Efficiency and Fairness of System-Optimal Routing with User Constraints*

Andreas S. Schulz
Sloan School of Management, Massachusetts Institute of Technology, Office E53-361, 77 Massachusetts Ave., Cambridge, Massachusetts 02139

Nicolás E. Stier-Moses
Graduate School of Business, Columbia University, Uris Hall, Room 418, 3022 Broadway Ave., New York, New York 10027

We study the route-guidance system proposed by Jahn, Möhring, Schulz, and Stier-Moses [Operations Research 53 (2005), 600–616] from a theoretical perspective. As system-optimal guidance is known to be problematic, this approach computes a traffic pattern that minimizes the total travel time subject to user constraints. These constraints are designed to ensure that routes suggested to users are not much longer than shortest paths for the prevailing network conditions. To calibrate the system, a certain measure—called normal length—must be selected. We show that when this length is defined as the travel time at equilibrium, the resulting traffic assignment is provably efficient and close to fair. To measure efficiency, we compare the output to the best solution without guidance and to user equilibria. To measure unfairness, we compare travel times of different users, and show that they do not differ too much. Inefficient or unfair traffic assignments cause users to travel too long or discourage people from accepting the system; either consequence would jeopardize the potential impact of a route-guidance system. © 2006 Wiley Periodicals, Inc. NETWORKS, Vol. 48(4), 223–234 2006

Keywords: selfish routing; price of anarchy; computational game theory; multicommodity flows; route guidance; traffic assignment

1. INTRODUCTION

Transportation authorities and users alike hope that route-guidance systems can help to mitigate the congestion generated by the ever-increasing amount of vehicular traffic. In particular, in-car navigation devices might be used not only to provide drivers with map information and current traffic conditions, but also to optimize the entire network. With this application in mind, Jahn et al. [13] introduced a route-guidance system that computes a traffic assignment that minimizes the total travel time subject to certain user constraints. These constraints are designed to overcome an inherent problem of system-optimal guidance. Indeed, it is well known that in a system-optimal flow some users may be assigned to considerably longer routes for the benefit of others. User constraints are intended to guarantee that no recommended route is significantly longer than that suggested to any other user with the same origin and destination. For the sake of algorithmic efficiency, Jahn et al. proposed performing this comparison based on normal path lengths instead of actual travel times. The *normal length* of a path is defined via some *a priori* estimate of the real travel time. Based on extensive computational studies on real-world instances, they concluded that the resulting constrained system-optimal flow has two desirable properties when the normal length is properly chosen: the total travel time in the network is close to that of the unconstrained system optimum, and individual users do not experience a notably larger travel time than others. In other words, the resulting traffic assignment is virtually efficient and fair.

In this article, we provide a theoretical framework for the work of Jahn et al. [13]. For our analysis, we rely on the price-of-anarchy concept [7, 15, 21, 23]. Although the *price of anarchy* was originally defined as the ratio of the total travel time of a Nash equilibrium to that of an ordinary system optimum, we adopt a more pragmatic perspective for the application considered here. Put in our terms, the original measure compares the user equilibrium to a system-optimal traffic flow, but the latter cannot realistically be used for route guidance because of its known unfairness [16]. Instead, we measure the price of anarchy with respect to a traffic pattern that can potentially be used in practice. In other words, we evaluate the efficiency of the route-guidance system with the

Received August 2004; accepted June 2006
Correspondence to: N. E. Stier-Moses; e-mail: ns2224@columbia.edu
*An extended abstract of a preliminary version appeared in the Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms [24].
DOI 10.1002/net.20133
Published online in Wiley InterScience (www.interscience.wiley.com).

help of the worst-case ratio of the total travel time of an equilibrium to that of a constrained system optimum. In addition, we compare the travel times experienced by different users between the same origin–destination (OD) pair analytically. A primary goal of any route-guidance system is to offer routes with similar travel times; otherwise, route recommendations would most likely be dismissed. We measure the *unfairness* of a flow as the worst-case ratio between travel times of users traveling between the same OD pair. Small unfairness has another desirable side effect. As the system assigns users to routes randomly, routes offered under similar circumstances (e.g., the same user performs the same trip every day) have similar travel times. That is, almost fair flows reduce the variance of latencies experienced by individual users.

Section 2 formally introduces constrained system optima and the price of anarchy. In addition, we study instances for which the corresponding equilibria have high total cost. Section 3 discusses the efficiency of solutions returned by the route-guidance system of Jahn et al. In Section 3.1, we concentrate on free-flow travel times as normal lengths, and argue that in this case user equilibria might be preferable over constrained system optima because they often feature smaller total cost. Section 3.2 analyzes the case in which normal lengths are defined as user equilibrium travel times. In contrast to the previous case, the cost of constrained system optima is guaranteed to be lower than that of user equilibria. This theoretical result plus the good performance in traffic assignments of real-world instances [13] suggests that this is an excellent choice for normal lengths. Moreover, the main result in Section 4 shows that the unfairness of constrained system optima is bounded above by a small constant. All results established here corroborate the conclusions drawn in [13]. Specifically, we can prove that constrained system optima are nearly efficient and fair, under the proper choice of normal lengths.

2. PRELIMINARIES

The road network is represented by a directed multigraph $G = (N, A)$ with two attributes for each arc $a \in A$: the normal length $\tau_a \geq 0$ gives an *a priori* estimate of the actual traversal time in the solution we seek; the latency function $\ell_a : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ maps the traffic flow x_a on arc a to its traversal time $\ell_a(x_a)$. The normal arc lengths have to be fixed in advance. Their proper choice will allow us to compute solutions that users of the route-guidance device are likely to accept; we refer the reader to Section 3 for details. The latency functions are assumed to be continuous and nondecreasing. These assumptions are naturally met by common latency functions (see, e.g., [6, 25]). We only consider latency functions that belong to a specified set $\mathcal{L}$. In practice, latency functions take a specific form, and the bounds one obtains without this restriction are unnecessarily pessimistic [23]. For some results, we will additionally assume that latency functions are affine; that is, they belong to $\mathcal{L}_{\text{aff}} := \{\ell : \ell(x) = qx + r \text{ for some } q, r \geq 0\}$. Although this may appear restrictive at first, congestion effects and counterintuitive phenomena can already occur, as evidenced by the apparent paradox discovered by Braess [4].

Vehicles are grouped according to their OD pairs $K \subseteq N \times N$. For each OD pair $k = (s_k, t_k) \in K$, let $\mathcal{P}_k$ be the set of directed (simple) paths in G from s_k to t_k, and let $d_k > 0$ be the demand rate associated with OD pair k. Let $\mathcal{P} := \bigcup_{k \in K} \mathcal{P}_k$ be the set of paths between all OD pairs. Because route-guidance systems eventually have to propose paths to the drivers, our formulation is path-based: a feasible flow x assigns a nonnegative value x_P to every path $P \in \mathcal{P}$ such that $\sum_{P \in \mathcal{P}_k} x_P = d_k$ for all $k \in K$. Note that flows are not required to be integral; they describe average rates. Furthermore, we define the latency of a path $P \in \mathcal{P}$ under a given flow x as $\ell_P(x) := \sum_{a \in P} \ell_a(x_a)$, where $x_a := \sum_{\{Q \in \mathcal{P} : a \in Q\}} x_Q$. We refer to the maximum latency of a flow-carrying path in $\mathcal{P}_k$ as $L_k(x) := \max\{\ell_P(x) : P \in \mathcal{P}_k, x_P > 0\}$, and to the shortest normal path length for OD pair k as $T_k := \min\{\tau_P : P \in \mathcal{P}_k\}$. Here, $\tau_P := \sum_{a \in P} \tau_a$ is the normal length of path P.

There are two aspects that define the quality of a flow. The fairness of the route assignment is of importance to the users, while the total travel time in the system is of importance to the traffic authority. Different users traveling between the same OD pair should experience the same travel time. If that was not the case, users would have an incentive to switch routes. We say that a flow with this property is *fair*. In a seminal contribution, Wardrop ([27], p. 345) stated a principle that formalizes this notion: "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." Traffic patterns satisfying this principle are called *user equilibria* [11], and will be denoted by f^{UE}. (Note that a flow can be fair without being a user equilibrium.) Although there may be multiple equilibria, the travel time that users experience is invariant across different equilibria [2]. In particular, all equilibria share the same total cost. We define the *unfairness* of a given flow x as the maximum ratio of the experienced travel times of two users sharing the same OD pair, that is, as $\max\{\ell_Q(x)/\ell_R(x) : Q, R \in \mathcal{P}_k, x_Q, x_R > 0,\ k \in K\}$. (Jahn et al. called it the *loaded unfairness* [13].) In Section 4, we bound the unfairness of constrained system optima by $\gamma(\mathcal{L})$. The value $\gamma(\mathcal{L})$, to be defined in that section, depends only on the set $\mathcal{L}$ of allowed latency functions; for example, it is $p + 1$ for polynomials of degree p.

It has been known since the classical work of Dupuit dating back to the middle of the 19th century, that equilibria can be inefficient [12]. Braess' famous paradox shows a similar situation for transportation networks [4]. Not only does this mean that there are solutions where the group is better off as a whole, but also that each user might benefit individually. Those solutions are usually called *social* (or *system*) *optimal*. Despite their efficiency, it is unrealistic to implement system optima in the context of transportation networks because they may be unfair [3]. Nevertheless, their study is important because they provide bounds on how efficient a system can possibly be, and they are relevant to centralized systems such as freight networks.

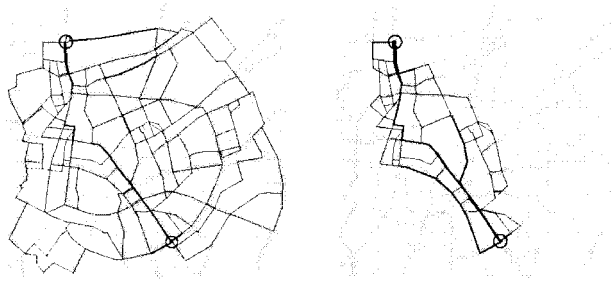

FIG. 1. Example without and with restrictions on path lengths.

The route-guidance system proposed by Jahn et al. [13] was designed to compute the most efficient traffic assignment among those that are not too unfair. As it is difficult to control the unfairness directly, the authors introduced an upper bound on the ratio of the normal length of different users traveling between the same OD pair. Let us describe this approach in more detail. Let $\varphi \geq 1$ quantify the tolerance of users to suboptimal paths. This factor is used to prevent users from being assigned to paths that are longer than φ times the length of a shortest path between their OD pair. In other words, users of OD pair $k \in K$ may only be assigned to paths in $\mathcal{P}_k^\varphi := \{P \in \mathcal{P}_k : \tau_P \leq \varphi T_k\}$. We call a path feasible when it belongs to $\mathcal{P}^\varphi := \bigcup_{k\in K} \mathcal{P}_k^\varphi$. Selecting the solution with minimum total travel time from all assignments of users to paths in $\mathcal{P}^\varphi$ is equivalent to solving the following minimum cost multicommodity flow problem with path constraints and a separable nonlinear objective function:

Problem CSO:

$$\begin{aligned} \min\ C(x) \ :=\ & \sum_{a\in A} \ell_a(x_a)x_a \\ \text{s.t.}\ & \sum_{P\in\mathcal{P}_k^\varphi} x_P = d_k && \text{for all } k \in K, \\ & \sum_{P:a\in P\in\mathcal{P}^\varphi} x_P = x_a && \text{for all } a \in A, \\ & x_P \geq 0 && \text{for all } P \in \mathcal{P}^\varphi. \end{aligned}$$

A *constrained system optimum* with tolerance factor φ, denoted by f^φ, is an optimal solution to this problem. A *system optimum* corresponds to the unconstrained case (when $\varphi = \infty$, path constraints disappear). We use f^{SO} to denote a system optimum. It is evident that the larger the factor φ, the larger is the feasible region. Consequently, $C(f^\varphi)$ is a nonincreasing function of φ, and $C(f^{\mathrm{SO}}) \leq C(f^\varphi)$ for all $\varphi \geq 1$. When $C(x)$ is convex and differentiable, Beckmann et al. [2] proved that a flow f^{SO} is a system optimum with respect to latency functions $\ell_a(x)$ if and only if it is a user equilibrium with respect to the modified latency functions $\ell_a^*(x) := \ell_a(x) + x\,\ell_a'(x)$. The latency functions ℓ_a^* include an extra term that accounts for the service degradation caused to the other users of arc a. As f^{SO} is at equilibrium with respect to ℓ_a^*, we denote the common travel time for all users of OD pair $k \in K$ by $L_k^*(f^{\mathrm{SO}})$.

Figure 1 demonstrates the effect of length constraints on the system optimum. Line thickness reflects arc capacity (light gray) and arc usage (black). The picture on the left displays the ordinary system optimum. The flow is distributed widely over the network to avoid high congestion on arcs and keep travel times low. In the picture on the right, the same demand is routed with the restriction that (free-flow) normal path lengths must not exceed the shortest normal path length by more than 10% (i.e., $\varphi = 1.1$).

In addition, Figure 2 presents a small numerical example. The instance has unit demand between two nodes that are connected by three arcs with latency functions $1+\varepsilon$, $1+x_2$, and $1+(x_3)^2$, respectively, for some $\varepsilon > 0$. The normal lengths are defined to be $1+\varepsilon$, 1, and 1 respectively, according to the free-flow travel time $\ell_a(0)$ of each arc (as in Section 3.1). A constrained system optimum for some $\varphi < 1+\varepsilon$ can only route flow on the lower two arcs; it therefore minimizes $(1+x_2)x_2+(1+(x_3)^2)x_3$ subject to $x_2+x_3 = 1$ and $x_2, x_3 \geq 0$. For $\varphi \geq 1+\varepsilon$, all arcs can be used, and then the problem is a regular multicommodity minimum cost flow problem with objective function $C(x)$.

2.1. The Price of Anarchy

In the context of transportation networks, Wardrop [27] was the first to formalize the notions of user equilibrium and social optimum, although these terms were coined a couple of decades later by Dafermos and Sparrow [11]. Beckmann et al. [2] characterized both solution concepts mathematically, which allowed them to prove existence, uniqueness, and also to provide algorithms for their computation. It was a long-standing open question to characterize the distance from optimality for a system for which no central coordination is imposed. For example, Mahmassani and Peeta ([16], p. 84) wrote:

> …the extent of the differences between SO [system optimum] and UE [user equilibrium] solutions, particularly in terms of overall system cost, is not known. This is very important for ATIS [Advanced Traveler Information Systems] because if the two solutions are not perceptibly different, coordinated cooperative SO route guidance imposed by a central controller may not be necessary, and descriptive information that is less complicated and simpler to disseminate to noncooperating drivers may be sufficient.

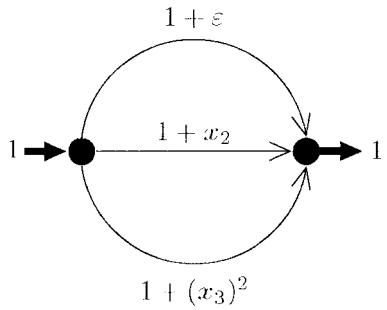

FIG. 2. A simple numerical example.

Luckily, this question has recently been answered for different systems, starting with the work of Koutsoupias and Papadimitriou [15]. They proposed to measure the inefficiency of equilibria in a network consisting of two nodes and multiple parallel arcs by computing the worst-case ratio of the social cost of an equilibrium to that of a system optimum. Later, a series of papers studied the efficiency of user equilibria in traffic networks, under less and less restrictive assumptions: affine latency functions [23], convex and differentiable latency functions [21], general latency functions in networks with side constraints [7], nonseparable symmetric latency functions [5], nonseparable asymmetric latency functions [17], and networks with a fixed congestion level and arbitrary latency functions [8]. For additional references, see also [22].

Results in [7, 8, 17] use variational inequalities to characterize user equilibria, which makes it possible to extend, refine, and simplify the initial bounds given in [21, 23]. Variational inequalities were proposed by Smith [26] and Dafermos [10] as a very powerful tool to compute user equilibria in very general instances such as those with nonseparable and asymmetric latency functions. In our case, a flow f^{UE} is a user equilibrium if and only if it satisfies the following variational inequality:

$$\sum_{a\in A} \ell_a\left(f_a^{\mathrm{UE}}\right) f_a^{\mathrm{UE}} \le \sum_{a\in A} \ell_a\left(f_a^{\mathrm{UE}}\right) x_a \text{ for all feasible flows } x. \tag{2.1}$$

Roughgarden proved that the cost $C(f^{\mathrm{UE}})$ of a user equilibrium is bounded from above by $\alpha(\mathcal{L})$ times the cost $C(f^{\mathrm{SO}})$ of a system optimum [21]. The constant $\alpha(\mathcal{L})$, defined in Section 2.2, depends only on the set of allowed latency functions; for example, it is 4/3 for affine, 1.626 for quadratic, and 1.896 for cubic functions, respectively. For polynomials of degree p, it grows asymptotically like $p/\ln p$. Therefore, although $\alpha(\mathcal{L})$ is not very large, users in real networks can still benefit from coordination, if done appropriately.

We measure the potential benefits of guiding users by comparing the cost of user equilibria to that of constrained system optima, in a similar fashion to the definition of the price of anarchy. Although the original definition relies on ordinary system optima, our notion is arguably more realistic in this context because ordinary system optima cannot be implemented in traffic networks because of their unfairness [16]. In the following definition, $f_{\mathcal{I}}^{\mathrm{UE}}$ and $f_{\mathcal{I}}^{\varphi}$ denote a user equilibrium and a constrained system optimum of an instance $\mathcal{I}$, respectively. To simplify notation, we drop the subindex $\mathcal{I}$ afterwards. The price of anarchy for a given tolerance factor φ and a given set $\mathcal{L}$ of allowed latency functions is defined as follows:

$$\alpha^{\varphi}(\mathcal{L}) := \sup_{\mathcal{I}\in \mathrm{inst}(\mathcal{L})} \frac{C\left(f_{\mathcal{I}}^{\mathrm{UE}}\right)}{C\left(f_{\mathcal{I}}^{\varphi}\right)}. \tag{2.2}$$

Here, $\mathrm{inst}(\mathcal{L})$ is the set of instances with latency functions drawn from $\mathcal{L}$. It is immediately clear that $\alpha^1(\mathcal{L}) \ge 1$ and that $\alpha^{\varphi}(\mathcal{L})$ is a nondecreasing function of φ. In addition, the previously mentioned bounds imply that

$$C(f^{\mathrm{UE}}) \le \alpha(\mathcal{L})\, C(f^{\mathrm{SO}}) \le \alpha(\mathcal{L})\, C(f^{\varphi}). \tag{2.3}$$

Equivalently, $\alpha^{\varphi}(\mathcal{L}) \le \alpha(\mathcal{L})$ for all $\varphi \ge 1$. Moreover, for instances with positive minimum normal length T_k for all OD pairs $k \in K$, a constrained system optimum with large tolerance is optimal in the unconstrained sense; that is, $C(f^{\varphi}) = C(f^{\mathrm{SO}})$ when φ is sufficiently large.

2.2. Tight Instances

For a specific instance, we refer to the ratio of the cost of a user equilibrium to that of a constrained system optimum as the *coordination ratio* of the instance. To help us understand what causes the inefficiency of equilibria, we now characterize instances with high coordination ratio. Although we work with ordinary system optima in this section, we will concentrate on arbitrary tolerance factors φ later. We call an instance *tight* when its coordination ratio $C(f^{\mathrm{UE}})/C(f^{\mathrm{SO}})$ matches the upper bound $\alpha(\mathcal{L})$. Here, f^{UE} and f^{SO} are a user equilibrium and a system optimum of the corresponding instance, respectively. Roughgarden and Tardos [23] presented an instance, inspired by the work of Pigou [18], that is tight for affine latency functions. Subsequently, Roughgarden [21] proved that the maximum coordination ratio among all instances with two parallel arcs and latency functions drawn from $\mathcal{L}$ matches the price of anarchy. This implies that the inefficiency of equilibria does not arise because networks are big and complex, and have multiple OD pairs, but merely from the selfishness of users.

Figure 3 illustrates Roughgarden's network, which consists of two nodes connected by two parallel arcs and a demand rate equal to d. For a given function $\ell \in \mathcal{L}$, the travel time of the top arc is $\ell(d)$ regardless of its flow, while the bottom arc has a travel time of $\ell(x)$. The user equilibrium f^{UE} assigns the entire demand d to the lower arc, whereas the system optimum f^{SO} routes $d - x^*$ on the top arc. Here, $x^* := \arg\max x(\ell(d) - \ell(x))$. The costs are $C(f^{\mathrm{UE}}) = \ell(d)d$ and $C(f^{\mathrm{SO}}) = \ell(d)(d - x^*) + \ell(x^*)x^*$, respectively. The supremum over d and ℓ of the ratio of the social costs precisely matches the following definition of $\alpha(\mathcal{L})$. Following

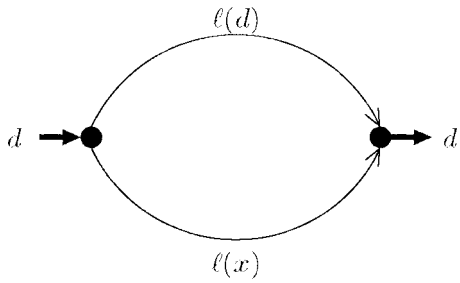

FIG. 3. Simple tight instance.

Correa et al. [7], we set $\beta(v,\ell) := \max\left\{\frac{x}{v}\left(1-\ell(x)/\ell(v)\right) : 0 \le x \le v\right\}$ and $\beta(\mathcal{L}) := \sup\{\beta(v,\ell) : v \ge 0 \text{ and } \ell \in \mathcal{L}\}$, which allows us to define $\alpha(\mathcal{L}) := (1-\beta(\mathcal{L}))^{-1}$. (We refer the reader to [8] for more details and intuition on these definitions.) This instance shows that the price of anarchy is at least $\alpha(\mathcal{L})$; the theorem below proves the reverse inequality. The proof we provide, which appeared in [7], will allow us to establish conditions that characterize instances that are tight in the unconstrained case.

Theorem 2.1 ([7, 21])**.** *Consider an instance with latency functions drawn from a set of continuous and nondecreasing latency functions $\mathcal{L}$. Then, $C(f^{\mathrm{UE}}) \le (1-\beta(\mathcal{L}))^{-1}\, C(f^{\mathrm{SO}})$.*

Proof ([7])**.** The claim follows from

$$C(f^{\mathrm{UE}}) \le \sum_{a\in A} \ell_a\left(f_a^{\mathrm{UE}}\right) f_a^{\mathrm{SO}} \le \sum_{a\in A} \beta\left(f_a^{\mathrm{UE}}, \ell_a\right) \ell_a\left(f_a^{\mathrm{UE}}\right) f_a^{\mathrm{UE}} + \sum_{a\in A} \ell_a\left(f_a^{\mathrm{SO}}\right) f_a^{\mathrm{SO}} \le \beta(\mathcal{L}) C(f^{\mathrm{UE}}) + C(f^{\mathrm{SO}}). \quad (2.4)$$

The first inequality is identical to (2.1); the other two follow directly from the definition of β. ∎

Observation 2.2. *Let $\mathcal{L}$ be a family of continuous and nondecreasing latency functions. An instance with latency functions drawn from the set $\mathcal{L}$ is tight if and only if the following three conditions are satisfied:*

for all $k \in K$ and $P \in \mathcal{P}_k$:

$$f_P^{\mathrm{SO}} > 0 \Rightarrow \ell_P(f^{\mathrm{UE}}) = L_k(f^{\mathrm{UE}}), \quad (2.5a)$$

$$\text{for all } a \in A: \quad f_a^{\mathrm{SO}} = \arg\max_{x\ge 0} x(\ell_a(f_a^{\mathrm{UE}}) - \ell_a(x)), \quad (2.5b)$$

$$\text{for all } a \in A: \quad \ell_a(f_a^{\mathrm{UE}}) f_a^{\mathrm{UE}} > 0 \Rightarrow \beta(f_a^{\mathrm{UE}}, \ell_a) = \beta(\mathcal{L}). \quad (2.5c)$$

Here, f^{SO} and f^{UE} are an arbitrary system optimum and user equilibrium, respectively.

Proof. An instance is tight if and only if all inequalities in the proof of the previous theorem are equalities. Equation (2.5a) follows from $\sum_{P\in\mathcal{P}} \ell_P(f^{\mathrm{UE}}) f_P^{\mathrm{UE}} = \sum_{P\in\mathcal{P}} \ell_P(f^{\mathrm{UE}}) f_P^{\mathrm{SO}}$, which is the flow-on-paths version of (2.1). The second inequality and the definition of $\beta_a(f_a^{\mathrm{UE}}, \ell)$ give (2.5b). Finally, the third inequality is equivalent to (2.5c). ∎

Let us make a few remarks related to Observation 2.2:

(1) When latency functions are differentiable, setting the derivative of the right-hand side of condition (2.5b) to zero, we see that $\ell_a^*(f_a^{\mathrm{SO}}) = \ell_a(f_a^{\mathrm{UE}})$. This implies that $\ell_P^*(f^{\mathrm{SO}}) = \ell_P(f^{\mathrm{UE}})$ for all $P \in \mathcal{P}$, and that $L_k^*(f^{\mathrm{SO}}) = L_k(f^{\mathrm{UE}})$ for all $k \in K$. For example, when latencies are affine, any user equilibrium and system optimum of a tight instance must satisfy $f_a^{\mathrm{SO}} = f_a^{\mathrm{UE}}/2$ for all arcs $a \in A$, with the exception of those that have constant travel time.

(2) Under differentiability, condition (2.5a) is implied by the optimality of f^{SO} and remark (1).

(3) For arcs with strictly increasing latency functions, condition (2.5b) implies that either $f_a^{\mathrm{SO}} < f_a^{\mathrm{UE}}$ or $f_a^{\mathrm{SO}} = f_a^{\mathrm{UE}} = 0$.

(4) Assume that $\mathcal{L}$ is closed under addition (i.e., if $\ell \in \mathcal{L}$ and $r \ge -\ell(0)$, then $\ell + r \in \mathcal{L}$) and that $\beta(\mathcal{L}) > 0$ (otherwise, the price of anarchy is 1 and all instances are tight). If an arc a carries flow in a user equilibrium, it must satisfy $\ell_a(0) = 0$.

Proof. Suppose $\ell_a(0) > 0$. Let x^* be a maximizer of $x(\ell_a(f_a^{\mathrm{UE}}) - \ell_a(x))$ in the definition of $\beta(f_a^{\mathrm{UE}}, \ell_a)$. Because $0 < \ell_a(0) \le \ell_a(f_a^{\mathrm{UE}})$, condition (2.5c) implies that $\beta(f_a^{\mathrm{UE}}, \ell_a) > 0$, meaning that $\ell_a(x^*) < \ell_a(f_a^{\mathrm{UE}})$. Therefore,

$$\frac{x^*}{f_a^{\mathrm{UE}}}\left(1 - \frac{\ell_a(x^*)}{\ell_a(f_a^{\mathrm{UE}})}\right) < \frac{x^*}{f_a^{\mathrm{UE}}}\left(1 - \frac{\ell_a(x^*) - \ell_a(0)}{\ell_a(f_a^{\mathrm{UE}}) - \ell_a(0)}\right).$$

This shows that $\beta(f_a^{\mathrm{UE}}, \ell_a) < \beta(f_a^{\mathrm{UE}}, \ell_a - \ell_a(0))$, which is a contradiction to condition (2.5c). ∎

The following lemma gives a necessary condition for an instance to be tight.

Lemma 2.3. *Let $\mathcal{L}$ be a family of continuous and nondecreasing latency functions that are either constant or strictly increasing, with $\beta(\mathcal{L}) > 0$. If $\ell_a(0) = 0$ for all $a \in A$, then the instance cannot be tight.*

Proof. To prove the claim, suppose that $\ell_a(0) = 0$ for all $a \in A$. Given the assumptions, there are two classes of arcs: those with travel times identically equal to 0 (we refer to them as 0-*latency arcs*), and those with strictly increasing latency functions. Consider a user equilibrium, a system-optimal flow, and an OD pair $k \in K$. Let us define the set $C_k := \{i \in N :$ there is a path from s_k to i using 0-latency arcs$\}$. If $t_k \in C_k$, both flows route the demand of OD pair k along a 0-latency path. When this happens for all OD pairs, the instance cannot be tight. Therefore, consider an OD pair $k \in K$ for which $t_k \notin C_k$. Thus, C_k defines an s_k-t_k-cut. Note that all the flow that reaches nodes in C_k has to follow paths that, up to the last node in C_k, consist of 0-latency arcs. Hence, a forward arc of the cut cannot be a 0-latency arc, and a backward arc cannot carry (user or system optimal) flow for OD pair k. [A forward (resp. backward) arc of a cut $(S, \bar{S})$ is an arc with tail (resp. head) in S and head (resp. tail) in $\bar{S}$.] The former is obvious; to see the latter, if a backward arc a carried flow, a would be a 0-latency arc because all flow reaching C_k must use paths with zero travel time. Thus, its tail would belong to C_k, too. Hence, there is no flow of OD pair k entering C_k, and all flow exits C_k along non 0-latency arcs. This is a contradiction to remark (3) because the sum of the flow on forward arcs of the cut C_k must equal the demand d_k. ∎

The insight provided by this lemma can be used to improve the upper bounds on the price of anarchy. Instances for which $\ell_a(0) = 0$ correspond to situations when the fixed component of the cost—the so-called free-flow travel time—can be ignored because the major influence comes from congestion-dependent costs. As suggested by Lemma 2.3, those instances should not have high coordination ratios. This is formalized by a result in [8] that proves that the price of anarchy is much lower than $\alpha(\mathcal{L})$: it is bounded from above by 1, 1.185, 1.25, and 1.999 instead of 4/3, 1.626, 1.896, and 2.151 for affine, quadratic, cubic, and quartic functions, respectively.

3. EFFICIENCY

As mentioned in the introduction, Jahn et al. [13] considered two possible definitions for the normal length used in the route-guidance system: free-flow travel times and user equilibrium travel times. The free-flow travel time of an arc $a \in A$ represents the time $\ell_a(0)$ needed to traverse a when there is no traffic (e.g., late at night). The user equilibrium travel time of an arc is the traversal time $\ell_a\left(f_a^{\mathrm{UE}}\right)$ when the prevailing condition is a user equilibrium. Recall that normal lengths can only be static; for instance, it is not possible to consider travel times under the current solution with this methodology. The advantage of this simple model is that it allows for faster algorithms that produce good solutions when normal lengths are properly selected. Although normal lengths have to be set in advance, it is important to point out that users do not need to know the normal lengths; they are merely an artifact to select solutions without extended detours and are controlled by the traffic authority.

3.1. Free-Flow Travel Times as Normal Lengths

In this section, we assume that normal lengths are defined as travel times in the uncongested network; that is, $\tau_a = \ell_a(0)$ for all $a \in A$. The theoretical results presented next help to explain the conclusion derived from the computational study of real-world instances in [13]: the "constrained" price of anarchy is smaller than the ordinary price of anarchy, in which one compares user equilibria with normal system optima instead of constrained system optima. In fact, we will see that for small values of φ, constrained system optima can even be worse than user equilibria. This is because for small values of φ, only a small number of paths are available, and therefore, they are likely to be highly congested, making the total travel time rather high. For this reason, free-flow travel times are not the ideal option for normal lengths.

Let us study $\alpha^{\varphi}(\mathcal{L})$ as a function of φ to understand how the price of anarchy depends on the users' tolerance to unfairness. We start by proving a structural property that implies that the price of anarchy is subadditive. For this purpose, we introduce a construction that enables us to modify the tolerance factor of an instance without altering its coordination ratio too much. For a fixed tolerance factor φ, consider an instance $\mathcal{I}$ with large coordination ratio; that is, the instance satisfies

$$\frac{C(f^{\mathrm{UE}})}{C(f^{\varphi})} \geq \alpha^{\varphi}(\mathcal{L}) - \varepsilon, \tag{3.1}$$

for some $\varepsilon > 0$. We construct a new instance $\widetilde{\mathcal{I}}$, which is equal to $\mathcal{I}$ except for the following modifications. The origins in $\widetilde{\mathcal{I}}$ are new vertices $\tilde{s}_k$ for $k \in K$ (instead of s_k), which are connected to s_k with arcs of constant travel time M_k, specified below. The natural extension $\tilde{x}$ of a flow x to the new instance is defined as $\tilde{x}_{\tilde{P}} := x_P$ for $P \in \mathcal{P}_k$ and $k \in K$, where $\widetilde{P}$ starts at $\tilde{s}_k$ and then continues with the original path P. The extensions $\widetilde{f}^{\mathrm{UE}}$ and $\widetilde{f}^{\mathrm{SO}}$ of a user equilibrium f^{UE} and a system optimum f^{SO} of $\mathcal{I}$ are a user equilibrium and a system optimum for $\widetilde{\mathcal{I}}$, respectively. The next lemma establishes a relation between the constrained system optima of the two instances.

Lemma 3.1. *Consider a fixed $\tilde{\varphi}$ such that $1 < \tilde{\varphi} < \varphi$, and set $M_k := \frac{\varphi-\tilde{\varphi}}{\tilde{\varphi}-1} T_k$. If f^{φ} is a φ-constrained system optimum of $\mathcal{I}$, then its natural extension $\widetilde{f}^{\varphi}$ is a $\tilde{\varphi}$-constrained system optimum of $\widetilde{\mathcal{I}}$.*

Proof. All paths in $\mathcal{P}_k$ that carry flow under f^{φ} have a normal length between T_k and φT_k. After adding M_k to each of them, their lengths are between $M_k + T_k$ and $M_k + \varphi T_k = \tilde{\varphi}(M_k + T_k)$. It follows that $\widetilde{f}^{\varphi}$ is a $\tilde{\varphi}$-constrained system optimum. ∎

Observe that extending a flow x of $\mathcal{I}$ to a flow $\tilde{x}$ of $\widetilde{\mathcal{I}}$ changes its cost by a fixed amount M; that is, $C(\tilde{x}) = M + C(x)$ with $M := \sum_{k \in K} M_k d_k$. Moreover, for this choice of normal lengths, $T_k \leq \ell_P(x)$ for all $P \in \mathcal{P}_k$, which implies that

$$M = \tfrac{\varphi-\tilde{\varphi}}{\tilde{\varphi}-1} \sum_{k \in K} T_k d_k \leq \tfrac{\varphi-\tilde{\varphi}}{\tilde{\varphi}-1}\, C(x). \tag{3.2}$$

We can now prove that the price of anarchy cannot increase too fast.

Theorem 3.2. *The function $\alpha^{\varphi}(\mathcal{L})/(\varphi-1)$ is nonincreasing in φ.*

Proof. Consider the instance $\mathcal{I}$ with large coordination ratio that we selected in (3.1), and let f^{UE} be a user equilibrium and f^{φ} be a φ-constrained system optimum. Furthermore, their natural extensions to $\widetilde{\mathcal{I}}$ are referred to as $\widetilde{f}^{\mathrm{UE}}$ and $\widetilde{f}^{\varphi}$, respectively. We bound the price of anarchy of the new instance $\widetilde{\mathcal{I}}$ with that of the original instance $\mathcal{I}$:

$$\alpha^{\tilde{\varphi}}(\mathcal{L}) \geq \frac{C(\widetilde{f}^{\mathrm{UE}})}{C(\widetilde{f}^{\varphi})} = \frac{M + C(f^{\mathrm{UE}})}{M + C(f^{\varphi})} \geq \frac{C(f^{\mathrm{UE}})}{\frac{\varphi-1}{\tilde{\varphi}-1}\, C(f^{\varphi})}$$
$$\geq \tfrac{\tilde{\varphi}-1}{\varphi-1} \left(\alpha^{\varphi}(\mathcal{L}) - \varepsilon\right) \qquad \text{for all } \tilde{\varphi} < \varphi.$$

The inequalities follow from (2.2), (3.2), and (3.1), respectively. As ε can be made arbitrarily small, $\alpha^{\tilde{\varphi}}(\mathcal{L}) \geq \frac{\tilde{\varphi}-1}{\varphi-1}\, \alpha^{\varphi}(\mathcal{L})$ for all $\tilde{\varphi} < \varphi$. ∎

The last theorem implies that the price of anarchy is subadditive as a function of δ, where $\delta \geq 0$ is a modified tolerance factor defined as $\varphi - 1$.

Corollary 3.3. *The function $\alpha^{1+\delta}(\mathcal{L})$ is subadditive in δ.*

Although the price of anarchy increases when the users' tolerance to unfairness increases, Corollary 3.3 implies that it cannot grow too quickly. This especially suggests that to obtain efficient constrained system optima, users would need to tolerate relatively large deviations from their shortest paths, which is unlikely to happen. Indeed, the left chart of Figure 3 in [13] shows that constrained system optima can be considerably less efficient than user equilibria when φ is relatively close to 1.

3.1.1. Bad Instances. In this section, we extend the results presented in Section 2.2 to constrained system optima. We call an instance *tight* if $C(f^{\mathrm{UE}})/C(f^{\varphi})$ matches the upper bound $\alpha(\mathcal{L})$, where f^{UE} and f^{φ} are a user equilibrium and a constrained system optimum of the corresponding instance, respectively. We will use Observation 2.2 and Lemma 2.3 to show that, under mild assumptions, there cannot exist tight instances for the constrained case. Note that this does not prevent $\alpha^{\varphi}(\mathcal{L})$ from being equal to $\alpha(\mathcal{L})$ for some φ. Again, the following result hints that with free-flow travel times as normal lengths, constrained system optima need not be nearly as efficient as system optima.

Lemma 3.4. *Consider an instance with latency functions drawn from a set $\mathcal{L}$ of continuous and nondecreasing latency functions that are either strictly increasing or constant. Assume that $\mathcal{L}$ is closed under addition and that $\beta(\mathcal{L}) > 0$. Then the coordination ratio $C(f^{\mathrm{UE}})/C(f^{\varphi}) < \alpha(\mathcal{L})$ for all finite $\varphi \geq 1$.*

Proof. Suppose that the coordination ratio equals $\alpha(\mathcal{L})$. In this case, f^{φ} is a system optimum in the unconstrained sense because the cost of the system optimum is a lower bound on that of f^{φ}, and the coordination ratio cannot be larger than $\alpha(\mathcal{L})$. From remark (4), in Section 2.2, we know that $\ell_a(0) = 0$ for all arcs a with $f_a^{\mathrm{UE}} > 0$. Hence, there is a path joining each OD pair whose free-flow travel time is equal to zero. In other words, the normal length T_k has to be 0 for all $k \in K$, which implies that a path belongs to $\mathcal{P}_k^{\varphi}$ only if its normal length is zero. Therefore, $\ell_a(0) = 0$ for all arcs a with flow in f^{UE} or f^{φ}, contradicting Lemma 2.3. ∎

We now turn our attention to characterizing instances with large coordination ratio for fixed φ. We say that a path $P \in \mathcal{P}_k$ is *longest* if its normal length τ_P equals the maximum possible value φT_k. The following result shows that, when a constrained system optimum routes flow along paths that are not longest, we can make the instance worse by adding the Pigou subnetwork shown in Figure 3, which has a large coordination ratio. Note that the conditions required by the result are satisfied by the standard choices of $\mathcal{L}$ such as polynomials of a given degree.

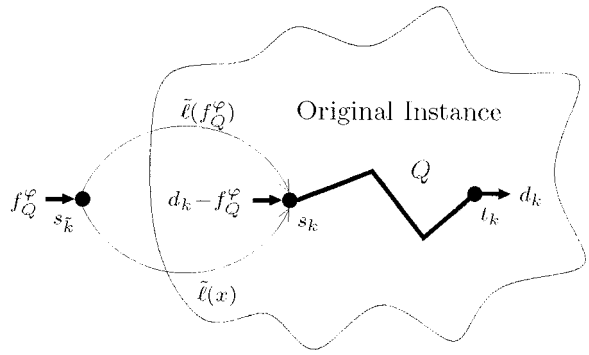

FIG. 4. Modified instance used in the proof of Theorem 3.5.

Theorem 3.5. *Consider a family $\mathcal{L}$ of differentiable latency functions that is closed under multiplication by nonnegative constants. Furthermore, assume that $\mathcal{L}$ is closed under scaling, that is, if $\ell \in \mathcal{L}$, $\ell^{\rho} : x \mapsto \ell(\rho x)$ belongs to $\mathcal{L}$ for all $\rho \geq 0$. Let $\varphi \geq 1$ and f^{φ} be a φ-constrained system optimum of a given instance with latency functions drawn from $\mathcal{L}$. If $C(f^{\mathrm{UE}})/C(f^{\varphi}) < \alpha(\mathcal{L})$ and f^{φ} routes flow along a path that is not longest, then the instance can be modified to increase the coordination ratio $C(f^{\mathrm{UE}})/C(f^{\varphi})$.*

Proof. Assume that for some $k \in K$ there is path $Q \in \mathcal{P}_k^{\varphi}$ that is not *longest* such that $f_Q^{\varphi} > 0$. We insert the network shown in Figure 3 at the source s_k, the origin of path Q. (This modification is illustrated in Fig. 4.) After the modification, two parallel arcs connect a new origin $s_{\tilde{k}}$ to s_k. Furthermore, f_Q^{φ} units of demand are reassigned from OD pair k to a new OD pair $\tilde{k}$ with terminals $s_{\tilde{k}}$ and t_k. We will now construct latency functions that will make the added network tight. Let $v > 0$ and $\ell \in \mathcal{L}$ be such that $\beta(v, \ell) = \beta(\mathcal{L})$. Consider the latency function $\tilde{\ell}$ defined as $\tilde{\ell}(x) := \ell(xv/f_Q^{\varphi})/M$, where $M > 0$ is a constant to be specified later. Note that the assumptions imply that $\tilde{\ell} \in \mathcal{L}$. It can be seen that $\beta(f_Q^{\varphi}, \tilde{\ell}) = \beta(\mathcal{L})$ as well. We let the latency functions of the new arcs be the constant $\tilde{\ell}(f_Q^{\varphi})$ and the function $\tilde{\ell}(x)$. After the modification, there are two possible extensions of Q: the path $Q_{\uparrow}$ (resp. $Q_{\downarrow}$) starts with the constant (resp. nonconstant) arc just added and continues along Q. Denoting the set of paths in the new instance by $\widetilde{\mathcal{P}}$ and setting M such that $\tilde{\ell}(f_Q^{\varphi}) + \tau_Q < \varphi T_k \leq \varphi T_{\tilde{k}}$, $\tau_{Q_{\uparrow}}$ and $\tau_{Q_{\downarrow}}$ are bounded from above by $\varphi T_{\tilde{k}}$. Thus, $Q_{\uparrow}$ and $Q_{\downarrow}$ belong to $\widetilde{\mathcal{P}}_{\tilde{k}}$.

The user equilibrium $\widetilde{f}^{\mathrm{UE}}$ of the new instance can be constructed as an extension of f^{UE}. We reassign (any) f_Q^{φ} units of flow from OD pair k to OD pair $\tilde{k}$, route them along the new nonconstant arc, and let them then follow their original paths. It is straightforward to see that $\widetilde{f}^{\mathrm{UE}}$ is feasible and at equilibrium. Similarly, the constrained system optimum $\widetilde{f}^{\varphi}$ of the new instance can be obtained easily from f^{φ}. Indeed, we distribute the flow originally in Q along the paths $Q_{\downarrow}$ and $Q_{\uparrow}$ in a way that satisfies $\tilde{\ell}^*_{Q_{\downarrow}}(\widetilde{f}^{\varphi}) = \tilde{\ell}^*_{Q_{\uparrow}}(\widetilde{f}^{\varphi})$. As $\widetilde{f}_a^{\varphi} = f_a^{\varphi}$ for the original arcs a, $\tilde{\ell}^*_P(\widetilde{f}^{\varphi}) = \tilde{\ell}^*_P(f^{\varphi})$ for all

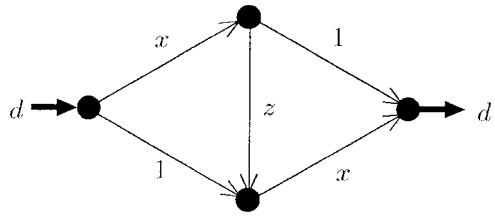

FIG. 5. Instance used in the proof of Lemma 3.7.

$P \in \widetilde{\mathcal{P}}_j^{\varphi}$ with $j \neq \tilde{k}$. Therefore, this extension is a constrained system optimum. Computing the total travel times of both flows and observing that the subgraph is tight, the coordination ratio $C(\widetilde{f}^{\mathrm{UE}})/C(\widetilde{f}^{\varphi})$ equals

$$\frac{C(f^{\mathrm{UE}}) + \tilde{\ell}\big(f_Q^{\varphi}\big)f_Q^{\varphi}}{C(f^{\varphi}) + \tilde{\ell}\big(f_Q^{\varphi}\big)f_Q^{\varphi}/\alpha(\mathcal{L})}.$$

Note that the last expression is a convex combination of $C(f^{\mathrm{UE}})/C(f^{\varphi})$ and $\alpha(\mathcal{L})$. As the former is smaller than the latter, the new instance has worse performance. ∎

The previous result implies that instances with the highest possible coordination ratio are not especially attractive to users either. Indeed, for such instances, a constrained system optimum routes all flow along longest paths with respect to normal lengths. Therefore, users would have to take paths that are at the limit of their tolerance to deviations from shortest paths. Fortunately, real-world instances seem not to exhibit these problems; their coordination ratios are typically not very high. According to the middle graph of Figure 3 in [13], which shows the distribution of experienced travel times, most users do not experience extreme values. The situation for free-flow travel times is similar.

Theorem 3.5 is actually still valid if latency functions are not differentiable. Although we cannot work with the modified latency functions $\tilde{\ell}^*$ anymore, we can use the system optimum of the instance shown in Figure 3 instead. The theorem also remains valid for user equilibrium normal lengths, which we will state explicitly in Section 3.2.

3.1.2. Bounds for the Price of Anarchy. In this section, we present upper and lower bounds for the function $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}})$. This will allow us to evaluate the performance of the route-guidance system when free-flow normal lengths are used. We start with an upper bound that improves on $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}}) \leq \alpha(\mathcal{L}_{\mathrm{aff}}) = 4/3$.

Theorem 3.6. *The price of anarchy* $\alpha^{\varphi}(\mathcal{L}_{aff}) \leq (2-\varphi)^{-1}$ *for all* $1 \leq \varphi < 2$. *In particular,* $\alpha^{1}(\mathcal{L}_{aff}) = 1$ *and* $\alpha^{\varphi}(\mathcal{L}_{aff}) < 4/3$ *for* $\varphi < 5/4$.

Proof. Consider a tolerance factor $1 \leq \varphi < 2$, and let f^{φ} and f^{UE} be a φ-constrained system optimum and a user equilibrium, respectively. We define the function $h : [0,1] \to \mathbb{R}$ by $h(z) := C(f^{\mathrm{UE}} + z(f^{\varphi} - f^{\mathrm{UE}}))$. Due to the convexity of $C(\cdot)$, $h(1) \geq h(0) + h'(0)$. To prove the claim, we verify that $h(0) + h'(0) \geq (2-\varphi)h(0)$ because then $C(f^{\varphi}) = h(1) \geq (2-\varphi)h(0) = (2-\varphi)C(f^{\mathrm{UE}})$, as required. Now,

$$\begin{aligned} h'(0) &= \sum_a \ell_a^*\big(f_a^{\mathrm{UE}}\big)\big(f_a^{\varphi} - f_a^{\mathrm{UE}}\big) \\ &= \sum_a \big[2\ell_a\big(f_a^{\mathrm{UE}}\big) - \ell_a(0)\big]\big(f_a^{\varphi} - f_a^{\mathrm{UE}}\big) \\ &\geq 2\left(\sum_k L_k(f^{\mathrm{UE}})d_k - \sum_k L_k(f^{\mathrm{UE}})d_k\right) \\ &\quad + \sum_k T_k d_k - \varphi \sum_k T_k d_k \\ &= (1-\varphi)\sum_k T_k d_k \geq (1-\varphi)C(f^{\mathrm{UE}}) = (1-\varphi)h(0). \end{aligned}$$

The first inequality follows from the fact that $\ell_P(f^{\mathrm{UE}}) = L_k(f^{\mathrm{UE}})$ for every $P \in \mathcal{P}_k$ such that $f_P^{\mathrm{UE}} > 0$, and $\ell_P(f^{\mathrm{UE}}) \geq L_k(f^{\mathrm{UE}})$ in general. Moreover, $\tau_P \leq \varphi T_k$ for every P such that $f_P^{\varphi} > 0$, and $T_k \leq \tau_P$ in general. ∎

We now offer a lower bound on $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}})$ by providing an instance with high coordination ratio. Although the instance shown in Figure 3 can be used, a stronger bound can be given with a collection of instances based on the Braess Paradox network [4].

Lemma 3.7. *The price of anarchy* $\alpha^{\varphi}(\mathcal{L}_{aff}) \geq 1 + \big(3 + \frac{2}{\varphi-1}\big)^{-1}$.

Proof. Consider the network depicted in Figure 5, where $z \geq 0$ is a constant and the demand between the single OD pair is $d \geq 0$. Maximizing the coordination ratio over z and d, we obtain the claim. ∎

Figure 6 summarizes the bounds for $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}})$. The main conclusion is that $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}})$ is close to 1 when φ is in the proximity of 1. Therefore, it is not necessary to compute the exact value of the price of anarchy to conclude that the total travel time of user equilibria is, in the worst-case, just a little bit

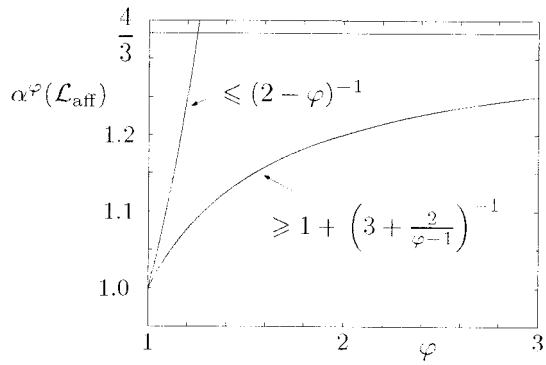

FIG. 6. Bounds for $\alpha^{\varphi}(\mathcal{L}_{\mathrm{aff}})$.

higher than that of constrained system optima. This adds more support to the conclusion that we drew before: if free-flow travel times are used, central coordination is not beneficial, and the route-guidance system does not deliver solutions of significantly improved quality. To further explain why these normal lengths do not work, note that free-flow travel times are demand-independent; by definition, they are based on the situation in which the network is not loaded. Hence, had they delivered consistently good solutions, it would have been regardless of the real demand. A posteriori, this seems unlikely because uncongested and congested networks operate under very different regimes. In addition, when $\varphi \approx 1$ most paths are not feasible, leaving the few feasible ones very congested, and the total travel time too high. Because of this negative result for affine cost functions, we will not compute the price of anarchy for more general sets $\mathcal{L}$.

3.2. User Equilibrium Travel Times as Normal Lengths

In this section, we assume that normal lengths are set equal to the travel times experienced in a user equilibrium. Based on superior empirical performance compared to free-flow travel times, Jahn et al. [13] concluded that these normal lengths are the "correct" choice. We now provide a theoretical foundation for their findings. The improvement compared to free-flow normal lengths comes from the fact that user equilibrium normal lengths depend on the actual demand.

A user equilibrium f^{UE} is a feasible solution to (CSO) because all paths used in f^{UE} are feasible. Therefore, the constrained system optimum f^{φ} satisfies

$$C(f^{\varphi}) \leq C(f^{\mathrm{UE}}) \quad \text{for all } \varphi \geq 1. \tag{3.3}$$

As in the previous section, we obtain a lower bound on the function $\alpha^{\varphi}(\mathcal{L})$ by providing an appropriate instance. However, in this case, the lower bound matches the upper bound.

Lemma 3.8. *The price of anarchy $\alpha^{\varphi}(\mathcal{L}) = \alpha(\mathcal{L})$ for all $\varphi \geq 1$.*

Proof. In the equilibrium of the instance depicted in Figure 3, travel times along both paths equal $\ell(d)$. Hence, regardless of the value of φ, the system optimum is a φ-constrained system optimum. The claim follows by taking the supremum over $\ell \in \mathcal{L}$. ∎

This lemma implies that the price of anarchy is the same, regardless of whether it is defined with respect to the system optimum or the constrained system optimum. Note that Lemma 3.8 is a worst-case statement. For realistic instances the two solutions typically differ, as shown in Figure 9 of [13], which we reproduce here, for convenience. Figure 7 shows, for a variety of realistic instances, the tradeoff between the unfairness of the considered solutions and the ratio of their cost to that of a system optimum. The left part of the diagram corresponds to system optima (SO), the lower part to user equilibria (UE), and the circled data-points represent constrained system optima with $\varphi = 1.02$.

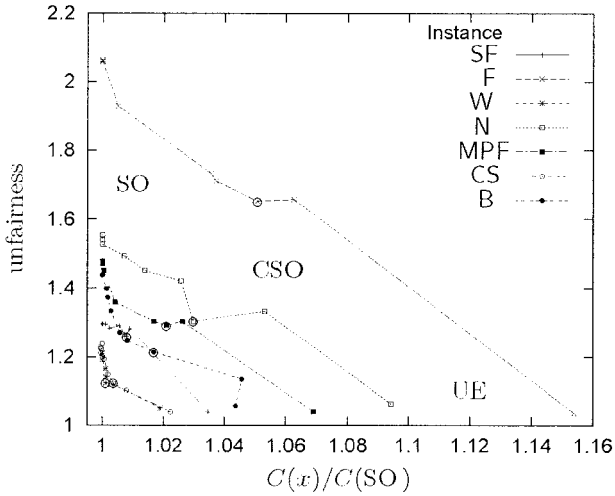

FIG. 7. Tradeoff between efficiency and unfairness.

The two bounds presented in (2.3) and (3.3) can be tight. For example, the proof of Lemma 3.8 describes an instance for $\varphi = 1$, satisfying $C(f^{\mathrm{SO}}) = C(f^{1}) = C(f^{\mathrm{UE}})/\alpha(\mathcal{L})$. On the other hand, if we add a small constant $\varepsilon > 0$ to the travel time of the first arc, the constrained system optimum coincides with the user equilibrium. Therefore, $C(f^{\mathrm{UE}}) = C(f^{1}) \approx \alpha(\mathcal{L})C(f^{\mathrm{SO}})$.

Theorem 3.5, proved before for free-flow normal lengths, is also valid when normal lengths are set to user equilibrium travel times. It is enough to note that at equilibrium the lengths of the two new arcs equal $\tilde{\ell}(f_Q^{\varphi})$; therefore, $Q_{\downarrow}$ and $Q_{\uparrow}$ belong to $\widetilde{\mathcal{P}}^{\varphi}$.

Observation 3.9. *Consider a family $\mathcal{L}$ of differentiable latency functions that is closed under multiplication by constants. Furthermore, assume that $\mathcal{L}$ is closed under scaling. Let $\varphi \geq 1$ and f^{φ} be a φ-constrained system optimum of a given instance with latency functions drawn from $\mathcal{L}$. If f^{φ} routes flow along a path that is not longest, then the instance can be modified to increase the coordination ratio $C(f^{\mathrm{UE}})/C(f^{\varphi})$.*

However, constrained system optima for real-world instances do not typically route all users along longest paths. The middle graph of Figure 4 in [13] shows the distribution of experienced travel times. It can be seen that few users are routed along longest paths; the situation for user equilibrium travel times is similar.

4. FAIRNESS

As we mentioned earlier, a common argument against using a system optimum in the design of route-guidance devices for traffic assignment is that it generally assigns some drivers to unacceptably long paths to use shorter paths for most other drivers (see, e.g., [1, 14]). This section

presents results related to the unfairness of system optima and constrained system optima. In this section, we work with arbitrary normal lengths, unless otherwise stated.

The following theorem quantifies the severity of this effect by characterizing the unfairness of the system optimum. It turns out that there is a relation to earlier work that compared the maximum latency of a system optimum in a single-sink single-source network to the latency of a user equilibrium [19]. This work showed that for a given class of latency functions $\mathcal{L}$, this ratio is bounded from above by $\gamma(\mathcal{L})$. Here, $\gamma(\mathcal{L})$ is defined to be the smallest value that satisfies $\ell^*(x) \leq \gamma(\mathcal{L})\ell(x)$ for all $\ell \in \mathcal{L}$ and $x \geq 0$. For example, $\gamma(\{\text{polynomials of degree } p \text{ with nonnegative coefficients}\}) = p + 1$. The unfairness of a system optimum is, in fact, bounded by the same constant, even for general instances with multiple commodities [9, 20].

It is not difficult to extend the bound on the unfairness of system optima to constrained system optima. Notice that the following theorem does not assume any particular definition of normal lengths.

Theorem 4.1. *Let f^φ be a constrained system optimum in a multicommodity flow network with latency functions drawn from a family $\mathcal{L}$ of differentiable and nondecreasing latency functions. Then the unfairness of f^φ is bounded from above by $\gamma(\mathcal{L})$.*

Proof. Using the definitions of ℓ^* and $\gamma(\mathcal{L})$, it is clear that $\ell_a(x) \leq \ell_a^*(x) \leq \gamma(\mathcal{L})\ell_a(x)$ for all $x \geq 0$. The first-order optimality conditions of (CSO) imply that for a constant $L_k^*(f^\varphi)$, $\ell_P^*(f^\varphi) = L_k^*(f^\varphi)$ for all $P \in \mathcal{P}_k^\varphi$ such that $f_P^\varphi > 0$. Therefore, for all paths $P \in \mathcal{P}_k^\varphi$ carrying flow,

$$\frac{L_k^*(f^\varphi)}{\gamma(\mathcal{L})} \leq \ell_P(f^\varphi) \leq L_k^*(f^\varphi).$$

Consequently, $\ell_Q(f^\varphi)/\ell_R(f^\varphi) \leq \gamma(\mathcal{L})$ for all $Q, R \in \mathcal{P}_k^\varphi$ with positive flow. ∎

Correa et al. [9] presented an example that can be used to show that this bound is tight. Consider the instance shown in Figure 3 with $d = 1$ and $\ell(x) := x$. User equilibrium normal lengths are equal to 1 for both arcs; therefore, both paths are feasible regardless of the value of φ. This means that any constrained system optimum is an unconstrained system optimum, and its unfairness is $\gamma(\mathcal{L}_{\text{aff}}) = 2$. Nevertheless, in practice these bounds are loose, as the extensive experiments in [13] show. In particular, Figure 7 demonstrates that for polynomials of degree 4, which are typically used by transportation planners, the highest observed unfairness was approximately 2.1, whereas the previous theorem implies an unfairness of 5 in the worst case.

Note that Theorem 4.1 does not imply that the unfairness of φ-constrained system optima is nondecreasing as a function of φ. We now present two examples corresponding to the two definitions of normal lengths that we studied. The example using free-flow travel times is the one we presented in the introduction (see Figure 2). A constrained system optimum with $\varphi = 1$ can only route flow on the two bottom arcs; it therefore has an unfairness strictly larger than 1. For $\varphi \geq 1 + \varepsilon$, all arcs can be used, and it is easy to see that the value of unfairness approaches 1 when $\varepsilon \to 0$.

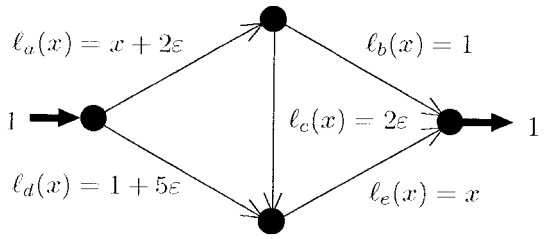

FIG. 8. The unfairness can decrease when φ increases (normal length = user equilibrium travel times).

For the case in which normal lengths are user equilibrium travel times, consider the instance shown in Figure 8. There are five arcs a, b, c, d, and e with latency functions $x + 2\varepsilon$, 1, 2ε, $1 + 5\varepsilon$, and x, respectively, where ε is a small positive number. The user equilibrium routes flow only along paths ab and ace; at equilibrium the path de is too long to carry flow. Therefore, the constrained system optimum f^1 can only use paths ab and ace, and its unfairness is $\frac{4+4\varepsilon}{3+6\varepsilon}$. If $\varphi \geq \frac{2+3\varepsilon}{2+2\varepsilon}$, the constrained system optimum f^φ can use all three paths. In this case, it routes the flow along ab and de, and its unfairness is $\frac{6+17\varepsilon}{6+11\varepsilon}$. For small enough values of ε, the unfairness of the constrained system optimum with $\varphi = 1$ is arbitrarily close to $4/3$ while the unfairness for a large enough tolerance factor φ is arbitrarily close to 1.

Finally, we show that in the case of affine latency functions, paths that are short with respect to free-flow normal lengths are also relatively short with respect to experienced travel times. To be more concrete, define the average travel time of a flow x between OD pair $k \in K$ as $\overline{C}_k(x) := \sum_{P \in \mathcal{P}_k^\varphi} \ell_P(x)x_P/d_k$.

Theorem 4.2. *Consider an instance with affine latency functions and arbitrary tolerance factor φ. Let $P \in \mathcal{P}_k^\varphi$ be a path satisfying $f_P^\varphi > 0$ and $\ell_P(0) \leq \varepsilon\,\overline{C}_k(f^\varphi)$ for some $\varepsilon \geq 0$. Then the experienced travel time $\ell_P(f^\varphi)$ is bounded from above by $(1 + \varepsilon/2)\overline{C}_k(f^\varphi)$.*

Proof. As latencies are affine functions and f^φ is at equilibrium with respect to latency functions ℓ^*, we have

$$\begin{aligned}\ell_P(f^\varphi) &= \ell_P^*(f^\varphi) - \sum_{a \in P} q_a f_a^\varphi \\ &= \sum_{Q \in \mathcal{P}_k^\varphi} \frac{f_Q^\varphi \ell_Q^*(f^\varphi)}{d_k} - (\ell_P(f^\varphi) - \ell_P(0)) \\ &\leq \sum_{Q \in \mathcal{P}_k^\varphi} 2\frac{f_Q^\varphi \ell_Q(f^\varphi)}{d_k} - \ell_P(f^\varphi) + \varepsilon\overline{C}_k(f^\varphi).\end{aligned}$$

Therefore, $2\ell_P(f^\varphi) \leq (2 + \varepsilon)\overline{C}_k(f^\varphi)$. ∎

5. SUMMARY AND CONCLUSION

This article presents a theoretical analysis of the route-guidance system proposed by Jahn et al. [13]. This route-guidance system aims at optimizing the efficiency of the traffic flow, while ensuring that all users are treated fairly. We have given bounds on the inefficiency and unfairness of the returned solutions. When the system uses user equilibrium travel times as normal lengths, constrained system optima are not much more costly than system-optimal solutions, and users with the same OD pair are assigned to paths of similar lengths.

In practice, network planners sometimes work with non-separable latency functions, so that the travel time of one arc also depends on the load of other arcs. Although these functions are more difficult to calibrate, they improve the predictive power because congestion levels on different arcs of the network are typically correlated. The most common examples are two-way streets and intersections. Theoretically, system optima and user equilibria can, and have been extended to that setting. Constrained system optima can be generalized without difficulty given that they have the same structure as system optima. Although bounds on the efficiency of user equilibria with the more general latency functions were previously given [5, 17], it is an interesting open question to determine the price of anarchy with respect to constrained system optima. It is clear that under user equilibrium normal lengths, (3.3) is still valid because any user equilibrium remains feasible for the constrained system optimum problem. Theorem 4.1 can also be extended to bound the unfairness of constrained system optima by generalizing the definition of $\gamma(\mathcal{L})$ to incorporate the more complicated derivatives of $C(x)$. Another interesting extension would be to incorporate structural insights of realistic networks. This has the potential of making worst-case bounds less pessimistic. As an example in this direction, Correa et al. [8] gave improved bounds on the price of anarchy that depend on the congestion level of the network.

Acknowledgments

The authors are grateful to two anonymous referees and an associate editor for their insightful comments, which helped to improve the presentation of this article.

REFERENCES

[1] G. Beccaria and A. Bolelli, Modelling and assessment of dynamic route guidance: The MARGOT project, Proc 3rd IEEE Vehicle Navigation & Information Systems Conf, Oslo, Norway, 1992, pp. 117–126.

[2] M.J. Beckmann, C.B. McGuire, and C.B. Winsten, Studies in the economics of transportation, Yale University Press, New Haven, CT, 1956.

[3] M.E. Ben-Akiva, M. Bierlaire, J. Bottom, H.N. Koutsopoulos, and R.G. Mishalani, Development of a route guidance generation system for real-time application, Proc 8th IFAC Symp on Transportation Systems, Chania, Greece, Elsevier Science, 1997, pp. 405–410.

[4] D. Braess, Über ein Paradoxon aus der Verkehrsplanung, Unternehmensforschung 12 (1968), 258–268. (An English translation appeared in Transportation Science 39 (2005), 446–450.)

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

[6] S. Cohen, "Flow variables," Concise Encyclopedia of Traffic & Transportation Systems, M. Papageorgiou (Editor), Pergamon Press, Oxford, 1991, pp. 139–143.

[7] J.R. Correa, A.S. Schulz, and N.E. Stier-Moses, Selfish routing in capacitated networks, Math Oper Res 29 (2004), 961–976.

[8] J.R. Correa, A.S. Schulz, and N.E. Stier-Moses, On the inefficiency of equilibria in congestion games, Proc 11th Conf on Integer Programming and Combinatorial Optimization, Lecture Notes in Computer Science, Vol. 3509, Springer, Berlin, 2005, pp. 167–181.

[9] J.R. Correa, A.S. Schulz, and N.E. Stier-Moses, Fast, fair, and efficient flows in networks, Oper Res. Forthcoming.

[10] S.C. Dafermos, Traffic equilibrium and variational inequalities, Transport Sci 14 (1980), 42–54.

[11] S.C. Dafermos and F.T. Sparrow, The traffic assignment problem for a general network, J Res U.S. Nat Bureau Stand 73B (1969), 91–118.

[12] J. Dupuit, On tolls and transport charges, Annales des Ponts et Chausse'es (1849), Reprinted in Int Econ Papers 11 (1962), 7–31.

[13] O. Jahn, R.H. Möhring, A.S. Schulz, and N.E. Stier-Moses, System-optimal routing of traffic flows with user constraints in networks with congestion, Oper Res 53 (2005), 600–616.

[14] D.E. Kaufman, R.L. Smith, and K.E. Wunderlich, An iterative routing/assignment method for anticipatory real-time route guidance, Proc IEEE Vehicle Navigation & Information Systems Conf, Vol. 2, Dearborn, MI, Society of Automotive Engineers, 1991, pp. 693–700.

[15] E. Koutsoupias and C.H. Papadimitriou, Worst-case equilibria, Proc 16th Ann Symp on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science, Vol. 1563, Springer, Berlin, 1999, pp. 404–413.

[16] H.S. Mahmassani and S. Peeta, Network performance under system optimal and user equilibrium dynamic assignments: Implications for advanced traveler information systems, Transport Res Rec 1408 (1993), 83–93.

[17] G. Perakis, The "price of anarchy" under nonlinear and asymmetric costs, Proc 10th Conf on Integer Programming and Combinatorial Optimization, Lecture Notes in Computer Science, Vol. 3064, Springer, Berlin, 2004, pp. 46–58.

[18] A.C. Pigou, The economics of welfare, Macmillan, London, 1920.

[19] T. Roughgarden, How unfair is optimal routing?, Proc 13th Ann ACM-SIAM Symp on Discrete Algorithms, San Francisco, CA, 2002, pp. 203–204.

[20] T. Roughgarden (2003), cited as personal communication in [9].

[21] T. Roughgarden, The price of anarchy is independent of the network topology, J Comput Syst Sci 67 (2003), 341–364.

[22] T. Roughgarden, Selfish routing and the price of anarchy, MIT Press, Cambridge, MA, 2005.

[23] T. Roughgarden and É. Tardos, How bad is selfish routing? J ACM 49 (2002), 236–259.

[24] A.S. Schulz and N.E. Stier-Moses, On the performance of user equilibria in traffic networks, Proc 14th Ann ACM-SIAM Symp on Discrete Algorithms, Baltimore, MD, 2003, pp. 86–87.

[25] Y. Sheffi, Urban transportation networks, Prentice-Hall, Englewood, NJ, 1985.

[26] M.J. Smith, The existence, uniqueness and stability of traffic equilibria, Transport Res 13B (1979), 295–304.

[27] J.G. Wardrop, Some theoretical aspects of road traffic research, Proc Institution of Civil Engineers, Part II, Vol. 1, 1952, pp. 325–378.