Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games

Lisa Fleischer Kamal Jain Mohammad Mahdian

July 30, 2004

Abstract

We prove the existence of tolls to induce multicommodity, heterogeneous network users that independently choose routes minimizing their own linear function of tolls versus latency to collectively form the traffic pattern of a minimum average latency flow. This generalizes both the previous known results of the existence of tolls for multicommodity, homogeneous users [1] and for single commodity, heterogeneous users [3].

Unlike previous proofs for single commodity users in general graphs, our proof is constructive - it does not rely on a fixed point theorem - and results in a simple polynomial-sized linear program to compute tolls when the number of different types of users is bounded by a polynomial.

We show that our proof gives a complete characterization of flows that are enforceable by tolls. In particular, tolls exist to induce any traffic pattern that is the result of minimizing an arbitrary function from $\mathbb{R}^{E(G)}$ to the reals that is nondecreasing in each of its arguments. Thus, tolls exist to induce flows with minimum average weighted latency, minimum maximum latency, and other natural objectives.

We give an exponential bound on tolls that is independent of the number of network users and the number of commodities. We use this to show that multicommodity tolls also exist when users are not from discrete classes, but instead define a general function that trades off latency versus toll preference.

Finally, we show that our result extends to very general frameworks. In particular, we show that tolls exist to induce the Nash equilibrium of general nonatomic congestion games to be system optimal. In particular, tolls exist even when 1) latencies depend on user type; 2) latency functions are nonseparable functions of traffic on edges; 3) the latency of a set S is an arbitrary function of the latencies of the resources contained in S. Our exponential bound on size of tolls also holds in this case; and we give an example of a congestion game that shows this is tight: it requires tolls that are exponential in the size of the game.

1 Introduction

We analyze when tolls on resource usage can induce users to behave in a way that maximizes some global objective, in systems where users selfishly select resources to meet their individual demands. We assume that the users (also known as the agents) are infinitesimally small, and therefore the action of a single user does not affect others considerably.

In the network setting, each edge has an associated latency function that is a nondecreasing function of the *congestion* of the edge: the number of users that use the edge. Without tolls, users seek a least latency path from their source to destination, where latency of a path is the sum of the latencies of the edges in a path [14]. The resulting flow is called a *Nash flow* or a *Wardrop equilibrium*. The network owner, on the other hand, may desire to maximize social welfare by minimizing average latency experienced by users, the *system optimal* flow. The Nash flow may be far from the system optimal flow [8, 12]. By placing tolls on the use of edges, the owner hopes to induce users to selfishly select a system optimal flow. With tolls, users seek to minimize some function of latency plus toll. Each user may have a different trade-off of latency for toll. For agent a, we can represent this trade-off as a latency multiplier, $\alpha(a)$ that converts latency into dollars.

This setting has been considered previously in the transportation and computer science literature. For the case when $\alpha(a) = 1$ for all agents a, it is well known that the Nash flow with marginal cost tolls is a system optimal flow [1, 9]. For distinct α, early work describes solutions that toll each user differently according to their aversion to latency [4, 13]. This is unsatisfying and hard to enforce, as it requires knowing each user's α value.

Three distinct attempts have been made to address this problem. Dial [5] shows that α-weighted marginal cost tolls induce a flow that minimizes the α-weighted average latency, even for multicommodity traffic. While this is a satisfying result, such a marginal cost toll result holds for this specific global objective function only, as it is a result of relation between the users objective functions and the gradient of the global objective function. Cole, Dodis, and Roughgarden [3], show that for the case when all agents have the same source and destination, then tolls exist so that the Nash flow with tolls minimizes average latency. They give an existential proof and pose as open questions both the existence of a constructive proof, and the existence of tolls in the multicommodity setting.

We generalize all of these results. We prove that for *any minimal congestion*, there exist tolls such that the Nash flow induced by *multicommodity*, heterogeneous users is the given congestion. This gives a complete characterization of flows that are enforceable by tolls. In particular, tolls exist to induce any traffic pattern that is the result of minimizing an arbitrary function from $\mathbb{R}^{E(G)}$ to the reals that is nondecreasing in each of its arguments. Thus, tolls exist to minimize average weighted latency flows, maximum latency flows, and other natural objectives.

Unlike the proof of Cole et al. [3], our proof is constructive and does not rely on a fixed point theorem. It is obtained using linear programming duality, and as a consequence, we get a simple polynomial time algorithm to compute the tolls for a bounded number of α types via linear programming. Our linear program (LP) is distinct from the one used in [3] in two important aspects: First, our LP gives a direct proof of the existence of tolls. The LP in [3] offers no such proof - its correctness relies on establishing the existence of tolls via a separate fixed point argument. Second, our LP does not assume any knowledge of the decomposition of the system optimal flow by an agent's α value. The constraints used in [3] do require this. This is a strong assumption, as there are many ways that a flow can be decomposed into paths, but perhaps only one of these decompositions corresponds to the set of paths used by users when the right set of tolls are imposed. Fleischer [6] gives an example to demonstrate that the correct decomposition may depend on α. A second consequence of the linear program approach we give is that we can compute a set of feasible tolls that minimize *any* linear objective function of tolls, including minimizing sum of tolls, or minimizing maximum toll.

We prove that any enforceable congestion can be enforced using tolls bounded by a value that is independent of the number of users and the number of commodities (but depends exponentially on the size of the network). We use this, together with a compactness argument, to show that tolls also exist when users are not from discrete classes, but instead define a general function that trades off latency versus toll preference.

We show that our results on the existence of tolls extend to more general nonatomic congestion games. For example, they hold in abstract resource allocation settings; they hold when latencies are arbitrary, nonseparable functions of resource use; they hold when latencies depend on user type; they hold when the latency of a set S is an arbitrary function of the latencies of the resources contained in S.

Two examples illustrate some uses of these generalizations: In a wireless network, latency at a link does not only depend upon the usage of that link but also depends upon the usage of the neighboring links, because of interference. This indicates that it is useful to consider nonseparable latency functions. It is also useful to consider latency functions that treat different commodity traffic differently: On the Internet some users may send TCP traffic and some may send UDP. These two types of traffic have different effects on system behavior.

Our exponential bound on size of tolls also holds in this case; and we give an example of a general congestion game that shows this is tight: it requires tolls that are exponential in the size of the game.

In this proceedings, Karakostas and Kolliopoulos also give a constructive proof to show that tolls exist to induce the minimum average latency multicommodity flow [7].

2 Problem Statement and Preliminaries

In this section we give a formal statement of the problem considered in this paper. We define the problem in two different models: the discrete model and the continuous model. The discrete model is a special case of the continuous model, where there are only a finite number of different types of agents. This model is simpler to understand; we will first prove our results in the

discrete setting, and then generalize it to the continuous setting using the existence result and an upper bound on the tolls that we prove in the discrete model.

Multicommodity networks. In both the discrete and the continuous model, we are given a *multicommodity network*, which consists of a directed graph G with vertex set V and edge set E, a latency function l_e for every $e \in E$, K commodities $\{(\textit{source}_i, \textit{dest}_i, d_i)\}_{i=1}^{K}$, and a parameter α_i (which could be a constant or a distribution) that represents the sensitivity of the ith commodity to latency. Each commodity i is specified by a triple $(\textit{source}_i, \textit{dest}_i, d_i)$, which means that d_i units of flow need to be routed from the vertex $\textit{source}_i \in V$ to the vertex $\textit{dest}_i \in V$ using the edges of G. Let $\mathcal{P}_i$ denote the collection of all paths from $\textit{source}_i$ to $\textit{dest}_i$ in G, and $\mathcal{P} := \cup_i \mathcal{P}_i$. We assume, without loss of generality, that $\sum_i d_i = 1$. With a slight abuse of notation, we sometimes denote the multicommodity network by G too.

The discrete model. In this model, a *multicommodity flow* for the graph G and commodities $\{(\textit{source}_i, \textit{dest}_i, d_i)\}$ is represented by a vector of nonnegative values (f_p^i) for every $i = 1, \ldots, K$ and $p \in \mathcal{P}_i$. Such a flow is feasible if for every i, $\sum_{p \in \mathcal{P}_i} f_p^i = d_i$. Intuitively, this means that the ith commodity sends f_p^i units of flow along the path p.

A *congestion* is defined as a vector $(g_e)_{e \in E} \in \mathbb{R}^E$. Every flow f corresponds to a congestion defined as $f_e = \sum_i \sum_{p \in \mathcal{P}_i : e \in p} f_p^i$. This is called the congestion induced by f. We say that a congestion g is *feasible* for the commodities $\{(\textit{source}_i, \textit{dest}_i, d_i)\}$ if there is a feasible multicommodity flow whose induced congestion on every edge e is less than or equal to g_e.

Initially, we assume that every edge $e \in E$ has a non-decreasing continuous *latency function* $l_e : [0, 1] \mapsto \mathbb{R}^+$ associated with it. This function specifies how much latency each commodity using e will suffer given the congestion of e (i.e., the total amount of flow that passes through e). More precisely, if (f_e) is the congestion induced by a flow f, then the latency observed on a path p is $l_p(f) := \sum_{e \in p} l_e(f_e)$. In Section 6, we look at more general functions for edge latency and path latency.

We assume that the flow is composed of infinitesimally small agents that behave selfishly. In the absence of tolls, each agent of the i'th commodity wants to get from $\textit{source}_i$ to $\textit{dest}_i$ using a path that minimizes her total latency. The selfish nature of the agents and the lack of coordination between them causes inefficiency in the system (see, for example, Braess's paradox [11]). In order to overcome this, a central authority sets tolls on the edges of the network, to direct the selfish behavior of the agents toward a social optimum. Formally, we denote the toll on an edge e by τ_e. An agent that uses a path p has to pay a toll of $\tau_p := \sum_{e \in p} \tau_e$ and experiences a delay of $l_p(f) := \sum_{e \in p} l_e(f_e)$. We assume the cost observed by an agent of commodity i using a path $p \in \mathcal{P}_i$ is of the form $\alpha_i l_p(f) + \tau_p$, where α_i is a given positive number that indicates the sensitivity of agents of commodity i to the latency.[1]

These utility functions define a game between the agents, whose equilibrium is called a *Nash flow* (also known as a *Wardrop equilibrium*) in G with respect to tolls τ, or a Nash flow in G^τ. More precisely, the Nash flow in G^τ is a multicommodity flow f such that for every commodity i and every two paths $p, p' \in \mathcal{P}_i$ such that $f_p^i > 0$, we have $\alpha_i l_p(f) + \tau_p \leq \alpha_i l_{p'}(f) + \tau_{p'}$ (in words, all paths that agents of commodity i are using are required to be minimum cost paths with respect to the cost function of these agents).

The continuous model. The difference between the continuous model and the discrete model is that in the discrete model we assume that all agents of commodity i have the same sensitivity α_i to latency, while in the continuous model we allow the sensitivity of these agents to come from an arbitrary given distribution. To model this formally, we represent each infinitesimal agent of commodity i as a real number in $[0, d_i]$. The sensitivity of agents of commodity i to latency is given by a function $\alpha_i : [0, d_i] \mapsto \mathbb{R}^+$. We assume that agents are ordered by their sensitivity; in other words α_i's are nondecreasing functions.

A multicommodity flow is a collection (f^i) of Lebesgue-measurable functions $f^i : [0, d_i] \mapsto \mathcal{P}_i$, one for each commodity i. The amount of flow of commodity i on a path $p \in \mathcal{P}_i$ is defined as the Lebesgue measure of $\{a \in [0, d_i] : f^i(a) = p\}$, and denoted by f_p^i. The congestion induced by f on an edge

[1] Cole, Dodis, and Roughgarden [3] consider utilities of the form $\beta_i T + L$. Our model is obviously equivalent to theirs by setting $\alpha_i = 1/\beta_i$. We will consider latencies as perceived differently for different users. In order for us to compare utilities, it is useful to express them in the common currency of money.

e is defined as $f_e := \sum_i \sum_{p \in \mathcal{P}_i : e \in p} f_p^i$. The latency experienced on a path p is defined in the same way as in the discrete model. Given a toll τ_e on each edge e, a flow f is called a Nash flow in G^τ if for every commodity i and every agent $a \in [0, d_i]$, the minimum of the cost $\alpha_i(a) l_p(f) + \tau_p$ over paths $p \in \mathcal{P}_i$ is achieved at $p = f^i(a)$ (in words, each agent uses a min cost path with respect to her sensitivity to latency, the current congestion, and tolls).

Notice that the discrete model is essentially equivalent to the continuous model when α_i's are step functions with a bounded number of steps.

It is known that a Nash flow always exists and is essentially unique (under mild conditions on the latency functions). [3] gives details and further references.

Enforceable congestions. Given a multicommodity network G, we call a congestion g *enforceable*, if there is a set of nonnegative tolls τ such that the congestion induced by the Nash flow in G^τ is g. Cole, Dodis, and Roughgarden [3] proved that in the case of networks with a single source, the optimal congestion, i.e., the congestion that minimizes the average latency of all agents is enforceable, and asked whether the same result holds for multicommodity flows. In this paper, we settle this question affirmatively, by giving a characterization of the set of all enforceable congestions. Our results even hold for the general class of *congestion games*, which is an important and extensively-studied class of games defined by Rosenthal [10].

Linear Programming preliminaries. In this paper we make strong use of linear programming duality. There are many basic reference texts on this subject, for example [2]. We briefly review some of the basics that we use here. A linear program defined by data matrices P and C and data vectors a, p, c with variable vector x of the form $\min ax; Px \leq p; Cx = c; x \geq 0$ has a *linear program dual* of the form $\max c^\mathsf{T} z - p^\mathsf{T} t; C^\mathsf{T} z - P^\mathsf{T} t \leq a; t \geq 0$. (Linear programs may have many different forms. This is just for example.) Solutions x and z, t are said to be *complementary* if $x_j > 0$ implies that $C_j z - P_j t = a_j$ (conversely, $C_j z - P_j t < a_j$ implies $x_j = 0$); $t_i > 0$ implies that $P_i x = p_i$; and $z_i > 0$ implies that $C_i x = c_i$.

FACT 2.1. *If both a linear program and its dual have feasible solutions, then they both have optimal solutions, and every pair of optimal solutions of the primal and the dual are complementary. Conversely, if x is a feasible solution to the primal and (t, z) is a feasible solution to the dual, and x and (t, z) are complementary, then both are optimal.*

3 Existence of optimal tolls in the discrete model

In this section, we prove that in the discrete model, it is possible to find tolls that enforce the optimal congestion. The proof is based on complementary slackness conditions applied to a pair of linear programs defined below.

Assume g is a congestion that we would like to enforce. Given this congestion, we define the linear program P_g as follows:

$$\text{minimize} \quad \sum_i \alpha_i \sum_{p \in \mathcal{P}_i} l_p(g) f_p^i \tag{3.1}$$

subject to

$$\forall e \in E: \quad \sum_i \sum_{p \in \mathcal{P}_i : e \in p} f_p^i \leq g_e \tag{3.2}$$

$$\forall i: \quad \sum_{p \in \mathcal{P}_i} f_p^i = d_i \tag{3.3}$$

$$\forall i \; \forall p \in \mathcal{P}_i: \quad f_p^i \geq 0 \tag{3.4}$$

The dual D_g of the above program is the following:

$$\text{maximize} \quad \sum_i d_i z_i - \sum_{e \in E} g_e t_e \tag{3.5}$$

subject to

$$\forall i \; \forall p \in \mathcal{P}_i: \quad z_i - \sum_{e \in p} t_e \leq \alpha_i l_p(g) \tag{3.6}$$

$$\forall e \in E: \quad t_e \geq 0 \tag{3.7}$$

Let $\hat{f}$ and $(\hat{t}, \hat{z})$ be optimal solutions to these respective programs. Complementary slackness implies that if $\hat{f}_p^i > 0$ then $\hat{z}_i = \sum_{e \in p} \hat{t}_e + \alpha_i l_p(g)$. This means that $\hat{z}_i$ represents the cost of all paths used by commodity i, so that $\hat{f}$ is a Nash flow.

We define the concept of *minimality* of a congestion as follows:

DEFINITION 1. *A feasible congestion g is minimal if and only if the linear program P_g has an optimal solution in which for every $e \in E$, the inequality (3.2) is tight.*

We now prove the following theorem, that characterizes the set of all enforceable congestions.

THEOREM 3.1. *A feasible congestion g is enforceable if and only if it is minimal.*

Proof. First, we prove the "if" part. By minimality of g and LP duality, there is an optimal solution f for P_g such that for every $e \in E$, the inequality (3.2) is tight (in other words, the congestion induced by f is g), and a corresponding complementary optimal solution (t, z) for D_g. Now, we prove, using the complementarity slackness conditions, that the flow f is a Nash flow in G^t. Fix a commodity i, and consider a path $p \in \mathcal{P}_i$ with nonzero flow (i.e., $f_e^i > 0$). By the primal complementarity slackness condition, for every such p we have $\alpha_i l_p(g) + \sum_{e \in p} t_e = z_i$. This means that the utility of the agents of commodity i using p is the same value z_i for all $p \in \mathcal{P}_i$. Also, for any other path $p \in \mathcal{P}_i$, by inequality (3.6) we have $\alpha_i l_p(g) + \sum_{e \in p} t_e \geq z_i$. Therefore, agents do not have an incentive to switch their paths. Thus, f is a Nash flow in G^t, and the congestion induced by f is g. Therefore, g is enforceable.

Conversely, assume that a congestion g is enforceable. This means that there is a multicommodity flow f and tolls τ such that f is a Nash flow in G^τ, and the congestion induced by it is g. Since f is a Nash flow, for every i, all the agents of type i should have the same utility. This means that for every $p \in \mathcal{P}_i$ such that $f_p^i > 0$, the value $\alpha_i l_p(g) + \tau(p)$ is the same. Let us call this value z_i. Since no agent has an incentive to change her path, for every path $p \in \mathcal{P}_i$ we must have $\alpha_i l_p(g) + \tau(p) \geq z_i$. Thus, if we consider f and (τ, z) as the solutions of the programs P_g and D_g, then they are both feasible solutions, and they satisfy the complementarity slackness conditions. Thus, f is an optimal solution for P_g, and we also know that for every e, inequality (3.2) is tight. Hence, g is minimal. □

We now show that the above theorem answers affirmatively the question asked by Cole, Dodis, and Roughgarden [3] regarding the enforceability of optimal congestion. We call a congestion g *optimal*, if g minimizes $\sum_e l_e(g) g_e$ over the set of all feasible congestions. Notice that $\sum_e l_e(g) g_e$ is equal to the average latency that the agents suffer in the network.

COROLLARY 3.1. *For every multicommodity network in the discrete setting, there are tolls that enforce an optimal congestion g^*.*

Proof. We call a congestion g *minimally feasible* if it is feasible, and for every congestion g' such that $g'_e \leq g_e$ for every $e \in E$ and $g'_e < g_e$ for at least one edge e, g' is not feasible. Take an optimal congestion g. We can turn this congestion into a minimally feasible congestion as follows: Let $g^{(0)} := g$. Consider the edges of the graph in an arbitrary order $e_1, e_2, \ldots$, and for each edge e_i, let $g^{(i)}$ be the congestion that is the same as $g^{(i-1)}$ everywhere except possibly on e_i, and $g_{e_i}^{(i)}$ is the minimum amount for which $P_{g^{(i)}}$ has a feasible solution. Let g^* be the final congestion. By this definition, g^* is minimally feasible. In other words, every feasible and therefore every optimal solution of P_{g^*} makes inequalities (3.2) tight for every edge e. Thus, g^* is minimal. Hence, by Theorem 3.1, g^* is enforceable. On the other hand, since latency functions are nondecreasing, $\sum_e l_e(g^*) g_e^* \leq \sum_e l_e(g) g_e$, and hence g^* is also optimal. □

Notice that the above proof works even if we define the optimal flow as a flow that minimizes an arbitrary nondecreasing function of congestion on the edges. This is formulated in the following corollary, whose proof is essentially the same as the proof of Corollary 3.1.

COROLLARY 3.2. *Let $w : \mathbb{R}^{E(G)} \mapsto \mathbb{R}$ be an arbitrary function that is nondecreasing in each of its arguments. Then there are tolls τ_e that enforce a congestion f that minimizes $w(f)$ over the set of all feasible congestions.*

The above corollary can be useful in certain applications. For example, by enforcing a flow f that minimizes $\max_i \min_{p \in \mathcal{P}_i} l_p(f)$, we can ensure that in the resulting Nash flow an emergency vehicle (in other words, an agent who only cares about the delay) can get from every *source*$_i$ to the corresponding *dest*$_i$ in the shortest possible time in the worst case.

An alternative (and arguably better in certain applications) way to define an optimal flow is to consider the weighted average of the latencies suffered by the agents, where the weight of an agent is equal to her sensitivity to latency. More precisely, we say that a flow f is *weighted optimal* if it minimizes

$\sum_i \alpha_i \sum_{p\in\mathcal{P}_i} l_p(f) f_p^i$ over the set of all feasible flows. The next corollary shows that minimal weighted flows are also enforceable. Notice that this statement says that not only the congestion induced by the flow, but also the flow itself is enforceable.

COROLLARY 3.3. *For every multicommodity network in the discrete setting, there are tolls that enforce a weighted optimal flow f^*.*

Proof. Among all weighted optimal flows, take a flow f^* such that $\sum_e f_e^*$ is the smallest. By Theorem 3.1 it is enough to show that this flow is minimal. Assume it is not. Therefore there is an optimal solution f for P_{f^*} for which inequality (3.2) is not tight for some edges. We have

$$\begin{aligned}\sum_i \alpha_i \sum_{p\in\mathcal{P}_i} l_p(f) f_p^i &\leq \sum_i \alpha_i \sum_{p\in\mathcal{P}_i} l_p(f^*) f_p^i \\ &\leq \sum_i \alpha_i \sum_{p\in\mathcal{P}_i} l_p(f^*) f^{*i}{}_p, \quad (3.8)\end{aligned}$$

where the first inequality follows from inequality (3.2) and the fact that latency functions are nondecreasing, and the second inequality is a consequence of the optimality of f for the linear program P_{f^*}. Equation (3.8) shows that f is also a weighted optimal flow. Also we know that $f_e \leq f_e^*$ for every edge e and $f_e < f_e^*$ for some edges. This contradicts with the assumption that f^* is the weighted optimal flow with the minimum value of $\sum_e f_e^*$. □

The argument in the proof of Corollary 3.1 can be used to show that *every* feasible congestion is enforceable in the following weaker sense: We say that a set of tolls τ *weakly enforces* a congestion g, if there is a congestion $g' \leq g$ that is enforced by τ.

COROLLARY 3.4. *Every feasible congestion g is weakly enforceable.*

Proof. As in the proof of the previous corollary, we start from the congestion g and consider the edges of the graph in an arbitrary order. For each edge in this order, we decrease the amount of congestion on that edge to the minimum amount for which the congestion is still feasible. Let g' denote the resulting congestion. Clearly, g' is minimally feasible, and therefore by Theorem 3.1 it is enforceable. Since $g' \leq g$, the corollary follows. □

It is also worth mentioning that if we allow negative tolls (i.e., if we can pay agents for using an edge), then every congestion is enforceable. This can be proved by changing inequality (3.2) in P_g to equality and using the argument in the proof of Theorem 3.1.

Polynomial time computation of tolls. The linear programs P_g and D_g give a polynomial-time algorithm to compute tolls that induce an optimal congestion (or in general, any enforceable congestion) in polynomial time. Although these linear programs have exponential size, they can be written as polynomial-size programs in the standard way: For P_g, we use variables f_e^i for every commodity i and edge e instead of f_p^i's, and write flow conservation constraint for every vertex and every commodity and the capacity constraint on every edge. Taking the dual of this program gives us a polynomial-size program equivalent to D_g, where tolls τ_e come from the dual variables corresponding to the capacity constraint in P_g.

After writing P_g and D_g as polynomial-size programs, we can solve them using an LP solver to compute optimal tolls and a corresponding Nash flow. Furthermore, by solving D_g once and computing the value of the objective function, we can add an inequality to this program so that the resulting set of inequalities give a complete characterization of the polytope of tolls that enforce g. This can be used to compute tolls that enforce g and are optimal with respect to another objective, for example, minimizing sum of tolls, or minimizing maximum toll.

Cole, Dodis, and Roughgarden [3] gave a different, although similar, linear program for computing tolls (In [3] this program is stated in the case of single-commodity networks, but it is easy to see that the same program works for multicommodity networks too). However, this program requires the knowledge of the flow pattern of different commodities in the Nash flow to be induced. This is a strong assumption, as there are many ways that a flow can be decomposed into paths, but perhaps only one of these decompositions corresponds to the set of paths used by users when the right set of tolls are imposed. Fleischer [6] gives an example to demonstrate that the Nash flow pattern may depend on α. Furthermore, as stated in [3], their linear program does not prove the existence of optimal tolls.

4 An exponential bound on the tolls

The following theorem gives a bound on the maximum value of tolls needed to enforce a given congestion. This bound is exponential in the number of edges of the graph, but it is important that it is independent of the number of commodities or types of agents. We will use this result in the next section in the proof of the existence of tolls in the continuous model. As we will see in Section 5, this bound also holds for more general congestion games.

We denote the maximum of α_i's by $\alpha_{\max}$. Also, let $l_{\max}$ denote $\max_{e \in E(G)} l_e(1)$.

THEOREM 4.1. *Let G be a multicommodity network, and g be an enforceable congestion in G. Then g is enforceable with tolls t satisfying $t_e \leq T$ for all $e \in E$, where T is a number that depends only on the number of edges in the graph, $l_{\max}$, and $\alpha_{\max}$, and not on the number of commodities.*

Proof. Consider a basic feasible solution (t, z) of the dual program D_g. This program has $K + m$ variables, where K is the number of commodities and m is the number of edges of G. Therefore, there should be a set of $K + m$ inequalities that are tight in (t, z), giving us $K + m$ equations with a unique solution of (t, z). Each z_i should be present in at least one of these tight inequalities, for otherwise the solution will not be unique. Therefore, we can use this equation to eliminate z_i from the set of our equations. After eliminating all z_i, we get m equations, each of the form $t_e = 0$ or of the form $\sum_{e \in p} t_e + \alpha_i l_p(g) = \sum_{e \in p'} t_e + \alpha_j l_{p'}(g)$. We can write these equations as a matrix equation $At = b$, where A is a matrix of $+1$'s and -1's, and b is a vector whose entries are of the form $\alpha_i l_p(g) - \alpha_j l_{p'}(g)$, and therefore are all at most $\alpha_{\max} m l_{\max}$. The collection of all $m \times m$ matrices with ± 1 entries is finite. Let S denote the maximum possible entry in the inverse of a matrix from this collection. Clearly, S is finite and only depends on m. Also, we have $t = A^{-1}b$, and therefore for every e, $t_e \leq m^2 S \alpha_{\max} l_{\max}$. This completes the proof of the theorem. □

5 Existence of optimal tolls in the continuous model

In this section we use the results of Sections 3 and 4 to show that in the continuous setting optimal tolls exist. The idea of the proof is to estimate continuous α_i's by a sequence of step functions. For each step function we can find the optimal tolls using Corollary 3.1. This is stated in the following lemma.

LEMMA 5.1. *Assume that for every i, the function α_i is a step function with a bounded number of steps. Then there are tolls $\{\tau_e\}$ that enforce an optimal congestion in this network.*

Proof. Let r_i denote the number of steps in the function α_i. Replace each commodity i with r_i commodities, each corresponding to one of the steps of α_i. Each of these commodities has a constant value of sensitivity to latency which is equal to the value of α_i in the corresponding step. Also, the demand for each of these commodities is equal to the length of the corresponding step in α_i. It is easy to see that the network constructed in this way is equivalent to the original network, in the sense that for any set of tolls, a Nash flow in the original network corresponds to a Nash flow in the constructed network. Thus, we can use Corollary 3.1 to find a set of tolls for this network, and therefore for the original network, that enforce an optimal congestion. □

The following lemma shows that no matter what α_i's are, we can represent a Nash flow concisely.

LEMMA 5.2. *For every network and every set of tolls in the continuous model, there is a Nash flow f such that for every commodity i and every path $p \in \mathcal{P}_i$, the set $\{a \in [0, d_i] : f^i(a) = p\}$ is a connected set.*

Proof Sketch. We show that for every two agents $a, b \in [0, d_i]$, if $a < b$, then the latency of the path $f^i(a)$ is greater than or equal to the latency of the path $f^i(b)$. This is true, since otherwise b has an incentive to switch to the path $f^i(a)$. Using this fact and Lebesgue-measurability of f^i, we can change f^i to get a flow that is still a Nash flow and also satisfies the condition of the lemma. □

THEOREM 5.1. *For every multicommodity network in the continuous model, there is a set of tolls that enforce an optimal congestion.*

Proof Sketch. For each commodity i, we estimate the function α_i by a sequence $\alpha_i^1, \alpha_i^2, \ldots$ of step functions.

Define a network G^k by replacing the function α_i by its k'th estimate α_i^k for every commodity i. By Lemma 5.1 for each k there is set of tolls τ^k that enforce an optimal congestion in G^k. Let $f^{(k)}$ denote the Nash flow in the network G^k with respect to tolls τ^k. We can assume that $f^{(k)}$'s satisfy the condition of Lemma 5.2, and therefore each of these flows can be represented by giving the end points of the intervals on which the flow is constant. This means that each $f^{(k)}$ can be given by a sequence of at most $|\mathcal{P}|$ real numbers in $[0, 1]$. Also, by Theorem 4.1 in the previous section, we can assume that all tolls in τ^k are bounded by a constant T, independent of k. Therefore, $(\tau^k, f^{(k)})$ belongs to a compact set. This means that there is a subsequence $k_1, k_2, \ldots$, such that $(\tau^k, f^{(k)})$ on this subsequence tends to some (τ, f). It is not hard to show that τ enforces the flow f in the original network. □

6 General Congestion Games

In the proof of Theorem 3.1 we did not use much of the structure of the network. In this section we show that similar results are true for a general class of congestion games. First, we discuss a simple setting, which is essentially the setting of general congestion games (originally defined by Rosenthal [10]) with infinitesimally small agents.

Consider a game which has N different kinds of users and M different resources. We want to toll resources so that we can enforce a certain usage of resources. Users have certain usage requirements and they are sensitive to both latencies and tolls. There is an infinite number of users of each kind, each having an infinitesimally small effect on the game. The i-th kind is described by the following parameters:

- total volume of the users, d_i.
- a latency sensitivity constant, α_i. This constant specifies the monetary value of one unit of latency for a user of type i.
- a collection $\mathcal{S}_i$ of subsets of the resources. Each set in $\mathcal{S}_i$ is a combination of resources that can satisfy a user of type i. If a user picks a set containing j, then we say that she is using the resource j. For example, in the multicommodity network game described in earlier sections the set of resources is the set of edges of the graph, and $\mathcal{S}_i$ is the set of all paths from *source*$_i$ to *dest*$_i$.

Usage of a resource is the total volume of users using that resource (i.e., picking sets containing the resource). Each resource j is characterized by its latency function $l_j : \mathbb{R}_+ \to \mathbb{R}_+$, which is a non-decreasing function of the total usage of j. A *usage vector* is a vector in $\mathbb{R}_+^M$ specifying the usage for every resource. A usage vector v is *feasible* if there exist a way to satisfy every user without using any resource j more than v_j. A usage vector is *minimally feasible* if decreasing any component by any positive amount makes it infeasible.

Our objective is to set tolls on the resources in order to induce a given usage vector. Let τ_j denote the toll on resource j. Users of the i'th kind seek to pick a set $S \in \mathcal{S}_i$ that minimizes $\alpha_i \sum_{j \in S} l_j(v_j) + \sum_{j \in S} \tau_j$, where v is the current usage vector. The Nash equilibrium of this game is defined in the same way as in Section 2. We say that a usage vector v is *enforceable*, if there are tolls τ such that v is the usage vector induced by a Nash equilibrium in the game resulting from the tolls τ.

THEOREM 6.1. *Suppose $v \in \mathbb{R}_+^M$ is a minimally feasible usage vector. Then there exist nonnegative tolls that enforce v.*

Proof. Let x_{iS} be the volume of users of the i-th kind that have chosen the set S. Let l_{iS} denote the quantity $\alpha_i \sum_{j \in S} l_j(v_j)$. Consider the following linear program with x_{iS} as variables.

$$\begin{aligned} \text{minimize} \quad & \sum_i \sum_{S \in \mathcal{S}_i} l_{iS} x_{iS} \qquad (6.9) \\ \text{subject to} \quad & \forall i : \sum_{S \in \mathcal{S}_i} x_{iS} \geq d_i \\ & \forall j : \sum_i \sum_{S \in \mathcal{S}_i |\ j \in S} x_{iS} \leq v_j \\ & \forall i, S \in \mathcal{S}_i : \ x_{iS} \geq 0 \end{aligned}$$

The first set of constraints tells us that the all the demands are met. The second set of constraints makes sure that we do not exceed the usage given by v. Minimality of v implies that these constraints are tight in any feasible solution. This means that every feasible

solution of the above program represents a situation in the game where v is the usage vector and hence l_{iS} is the total monetary value of the latency of resources in S for a user of type i.

The dual of the above program will give us the tolls to enforce v. The dual can be written as follows, with τ_j and z_i as the dual variables corresponding to the jth resource and the ith type of users, respectively.

$$\begin{aligned} \text{maximize} \quad & \sum_i d_i z_i - \sum_j v_j \tau_j \qquad (6.10) \\ \text{subject to} \quad & \forall i, S \in \mathcal{S}_i : \; z_i \leq l_{iS} + \sum_{j \in S} \tau_j \\ & \forall i : \; z_i \geq 0 \\ & \forall j : \; \tau_j \geq 0 \end{aligned}$$

We interpret the dual variable τ_j as the toll on resource j. The right-hand side of the first set of constraints is the total cost for users of type i to choose S. Since z_i appears with positive coefficient in the dual objective function, at least one constraint for z_i must be tight. This implies that z_i is actually the cheapest cost for satisfying a user of type i. By complementary slackness condition, for any optimal primal solution x and optimal dual solution (g, τ), whenever x_{iS} is positive the corresponding constraint in the dual must be tight. This means that whenever users of kind i are choosing S to satisfy themselves their cost of doing so is z_i, which as argued is the cheapest cost. Since each user is infinitesimally small, changing the strategy for any user does not change the latencies. Hence choosing the cheapest S is a best response strategy for every infinitesimally small user. This implies that x is a Nash equilibrium for the tolls τ_j, inducing the usage vector v. □

In fact, it is not difficult to argue that whenever we have a Nash equilibrium satisfying the primal LP (6.9), the tolls will satisfy the dual LP (6.11) and they will form a primal-dual optimal pair.

The definition of *weakly enforcing* and the proof of the following corollary is similar to the ones in Section 3.

COROLLARY 6.1. *Suppose $v \in \mathbb{R}_+^M$ is a feasible usage vector. Then v can be weakly enforced via tolls.*

It can be easily observed that the proof of Theorem 6.1 did not use many of the assumptions of the model. In the following, we describe three increasingly more general models in which our results still hold. As mentioned below, these generalizations are useful in certain practical applications.

1. Different types of users may experience different latencies for a resource with the same congestion. In natural settings, users may intend to use a resource differently. For example, on the Internet, UDP traffic and TCP traffic might be affected differently by congestion, or in a road, a motorbike and a big truck experience different latencies in the same traffic. So we can assume that latency is a function which may assign different latencies to different kinds of users. Formally $l_j : \mathbb{R}_+ \to \mathbb{R}_+^N$. Theorem 6.1 and Corollary 6.1 hold for this generalization. In fact, now we can pull α_i into l_{ji}, where l_{ji} is the latency function of j for i. So we do not need α_i's; instead, latency functions themselves converts the latencies into monetary values.

2. Latency functions may be nonseparable functions of the usage of resources. For example, in wireless networks, because of interference, latency on a link is not only a function of the traffic on the link but also a function of the traffic on the neighboring links. In road networks, congestion on a road depends on traffic on adjacent roads. Our model permits latencies to be a general function of the usage of all the resources. Formally, $l_j : \mathbb{R}_+^M \to \mathbb{R}_+^N$. Theorem 6.1 and Corollary 6.1 hold for this generalization.

3. We assumed that the latency of a set S is the sum of latencies of the resources in it. This assumption is also not necessary. Our results hold even if we allow each type of user to have an arbitrary function $l_i : \mathcal{S}_i \times \mathbb{R}_+^M \mapsto \mathbb{R}_+$ that for every set $S \in \mathcal{S}_i$ and every usage vector $v \in \mathbb{R}_+^M$, gives the monetary value of the latency experienced by i, if she picks S and the current usage vector is v. Furthermore, we could allow $\mathcal{S}_i$'s to be collections of *fractional* sets of resources.

Bounds on Generalized Congestion Games. The exponential bound on tolls given in Theorem 4.1 also holds for generalized congestion games. The proof generalizes easily to this setting. Therefore, tolls exist to enforce usage patterns of generalized congestion games also in the continuous setting analogous to the continuous model for network games described in Section 5.

Furthermore, as the following example shows, the

bound in Theorem 4.1 cannot be improved significantly in general congestion games.

EXAMPLE 1. *Consider an abstract congestion game consisting of k types of agents, and $2(k+1)$ resources called $a_0, \ldots, a_k, b_0, \ldots, b_k$. All agents have the same sensitivity to latency. Agents of the i'th type have strategy set $\mathcal{S}_i = \{\{a_{i-1}, b_{i-1}\}, \{a_i\}, \{b_i\}\}$. The latency of the resources a_0 and b_0 is always one, while the latency of all other resources is always zero. The congestion g that we would like to enforce is the following: the congestion of a_0, b_0, a_k, and b_k are 1/3, and the congestion of all other resources is 2/3. It is easy to see that in order to enforce this congestion, we must have $\tau_{a_i} = \tau_{b_i} = \tau_{a_{i-1}} + \tau_{b_{i-1}}$ for every $i > 1$, and $\tau_{a_1} = \tau_{b_1} = 2$. Therefore, we need tolls exponential in the number of commodities in order to enforce g in this game.*

References

[1] M. Beckman, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation.* Yale University Press, 1956.

[2] Vasek Chvatal. *Linear Programming*. W H Freeman & Co., 1983.

[3] R. Cole, Y. Dodis, and T. Roughgarden. Pricing network edges for heterogeneous selfish users. In *STOC 2003*, pages 521–530, 2003.

[4] S. C. Dafermos. Toll patterns for multiclass-user transportation networks. *Transportation Sci.*, 7:211–223, 1973.

[5] R. B. Dial. Network-optimized road pricing: Part i: A parable and a model. *Operations Research*, 47(1):54–64, 1999.

[6] L. Fleischer. Linear taxes suffice. In *Proc. of ICALP*, 2004. To appear.

[7] G. Karakostas and S. Kolliopoulos. Edge pricing of multicommodity networks for heterogeneous selfish users. In this proceedings.

[8] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. *Lecture Notes in Computer Science*, 1563:404–413, 1999.

[9] A. C. Pigou. *The Economics of Welfare*. Macmillan, 1920.

[10] R.W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *Int. J. Game Theory*, 2:65–67, 1973.

[11] T. Roughgarden. *Selfish Routing*. PhD thesis, Cornell University, 2002.

[12] Tim Roughgarden and Eva Tardos. How bad is selfish routing? In *IEEE Symposium on Foundations of Computer Science*, pages 93–102, 2000.

[13] M. J. Smith. The marginal cost taxation of a transportation network. *Trans. Res. Ser. B*, 13:237–242, 1979.

[14] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proc. Institute of Civil Engineers, Pt. II*, volume 1, pages 325–378. 1952.