SELFISH ROUTING IN CAPACITATED NETWORKS

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

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

ABSTRACT. 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 extensions of recent 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.

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 in which 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 in which 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) 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 and Knight, became accepted as a sound and simple behavioral principle to describe the spreading of...
trips over alternate routes due to congested conditions (Florian 1999). 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:

The average journey time is a minimum.

Not surprisingly, the total (or equivalently, average) travel time is generally not minimized by the user equilibrium, since users do not pay for their external costs (Dupuit 1849; Pigou 1920; Knight 1924).\(^1\) Hence, the recent result that user equilibria are near optimal (Roughgarden and Tardos 2002a) 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 (2002) 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 note 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 (2002a) and Roughgarden (2002) 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. Moreover, we provide simple proofs of these results, which are, in addition, valid for general nondecreasing functions (nondifferentiable, nonconvex and just lower semicontinuous).

The 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 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 Section 3. Here, 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 Section 4. While the previous sections still assume continuous functions, we take a separate look at lower semicontinuous travel cost functions in Section 5. Eventually, 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 telecommunication 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 \(P_k\) be the set of directed (simple) paths in \(D\) connecting

\(^1\)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).
the corresponding origin with its destination, and let \( \mathcal{P} := \bigcup \mathcal{P}_k \). Furthermore, a nonnegative, nondecreasing and continuous latency function \( \ell_a \) 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 Section 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 \( \ell^f_a := \ell_a(f_a) \) plays an important role; here, \( f \) is a given feasible flow. For a feasible flow \( x \), \( C^f(x) := \sum_{a \in A} \ell^f_a x_a \). Notice 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:

\[
\begin{align*}
\text{min} & \quad \sum_{a \in A} \ell_a(f_a) f_a \\
\text{s.t.} & \quad \sum_{P \ni a} f_P = f_a \quad \text{for all } a \in A, \\
& \quad \sum_{P \in \mathcal{P}_k} f_P = d_k \quad \text{for all } k \in K, \\
& \quad f_P \geq 0 \quad \text{for all } P \in \mathcal{P}.
\end{align*}
\]

Roughgarden (2002) 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

\[
\sum_{a \in P} \ell^*_a(f^*_a) \leq \sum_{a \in Q} \ell^*_a(f^*_a) \quad \text{for all } k \in K \text{ and all paths } P, Q \in \mathcal{P}_k \text{ such that } f^*_P > 0.
\]

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

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

In turn, (2.6) can be interpreted as the optimality conditions of a convex min-cost multicommodity flow problem like (2.1) – (2.4) in which (2.1) is replaced by \( \sum_{a \in A} \int_0^{f_a} \ell_a(x) dx \) (Beckmann, McGuire, and Winsten 1956). (Note that Roughgarden’s 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). Here, 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

\[
C^f(f) \leq C^f(x) \quad \text{for all flows } x.
\]

Note that this inequality is a direct consequence of the fact that in equilibrium, users travel on shortest paths with respect to arc costs \( \ell^*_a \).

We are now ready to give a different proof of the main result of Roughgarden and Tardos (2002a) 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 (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 3.5) also provide a worst-case bound on the degradation of the total (and therefore average) travel time that can possibly be caused by Braess’s Paradox.

**Theorem 2.1** (Roughgarden and Tardos 2002a). Let $f$ be a user equilibrium and let $f^*$ be a system optimum for an instance of (2.1) – (2.4) with linear latency functions. Then $C(f) \leq \frac{4}{3} C(f^*)$.

**Proof.** Let $x$ be a feasible flow. From Condition (2.7), $C(f) \leq C^f(x)$. 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 since $(x_a - f_a / 2)^2 \geq 0$. It follows that $\frac{3}{4} C(f) \leq C(x)$ for any feasible flow $x$. Hence, $C(f) \leq \frac{4}{3} C(f^*)$. $\square$

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 3.2 in lieu of Condition (2.7). Moreover, Corollary 4.3 further generalizes this worst-case bound of $4/3$ to travel cost functions $\ell$ satisfying $\ell(c x) \geq c \ell(x)$ 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 $\ell_a$ relate the average travel times to the volumes $f_a$ of traffic 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).\(^2\) Branston (1976) and Larsson and Patriksson (1995) argue that functions of that 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) noted that in the basic model described in Section 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.” Hearn and others, in particular Larsson and Patriksson (1994, 1995, 1999) and, most recently, Marcotte, Nguyen, and Schoeb (2003), have therefore advocated the inclusion of arc flow capacities as an obvious way of improving the quality of traffic assignment models.

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, Janson, and Eash (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 uncaptacitated problem with travel time functions adjusted by the corresponding optimal shadow prices. The solution of a capacitated problem can therefore

\(^2\)Interestingly, the widely popular link delay formula proposed by the Bureau of Public Roads (1964) includes a capacity parameter.
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; check out Bernstein and Smith (1994) for references.

It also 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 used link flow capacity constraints to model congestion effects (Charnes and Cooper 1961), and that several authors discussed the consequences of including capacities on existing algorithms for the uncapacitated case (Daganzo 1977a; Daganzo 1977b; Hearn 1980; Hearn and Ribera 1980; Hearn and Ribera 1981; Larsson and Patriksson 1994; Larsson and Patriksson 1995; Larsson and Patriksson 1999).

A solution to an explicitly capacitated traffic assignment problem will, in the user equilibrium case, no longer comply with Wardrop’s first principle. Hence, 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 $\infty$). Moreover, we call a flow $f$ feasible if it satisfies all upper bound constraints $f_a \leq c_a$, for $a \in A$. For convenience, we henceforth assume that we only consider instances that possess a feasible flow. A path $P \in P$ 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 $P \in P_k$, then $\ell_P(f) \leq \min\{\ell_Q(f) : Q \in P_k, f_Q \text{ unsaturated}\}$.

In the uncapacitated case, Definition 3.1 is obviously equivalent to Wardrop’s first principle, since all paths are unsaturated. In particular, all used paths in $P_k$ 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 $L_k(f) := \max\{\ell_P(f) : P \in P_k, f_P > 0\}$, 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 $P_k$ 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, Multiplicity, 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 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 $\{v, w \in [0, 2] : v + w \leq 2\}$ 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 at least one of the following two conditions hold:

(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 onto 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 $2 \frac{1}{2} (\frac{1}{2} + 1) = \frac{3}{2}$. Clearly, the ratio of the two values goes to infinity when $M \to \infty$.

It is worth mentioning that our definition of a capacitated user equilibrium includes solutions that Marcotte, Nguyen, and Schoeb (2003) 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 one:

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

(3.1)

While both versions 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 instance, the problem alluded to by Marcotte et al. is obviously eliminated by (3.1). Since Definition 3.1 is more comprehensive than the principle described by (3.1), we chose to go for the broader

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{figure2.png}
\caption{Instance with arbitrarily bad equilibria.}
\end{figure}
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, satisfies (3.1), too.

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

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

s.t.
\[
\sum_{P \ni a} f_P = f_a \quad \text{for all } a \in A, \tag{3.3}
\]
\[
\sum_{P \in \mathcal{P}_k} f_P = d_k \quad \text{for all } k \in \mathcal{K}, \tag{3.4}
\]
\[
f_a \leq c_a \quad \text{for all } a \in A, \tag{3.5}
\]
\[
f_P \geq 0 \quad \text{for all } P \in \mathcal{P}. \tag{3.6}
\]

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 Beckmann user equilibria is given by \{0 \leq v \leq 1, w = 1\}, as illustrated in Figure 3. Note that a Beckmann 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.

Similarly to Condition (2.5), first-order optimality conditions imply that a flow \( f \) is a Beckmann user equilibrium if and only if

\[
\text{for all feasible directions } h : \sum_{a \in A} h_a \ell_a(f_a) \geq 0. \tag{3.7}
\]

Lemma 3.2. A feasible flow \( f \) is a Beckmann user equilibrium of a network with arc capacities if and only if

\[
C^f(f) \leq C^f(x) \quad \text{for all feasible flows } x. \tag{3.8}
\]

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 (3.7) is equivalent to

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

which is just (3.8). \( \square \)
Like its counterpart (2.7) for uncapacitated networks, Condition (3.8) is crucial for proving results on the efficiency of (Beckmann) user equilibria. First, however, let us show that a Beckmann user equilibrium is indeed a capacitated user equilibrium in the sense of (3.1) and hence Definition 3.1.

**Lemma 3.3.** If $f$ is a Beckmann user 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 $\ell_R(f^\varsigma) < \ell_Q(f)$, where

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

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

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

Since latency functions are continuous and nondecreasing, it follows that $\ell_R(f) - \ell_Q(f) < 0$ and so we have a contradiction to (3.7).

It is interesting to note that the model (3.2) – (3.6) has been used before without the formal introduction of the concept of a capacitated user equilibrium; see Daganzo 1977a; Daganzo 1977b; Hearn 1980; Hearn and Ribera 1980; Hearn and Ribera 1981; Larsson and Patriksson 1995; Larsson and Patriksson 1999, among others. However, Hearn (1980) noted that a Beckmann user 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) – (3.6). This point of view facilitates an alternative proof of (3.8). In fact, let $f$ be a Beckmann user equilibrium and $x$ be any feasible flow. Then

$$C^f(x) = \sum_{a \in A} \ell_a(f_a)x_a = \sum_{a \in A} (\ell_a(f_a) + \gamma_a)x_a - \sum_{a \in A} \gamma_a c_a$$

Here, the first equality follows from complementary slackness. The first inequality uses (2.7) for uncapacitated user equilibria, while the second one makes use of the feasibility of $x$.

Equipped with Lemma 3.2, we can straightforwardly extend Theorem 2.1 to networks with capacities. However, we instead want to go after instances with more general travel cost functions.

### 3.3. The Efficiency of Beckmann User Equilibria.

We now present upper bounds on the inefficiency of any Beckmann equilibrium. Recall from Section 3.1 that an arbitrary capacitated user equilibrium can be arbitrarily inefficient (which is in contrast to the situation in networks without capacities). We first focus on a bicriteria result a’la Roughgarden and Tardos (2002a, Theorem 3.1).

**Theorem 3.4.** Consider an instance of the capacitated traffic assignment model (3.3) – (3.6) with continuous and nondecreasing latency functions. If $f$ is a Beckmann user equilibrium for that
instance and $x$ is a feasible flow for the same network but with demands and capacities doubled, then $C(f) \leq C(x)$.

**Proof.** Like Roughgarden and Tardos, we start by modifying the original latency functions $\ell_a$. Namely,

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

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

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

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

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

Since $x/2$ is feasible for the original instance, Condition (3.8) implies that $C^f(x/2) \geq C(f)$. Eventually, putting the three inequalities together yields

$$
C(f) = 2C(f) - C(f) \leq 2C^f(x/2) - C(f) = C^f(x) - C(f) \leq \overline{C}(x) - C(f) \leq C(x).
$$

Note that the theorem is still 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 Beckmann user 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 instance, $\mathcal{L}$ could be the polynomials of degree at most $n$. For every function $\ell \in \mathcal{L}$ and every value $v \geq 0$, let us define:

$$
\nu(v, \ell) := \frac{1}{v \ell(v)} \max_{x \geq 0} \left\{ x(\ell(v) - \ell(x)) \right\},
$$

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

**Theorem 3.5.** Let $\mathcal{L}$ be a family of continuous, nondecreasing latency functions. Consider an instance of the capacitated traffic assignment model (3.3) - (3.6) with latency functions drawn from $\mathcal{L}$. Then, the ratio of the total travel time of a Beckmann user equilibrium $f$ to that of a system optimum $f^*$ is bounded from above by $(1 - \nu(\mathcal{L}))^{-1}$, i.e.,

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

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

$$
C^f(x) \leq \sum_{a \in A} \nu(f_a, \ell_a)f_a + \sum_{a \in A} \ell_a(x_a)x_a \leq \nu(\mathcal{L})C(f) + C(x).
$$

From Lemma 3.2, $C(f) \leq C^f(x)$, and the claim follows by applying (3.10) to $x = f^*$.

In spite of the simplicity of its proof, the power and flexibility of Theorem 3.5 will become evident when we relate it to the main result of Roughgarden (2002) next and demonstrate several further implications in Section 4. The key was to get the definition of $\nu(\mathcal{L})$ “right.”
Hence, the ratio between the total latency of the user equilibrium and that of the system optimum is constant latency.

Theorem 3.5 not only implies Roughgarden’s main result (Roughgarden 2002, Theorem 3.8) but also extends it to functions that are not necessarily differentiable and that do not necessarily satisfy the independence of the network topology property highlighted by Roughgarden. Let us assume that the value \( \lambda \) of a latency function \( \ell \) is differentiable and \( x \) is a convex function of \( \lambda \). 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 \( (1 - \nu(\mathcal{L}))^{-1} \). Therefore, Theorem 3.5 not only implies Roughgarden’s main result (Roughgarden 2002, Theorem 3.8) but also extends it to functions \( \ell \) 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 3.5 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 \( \nu(\mathcal{L}) \) 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 since it is easy and does not provide further insights.) Consider the network depicted in Figure 4 (Pigou 1920; Roughgarden 2002), 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:

\[
C(f^*) = \min_{0 \leq x \leq v} \{ x \ell(x) + \ell(v)(v-x) \} = v \ell(v) - \max_{0 \leq x \leq v} \{ x(\ell(v) - \ell(x)) \}.
\]

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

\[
\frac{C(f)}{C(f^*)} = \left(1 - \frac{\max_{x \geq 0} \{ x(\ell(v) - \ell(x)) \}}{v \ell(v)}\right)^{-1} = (1 - \nu(\mathcal{L}))^{-1}.
\]
4. Computing \((1 - \nu(\mathcal{L}))^{-1}\)

Since the results in the last subsection generalize the results by Roughgarden (2002), 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 \(\ell\) satisfying \(\ell(c x) \geq s(c) \ell(x)\) for all \(c \in [0,1]\), for some real function \(s\). Then

\[\nu(\mathcal{L}) \leq \sup_{0 \leq x \leq 1} \{ x(1 - s(x)) \} .\]

**Proof.** Recall from (3.9) that \(\nu(v, \ell)\) is defined as

\[\nu(v, \ell) = \max_{0 \leq x \leq v} \left\{ \frac{x}{v} \left( 1 - \frac{\ell(x)}{\ell(v)} \right) \right\} .\]

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

\[\sup_{0 \leq x \leq v} \left\{ \frac{x}{v} \left( 1 - \frac{x}{v} \right) \right\} = \sup_{0 \leq x \leq 1} \{ x(1 - s(x)) \} ,\]

which implies the claim. \(\square\)

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

\[C(f) \leq C(f^*) - |A| D \inf_{0 \leq x \leq 1} \{ x s(x) \} ,\]

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

**Proof.** In this case, it is easy to see that

\[\ell(v) \nu(v, \ell) \leq \sup_{0 \leq x \leq v} \left\{ - \frac{x}{v} s \left( \frac{x}{v} \right) \right\} = - \inf_{0 \leq x \leq 1} \{ x s(x) \} .\]

If we plug this into (3.10) with \(\ell = \ell_a\) and \(v = f_a\), we obtain,

\[C(f) \leq \sum_{a \in A} \nu(a, \ell_a) \ell_a f_a + \sum_{a \in A} \ell_a(f^*_a) f_a^* \leq C(f^*) - |A| D \inf_{0 \leq x \leq 1} \{ x s(x) \} .\]

\(\square\)

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 Beckmann user 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 \(\{ \ell(\cdot) : \ell(c x) \geq c \ell(x) \text{ for } c \in [0,1] \}\), then \((1 - \nu(\mathcal{L}))^{-1} \leq 4/3\).

**Proof.** Use Lemma 4.1 and note that \(\sup_{0 \leq x \leq 1} \{ x(1 - x) \} = 1/4\). \(\square\)

**Corollary 4.4.** If the set \(\mathcal{L}\) of continuous and nondecreasing latency functions is contained in the set \(\{ \ell(\cdot) : \ell(c x) \geq c^n \ell(x) \text{ for } c \in [0,1] \}\), then

\[ (1 - \nu(\mathcal{L}))^{-1} \leq \frac{(n + 1)^{1+1/n}}{(n + 1)^{1+1/n} - n} . \]
Proof. Use Lemma 4.1 and note that $\sup_{0 \leq x \leq 1} \{x(1 - x^n)\} = \frac{n}{(n + 1)^{1+1/n}}$. \hfill \Box

Finally, the following result comprises the case in which latency functions are logarithmic (i.e., $\ell(x) = \log(1 + x)$). The Beckmann user 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 $\{\ell(\cdot) : \ell(cx) \geq \log_b(c) + \ell(x) \text{ for } c \in [0,1]\}$, then

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

Proof. Use Lemma 4.2 and note that $\inf_{0 \leq x \leq 1} \{x \log_b(x)\} = -\frac{1}{e \ln b}$. \hfill \Box

5. Lower Semicontinuous Travel Cost Functions

Traffic assignment models customarily depend on the assumption of continuous travel cost functions. However, Bernstein and Smith (1994) have 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, whereas no solution satisfies Definition 3.1 or Condition (3.1).)

We will now sketch that, under minor modifications, Theorem 3.5 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 Nesterov and de Palma underline the importance of this class of travel cost functions. In particular, a Beckmann user equilibrium always exists and it is a Nash equilibrium. Hence, in this more general setting, Theorem 3.5 still provides a bound on the inefficiency of certain Nash equilibria.

Recall that a real function $\ell$ 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, $\liminf \ell(x_n) = \lim_{n \to \infty} \inf \{\ell(x_m) : m \geq n\}$. In fact, if $\ell$ is nondecreasing and lower semicontinuous, then $\ell(x) = \lim_{y \to x} \ell(y)$, and the limit always exists.

For a feasible (arc) flow $x$, we redefine $C^f(x)$ to be the standard inner product between $\nabla f$ and $x$, i.e., $C^f(x) := \langle \nabla f, x \rangle$, where $\nabla f$ is a subgradient of $\sum_{a \in A} \int_a^{x_a} \ell_a(x) dx$ at $f$ satisfying the optimality conditions for (3.2) – (3.6). In other words, $\nabla f$ satisfies a condition similar to (3.8), namely $\langle \nabla f, x - f \rangle \geq 0$ for all feasible flows $x$. Moreover, note that $\lim_{y \to f} \ell_a(y) \leq \langle \nabla f, \ell_a \rangle$, for all $a \in A$. The first of these inequalities together with the lower semicontinuity of $\ell$ implies that $C(f) \leq C^f(f)$. To proceed as we did in the proof of Theorem 3.5, we also need a slight technical change in the definition of $\nu(v, \ell)$, which should now be defined as $\nu(v, \ell) :=$
\begin{align*}
\frac{1}{\nu(\ell)} \max_{x \geq 0} \{ x (\ell(v^+) - \ell(x)) \}. \quad \text{Here, } \ell(v^+) = \lim_{y \searrow x} \ell(y). \quad \text{After these preparations, we can complete the proof. Let } x \text{ be a feasible flow. We derive}
\end{align*}

\begin{align*}
C^f(x) = \langle \nabla f, x \rangle & \leq \sum_{a \in A} \ell_a(f_a^+)^+ x_a \leq \sum_{a \in A} \nu(f_a, \ell_a) \ell_a(f_a) f_a + \sum_{a \in A} \ell_a(x_a) x_a \leq \nu(\mathcal{L}) C(f) + C(x) .
\end{align*}

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

\begin{align*}
\ell_b(f_b) = \begin{cases} 
\frac{1}{2} & \text{if } 0 \leq f_b < \frac{1}{2}, \\
\frac{3}{4} f_b + \frac{1}{3} & \text{if } \frac{1}{2} \leq f_b < 1, \\
\frac{4}{3} & \text{if } f_b \geq 1.
\end{cases}
\end{align*}

The Beckmann user 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 \( 3/4 \). Since our previous definition of \( \nu(\mathcal{L}) \) assumes that latencies are lower semicontinuous, let us consider a more pessimistic notion, for which we can still show that an analog of Theorem 3.5 does not hold. So let \( \nu(v, \ell) := \frac{1}{\nu(\ell)} \sup_{x \geq 0} \{ x (\ell(v^+) - \ell(x^-)) \} \), where \( \ell(x^-) = \lim_{y \searrow x} \ell(y) \). In the example, \( \nu(\mathcal{L}) = \sup_{\ell, v} \nu(v, \ell) = 5/12 \). Hence, \( (1 - \nu(\mathcal{L}))^{-1} C(f^*) = \frac{12}{7} \frac{3}{4} = \frac{9}{7} < \frac{4}{3} = C(f) \). Consequently, Theorem 3.5 (or reasonable extensions 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, Gershwin, Little, and Ross 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, Charbonnier, and Farges 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.

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. Moreover, an equilibrium of that quality, namely the Beckmann user 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 (2002b). Consequently, their findings also hold when the elements of the ground set have capacities and the cost functions satisfy the weaker assumptions made in the paper at hand.

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.
References


