Selfish routing in the presence of side constraints

George Karakostas* and Stavros G. Kolliopoulos**

Department of Computing and Software, McMaster University, Hamilton, Ontario, L8S 4K1, Canada

Abstract. The natural approach for describing network flow problems is to introduce *side constraints* that capture restrictions of a logical or technological nature, e.g., capacity or budgetary constraints. We study the traffic equilibria arising from selfish routing of individual users in networks with side constraints.
For the problem without side constraints the classic Wardrop principle suggests that in equilibrium for each origin-destination pair the latency on all the used routes is equal or less to the latency on all unused routes. In this paper we use a natural extension of the Wardrop principle [13] in order to obtain results on selfish routing with side constraints. The extended Wardrop principle suggests that in equilibrium flow is routed along paths of minimal latency according to a modified function that essentially incorporates penalty terms for the side constraints. This approach provides a basis for the systematic application of relevant optimization theory to selfish routing problems.

1 Introduction

The natural approach for describing network flow problems is to introduce *side constraints* that capture restrictions of a logical or technological nature. These constraints describe often limitations on the availability of scarce resources (e.g., transportation or production capacities, investment capital available) which are shared by several activities. In this paper we study the traffic equilibria arising from selfish routing of individual users in networks with side constraints. We generalize previous results to this model using techniques from the theory of variational inequalities and complementarity problems.

Recently there has been a surge of interest in theoretical computer science on the consequences of selfish behavior in unregulated networks. Data networks such as the Internet motivated much of the ongoing research. We conform to standard usage in the transportation community and refer to *traffic networks.* We are given a directed network $G = (V, E)$ with a nonnegative latency (cost) function l on the paths describing the delay experienced by users wishing to travel on the path as a function of the total flow using edges of the path. A set I of origin-destination pairs is given, each corresponding to a different commodity

* (`karakos@mcmaster.ca`). Research supported by an NSERC Discovery grant.
** (`stavros@mcmaster.ca`). Research partially supported by NSERC Research Grant 227809-00.

with fixed demand (traffic) rate d_i. Individual users are thought as carrying each an infinitesimal amount of a commodity. Let f be a vector of path flows that satisfies the demands and $l(f)$ the corresponding vector of path latencies. The quality of a traffic assignment can be measured by the *social cost,* i.e., $\sum_{\text{path } P} f_P l_P(f)$. The problem as described has no side constraints and was first studied from a theoretical computer science perspective by Roughgarden and Tardos [19]. The solution concept deemed to capture the selfish behavior of users is that of a *traffic equilibrium,* i.e., a feasible flow f^* that solves the following variational inequality

$$\langle l(x), f - x \rangle \geq 0, \quad \forall \text{ feasible } f. \tag{1}$$

The classic Wardrop's principle [23] suggests that f^* is an equilibrium if and only if for each origin-destination pair the travel times on all the used routes are the same and equal or less than the travel times on all unused routes. It is well-known that when l is continuous and the feasible set is compact and convex the variational inequality (1) has a solution [7].

To quantify the degradation in network performance due to lack of coordination Koutsoupias and Papadimitriou [9] introduced the notion of *coordination ratio* or *price of anarchy.* Many equilibria may exist for an instance, so one would like to measure the worst-case performance of uncoordinated users against the optimal solution achieved by a centralized coordinator. This ratio ρ is defined as

$$\rho = \frac{\text{Worst equilibrium social cost}}{\text{Optimal social cost}}.$$

For the problem without side constraints the coordination ratio is unbounded for general latency functions [19]. Roughgarden and Tardos [19] showed that the social cost of an equilibrium for selfish users with demand vector d is upper bounded by the latency of any feasible flow that satisfies demands $2d$. In the case of linear latency functions Roughgarden and Tardos [19] showed that the coordination ratio is at most $4/3$ and this is tight. Schulz and Stier Moses [20] (see also [3]) introduced a definition of equilibrium under capacity constraints. Under the definition in [20] the coordination ratio is unbounded due to an example flow which does not conform to the different definition of equilibrium we follow (cf. Definition 2 below), and to which the authors of [20] actually revert in order to get their positive results. Definition 2 is standard in the math programming and transportation communities where the study of traffic equilibria originated and has been pursued for decades. See, e.g., [21, 5, 14, 17] and more specifically [10, 13, 16] for side-constrained equilibria.

In this paper we study traffic equilibria under general side constraints through the Extended Wardrop Principle (cf. Thm. 1) [10, 13] that characterizes side-constrained equilibria in terms of equilibria under a *modified latency function* and *without* side constraints. This principle suggests that an equilibrium flow routes flow along paths of minimal modified latency. The advantage of this powerful approach is similar to the advantage of Lagrangean relaxation in optimization, which transforms a constrained optimization problem to an unconstrained one.

To our knowledge the Extended Wardrop Principle has not been used before in a theoretical computer science setting. We demonstrate the applicability of this approach in two cases where side constraints are present: (i) the case of linear latency functions with side constraints that define a convex set, and (ii) the case of general continuous nonnegative latency functions with side constraints of the form $g_j(f) \leq 0$, where for all j, $g_j(f)$ is convex and differentiable with nonnegative partial derivatives.

In the case of linear latency functions, we show how one can extend the 4/3 coordination ratio under capacity constraints of Schulz and Stier Moses [20] (see also [3]) using the Extended Wardrop Principle. Our extensions hold for very general descriptions of the constraints. Packing constraints such as $\sum_i a_i f_i \leq b_j$ with all $a_i \geq 0$, form only a special case among the side constraints we can handle. In turn, capacity constraints on the edges form a special case of packing constraints. The only assumption we impose on the side constraints is that they are described by a convex set D. We show that the coordination ratio is at most 4/3, which is known to be tight, therefore matching the result of Roughgarden and Tardos [19] for the case without side constraints. In a later version of [3] Correa, Schulz and Stier Moses observe that the 4/3 result applies to constraints expressed by any convex set [4]. Our proof was obtained independently and is substantially different as it relies on the generalization of Wardrop's principle to the setting with side constraints as given in [13]. We believe that making the connection with the classic Wardrop principle is a significant part of our contribution.

For the case of general continuous nonnegative latency functions we consider side constraints of the form $g_j(f) \leq 0$, where for all j, $g_j(f)$ is convex and differentiable with nonnegative partial derivatives. Using again the Extended Wardrop Principle we formulate the traffic equilibrium as a nonlinear complementarity problem. We use a classic transformation [22] to reduce the existence of a solution to the complementarity problem to the existence of a Brouwer fixed point for an appropriate continuous function. Our approach is inspired by the seminal work of Aashtiani and Magnanti [1] on traffic equilibria without side constraints. In order to apply the fixed-point approach for the side constrained complementarity problem, one needs to upper-bound the variables of the complementarity problem, thus showing that the possible solutions lie in a cube. We show that a method for doing this comes from the following observation: the existence of a solution for the complementarity problem with the side constraints (which is a traffic equilibrium for the side-constrained network) implies a solution for the complementarity problem *without the side constraints* (which is a traffic equilibrium for the unconstrained network). This framework can be used to derive existence results. As an application, we show that the recent result of Cole, Dodis and Roughgarden [2] on the existence of optimal taxes that steer the equilibrium of a single-commodity problem with heterogeneous users towards a minimum social cost solution can be interpreted as a proof that upper bounds the Lagrangean multipliers of capacity constraints. Casting the rather involved proof from [2] in this general framework makes it more intuitive. We believe that our method can find further applications of interest.

Constructive versions of the existential results can be easily established when standard assumptions allow the formulation of the corresponding convex optimization programs. We omit the details. An early version of some of our results appeared in [8].

2 The model

Let $G = (V, E)$ be a directed network, and $I \subseteq V \times V$ a set of origin-destination pairs (s_i, t_i). Let $\mathcal{P}_i$ be the set of all (acyclic) paths from s_i to t_i and $\mathcal{P} = \bigcup_{i \in I} \mathcal{P}_i$ the set of all origin-destination paths. $\Delta = [\delta_{eP}]$ is the edge-path incidence matrix for G, i.e., $\delta_{eP} = 1$ if $e \in P$, 0 otherwise. $(\Delta)_e$ denotes the row of the matrix corresponding to edge $e \in E$. In general we use f to denote a path flow vector in $\mathbb{R}^{|\mathcal{P}|}$. An edge flow representation can be obtained by setting $f_e = (\Delta)_e f$, for every $e \in E$. Suppose that in addition we are given

- a flow demand (rate) $d_i > 0$ for every $i \in I$;
- lower and upper bounds λ_P, μ_P for every path flow f_P; i.e., $\lambda_P \le f_P \le \mu_P$ (if $\lambda_P = \mu_P = 0$ then path P is not allowed to carry any flow);
- a travel cost function l which assigns to each flow $f \in \mathbb{R}^{|\mathcal{P}|}$ a vector of path travel costs $l(f) \in \mathbb{R}^{|\mathcal{P}|}$;
- a convex set $D \subseteq \mathbb{R}^{|\mathcal{P}|}$ which describes the additional side constraints imposed on the path flows. For example, we can impose constraints on individual edges; such constraints could be edge capacity constraints of the form $f_e \le u_e$, where u_e is the capacity of edge e, or budgetary constraints of the form $\sum_{P \in \mathcal{P}_i} (\sum_{e \in P} c_e) f_P \le B_i$, where c_e is the per unit of flow cost for edge e and B_i is the budget for origin-destination pair i.

Let

$$K = K(\lambda, \mu, d) := \{f : \lambda_P \le f_P \le \mu_P, \forall P \ \wedge \ \sum_{P \in \mathcal{P}_i} f_P = d_i, \forall i\}$$

be the set of all flows that satisfy the path flow bounds and the demands (note that they do not need to satisfy the side constraints).

Definition 1 *A flow f is called* feasible *iff $f \in K \cap D$.*

An equilibrium in this setting (see [21], [10]) is a feasible flow $f^* \in K \cap D$ such that

$$\langle l(f^*), f - f^* \rangle \ge 0, \quad \forall f \in K \cap D. \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. If $l(x) = F'(x)$, where F is convex and Gâteaux differentiable with derivative F', then (2) is equivalent to solving $\min\{F(x) : x \in K \cap D\}$. This is the assumption behind the definition of *Beckmann User Equilibria*, defined in [20], but we will not need it here.

Note that (2) satisfies the general definition of an equilibrium: there is no way for any user to profit from any change of the way his flow is routed, without violating (2).

Definition 2 *A flow $f^* \in K \cap D$ that satisfies* (2) *is called a traffic equilibrium.*

In the model without upper and lower path flow bounds ($\lambda = 0, \mu = +\infty$) and without side constraints ($D = \mathbb{R}^{|\mathcal{P}|}$), these traffic equilibria satisfy Wardrop's principle, i.e., no user has incentives to change his path flows. In that setting, Wardrop's principle implies that the travel cost for all paths $P \in \mathcal{P}_i$ with positive flow $f_P > 0$ is the same and equal or smaller than the travel cost of any path with zero flow (see [19] and the references therein). This is the model studied from the coordination ratio perspective by Roughgarden and Tardos [19] and their analysis relies heavily on Wardrop's principle.

Wardrop's principle has been extended first to the model with lower and upper path flow bounds by Maugeri [12], and more recently to the more general model defined above by Maugeri et al. [13] (see also [10]):

Definition 3 (Extended Wardrop principle)[13] *We say that $f^* \in K \cap D$ satisfies the Extended Wardrop Principle iff there exists $t(f^*) \in \mathbb{R}^{|\mathcal{P}|}$ such that*

$$\langle t(f^*), f - f^* \rangle \geq 0, \quad \forall f \in D, \tag{3}$$

and, with $\bar{l} := l(f^*) - t(f^*)$, the following holds for all $i \in I$ and $P_1, P_2 \in \mathcal{P}_i$,

$$\bar{l}_{P_1} < \bar{l}_{P_2} \Rightarrow f^*_{P_1} = \mu_{P_1} \text{ or } f^*_{P_2} = \lambda_{P_2}. \tag{4}$$

Note that the second condition in the definition above is the usual Wardrop's principle but for the *modified travel cost function* $\bar{l}$. The second condition is shown in [13] to be equivalent to the following inequality:

$$\langle l(f^*), f - f^* \rangle \geq \langle t(f^*), f - f^* \rangle, \quad \forall f \in K. \tag{5}$$

The crucial observation for extending the methods of [19] in the general setting with side constraints is that conditions (3) and (5) are *decoupling* the feasibility region: the first is a condition on D and the second is a condition on K.

Let $intD$ be the interior of D (recall that D is a convex set). Maugeri et al. [13] prove the following theorem that connects traffic equilibria to the Extended Wardrop Principle:

Theorem 1. [13] *Every feasible flow, which satisfies the Extended Wardrop Principle is a traffic equilibrium flow. If $K \cap intD \neq \emptyset$, then every traffic equilibrium flow satisfies the Extended Wardrop Principle.*

Therefore, in order to be able to study traffic equilibria using the Extended Wardrop Principle, we make the following assumption:

Assumption 1 *We assume that $K \cap intD \neq \emptyset$.*

3 Coordination ratio for linear latency functions

We study the coordination ratio for the case of linear travel cost functions $l_e(f_e) = a_e f_e + b_e$, where $a_e, b_e \geq 0$ and $f_e := \sum_{P \ni e} f_P$, for all edges $e \in E$. In the case of the *additive model* used in [19], the total cost for a flow f is $C(f) := \sum_{e \in E} f_e l_e(f_e) = \sum_{P \in \mathcal{P}} f_P l_P(f)$, where $l_P(f) := \sum_{e \in P}(a_e f_e + b_e)$. We extend the analysis of [19] to incorporate the existence of side constraints that define the convex set D. The easy proof of the following fact is in the Appendix.

Fact 1 *Let a routing problem be specified on a network $G = (V, E)$ by the pair of convex sets (K, D) with $K = K(\lambda, \mu, d)$. For any $\delta > 0$ convex sets K', D' can be defined so that (K', D') is a routing problem and f is feasible for (K, D) iff f/δ is feasible for (K', D'). Moreover $K \cap intD \neq \emptyset$ iff $K' \cap intD' \neq \emptyset$.*

We introduce the notation $[x_i]$ to denote the vector whose ith coordinate is equal to x_i. Similarly for the matrix $[x_{ij}]$. The dimensions will be clear from the context. We will use the following characterization of the solutions to the minimization of a convex, differentiable function $F : \mathbb{R}^m \to \mathbb{R}$ over $K \cap D$ [13], [15]:

Theorem 2. [13, 15] *Under Assumption 1, $\bar{x} \in \mathbb{R}^m$ minimizes the convex differentiable function F over $K \cap D$ iff there is a vector $t(\bar{x}) \in \mathbb{R}^m$ such that $\langle t(\bar{x}), x - \bar{x} \rangle \geq 0,\ \forall x \in D$ and*[1] *$\langle \nabla F(\bar{x}), x - \bar{x} \rangle \geq \langle t(\bar{x}), x - \bar{x} \rangle,\ \forall x \in K$.*

Considering the fact that

$$\frac{\partial}{\partial f_P} C(f) = \sum_{e \in P} (2a_e f_e + b_e) \ \ \forall P \in \mathcal{P},$$

Theorem 2 takes the following form in our case:

Theorem 3. *Under Assumption 1, $\bar{f}$ minimizes the convex differentiable function $C(f)$ over $K \cap D$ iff there is a vector $t(\bar{f}) \in \mathbb{R}^{|\mathcal{P}|}$ such that*

$$\langle t(\bar{f}), f - \bar{f} \rangle \geq 0,\ \forall f \in D \tag{6}$$

and

$$\langle [\sum_{e \in P} (2a_e f_e + b_e)], f - \bar{f} \rangle \geq \langle t(\bar{f}), f - \bar{f} \rangle,\ \forall f \in K. \tag{7}$$

From Theorem 1, we have the following characterization of a traffic equilibrium for the given instance:

Theorem 4. *Under Assumption 1, f^* is a traffic equilibrium iff there exists $t(f^*) \in \mathbb{R}^{|\mathcal{P}|}$ such that*

$$\langle t(f^*), f - f^* \rangle \geq 0, \quad \forall f \in D, \tag{8}$$

and

$$\langle [\sum_{e \in P} (a_e f_e^* + b_e) - t_P(f^*)], f - f^* \rangle \geq 0,\ \forall f \in K. \tag{9}$$

---

[1] If F is simply continuous but not differentiable at some point of K, we can replace the left-hand side of the inequality with the directional derivative at $\bar{x}$ ([18], Ch. 23).

We can combine Theorems 3 and 4 to prove the following (proof in the appendix):

Lemma 1. *If $\bar{f}$ is a traffic equilibrium for a given instance, then $\bar{f}/2$ is an optimum flow for the instance scaled by $1/2$.*

Then we can prove the following (proof in the appendix)

Theorem 5. *Under Assumption 1, the coordination ratio for linear latency functions is at most $\frac{4}{3}$.*

We remark that essentially the same proof can be used to show a $4/3$ coordination ratio for the alternative model where the latency of a path P is given by $l_P(f) = a_P f_P + b_P$ where $a_P, b_P \geq 0$.

4 General latency functions

In this section we study existence and uniqueness of solutions, and coordination ratios for latency functions that need to satisfy only mild assumptions. The basis for our analysis is the formulation of the Extended Wardrop Principle (Definition 3) as a nonlinear complementarity problem. This formulation has been studied by Larsson and Patriksson [10],[11], but here we study it in terms of the existence and uniqueness of traffic equilibria following the approach used by Aashtiani and Magnanti [1]. The latter authors studied the case of general latency functions but without side constraints. Their elegant approach can be easily extended to deal with broad families of side constraints (which include the capacity and budgetary constraints that motivated our work), but it is conceivable that it could be extended to even more general settings.

We start by assuming that the convex set D of Section 2 is defined by a set of J side constraints $\{g_j(f) \leq 0,\ j = 1, \ldots, J\}$ with each $g_j : \mathbb{R}^{|\mathcal{P}|} \to \mathbb{R}$ being a convex and differentiable function. We assume a suitable regularity condition (such as Slater's condition which requires the existence of $\bar{f} \in K$ such that $g_j(\bar{f}) < 0$ for all $j \in J$) so that the Kuhn-Tucker conditions hold. The Slater condition implies that $K \cap intD \neq \emptyset$. Also we will assume that

$$\frac{\partial g_j}{\partial f_P}(f) \geq 0,\ \forall j, \forall f.$$

While this is a limitation to the kinds of side constraints studied here, it covers many important ones such as linear capacity and budgetary constraints.

In what follows, we assume that the upper and lower bounds for the path flows that define K are the trivial ones, namely $\lambda_P = 0, \mu_P > \max_i d_i$. We also assume that Assumption 1 holds. Then Theorem 1 implies that any traffic equilibrium f^* is characterized exactly by Wardrop's principle (Definition 3). Under the assumptions above, [10] show that this principle translates into the following:

There are J real numbers $b_j,\ j = 1, \ldots, J$ such that:

$$b_j \geq 0,\ b_j g_j(f^*) = 0,\ j = 1, \ldots, J \qquad (a)$$

$$\langle L(f^*) + \sum_{j=1}^{J} b_j \nabla g_j(f^*), f - f^* \rangle \geq 0, \forall f \in K \qquad (b)$$

where $L(f^*) = [l_{P_1}(f^*), \ldots, l_{P_{|\mathcal{P}|}}(f^*)]^T$ is the vector of the latencies for all paths for f^*.

Let $F(x)$ be a vector-valued function from $\mathbb{R}^n$ to itself. The nonlinear complementarity problem is to find a vector x that satisfies the system:

$$\langle x, F(x) \rangle = 0, \quad F(x) \geq 0, \quad x \geq 0.$$

Following [1] (Proposition 4.1, p. 219), we can show that, under certain conditions, the given characterization of a traffic equilibrium is equivalent to the solution of a nonlinear complementarity problem.

Theorem 6. *Suppose that, for all paths $l_P(f) > 0$ when $f_P \neq 0$. Then $f^* \in K \cap D$ is a traffic equilibrium iff there are vectors $b^* \in \mathbb{R}^J, u^* \in \mathbb{R}^I$ such that (f^*, b^*, u^*) is a solution to the following complementarity problem:*

$$f_P(l_P(f) + \sum_{j=1}^{J} b_j \frac{\partial g_j}{\partial f_P}(f) - u_i) = 0, \quad \forall i \in I,\ \forall P \in \mathcal{P}_i \qquad (10)$$

$$l_P(f) + \sum_{j=1}^{J} b_j \frac{\partial g_j}{\partial f_P}(f) \geq u_i, \quad \forall i \in I,\ \forall P \in \mathcal{P}_i \qquad (11)$$

$$u_i(\sum_{P \in \mathcal{P}_i} f_P - d_i) = 0, \quad \forall i \in I \qquad (12)$$

$$\sum_{P \in \mathcal{P}_i} f_P \geq d_i, \quad \forall i \in I \qquad (13)$$

$$b_j g_j(f) = 0, \quad \forall j = 1, \ldots, J \qquad (14)$$

$$-g_j(f) \geq 0, \quad \forall j = 1, \ldots, J \qquad (15)$$

$$f_P, b_j, u_i \geq 0, \quad \forall P, j, i \qquad (16)$$

Proof. Since $K \cap intD \neq \emptyset$, by Theorem 1 it suffices to show that any (f^*, b^*) that satisfies the equilibrium conditions (a), (b), solves the complementarity problem, and vice versa. Maugeri et al. show (Theorem 1 in [13]) that condition (5) is equivalent to the following:

$$\forall i,\ \exists \gamma_i \text{ such that } \left\{ \begin{array}{l} l_P(f^*) + \sum_{j=1}^{J} b_j \frac{\partial g_j}{\partial f_P}(f^*) < \gamma_i \Rightarrow f_P^* = \mu_P \\ l_P(f^*) + \sum_{j=1}^{J} b_j \frac{\partial g_j}{\partial f_P}(f^*) > \gamma_i \Rightarrow f_P^* = \lambda_P \end{array} \right\}, \forall P \in \mathcal{P}_i. \qquad (17)$$

Since, by our assumption on the values of λ_P, μ_P, we have $f_P^* < \mu_P, \forall P$, it must be that $l_P(f^*) + \sum_{j=1}^{J} b_j \frac{\partial g_j}{\partial f_P}(f^*) \geq \gamma_i,\ \forall i \in I, \forall P \in \mathcal{P}_i$. Hence (17) implies that $f_P^*(l_P(f^*) + \sum_{j=1}^J b_j \frac{\partial g_j}{\partial f_P}(f^*) - \gamma_i) = 0$. Also, due to the hypotheses in the statement of the theorem, we can increase the γ_i's until we have $\gamma_i \geq 0, \forall i$ without violating (17). And the fact that $f^* \in K$ implies that $\sum_{P\in\mathcal{P}_i} f_P = d_i,\ \forall i \in I$. Therefore, if we also recall that condition (a) must also hold, we see that (f^*, b^*, u^*) with $b_j^* := b_j, u_i^* := \gamma_i$ is a solution to the complementarity problem.

For the opposite direction, let (f^*, b^*, u^*) be a solution to the complementarity problem. Notice that condition (a) holds. First of all, we show that $f^* \in K$. For, suppose that $\sum_{P\in\mathcal{P}_i} f_P^* > d_i \geq 0$ for some i. Then (12) implies that $u_i^* = 0$ and there is path $P \in \mathcal{P}_i$ such that $f_P^* > 0$. Since $f_P^* \neq 0$, the hypotheses of the theorem imply that $l_P(f^*) + \sum_{j=1}^J b_j^* \frac{\partial g_j}{\partial f_P}(f^*) > 0 = u_i^*$ and therefore (10) implies that $f_P^* = 0$, a contradiction. Hence $\sum_{P\in\mathcal{P}_i} f_P^* = d_i$, i.e., $f^* \in K$. All that remains is to see that for $\gamma_i := u_i^*$, condition (17) is satisfied, i.e., condition (b) is also satisfied and therefore f^*, b^* satisfy the Extended Wardrop Principle, and f^* is a traffic equilibrium. □

In what follows, let $[x]^+ := \max\{0, x\}$. We define the continuous mapping ϕ from $\{0 \leq f \leq K_1 e\} \times \{0 \leq b \leq Be'\} \times \{0 \leq u \leq K_2 e''\} \subseteq \mathbb{R}^{|\mathcal{P}|+J+I}$ to itself as follows:

$$\phi_P(f, b, u) = \min\{K_1, [f_P + u_i - l_P(f) - \sum_j b_j \frac{\partial g_j}{\partial f_P}(f)]^+\}, \quad \forall i, \forall P \in \mathcal{P}_i$$

$$\phi_j(f, b, u) = \min\{B, [b_j + g_j(f)]^+\}, \quad \forall j$$

$$\phi_i(f, b, u) = \min\{K_2, [u_i + d_i - \sum_{P\in\mathcal{P}_i} f_P]^+\}, \quad \forall i$$

where $K_1 > \max_i d_i$, $K_2 > \max_P \max_{0\leq f\leq K_1 e}\{l_P(f) + B\sum_j \frac{\partial g_j}{\partial f_P}(f)\}$ and B is a positive number which will be defined later. The mapping ϕ has at least one fixed point. The assumption that follows relates the possibility that some $b_j^* = B$ with the fact that f^* *is not* an equilibrium for the problem *without side constraints.*

Assumption 2 *There is $B > 0$ such that for every fixed point (f^*, u^*, b^*) of ϕ with $b_j^* = B$ for some j, there is a flow $f_0 \in K$ such that $\langle L(f^*) + \sum_{j=1}^J b_j^* \nabla g_j(f^*), f_0 - f^*\rangle < 0$.*

Theorem 7. *Under Assumption 2, and if $l_P(f) \geq 0,\ \forall f, P$ and continuous, and $\frac{\partial g_j}{\partial f_P}(f),\ \forall j, f$ are continuous positive functions, every fixed point of ϕ is a solution of the complementarity problem of Theorem 6.*

Proof. Let B be as in Assumption 2 and (f^*, u^*, b^*) be a fixed point of ϕ. Then one can show as in Theorem 5.3 of [1] that

$$f_P^* = [f_P^* + u_i^* - l_P(f^*) - \sum_j b_j^* \frac{\partial g_j}{\partial f_P}(f^*)]^+,\ \forall P$$

$$u_i^* = [u_i^* + d_i - \sum_{P\in\mathcal{P}_i} f_P]^+,\ \forall i$$

$$b_j^* = \min\{B, [b_j^* + g_j(f^*)]^+\},\ \forall j$$

Note that (f^*, u^*) is a solution of the complementarity problem

$$\begin{aligned} f_P(l_P(f) + \sum_{j=1}^{J} b_j^* \frac{\partial g_j}{\partial f_P}(f) - u_i) &= 0, \ \ \forall i \in I, \ \forall P \in \mathcal{P}_i \\ l_P(f) + \sum_{j=1}^{J} b_j^* \frac{\partial g_j}{\partial f_P}(f) &\geq u_i, \ \ \forall i \in I, \ \forall P \in \mathcal{P}_i \\ u_i(\sum_{P \in \mathcal{P}_i} f_P - d_i) &= 0, \ \ \forall i \in I \\ \sum_{P \in \mathcal{P}_i} f_P &\geq d_i, \ \ \forall i \in I \\ f_P, u_i &\geq 0, \ \ \forall P, j, i \end{aligned}$$

which is equivalent to the following variational inequality:

$$\langle L(f) + \sum_{j=1}^{J} b_j^* \nabla g_j(f), \bar{f} - f \rangle \geq 0, \ \forall \bar{f} \in K.$$

Hence for $f := f^*$ we have that

$$\langle L(f^*) + \sum_{j=1}^{J} b_j^* \nabla g_j(f^*), f - f^* \rangle \geq 0, \ \forall f \in K. \tag{18}$$

If there is j such that $b_j^* = B$, then Assumption 2 implies that there is $f_0 \in K$ such that $\langle L(f^*) + \sum_{j=1}^{J} b_j^* \nabla g_j(f^*), f_0 - f^* \rangle < 0$ which contradicts (18). Therefore $b_j^* < B$, $\forall j$, and $b_j^* = [b_j^* + g_j(f^*)]^+$, $\forall j$. From [22] we get that (f^*, u^*, b^*) is a solution of the complementarity problem of Theorem 6.

□

Theorem 7 implies that if its assumptions hold, there is at least one solution to the complementarity problem of Theorem 6.

4.1 Uniqueness for linear edge side constraints

In many cases, the latency of a path is just the sum of the latencies of the path edges, i.e., $l_P(f) = \sum_{e \in P} l_e(f) = (\Delta^T)_P \cdot \kappa(f)$, where Δ is the edge-path incidence matrix for G and $\kappa(f) \in \mathbb{R}^{|E|}$ is the vector of edge latencies. This *additive model* (which we also used in Section 3) is particularly useful if it is more convenient to talk about edge flows instead of path flows: if $f_P, \forall P$ are the path flows, then $F_e, \forall e \in E$ with $F = \Delta f$ are the edge flows. Here we show that if the l_e's are strictly monotone, and the side constraints are linear functions of the *edge* flows, then the solution of the complementarity problem (if one exists) is unique.

As mentioned above, for every j the side constraint g_j has the following form:

$$g_j(f) = \sum_{e \in E} a_{je} F_e - c_j = \sum_{e \in E} a_{je} (\sum_{P \ni e} f_P) - c_j$$

where $a_{je} = 0$ if g_j doesn't depend on the flow for edge e. Let $A := [a_{je}]$ be the coefficient matrix for the side constraints. Then the Jacobian matrix of the vector-valued function $g(f) = [g_1(f), \ldots, g_J(f)]^T$ is $[k_{jp}] = A\Delta$ with $k_{jp} = \frac{\partial g_j(f)}{\partial f_P}$.

Recall that a function $F : C \to \mathbb{R}^n, C \subset \mathbb{R}^n$ is *strictly monotone on* C iff for every $x, y \in C$ with $x \neq y$, $(x - y)(F(x) - F(y)) > 0$. We prove the equivalent of Theorem 6.2 in [1] for our case (see Appendix):

Theorem 8. *If the functions l_e are strictly monotone, then the edge flow induced by any solution of the complementarity problem (if such solutions exist) is unique.*

4.2 Application: edge taxation for heterogeneous selfish users

We cast some recent results by Cole et al. [2] in the framework of Aashtiani and Magnanti [1] as extended here.

Suppose that in the given network, the users want to send flow from a node s to a node t, and experience not only the latency costs, but a tax imposed to them on every edge they go through. Given an infinite number of users with a collective flow of one unit, there is a distribution function that assigns each user to one of I categories (hence the heterogeneity), such that the users in category i weigh their taxation burden with a weight $a(i)$, i.e., these users see a cost $\bar{l}_e(f_e) = l_e(f_e) + a(i)\tau_e$ per flow unit, and try selfishly to route their flow through the cheapest $s - t$ path. We assume that $0 < a(0) < a(1) < \ldots < a(|I|) \leq 1$, and let $d_0, d_1, \ldots, d_{|I|}$ be the total flow that falls in category $0, 1, \ldots, |I|$. Note that although all categories have the same source s and sink t, their path sets $\mathcal{P}_i$ are treated separately.

If $\hat{f}$ is a flow that minimizes the social cost $\sum_e f_e l_e(f_e)$, then define the following convex sets K, D:

$$K = \{f : \sum_{P \in \mathcal{P}_i} f_P = d_i,\ i = 0, \ldots, |I|\}$$

$$D = \{f : \sum_{P \ni e} f_P \leq \hat{f}_e,\ \text{for all } e \text{ such that } \hat{f}_e > 0\}$$

Note that D is essentially defined by a set of artificial capacity constraints for the given network. If we can apply the results of the previous sections to show that there is a unique equilibrium (f, b, u) for the latency costs in $K \cap D$, we can use the b_j's to define the taxes τ_e so that f is also the unique equilibrium in K for the costs with taxation.[2] Then the equilibrium state in this system achieves the minimum social cost as well.

As described in [2], one can set the taxes on edges with $\hat{f}_e = 0$ to $\tau_e > n(T + l_{\max}/a(0))$, where n is the number of vertices, $l_{\max} = \max_e l_e(1)$ and T

[2] Note that f need not be an equilibrium flow over K under the original latency function.

is an upper bound on the edge taxes. The existence of T is in fact what allows the proof of existence of optimal taxes to go through, in much the same way as in Theorem 7. Also the same arguments allow us to add dummy edges that are never going to be used by the users (because they are too heavily taxed), by setting their taxes to $\tau_e > n(T + l_{\max}/a(0))$. By introducing such a dummy edge (s,t), we can make sure that the Slater conditions hold (by sending all flow through the dummy edge, all constraints of D are strictly satisfied). This implies that $K \cap int D \neq \emptyset$.

We define a set of new path latency functions

$$l'_P(f) := \frac{l_P(f)}{a(i)} = \frac{\sum_{e \in P} l_e(f)}{a(i)}$$

and assume that the functions l_e are strictly monotone (and non-negative). Then Theorem 6 shows that the traffic equilibria are exactly the solutions to the complementarity problem for the l'. Set $B := T$ for any $T \geq 3n^3 l_{\max}/a(0)$, and let (f^*, u^*, b^*) be a fixed point of ϕ. If $b^*_e = B$ for some edge $e = (v,w)$, let P_1 be the shortest path from s to w where the edge lengths are given by b^*. P_1 has under the latency function $l(f)' + [\sum_{e \in P} b^*_e]$ a shorter travel time than any path P_2 with $f^*_{P_2} > 0$ that uses e. This means that f^* cannot be an equilibrium in K, i.e., for the problem *without* the capacity constraints. More precisely, if we set

$$f_{0_P} := \begin{cases} f^*_{P_1} + \delta & P = P_1 \\ f^*_{P_2} - \delta & P = P_2 \\ f^*_P, & \forall P \neq P_1, P_2 \end{cases}$$

Lemma 3.3 in [2] shows that with $\delta > 0$

$$\langle L(f^*) + \sum_{j=1}^{J} b^*_j \nabla g_j(f^*), f_0 - f^* \rangle = \delta(l'_{P_1}(f^*) + \sum_{e \in P_1} b_e - l'_{P_2}(f^*) - \sum_{e \in P_2} b_e) < 0.$$

Therefore Assumption 2 holds for these particular B, f_0, and Theorem 7 implies the existence of a solution (f^*, b^*, u^*) for the complementarity problem (10)-(16). If we set the taxes of edges with $\hat{f}_e > 0$ to b^*_e, f^* is also a traffic equilibrium for the cost function $\bar{l}$ defined above.

Note that the difference between the problem addressed in [2] and the capacitated network is that in the former the capacity constraints are not imposed by the instance itself, but by us, and therefore one has to 'lure' the users into moving towards the desired equilibrium by imposing taxes (modified cost), while in the latter the capacities are part of the instance and hence one should argue for the initial latency costs. This also explains the insistence that the traffic equilibria in the first case should induce *identical* edge flows: without this restriction, the users can achieve a second equilibrium without the exact same edge flows for the same set of taxes. This equilibrium may not obey our (artificial) side constraints (i.e., is not in $K \cap D$), but the users do not care about this since they are only required to induce flows feasible in K.

Acknowledgements. The authors thank Tim Roughgarden for useful discussions on a draft of this paper and Andreas Schulz for bringing [4] to our attention and useful comments.

References

1. H. Z. Aashtiani and T. L. Magnanti. *Equilibria on a congested transportation network.* SIAM J. on Algebraic and Discrete Methods Vol. 2(3), pp. 213-226, 1981.
2. R. Cole, Y. Dodis, and T. Roughgarden. *Pricing network edges for heterogeneous selfish users.* In Proceedings of 35th ACM Symposium on Theory of Computing (STOC), pp. 521–530, 2003.
3. J. R. Correa, A. S. Schulz and N. E. Stier Moses. *Selfish routing in capacitated networks.* Working Paper 4319-03, MIT Sloan School of Management, June 2003, Cambridge, MA.
4. J. R. Correa, A. S. Schulz and N. E. Stier Moses. *Selfish routing in capacitated networks.* To appear in MOR, February 2004.
5. S. C. Dafermos. *Traffic equilibria and variational inequalities.* Transportation Science, 14, 42–54, 1980.
6. S. C. Dafermos and F. T. Sparrow. *The traffic assignment problem for a general network.* Journal of Research of the National Bureau of Standards, Series B, 73B (2):91-118, 1969.
7. P. Hartman and G. Stampacchia. *On some nonlinear elliptic differential functional equations.* Acta Mathematica, 115, 153–188, 1966.
8. G. Karakostas and S. G. Kolliopoulos. *Selfish routing in the presence of side constraints.* Department of Computing and Software, McMaster University, TR CAS-03-13-GK, November 2003, Hamilton, Ontario.
9. E. Koutsoupias and C. Papadimitriou. *Worst-Case Equilibria.* In Proceedings of 16th Symposium on Theoretical Aspects of Computer Science, LNCS 1563, pp. 404–413, 1999.
10. T. Larsson and M. Patriksson. *Equilibrium characterizations of solutions to side constrained asymmetric traffic assignment models.* Le Matematiche 49, pp. 249-280, 1994.
11. T. Larsson and M. Patriksson. *Side Constrained Traffic Equilibrium Models - Analysis, computation and applications.* Transportation Research 33B, pp. 233–264, 1999.
12. A. Maugeri. *Optimization problems with side constraints and generalized equilibrium principles.* Le Matematiche 49, pp. 305-312, 1994.
13. A. Maugeri, W. Oettli, and D. Schläger. *A flexible form of Wardrop's principle for traffic equilibria with side constraints.* Rendiconti del Circolo Matematico di Palermo, Serie II, Supplement 48, pp. 185-193, 1997.
14. A. Nagurney. *Network Economics, A Variational Inequality Approach.* Kluwer, Dordrecht 1993.
15. W. Oettli. *Kolmogorov conditions for vectorial optimization problems.* OR-Spektrum 17, pp. 227–229, 1995.
16. W. Oettli. *Necessary and sufficient conditions of Wardrop type for vectorial traffic equilibria*, In Equilibrium problems: nonsmooth optimization and variational inequality models, edited by F. Giannessi, A. Maugeri and P. M. Pardalos. Kluwer, pp. 223–229, 2001.

17. M. Patriksson. *The traffic assignment problem: models and methods.* Topics in Transportation Series, VSP, Utrecht, The Netherlands, 1994.
18. R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.
19. T. Roughgarden and É. Tardos. *How bad is selfish routing?* Journal of the ACM 49:2, pp. 236-259, 2002.
20. A. Schulz and N. Stier Moses. *On the Performance of User Equilibria in Traffic Networks.* In Proceedings of 14th ACM/SIAM Symposium on Discrete Algorithms (SODA), pp. 86-87, 2003.
21. M. J. Smith. *The existence, uniqueness and stability of traffic equilibria.* Transportation Research Vol. 13B, pp. 295-304, 1979.
22. M. J. Todd. *The computation of fixed points and applications.* Lecture Notes in Economics and Mathematical Systems, 124, Springer-Verlag, Berlin, 1976.
23. J. G. Wardrop. *Some theoretical aspects of road traffic research.* In Proceedings of the Institute of Civil Engineers, Pt. II, vol. 1, pp. 325–378, 1952.

Appendix

Proof of Fact 1. The set K' is defined as $K'(\lambda/\delta, \mu/\delta, d/\delta)$. The only requirement for the set D' is that it is convex. Let $D' = \{y | \delta y \in D\}$. Any convex combination of two elements of D' belongs to D' hence D' is indeed convex. Moreover by construction f is feasible for (K, D) iff f/δ is feasible for (K', D'). The second statement follows because $f \in K \cap intD$ iff $f/\delta \in K' \cap intD'$. $\square$

Proof of Lemma 1. Fact 1 implies that if f is feasible for the given instance (K, D), $f/2$ is feasible for the instance scaled by $1/2$, (K', D'), and vice-versa, if f' is feasible for the scaled instance, then $2f'$ is feasible for the initial instance. Hence by Theorem 4 $\langle t(\bar{f}), f - \bar{f} \rangle \geq 0,\ \forall f \in D \Rightarrow \langle t(\bar{f}), \frac{f}{2} - \frac{\bar{f}}{2} \rangle \geq 0,\ \forall \frac{f}{2} \in D'$ for some vector $t(\bar{f})$, and $\langle [\sum_{e \in P}(a_e \bar{f}_e + b_e) - t_P(\bar{f})], f - \bar{f} \rangle \geq 0,\ \forall f \in K \Rightarrow \langle [\sum_{e \in P}(2a_e \frac{\bar{f}_e}{2} + b_e) - t_P(\bar{f})], \frac{f}{2} - \frac{\bar{f}}{2} \rangle \geq 0, \forall \frac{f}{2} \in K'$. Theorem 3 now implies that $\bar{f}/2$ is optimal for the scaled instance (K', D'). $\square$

Proof of Theorem 5. Let $\bar{f}$ be a traffic equilibrium for the given instance (K, D), and f^* be the flow of minimum total cost. Also, if f is a flow feasible for (K, D), then we set $f_{1+\delta} := (1+\delta)f$, for any $\delta > 0$. Note that because of Fact 1, $f_{1+\delta}$ is feasible for the instance scaled by $1/(1+\delta)$.

From the convexity of $C(f)$, we get that

$$C(f) \geq C(f^*) + \langle \nabla C(f^*), f - f^* \rangle,\ \forall f \in \mathbb{R}^{|\mathcal{P}|}$$

which implies that

$$C(f_{1+\delta}) \geq C(f^*) + \langle [\sum_{e \in P}(2a_e f_e^* + b_e)], f_{1+\delta} - f^* \rangle. \tag{19}$$

Claim. Let f be any feasible flow for the given instance. Then

$$\langle [\sum_{e \in P}(2a_e f_e^* + b_e) - t_P(f^*)], f \rangle \geq \langle [\sum_{e \in P}(2a_e f_e^* + b_e) - t_P(f^*)], f^* \rangle$$

where $t(f^*)$ is the vector from Theorem 3 for f^*.

Proof of claim: From the discussion after Definition 3, we have that condition (7) of Theorem 3 is equivalent to

$$\bar{c}_{P_1} < \bar{c}_{P_2} \Rightarrow f^*_{P_1} = \mu_{P_1} \text{ or } f^*_{P_2} = \lambda_{P_2},\ \forall P_1, P_2 \in \mathcal{P}_i, \forall i$$

where $\bar{c}_P := \sum_{e\in P}(2a_e f^*_e + b_e) - t_P(f^*)$. Hence $\bar{c}$ defines an ordering of the paths, so that the cheapest paths are saturated by f^*, i.e., $f^*_P = \mu_P$, the most expensive have the minimum possible flow, i.e., $f^*_P = \lambda_P$, and in between there may be paths $P_1, P_2, \ldots$ with flow values $\lambda_{P_1} \le f^*_{P_1} \le \mu_{P_1},\ \lambda_{P_2} \le f^*_{P_2} \le \mu_{P_2}, \ldots$ Note that for the latter it must be the case that $\bar{c}_{P_1} = \bar{c}_{P_2} = \ldots$, otherwise condition (7) would not hold for f^*. Since for every flow $f \in K$ $\lambda_P \le f_P \le \mu_P,\ \forall P$ and path flows are non-negative numbers, one can reduce the value of $\sum_P \bar{c}_P f_P$ by moving flow from the more expensive paths to the cheaper ones, while keeping the flow within K. It is obvious that the smallest value is achieved by f^*, and the claim holds. □

For $f := \frac{f_{1+\delta}}{1+\delta}$, the claim above implies that

$$\langle[\sum_{e\in P}(2a_e f^*_e + b_e) - t_P(f^*)], f_{1+\delta}\rangle \ge (1+\delta)\langle[\sum_{e\in P}(2a_e f^*_e + b_e) - t_P(f^*)], f^*\rangle$$

or, equivalently,

$$\langle[\sum_{e\in P}(2a_e f^*_e + b_e)], f_{1+\delta} - f^*\rangle \ge \delta\langle[\sum_{e\in P}(2a_e f^*_e + b_e) - t_P(f^*)], f^*\rangle + \langle t(f^*), f_{1+\delta} - f^*\rangle \tag{20}$$

By combining (19) with (20) we get that

$$\begin{aligned} C(f_{1+\delta}) &\ge C(f^*) + \delta\langle[\sum_{e\in P}(2a_e f^*_e + b_e) - t_P(f^*)], f^*\rangle + \langle t(f^*), f_{1+\delta} - f^*\rangle \\ &= C(f^*) + \delta\langle[\sum_{e\in P}(2a_e f^*_e + b_e)], f^*\rangle + (1+\delta)\langle t(f^*), \frac{f_{1+\delta}}{1+\delta} - f^*\rangle \\ &\ge C(f^*) + \delta\langle[\sum_{e\in P}(2a_e f^*_e + b_e)], f^*\rangle \end{aligned} \tag{21}$$

where the last inequality holds because $\frac{f_{1+\delta}}{1+\delta}$ is a feasible flow for the given instance, i.e., $\frac{f_{1+\delta}}{1+\delta} \in K \cap D$, and condition (6) must hold for f^*.

Since $\bar{f}$ is a traffic equilibrium for the given instance, $\bar{f}/2$ is an optimal flow for the scaled by $1/2$ instance (Lemma 1). Then we can apply (21) for the scaled instance, with $f^* := \bar{f}/2$ and $f_{1+\delta} :=$ any $f \in K \cap D$: $C(f) \ge C(\frac{\bar{f}}{2}) + \langle[\sum_{e\in P}(2a_e \frac{\bar{f}_e}{2} + b_e)], \frac{\bar{f}}{2}\rangle = C(\frac{\bar{f}}{2}) + \frac{1}{2}\langle[\sum_{e\in P}(a_e \bar{f}_e + b_e)], \bar{f}\rangle = C(\frac{\bar{f}}{2}) + \frac{1}{2}C(\bar{f})$. By setting $f := f^*$, we get the following lower bound for the cost of the optimum solution of the given instance:

$$C(f^*) \ge C(\frac{\bar{f}}{2}) + \frac{1}{2}C(\bar{f}) \tag{22}$$

Also $C(\frac{\bar{f}}{2}) = \sum_e(\frac{1}{4}a_e \bar{f}_e^2 + \frac{1}{2}b_e \bar{f}_e) \ge \frac{1}{4}\sum_e(a_e \bar{f}_e^2 + b_e \bar{f}_e) = \frac{1}{4}C(\bar{f})$ which, in combination with (22), implies $C(f^*) \ge \frac{3}{4}C(\bar{f})$. □

Proof of Theorem 8. In addition to A, Δ, we will use the matrix $\Gamma := [\gamma_{Pi}]$, with $\gamma_{Pi} = 1$ if $P \in \mathcal{P}_i$, and 0 otherwise. Let $(f_1, b_1, u_1), (f_2, b_2, u_2)$ be two solutions of the complementarity problem. As in [1], we use the fact that if x_1, x_2 are two solutions of the complementarity problem

$$x^T F(x) = 0, F(x) \geq 0, x \geq 0$$

then $(x_1 - x_2)^T (F(x_1) - F(x_2)) \leq 0$. In our case, this implies that

$$\begin{aligned}
0 \geq &(f_1 - f_2)^T (\Delta^T \kappa(\Delta f_1) + \Delta^T A^T b_1 - \Gamma u_1 - \Delta^T \kappa(\Delta f_2) - \Delta^T A^T b_2 + \Gamma u_2) + \\
&+ (u_1 - u_2)^T (\Gamma^T f_1 - d - \Gamma^T f_2 + d) + \\
&+ (b_1 - b_2)^T (-A\Delta f_1 + A\Delta f_2) \\
\Rightarrow &(f_1 - f_2)^T (\Delta^T \kappa(\Delta f_1) - \Delta^T \kappa(\Delta f_2)) \leq 0 \\
\Rightarrow &(\Delta f_1 - \Delta f_2)^T (\kappa(\Delta f_1) - \kappa(\Delta f_2)) \leq 0
\end{aligned}$$

which implies that $\Delta f_1 = \Delta f_2$ because the functions l_e are strictly monotone.