimum. It is important to point out that Roughgarden's analysis only works for latency functions that are *nondecreasing*, *differentiable*, and their respective products with the identity function are *convex*.

In this paper, we extend the work of Roughgarden and Tardos (2002) and Roughgarden (2003) to network models that are more realistic. We introduce and analyze user equilibria in *capacitated* networks with more general classes of latency functions. In contrast to networks without capacities, 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, we prove that adding capacities does not change the worst ratio between the *best* user equilibrium and the system optimum, given an arbitrary but fixed class of allowable latency functions. In other words, while Roughgarden showed that the worst ratio of the total latency of a user equilibrium to that of a corresponding system optimum does not depend on the topology of the network, we establish that this ratio is also independent of arc capacities, as long as one considers the best equilibrium. Practically all results remain actually true for more general side constraints. For simplicity of presentation, we restrict ourselves to capacity constraints (see §6 for additional details). Moreover, we provide simple proofs of these results, which are, in addition, valid for general nondecreasing functions (not necessarily differentiable or convex, and just lower semicontinuous).

This paper is organized as follows. Section 2 introduces the specifics of the basic model together with the obligatory notation. It also features a new, simpler proof of the original result of Roughgarden and Tardos (2002) that helps to set the stage for the subsequent discourse. In fact, the model with arc capacities and more general latency functions is the subject of §3. There, we also discuss the relevance of network models with capacities and less restricted families of travel cost functions. Applications to specific classes of latency functions are discussed in §4. While the previous sections still assume continuous functions, we take a separate look at lower semicontinuous travel cost functions in §5. Section 6 contains our concluding remarks.

2. The basic model. We consider a directed network $D = (N, A)$ and a set $K \subseteq N \times N$ of origin-destination (OD) pairs. For each $k \in K$, a flow of rate d_k must be routed from the origin to the destination. In the context of traffic or other communication networks, such demands are typically assumed to be arbitrarily divisible; in fact, the route decision of a single individual has only an infinitesimal impact on other users. For $k \in K$, let $\mathcal{P}_k$ be the set of directed (simple) paths in D connecting the corresponding origin with its destination, and let $\mathcal{P} := \bigcup \mathcal{P}_k$. Furthermore, a nonnegative, nondecreasing, and continuous

latency function ℓ_a with values in $\mathbb{R}_{\geq 0} \cup \{\infty\}$ maps the flow on arc a to the time needed to traverse a. (We drop the continuity assumption later; see §5.) A path flow is a nonnegative vector $f = (f_P)_{P \in \mathcal{P}}$ that meets the demand, i.e., $\sum_{P \in \mathcal{P}_k} f_P = d_k$ for $k \in K$. Given a path flow, the corresponding arc flow is easily computed as $f_a = \sum_{P \ni a} f_P$ for each $a \in A$. For a flow f, the travel time along a path P is $\ell_P(f) := \sum_{a \in P} \ell_a(f_a)$. Hence, the flow's total travel time is $C(f) := \sum_{P \in \mathcal{P}} \ell_P(f) f_P = \sum_{a \in A} \ell_a(f_a) f_a$. The cost function C^f with constant latencies $\ell_a^f := \ell_a(f_a)$ plays an important role; here, f is a given feasible flow. For a feasible flow x, $C^f(x) := \sum_{a \in A} \ell_a^f x_a$. Note that $C^f(f) = C(f)$.

A system optimum f^* is an optimal solution to the following nonlinear min-cost multicommodity flow problem with separable objective function:

$$\min \quad \sum_{a \in A} \ell_a(f_a) f_a \tag{2.1a}$$

$$\text{s.t.} \quad \sum_{P \ni a} f_P = f_a \quad \text{for all } a \in A, \tag{2.1b}$$

$$\sum_{P \in \mathcal{P}_k} f_P = d_k \quad \text{for all } k \in K, \tag{2.1c}$$

$$f_P \geq 0 \quad \text{for all } P \in \mathcal{P}. \tag{2.1d}$$

Roughgarden (2003) assumed that $\ell_a(x)$ is differentiable and $\ell_a(x)x$ is convex, for each arc $a \in A$. If that is the case, a flow f^* is optimal if and only if

$$\sum_{a \in P} \ell_a^*(f_a^*) \leq \sum_{a \in Q} \ell_a^*(f_a^*) \quad \text{for all } k \in K \text{ and all paths } P, Q \in \mathcal{P}_k \text{ such that } f_P^* > 0. \tag{2.2}$$

Here, $\ell_a^*(f_a) := \ell_a(f_a) + \ell_a'(f_a) f_a$. In other words, f^* is optimal if and only if the marginal travel cost of any used path is not greater than that of any other path. It is no accident that this condition closely resembles that of a user equilibrium. In fact, the difference between private cost and social cost is $\ell_a'(f_a) f_a$; hence, a flow f is in equilibrium if and only if

$$\sum_{a \in P} \ell_a(f_a) \leq \sum_{a \in Q} \ell_a(f_a) \quad \text{for all } k \in K \text{ and all paths } P, Q \in \mathcal{P}_k \text{ such that } f_P > 0. \tag{2.3}$$

In turn, (2.3) can be interpreted as the optimality conditions of a convex min-cost multicommodity flow problem like (2.1), where (2.1a) is replaced by $\sum_{a \in A} \int_0^{f_a} \ell_a(x)\, dx$ (Beckmann et al. 1956). (Note that Roughgarden's (2003) assumptions on latency functions are not required for that to be true; indeed, continuity and monotonicity suffice.) In particular, a user equilibrium always exists, its total latency is unique, and it can be computed efficiently using standard procedures. For an extensive discussion of algorithmic techniques and related aspects, we refer the reader to Magnanti (1984), Sheffi (1985), Patriksson (1994), and Florian and Hearn (1995). There, we are particularly interested in an equivalent characterization in terms of a variational inequality problem due to Smith (1979) (see also Dafermos 1980). Accordingly, a flow f is a user equilibrium if and only if

$$C^f(f) \leq C^f(x) \quad \text{for all flows } x. \tag{2.4}$$

Note that this inequality is a direct consequence of the fact that in equilibrium, users travel on shortest paths with respect to arc costs ℓ_a^f.

We are now ready to give a different proof of the main result of Roughgarden and Tardos (2002) for linear latency functions $\ell_a(x) = q_a x + r_a$ with $q_a, r_a \geq 0$ for all $a \in A$. Note that linear travel time functions are sufficient for the occurrence of certain congestion phenomena. One interesting example is the so-called Braess Paradox (Braess 1968), which describes the fact that the addition of a link to a network can result in increased travel times for all users in an equilibrium state. The following result as well as our main result

on general latency functions and networks with capacities (Theorem 3.6) also provide a worst-case bound on the degradation of the total (and therefore average) travel time that can possibly be caused by the Braess Paradox.

THEOREM 2.1 (ROUGHGARDEN AND TARDOS 2002). *Let f be a user equilibrium and let f^* be a system optimum for an instance of* (2.1) *with linear latency functions. Then,* $C(f) \leq \frac{4}{3}C(f^*)$.

PROOF. Let x be a feasible flow. From condition (2.4), $C(f) \leq C^f(x)$. Furthermore,

$$C^f(x) = \sum_{a\in A}(q_a f_a + r_a)x_a \leq \sum_{a\in A}(q_a x_a + r_a)x_a + \frac{1}{4}\sum_{a\in A} q_a f_a^2 \leq C(x) + \frac{1}{4}C(f),$$

where the first inequality holds because $(x_a - f_a/2)^2 \geq 0$. It follows that $\frac{3}{4}C(f) \leq C(x)$ for any feasible flow x. Hence, $C(f) \leq \frac{4}{3}C(f^*)$. □

Let us make a remark that simultaneously is a preview: exactly the same proof works for networks with capacities on arcs. In fact, one can use Lemma 3.3 in lieu of condition (2.4). Moreover, Corollary 4.3 further generalizes this worst-case bound of 4/3 to travel cost functions ℓ satisfying $\ell(cx) \geq c\ell(x)$ for $c \in [0, 1]$ (with the only restriction that they are nonnegative, nondecreasing, and lower semicontinuous). This includes, among others, concave functions.

3. Networks with capacities. The link performance functions ℓ_a relate the average travel times to the traffic rates f_a on the links $a \in A$. To account for congestion effects, these functions are typically nonlinear, positive, and strictly increasing with flow (Patriksson 1994, p. 29). In practice, the most frequently used functions are polynomials whose degrees and coefficients are determined from real-world data through statistical methods (Patriksson 1994, p. 70). Branston (1976) and Larsson and Patriksson (1995) argue that functions of this kind are unrealistic in the sense that the resulting travel times are finite whenever the arc flows are finite, so that the arcs are actually assumed to be able to carry arbitrarily large volumes of traffic flows; in practice, however, road links have some finite limits on traffic flows. Moreover, they point out that travel times predicted in the overloaded range do not have a real meaning. In connection with this deficiency, Hearn (1980) notes that in the basic model described in §2, "the predicted flow on some links will be far lower or far greater than the traffic engineer knows they should be *if all assumptions of the model are correct*" (p. 1). Hearn and others, in particular Larsson and Patriksson (1994, 1995, 1999), and most recently, Marcotte et al. (2004), have therefore advocated the inclusion of arc flow capacities as an obvious way of improving the quality of traffic assignment models. Interestingly, the widely popular link delay formula proposed by the Bureau of Public Roads (1964) includes a capacity parameter.

A frequently used way to (implicitly) incorporate capacities is to employ volume delay formulas that tend to infinity as the arc flow approaches the arc capacity; see, e.g., Branston (1976) for a discussion. Boyce et al. (1981) have empirically found that asymptotic travel time functions yield unrealistically high travel times and devious rerouting of trips. In addition, Larsson and Patriksson (1995) criticize the inherent numerical ill conditioning of this approach. They go on to exalt the extension of the basic model by including explicit arc capacities as an interesting alternative to the use of asymmetric traffic assignment models, where such extensions are made through the development of complex travel cost functions, which, in practical applications, are difficult to calibrate. In fact, the link flow pattern found by solving a capacitated model may also be found by solving the corresponding uncapacitated problem with travel time functions adjusted by the corresponding optimal shadow prices. The solution of a capacitated problem can therefore be used as a tool for guiding the traffic engineers in correcting the travel time functions so as to bring the flow

pattern into agreement with the anticipated results (Hearn 1980). In a related application, the introduction of capacities can be used to derive tolls for the reduction of flows on overloaded links (Hearn and Ramana 1998); see Bernstein and Smith (1994) for additional references.

It is worth mentioning that some traffic control policies give rise to link flow capacity constraints (Yang and Yagar 1994), that some of the first mathematical models of traffic assignment problems use link flow capacity constraints to model congestion effects (Charnes and Cooper 1961), and that several authors discuss the consequences of including capacities on existing algorithms for the uncapacitated case (Daganzo 1977a, b; Hearn 1980; Hearn and Ribera 1980, 1981; Larsson and Patriksson 1994, 1995, 1999).

A solution to an explicitly capacitated traffic assignment problem will, in the user equilibrium case, no longer comply with Wardrop's (1952) first principle. Hence, following Jorgensen (1963) and Larsson and Patriksson (1995), let us first extend the notion of a user equilibrium to networks with arc capacities. Before we do so, we formally associate a nonnegative capacity c_a with each arc $a \in A$ (which may be ∞). Moreover, we call a flow f feasible if it satisfies all upper bound constraints $f_a \le c_a$ for $a \in A$. For convenience, we henceforth assume that we only consider instances that possess a feasible flow. A path $P \in \mathcal{P}$ is said to be *unsaturated* with respect to a given feasible flow f if and only if $f_a < c_a$ for all arcs $a \in P$. Otherwise, it is called *saturated*.

Definition 3.1. A flow f represents a (*capacitated*) *user equilibrium* if no OD pair has an unsaturated path with strictly smaller cost than any path used for that pair. That is, if $f_P > 0$ for $P \in \mathcal{P}_k$, then $\ell_P(f) \le \min\{\ell_Q(f)\colon Q \in \mathcal{P}_k \text{ unsaturated}\}$.

In the uncapacitated case, Definition 3.1 is obviously equivalent to Wardrop's first principle, because all paths are unsaturated. In particular, all used paths in $\mathcal{P}_k$ are of equal (and actually minimal) latency. In contrast, the flow-carrying paths between the same OD pair in a capacitated user equilibrium can have different latencies (and are therefore not necessarily of minimal length). If we define $L_k(f) := \max\{\ell_P(f)\colon P \in \mathcal{P}_k, f_P > 0\}$, a user equilibrium f satisfies the following conditions: If $\ell_P(f) > L_k(f)$, then $f_P = 0$; if $\ell_P(f) < L_k(f)$, then P is saturated. In other words, we can partition $\mathcal{P}_k$ into three sets: paths that are short and saturated, paths that have a common length equal to $L_k(f)$, and longer paths without flow.

3.1. Inefficiency, nonuniqueness, and nonconvexity of capacitated 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 restrict the route choice for the remaining users. Figure 1 provides an example with two commodities. The nodes on the left represent one OD pair, while the nodes on the right form

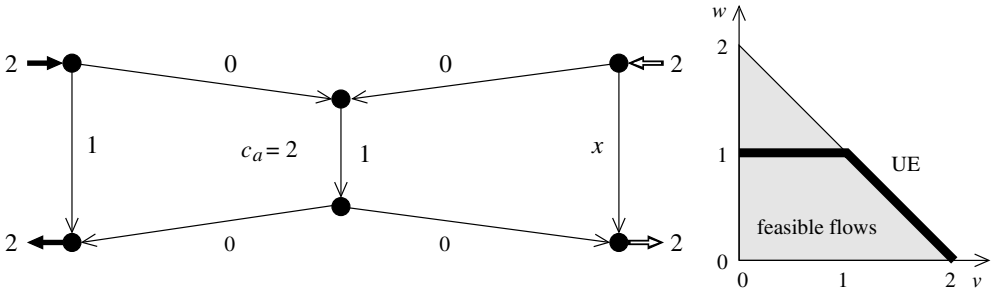

Figure 1. Example showing that the set of user equilibria can be nonconvex. The instance is on the left-hand side. The graph on the right depicts the space of flows. The heavy solid line represents the set of user equilibria (UE).

the other OD pair. The demand rate is 2 in both cases. Arc labels indicate the corresponding latency functions; the arc in the center is the only one with finite capacity. Every user has two options: the route that goes through the center and the alternative at the side. One can represent any feasible flow in this network using two variables. Let v and w denote the flow that is routed through the common arc corresponding to the left and the right OD pair, respectively. Thus, the set of all feasible flows is in one-to-one correspondence with $\{v, w \in [0, 2]\colon\ v + w \leq 2\}$ because the capacity constraint must be obeyed and the flow on the four paths must be nonnegative. According to Definition 3.1, a feasible flow is a capacitated user equilibrium if and only if one of the following two conditions holds:

(i) $w = 1$, i.e., the travel times along both paths for the OD pair on the right are the same,

(ii) $v + w = 2$ and $w < 1$, i.e., the common arc is used up to capacity and the alternative path for the OD pair on the right has higher cost.

Consequently, multiple equilibria with different total travel times can exist. This example additionally shows that the space of equilibria is in general not convex. Indeed, the right part of Figure 1 shows that the projection of the space of flows into the v, w-plane is nonconvex.

Moreover, the price of anarchy (Papadimitriou 2001), the ratio of the cost of the worst user equilibrium to that of the system optimum, is in general unbounded, too. For that, consider the single commodity instance shown in Figure 2. Arc labels again represent the corresponding latency functions; two arcs have finite capacity. The flow that routes $1/2$ on the only path consisting of three arcs and $1/2$ on the arc with constant cost M is a capacitated user equilibrium. Its total travel time is $\frac{1}{2}\left(\frac{1}{2} + 0 + \frac{1}{2}\right) + \frac{1}{2}M = \frac{1}{2}(M + 1)$. On the other hand, the system-optimal flow, which incidentally happens to be another capacitated user equilibrium, routes $1/2$ on each of the two paths with two arcs. Its total travel time is $2\frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{2}$. Clearly, the ratio of the two values goes to infinity when $M \to \infty$.

It is worth mentioning that the definition of a capacitated user equilibrium includes solutions that Marcotte et al. (2004) consider "less natural" because drivers could contribute to the saturation of a shorter path by using a longer path that shares the same bottleneck arc with the shorter one. Actually, an alternative extension of the uncapacitated equilibrium concept is the following:

(3.1) No arbitrarily small bundle of drivers on a common path can strictly decrease its cost by switching to another path.

This definition coincides with the definition of a user equilibrium in Bernstein and Smith (1994) if, in lieu of working with explicit arc capacities, one assumes that latency functions jump to $+\infty$ as soon as arc capacities are exceeded.

While Definition 3.1 and (3.1) are equivalent for uncapacitated networks with continuous and monotone travel cost functions, this is not necessarily the case when some arc capacities are finite. For example, the problem alluded to by Marcotte et al. is obviously eliminated by (3.1). Because Definition 3.1 is more comprehensive than the principle described by

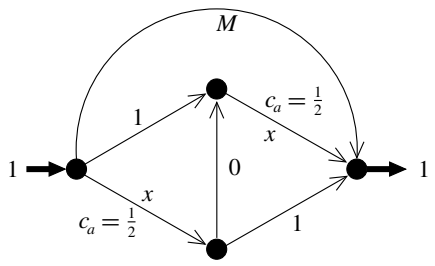

FIGURE 2. Instance with arbitrarily bad equilibria.

(3.1), we chose to go for the broader notion. Nonetheless, the examples given in Figures 1 and 2 are also valid for the more restricted concept. Moreover, the particular equilibrium that we single out next to overcome the difficulty of characterizing the best capacitated user equilibrium in a not necessarily convex space, also satisfies (3.1).

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

$$\min \quad \sum_{a\in A}\int_0^{f_a} \ell_a(x)\,dx \tag{3.2a}$$

$$\text{s.t.} \quad \sum_{P\ni a} f_P = f_a \quad \text{for all } a\in A, \tag{3.2b}$$

$$\sum_{P\in\mathcal{P}_k} f_P = d_k \quad \text{for all } k\in K, \tag{3.2c}$$

$$f_a \le c_a \quad \text{for all } a\in A, \tag{3.2d}$$

$$f_P \ge 0 \quad \text{for all } P\in\mathcal{P}. \tag{3.2e}$$

As this amounts to minimizing a convex function over a nonempty polytope, the set of optimal flows is nonempty and convex. For the example in the previous section, the set of all BMW equilibria is given by $\{0 \le v \le 1, w = 1\}$, as illustrated in Figure 3. Note that a BMW 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. Let us first show that a BMW equilibrium is indeed a capacitated user equilibrium in the sense of (3.1) and hence Definition 3.1.

LEMMA 3.2. *If f is a BMW equilibrium, then it is a capacitated user equilibrium.*

PROOF. To show that (3.1) holds, suppose to the contrary that there are two paths Q, $R\in\mathcal{P}_k$ for some OD pair k with $f_Q > 0$ such that $\ell_R(f^\varepsilon) < \ell_Q(f)$, where

$$f_P^\varepsilon = \begin{cases} f_Q - \varepsilon & \text{if } P = Q,\\ f_R + \varepsilon & \text{if } P = R,\\ f_P & \text{otherwise,}\end{cases} \qquad \text{for } P\in\mathcal{P}$$

is a feasible flow for all $0 < \varepsilon \le \bar{\varepsilon}$ for some $\bar{\varepsilon}$. Now, keep $x := f^{\bar{\varepsilon}}$ fixed and consider

$$\sum_{a\in A}(x_a - f_a)\ell_a(f_a) = \sum_{P\in\mathcal{P}}(x_P - f_P)\ell_P(f) = \bar{\varepsilon}(\ell_R(f) - \ell_Q(f)).$$

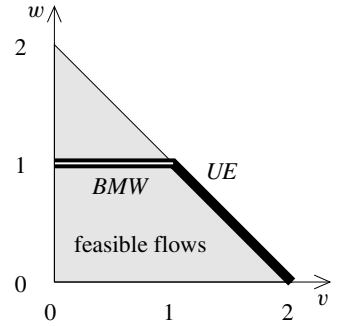

FIGURE 3. Convexity of BMW equilibria.

Because latency functions are continuous and nondecreasing, it follows that $\ell_R(f) - \ell_Q(f) < 0$ and so we have a contradiction to (3.3) below. □

Similarly to condition (2.2), first-order optimality conditions imply that a flow f is a BMW equilibrium if and only if

$$\sum_{a\in A} h_a \ell_a(f_a) \geq 0 \quad \text{for all feasible directions } h. \tag{3.3}$$

LEMMA 3.3. *A feasible flow f is a BMW equilibrium of a network with arc capacities if and only if*

$$C^f(f) \leq C^f(x) \quad \textit{for all feasible flows } x. \tag{3.4}$$

PROOF. Let x be any feasible flow. Hence, $x - f$ is a feasible direction at f (and all feasible directions can be obtained in this way). Therefore, condition (3.3) is equivalent to

$$\sum_{a\in A} (x_a - f_a)\ell_a(f_a) \geq 0,$$

which is just (3.4). □

It is interesting to note that the model (3.2a)–(3.2e) has been used before without the formal introduction of the concept of a capacitated user equilibrium. The model (3.2) has been used before (see Daganzo 1977a, b; Hearn 1980, Hearn and Ribera 1980, 1981; Larsson and Patriksson 1995, 1999, among others). In particular, Hearn (1980) noted that a BMW equilibrium is an *uncapacitated* user equilibrium with respect to latencies $\ell_a(\cdot) + \gamma_a$, where $\gamma_a \geq 0$ is the shadow price (Karush-Kuhn-Tucker multiplier) of the capacity constraint $x_a \leq c_a$ for arc $a \in A$ in an optimal solution to (3.2). This point of view facilitates an alternative proof of (3.4). In fact, let f be a BMW equilibrium and x be any feasible flow. Then,

$$\begin{aligned}
C^f(f) &= \sum_{a\in A} \ell_a(f_a) f_a + \sum_{a\in A} \gamma_a(f_a - c_a) \\
&= \sum_{a\in A} (\ell_a(f_a) + \gamma_a) f_a - \sum_{a\in A} \gamma_a c_a \\
&\leq \sum_{a\in A} (\ell_a(f_a) + \gamma_a) x_a - \sum_{a\in A} \gamma_a c_a \\
&= \sum_{a\in A} \ell_a(f_a) x_a + \sum_{a\in A} \gamma_a(x_a - c_a) \\
&\leq C^f(x).
\end{aligned}$$

Here, the first equality follows from complementary slackness. The first inequality uses (2.4) for uncapacitated user equilibria, while the second one makes use of the feasiblity of x.

Like its counterpart (2.4) for uncapacitated networks, condition (3.4) is crucial for proving results on the efficiency of BMW equilibria. But let us first discuss another good reason for paying attention to BMW equilibria. Suppose one would forgo explicit arc capacities and would instead incorporate barrier terms in the latency functions. More specifically, let $\mu \in \mathbb{R}_{\geq 0}$ be a penalty parameter and consider the modified latency functions $\ell_a^\mu(x_a) := \ell_a(x_a) + \mu/(c_a - x_a)$ for all arcs a with finite capacities, with the understanding that the barrier term equals $+\infty$ for $x_a \geq c_a$. The next lemma essentially shows that in the limit (for $\mu \to 0$), selfish users behave like they would at a BMW equilibrium.

LEMMA 3.4. *Let (μ_i) be a parameter sequence with $0 < \mu_{i+1} < \mu_i$ for $i = 0, 1, \ldots,$ and $\mu_i \to 0$. Let (f^i) be the corresponding sequence of user equilibria in the network without capacities but with modified latencies $\ell_a^{\mu_i}$. Every limit point of the sequence (f^i) is a BMW equilibrium of the original instance (i.e., with capacities).*

PROOF. According to Beckmann et al. (1956), each user equilibrium f^i minimizes the following objective function, subject to (3.2b), (3.2c), and (3.2e):

$$\sum_{a\in A}\int_0^{f_a}\left(\ell_a(x)+\frac{\mu_i}{c_a-x}\right)dx. \tag{3.5}$$

Hence, f^i also minimizes

$$\sum_{a\in A}\int_0^{f_a}\ell_a(x)\,dx-\mu_i\sum_{a\in A}\ln(c_a-f_a), \tag{3.6}$$

which differs from (3.5) by a constant. As the second term in (3.6) is a barrier function as well, it follows that each limit point of (f^i) is an optimal solution of the original problem (3.2) (see, e.g., Bertsekas 1999, Proposition 4.1.1). □

Although each convergent subsequence converges to a BMW equilibrium, it is not true that the coordination ratio of an instance in which capacities are enforced by using modified latency functions approaches the coordination ratio of a BMW equilibrium of the capacitated instance. In other words, if the subsequence $(\bar{f}^i)$ of user equilibria converges to the BMW equilibrium f, then in general

$$\frac{C^{\mu_i}(\bar{f}^i)}{C^{\mu_i}(\bar{f}^{i,*})}\overset{\mu_i\to 0}{\not\longrightarrow}\frac{C(f)}{C(f^*)}. \tag{3.7}$$

Here, f^* and $\bar{f}^{i,*}$ are system-optimal solutions corresponding to the instances for which f and $\bar{f}^i$ are user equilibria, respectively, and C^μ is the cost with respect to latencies ℓ_a^μ. In fact, consider a network of two parallel arcs connecting a single origin with a single destination of demand rate 2. One of the arcs has unit latency and unit capacity while the second arc has latency equal to 2 and infinite capacity. Both the BMW equilibrium and the system optimum route one unit of flow along each arc. The total travel time of both solutions is 3. If we try to enforce the capacity of the first arc with the help of a barrier term, its latency becomes $1+\mu(1-x)^{-1}$. The corresponding user equilibrium is $(1-\mu, 1+\mu)$; here, the first coordinate refers to the capacitated arc. As both latency functions evaluate to 2, the total travel time is 4. The optimal flow is $(1-\sqrt{\mu}, 1+\sqrt{\mu})$, and its total travel time is $3-\mu+2\sqrt{\mu}$. While the sequence $\{(1-\mu, 1+\mu)\}$ converges to the capacitated user equilibrium $(1, 1)$ for $\mu\to 0$, the corresponding sequence of total travel times remains constant at 4. Hence, the left-hand side of (3.7) converges to $4/3$ and not to 1, the value of the right-hand side.

3.3. The efficiency of BMW equilibria. We now present upper bounds on the inefficiency of any BMW equilibrium. Recall from §3.1 that an arbitrary capacitated user equilibrium can be arbitrarily inefficient (in contrast to the situation in networks without capacities). We first focus on a bicriteria result similar to Theorem 3.1 in Roughgarden and Tardos (2002).

THEOREM 3.5. *Consider an instance of the capacitated traffic assignment model* (3.2b)–(3.2e) *with continuous and nondecreasing latency functions. If f is a BMW equilibrium for that instance and x is a feasible flow for the same network but with demands and capacities doubled, then $C(f)\le C(x)$.*

PROOF. Like Roughgarden and Tardos (2002), we start by modifying the original latency functions ℓ_a. Namely,

$$\bar{\ell}_a(x_a):=\begin{cases}\ell_a(f_a) & \text{if } x_a\le f_a,\\ \ell_a(x_a) & \text{if } x_a\ge f_a.\end{cases}$$

The increase of the cost of x with respect to the new latencies is bounded by the following expression:

$$\bar{C}(x) - C(x) = \sum_{a \in A} (\bar{\ell}_a(x_a) - \ell_a(x_a)) x_a \le \sum_{a \in A} \ell_a(f_a) f_a = C(f),$$

where the inequality follows directly from the definition of $\bar{\ell}(\cdot)$. Using $\bar{\ell}_P(x) \ge \bar{\ell}_P(0) = \ell_P(f)$ for any path P, we also obtain

$$\bar{C}(x) = \sum_{P \in \mathcal{P}} \bar{\ell}_P(x) x_P \ge \sum_{P \in \mathcal{P}} \ell_P(f) x_P = C^f(x).$$

Because $x/2$ is feasible for the original instance, condition (3.4) implies that $C^f(x/2) \ge C(f)$. Eventually, putting the three inequalities together yields

$$C(f) = 2C(f) - C(f) \le 2C^f(x/2) - C(f) = C^f(x) - C(f) \le \bar{C}(x) - C(f) \le C(x). \quad \square$$

Note that the theorem remains true for capacities less than twice the original capacities, so long as the new instance still has a feasible solution for twice the demand.

We now turn our attention to the main result, a direct bound on the inefficiency of any BMW equilibrium. We shall continue to assume that latency functions are just continuous and nondecreasing. Let $\mathcal{L}$ be a family of latency functions of that kind. For example, $\mathcal{L}$ could be the polynomials of degree at most n. For every function $\ell \in \mathcal{L}$ and every value $v \ge 0$, let us define

$$\beta(v, \ell) := \frac{1}{v\ell(v)} \max_{x \ge 0} \{x(\ell(v) - \ell(x))\}, \tag{3.8}$$

where by convention $0/0 = 0$. It is obvious that $\beta(v, \ell) \ge 0$ and because $x(\ell(v) - \ell(x)) \le 0$ for $x > v$, we could have restricted the maximum to the interval $[0, v]$. In addition, let us define $\beta(\ell) := \sup_{v \ge 0} \beta(v, \ell)$ and $\beta(\mathcal{L}) := \sup_{\ell \in \mathcal{L}} \beta(\ell)$. Note that $\beta(\mathcal{L}) \le 1$.

THEOREM 3.6. *Let $\mathcal{L}$ be a family of continuous, nondecreasing latency functions. Consider an instance of the capacitated traffic assignment model* (3.2b)–(3.2e) *with latency functions drawn from $\mathcal{L}$. Then, the ratio of the total travel time of a BMW equilibrium f to that of a system optimum f^* is bounded from above by* $(1 - \beta(\mathcal{L}))^{-1}$, *i.e.*,

$$C(f) \le \frac{1}{1 - \beta(\mathcal{L})} C(f^*).$$

PROOF. Let x be a feasible flow. By definition $C^f(x) = \sum_{a \in A} \ell_a(f_a) x_a$; hence,

$$C^f(x) \le \sum_{a \in A} \beta(f_a, \ell_a) \ell_a(f_a) f_a + \sum_{a \in A} \ell_a(x_a) x_a \le \beta(\mathcal{L}) C(f) + C(x). \tag{3.9}$$

From Lemma 3.3, $C(f) \le C^f(x)$, and the claim follows by applying (3.9) to $x = f^*$. $\square$

In spite of the simplicity of its proof, the power and flexibility of Theorem 3.6 will become evident when we relate it to the main result of Roughgarden (2003) next and demonstrate several further implications in §4. The key was to get the definition of $\beta(\mathcal{L})$ "right."

3.4. The parameter $\beta(\mathcal{L})$ and the anarchy value $\alpha(\mathcal{L})$. Let $\mathcal{L}$ be a given family of latency functions. We now relate $\beta(\mathcal{L})$ to the "anarchy value" $\alpha(\mathcal{L})$ introduced by Roughgarden (2003). To do so, we have to assume that, in addition to being continuous and monotone, ℓ is differentiable and $x\ell(x)$ is convex for all $\ell \in \mathcal{L}$ (i.e., the setting of Roughgarden). The *anarchy value* $\alpha(\ell)$ of a latency function ℓ is

$$\alpha(\ell) := \sup_{v > 0:\ \ell(v) > 0} \left[\lambda \frac{\ell(\lambda v)}{\ell(v)} + (1 - \lambda) \right]^{-1},$$

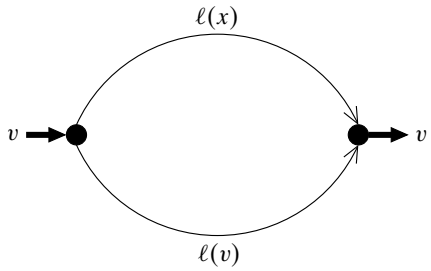

FIGURE 4. Tight instance for the price of anarchy.

where $\lambda \in [0, 1]$ solves $\ell^*(\lambda v) = \ell(v)$. By rearranging terms,

$$\alpha(\ell) = \left[1 - \sup_{v>0:\ \ell(v)>0} \lambda\left(\frac{\ell(v) - \ell(\lambda v)}{\ell(v)}\right)\right]^{-1},$$

and we can prove that $\alpha(\ell) = (1 - \beta(\ell))^{-1}$. Indeed, if we use $x = \lambda v$ in the definition of $\beta(\ell)$, it is clear that $\alpha(\ell) \le (1 - \beta(\ell))^{-1}$. For the other inequality, consider a given v. Because $x(\ell(v) - \ell(x))$ is concave and its value in 0 and v is zero, there is a point $x^* \in (0, v)$ that attains the maximum. From the differentiability of ℓ, $(x(\ell(v) - \ell(x)))'$ evaluated at $x = x^*$ equals zero. Therefore, $\lambda = x^*/v$ satisfies $\ell^*(\lambda v) = \ell(v)$, as required.

Hence, the anarchy value $\alpha(\mathcal{L}) := \sup_{\ell \in \mathcal{L}} \alpha(\ell)$ of a class $\mathcal{L}$ is equal to $(1 - \beta(\mathcal{L}))^{-1}$. Therefore, Theorem 3.6 not only implies Roughgarden's main result (Roughgarden 2003, Theorem 3.8) but also extends it to functions ℓ that are not necessarily differentiable and that do not necessarily satisfy that $x\ell(x)$ is a convex function of x. Moreover, it does not matter if arcs have finite capacities.

We conclude this section by showing that the bound given in Theorem 3.6 is tight. In fact, if $\mathcal{L}$ contains the constant functions, this bound is attained by a single-commodity network consisting of two parallel arcs, which essentially reflects the *independence of the network topology* property highlighted by Roughgarden. Let us assume that the value $\beta(\mathcal{L})$ is achieved for $\ell \in \mathcal{L}$ and $v > 0$. (Although we could use a convergent sequence if the supremum is not attained, we omit this analysis because it does not provide further insights.) Consider the network depicted in Figure 4 (Pigou 1920, Roughgarden 2003), with two parallel links, one with latency $\ell(x)$ and the other with constant latency $\ell(v)$. A demand of v is to be routed. In this situation, the cost of the equilibrium f is $C(f) = v\ell(v)$, while the system optimum f^* can be evaluated as follows:

$$C(f^*) = \min_{0 \le x \le v} \{x\ell(x) + \ell(v)(v - x)\} = v\ell(v) - \max_{0 \le x \le v} \{x(\ell(v) - \ell(x))\}.$$

Hence, the ratio between the total latency of the user equilibrium and that of the system optimum is

$$\frac{C(f)}{C(f^*)} = \left(1 - \frac{\max_{x \ge 0}\{x(\ell(v) - \ell(x))\}}{v\ell(v)}\right)^{-1} = (1 - \beta(\mathcal{L}))^{-1}.$$

4. Computing $(1 - \beta(\mathcal{L}))^{-1}$. Because the results in the last subsection generalize the results by Roughgarden (2003), the bounds he obtains for linear functions, polynomials with positive coefficients, etc. apply here as well. In this section, we study bounds for more general latency functions. We start with two auxiliary lemmas.

LEMMA 4.1. *Let $\mathcal{L}$ be a family of continuous, nondecreasing latency functions ℓ satisfying $\ell(cx) \ge s(c)\ell(x)$ for all $c \in [0, 1]$, for some real function s. Then,*

$$\beta(\mathcal{L}) \le \sup_{0 \le x \le 1} \{x(1 - s(x))\}.$$

PROOF. Recall from (3.8) that $\beta(v, \ell)$ is defined as

$$\beta(v, \ell) = \max_{0 \le x \le v} \left\{ \frac{x}{v} \left(1 - \frac{\ell(x)}{\ell(v)} \right) \right\}.$$

Rewriting x as $v(x/v)$ and using the assumption, we can bound this expression from above by

$$\sup_{0 \le x \le v} \left\{ \frac{x}{v} \left(1 - s\left(\frac{x}{v} \right) \right) \right\} = \sup_{0 \le x \le 1} \{x(1 - s(x))\},$$

which implies the claim. □

LEMMA 4.2. *Let $\mathcal{L}$ be a family of continuous, nondecreasing latency functions ℓ satisfying $\ell(cx) \ge s(c) + \ell(x)$ for all $c \in [0, 1]$, for some real function s. Then,*

$$C(f) \le C(f^*) - |A| D \inf_{0 \le x \le 1} \{xs(x)\},$$

where $D = \sum_{k \in K} d_k$ is the total demand to be routed, f is a BMW equilibrium, and f^* is a system optimum.

PROOF. In this case, it is easy to see that

$$\ell(v)\beta(v, \ell) \le \sup_{0 \le x \le v} \left\{ -\frac{x}{v} s\left(\frac{x}{v} \right) \right\} = - \inf_{0 \le x \le 1} \{xs(x)\}.$$

If we plug this into (3.9) with $\ell = \ell_a$ and $v = f_a$, we obtain

$$C(f) \le \sum_{a \in A} \beta_a(f_a, \ell_a) \ell_a(f_a) f_a + \sum_{a \in A} \ell_a(f_a^*) f_a^* \le C(f^*) - |A| D \inf_{0 \le x \le 1} \{xs(x)\}. \quad \square$$

We now apply Lemmas 4.1 and 4.2 to specific classes of latency functions. Corollaries 4.3 and 4.4 extend Theorem 2.1. Indeed, the following corollary implies that the price of anarchy is 4/3 for all nonnegative concave functions and this bound still holds in networks with capacities (assuming that a BMW equilibrium is chosen). Corollary 4.4 generalizes Roughgarden's bound for polynomials of degree n with positive coefficients.

COROLLARY 4.3. *If the set $\mathcal{L}$ of continuous and nondecreasing latency functions is contained in the set $\{\ell(\cdot)\colon \ell(cx) \ge c\ell(x)$ for $c \in [0, 1]\}$, then $(1 - \beta(\mathcal{L}))^{-1} \le 4/3$.*

PROOF. Use Lemma 4.1 and note that $\sup_{0 \le x \le 1} \{x(1 - x)\} = 1/4$. □

COROLLARY 4.4. *If the set $\mathcal{L}$ of continuous and nondecreasing latency functions is contained in the set $\{\ell(\cdot)\colon \ell(cx) \ge c^n \ell(x)$ for $c \in [0, 1]\}$ for some positive number n, then*

$$(1 - \beta(\mathcal{L}))^{-1} \le \frac{(n+1)^{1+1/n}}{(n+1)^{1+1/n} - n}.$$

PROOF. Use Lemma 4.1 and note that $\sup_{0 \le x \le 1} \{x(1 - x^n)\} = n/(n+1)^{1+1/n}$. □

In particular, the price of anarchy in networks with quadratic or cubic latency functions is 1.626 and 1.896, respectively. It is 2.151 for nonnegative polynomials of degree 4.

Finally, the following result comprises the case in which latency functions are logarithmic (i.e., $\ell(x) = \log(1 + x)$). The BMW equilibrium offers an additive performance guarantee in this situation.

COROLLARY 4.5. *If the set $\mathcal{L}$ of continuous and nondecreasing latency functions is contained in the set $\{\ell(\cdot)\colon \ell(cx) \ge \log_b(c) + \ell(x)$ for $c \in [0, 1]\}$, then*

$$C(f) \le C(f^*) + \frac{|A| D}{e \ln b}.$$

PROOF. Use Lemma 4.2 and note that $\inf_{0 \le x \le 1} \{x \log_b(x)\} = -1/(e \ln b)$. □

5. Lower semicontinuous travel cost functions. Traffic assignment models customarily depend on the assumption of continuous travel cost functions. However, Bernstein and Smith (1994) pointed out that there are times when this assumption is not appropriate. In this situation, a more careful distinction between different versions of the equilibrium concept is essential. (These notions are equivalent to each other for continuous latency functions so that we have previously neglected to draw the fine line between them; however, recall the discussion on a behaviorally meaningful definition of a capacitated user equilibrium (Definition 3.1 versus (3.1)) at the end of §3.1.)

While we do not want to engage in a discussion of that circumstance here and rather refer the reader to the very informative papers by Bernstein and Smith (1994) and de Palma and Nesterov (1998), let us borrow the following example from Florian and Hearn (1995) to illustrate that it may happen that none of the equilibrium concepts takes effect, with the exception of the BMW equilibrium. As in Figure 4, two parallel arcs connect an OD pair with demand rate 2. The travel cost function for the first arc is $\ell_a(x_a) = x_a$; for the second arc, it is

$$\ell_b(x_b) = \begin{cases} x_b & \text{if } x_b < 1, \\ x_b + 1 & \text{if } x_b \geq 1. \end{cases}$$

Note that the solution $x_a = x_b = 1$ indeed is a BMW equilibrium, whereas no solution satisfies Definition 3.1 or condition (3.1).

We will now sketch that, under minor modifications, Theorem 3.6 still holds in the more general setting of latency functions that are just lower semicontinuous. (Note that we maintain the monotonicity assumption.) Bernstein and Smith as well as de Palma and Nesterov underlined the importance of this class of travel cost functions. In particular, a BMW equilibrium always exists and it is a Nash equilibrium. Hence, in this more general setting, Theorem 3.6 still provides a bound on the inefficiency of certain Nash equilibria.

A real function ℓ is lower semicontinuous if $\ell(x) \leq \liminf \ell(x_n)$ for all x in its domain and all sequences (x_n) with $\lim_{n\to\infty} x_n = x$. Here, $\liminf \ell(x_n) = \lim_{n\to\infty} \inf\{\ell(x_m)\colon m \geq n\}$. In fact, if ℓ is nondecreasing and lower semicontinuous, then $\ell(x) = \lim_{y\nearrow x} \ell(y)$, and the limit always exists. (Recall that $\lim_{y\nearrow x} \ell(y)$ represents the limit of $(\ell(y_i))$ with respect to any increasing sequence (y_i) that converges to x from below; $\lim_{y\searrow x} \ell(y)$ is similarly defined.)

For a feasible (arc) flow x, we redefine $C^f(x)$ to be the standard inner product between ∇_f and x, i.e., $C^f(x) := \langle \nabla_f, x \rangle$, where ∇_f is a subgradient of $\sum_{a\in A} \int_0^{f_a} \ell_a(x)\,dx$ at f satisfying the optimality conditions for (3.2). In other words, ∇_f satisfies a condition similar to (3.4), namely $\langle \nabla_f, x - f \rangle \geq 0$ for all feasible flows x. Moreover, note that $\lim_{y\nearrow f_a} \ell_a(y) \leq (\nabla_f)_a \leq \lim_{y\searrow f_a} \ell_a(y)$ for all $a \in A$. The first of these inequalities together with the lower semicontinuity of ℓ implies that $C(f) \leq C^f(f)$. To proceed as we did in the proof of Theorem 3.6, we also need a slight technical change in the definition of $\beta(v, \ell)$, which should now be defined as $\beta(v, \ell) := (1/v\ell(v)) \max_{x\geq 0}\{x(\ell(v^+) - \ell(x))\}$. Here, $\ell(v^+) = \lim_{y\searrow v} \ell(y)$. After these preparations, we can complete the proof. Let x be a feasible flow. We derive

$$C^f(x) = \langle \nabla_f, x \rangle \leq \sum_{a\in A} \ell_a(f_a^+) x_a \leq \sum_{a\in A} \beta(f_a, \ell_a) \ell_a(f_a) f_a + \sum_{a\in A} \ell_a(x_a) x_a \leq \beta(\mathcal{L}) C(f) + C(x).$$

Recall that $C(f) \leq C^f(f)$ by lower semicontinuity and $C^f(f) \leq C^f(x)$ from the optimality conditions. Therefore, the claim follows by replacing x with a system optimum f^*.

Let us eventually note that it appears difficult to extend our main result to families of latency functions that are not lower semicontinuous. Consider an instance consisting of two

nodes connected by arcs a and b (similar to the one depicted in Figure 4) with unit demand. Let the latencies be $\ell_a(f_a) = 1$ and

$$\ell_b(f_b) = \begin{cases} \frac{1}{2} & \text{if } 0 \le f_b < \frac{1}{2}, \\ \frac{2}{3}f_b + \frac{1}{3} & \text{if } \frac{1}{2} \le f_b < 1, \\ \frac{4}{3} & \text{if } f_b \ge 1. \end{cases}$$

The BMW equilibrium f routes all demand along arc b for a total cost of $4/3$. Although a system optimum cannot be attained, it can be approximated by a flow that routes $1/2 + \varepsilon$ along a and the rest along b. For $\varepsilon \to 0$, the total cost goes to $3/4$. Because our previous definition of $\beta(\mathcal{L})$ assumes that latencies are lower semicontinuous, let us consider a more pessimistic notion, for which we can still show that an analog of Theorem 3.6 does not hold. So let $\beta(v, \ell) := (1/v\ell(v^-)) \sup_{x \ge 0}\{x(\ell(v^+) - \ell(x^-))\}$, where $\ell(x^-) = \lim_{y \nearrow x} \ell(y)$. In the example, $\beta(\mathcal{L}) = \sup_{\ell, v} \beta(v, \ell) = 5/12$. Hence, $(1 - \beta(\mathcal{L}))^{-1} C(f^*) = 12/7 \cdot 3/4 = 9/7 < 4/3 = C(f)$. Consequently, Theorem 3.6 (or any reasonable extension thereof) does not hold for discontinuous functions in general.

6. Conclusion. While Wardrop (1952) had used the concept of Nash equilibrium to *describe* user behavior in traffic networks, it has been exploited in traffic management systems to *predict* and in proposals for route-guidance systems to *prescribe* user behavior (e.g., Prager 1954, Steenbrink 1974, Gartner et al. 1980, Boyce 1989). Yet, Nash equilibria in general and user equilibria in particular are known to be inefficient, and many experts have favored in principle the difficult-to-implement system optimum (Merchant and Nemhauser 1978, Henry et al. 1991), which guarantees that the total travel time is minimal. Our results provide an a posteriori justification for employing user equilibria in traffic assignment models. We have shown for a broader class of network models than considered before that the expense of working with user equilibria instead of system optima is limited. In actual fact, while we have confined the above presentation to capacity constraints, virtually all results apply to more general side constraints of the form $f \in X$, for some convex set $X \subseteq \mathbb{R}^A$. Capacity constraints just represent the simplest and arguably the most important class of side constraints. For further details see Stier-Moses (2004).

The introduction of side constraints 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. Moreover, an equilibrium of that quality, namely the BMW equilibrium, can be efficiently computed.

Let us finally remark that all results in this paper also carry on to the setting of *nonatomic congestion games* discussed by Roughgarden and Tardos (2004). Consequently, their findings also hold when side constraints are present (e.g., the elements of the ground set have capacities) and the cost functions satisfy the weaker assumptions made in the paper at hand. This comment also applies to models with nonseparable latency functions, where the latency of one arc may depend on the flow on other arcs as well. The price of anarchy for systems with symmetric nonseparable latency functions was studied in the context of nonatomic congestion games by Chau and Sim (2003), who also considered elastic demands. Subsequently, Perakis (2004) presented bounds on the price of anarchy for asymmetric nonseparable latency functions and fixed demand, which actually are also valid in the presence of side constraints.

Acknowledgments. This work was supported by the High Performance Computation for Engineered Systems (HPCES) program of the Singapore-MIT Alliance (SMA), by ONR Grant N00014-98-1-0317, and by a General Motors Innovation Grant. This research was done while the first author and the third author were with the Operations Research Center at the Massachusetts Institute of Technology.

References

Beckmann, M. J., C. B. McGuire, C. B. Winsten. 1956. *Studies in the Economics of Transportation.* Yale University Press, New Haven, CT.

Bernstein, D., T. E. Smith. 1994. Equilibria for networks with lower semicontinuous costs: With an application to congestion pricing. *Transportation Sci.* **28** 221–235.

Bertsekas, D. P. 1999. *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA.

Boyce, D. E. 1989. Contributions of transportation network modelling to the development of a real-time route guidance system. D. Batten, R. Thord, eds. *Transportation for the Future.* Springer, Berlin, Germany, 161–177.

Boyce, D. E., B. N. Janson, R. W. Eash. 1981. The effect on equilibrium trip assignment of different link congestion functions. *Transportation Res.* **15A** 223–232.

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

Branston, D. 1976. Link capacity functions: A review. *Transportation Res.* **10** 223–236.

Bureau of Public Roads. 1964. *Traffic Assignment Manual.* U.S. Department of Commerce, Urban Planning Division, Washington, DC.

Charnes, A., W. W. Cooper. 1961. Multicopy traffic network models. R. Herman, ed. *Theory of Traffic Flow.* Elsevier, Amsterdam, The Netherlands, 85–96.

Chau, C. K., K. M. Sim. 2003. The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. *Oper. Res. Lett.* **31** 327–334.

Dafermos, S. 1980. Traffic equilibrium and variational inequalities. *Transportation Sci.* **14** 42–54.

Dafermos, S. C., F. T. Sparrow. 1969. The traffic assignment problem for a general network. *J. Res. National Bureau of Standards* **73B** 91–118.

Daganzo, C. F. 1977a. On the traffic assignment problem with flow dependent costs—I. *Transportation Res.* **11** 433–437.

Daganzo, C. F. 1977b. On the traffic assignment problem with flow dependent costs—II. *Transportation Res.* **11** 439–441.

de Palma, A., Y. Nesterov. 1998. Optimization formulations and static equilibrium in congested transportation networks. CORE discussion paper 9861, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Downs, A. 1962. The law of peak-hour expressway congestion. *Traffic Quart.* **16** 393–409.

Dupuit, J. 1849. On tolls and transport charges. *Annales des Ponts et Chaussées*. Reprinted in *Internat. Econom. Papers* **11**(1962) 7–31.

Florian, M. 1999. Untangling traffic congestion: Application of network equilibrium models in transportation planning. *OR/MS Today* **26**(2) 52–57.

Florian, M., D. Hearn. 1995. Network equilibrium models and algorithms. M. O. Ball, T. L. Magnanti, C. L. Monma, G. L. Nemhauser, eds. *Network Routing. Handbooks in Operations Research and Management Science*, Vol. 8, Chap. 6. Elsevier, Amsterdam, The Netherlands, 485–550.

Gartner, N. H., S. B. Gershwin, J. D. C. Little, P. Ross. 1980. Pilot study of computer-based urban traffic management. *Transportation Res.* **14B** 203–217.

Hearn, D. W. 1980. Bounding flows in traffic assignment models. Technical report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL.

Hearn, D. W., M. V. Ramana. 1998. Solving congestion toll pricing models. P. Marcotte, S. Nguyen, eds. *Equilibrium and Advanced Transportation Modeling.* Kluwer Academic Publishers, Boston, MA, 109–124.

Hearn, D. W., J. Ribera. 1980. Bounded flow equilibrium problems by penalty methods. *Proc. IEEE Internat. Conf. Circuits Comput.*, Vol. 1, IEEE, New York, 162–166.

Hearn, D. W., J. Ribera. 1981. Convergence of the Frank-Wolfe method for certain bounded variable traffic assignment problems. *Transportation Res.* **15B** 437–442.

Henry, J. J., C. Charbonnier, J. L. Farges. 1991. Route guidance, individual. M. Papageorgiou, ed. *Concise Encyclopedia of Traffic and Transportation Systems.* Pergamon Press, Oxford, U.K., 417–422.

Jorgensen, N. O. 1963. Some aspects of the urban traffic assignment problem. Master's thesis, Institute of Transportation and Traffic Engineering, University of California at Berkeley, Berkeley, CA.

Knight, F. H. 1924. Some fallacies in the interpretation of social cost. *Quart. J. Econom.* **38** 582–606.

Kohl, J. G. 1841. *Der Verkehr und die Ansiedelungen der Menschen in ihrer Abhängigkeit von der Gestaltung der Erdoberfläche.* Arnold, Dresden, Germany.

Larsson, T., M. Patriksson. 1994. Equilibrium characterizations of solutions to side constrained asymmetric traffic assignment models. *Le Matematiche* **49** 249–280.

Larsson, T., M. Patriksson. 1995. An augmented Lagrangean dual algorithm for link capacity side constrained traffic assignment problems. *Transportation Res.* **29B** 433–455.

Larsson, T., M. Patriksson. 1999. Side constrained traffic equilibrium models—Analysis, computation and applications. *Transportation Res.* **33B** 233–264.

Magnanti, T. L. 1984. Models and algorithms for predicting urban traffic equilibria. M. Florian, ed. *Transportation Planning Models.* North-Holland, Amsterdam, The Netherlands, 153–185.

Marcotte, P., S. Nguyen, A. Schoeb. 2004. A strategic flow model of traffic assignment in static capacitated networks. *Oper. Res.* **52** 191–212.

Merchant, D. K., G. L. Nemhauser. 1978. A model and an algorithm for the dynamic traffic assignment problems. *Transportation Sci.* **12** 183–199.

Papadimitriou, C. H. 2001. Algorithms, games, and the Internet. *Proc. 33rd Annual ACM Sympos. Theory Comput.*, Heraklion, Greece, 749–753.

Patriksson, M. 1994. *The Traffic Assignment Problem: Models and Methods*. VSP, Utrecht, The Netherlands.

Perakis, G. 2004. The price of anarchy when costs are non-separable and asymmetric. D. Bienstock, G. Nemhauser, eds. *Integer Programming and Combinatorial Optimization*, *Lecture Notes in Computer Science*, Vol. 3064. Springer, Berlin, Germany, 46–58.

Pigou, A. C. 1920. *The Economics of Welfare.* Macmillan, London, U.K.

Prager, W. 1954. Problems of traffic and transportation. *Proc. Sympos. Oper. Res. Bus. Indust.* Midwest Research Institute, Kansas City, MO, 105–113.

Roughgarden, T. 2003. The price of anarchy is independent of the network topology. *J. Comput. System Sci.* **67** 341–364.

Roughgarden, T., É. Tardos. 2002. How bad is selfish routing? *J. ACM* **49** 236–259.

Roughgarden, T., É. Tardos. 2004. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games Econom. Behavior.* **47** 389–403.

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

Smith, M. J. 1979. The existence, uniqueness and stability of traffic equilibria. *Transportation Res.* **13B** 295–304.

Stier-Moses, N. E. 2004. Selfish versus coordinated routing in network games. Ph.D. thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Steenbrink, P. A. 1974. *Optimization of Transport Networks.* John Wiley and Sons, London, U.K.

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

Yang, H., S. Yagar. 1994. Traffic assignment and traffic control in general freeway-arterial corridor systems. *Transportation Res.* **28B** 463–486.