The Effectiveness of Stackelberg Strategies and Tolls for Network Congestion Games

Chaitanya Swamy*

Abstract

It is well known that in a network with arbitrary (convex) latency functions that are a function of edge traffic, the worst-case ratio, over all inputs, of the system delay caused due to selfish behavior versus the system delay of the optimal centralized solution may be *unbounded* even if the system consists of only two parallel links. This ratio is called the *price of anarchy* (PoA). In this paper, we investigate ways by which one can reduce the performance degradation due to selfish behavior. We investigate two primary methods (a) *Stackelberg routing strategies*, where a central authority, e.g., network manager, controls a fixed fraction of the flow, and can route this flow in any desired way so as to influence the flow of selfish users; and (b) *network tolls*, where tolls are imposed on the edges to modify the latencies of the edges, and thereby influence the induced Nash equilibrium. We obtain results demonstrating the effectiveness of both Stackelberg strategies and tolls in controlling the price of anarchy.

For Stackelberg strategies, we obtain the first results for nonatomic routing in graphs more general than parallel-link graphs, and strengthen existing results for parallel-link graphs. (i) In series-parallel graphs, we show that Stackelberg routing reduces the PoA to a constant (depending on the fraction of flow controlled). (ii) For general graphs, we obtain latency-class specific bounds on the PoA with Stackelberg routing, which give a continuous trade-off between the fraction of flow controlled and the price of anarchy. (iii) In parallel-link graphs, we show that for *any* given class $\mathcal{L}$ of latency functions, Stackelberg routing reduces the PoA to at most $\alpha + (1-\alpha)\cdot\rho(\mathcal{L})$, where α is the fraction of flow controlled and $\rho(\mathcal{L})$ is the PoA of class $\mathcal{L}$ (when $\alpha = 0$).

For network tolls, motivated by the known strong results for nonatomic games, we consider the more general setting of *atomic splittable* routing games. We show that tolls inducing an optimal flow always exist, even for general asymmetric games with heterogeneous users, and can be computed efficiently by solving a *convex program*. Furthermore, we give a complete characterization of flows that can be induced via tolls. These are the first results on the effectiveness of tolls for atomic splittable games.

*cswamy@math.uwaterloo.ca. Dept. of Combinatorics and Optimization, Univ. Waterloo, Waterloo, ON N2L 3G1. Research supported partially by NSERC grant 32760-06. Work done while the author was a postdoctoral scholar at Caltech, CA 91125.

1 Introduction

We consider the problem of optimizing the performance of a network in the presence of selfish, noncooperative, uncoordinated traffic (users). A popular way of modeling such selfish behavior is by means of a noncooperative game played between the selfish agents, and by viewing the equilibria, typically Nash equilibria, of the game as outcomes of selfish behavior. It is well known that selfish behavior and the lack of coordination often leads to a degradation in performance quality, and in recent years there has been considerable interest and progress in quantifying the performance loss in various settings in terms of the *price of anarchy* of the corresponding induced game; see e.g., [18, 26, 23]. The price of anarchy (abbreviated PoA) of a game is the ratio between the cost of the Nash equilibrium solution (i.e., the outcome of selfish behavior) and the socially optimum solution. In this paper, we investigate ways of reducing the price of anarchy in network congestion games (which we also refer to as network routing games).

A *network routing game* is determined by a directed network $G = (V, E)$ with nonnegative, monotone nondecreasing latency or delay functions $\ell_e(x)$ on the edges, a source and sink $s, t \in V$, and given volume of flow that needs to be routed from s to t, which is split into many users. To analyze selfish behavior, we focus on the concept of a *Nash equilibrium*, which is a combination of users' strategies where no individual user can reduce her cost by changing her strategy (when the other users' strategies stay fixed). In the context of congestion games, the cost of a user is measured in terms of the congestion experienced by it, and the system cost is the aggregate of the user costs.

In the *nonatomic* routing game, there are an infinite number of users, each controlling an infinitesimal amount of flow (traffic). The strategy space of each user

is the collection of directed paths in G from s to t. The cost associated with selecting a path P is the delay of P, which is the sum of the delays of the edges in P. In the *atomic, splittable* routing game, there are a finite number of users. Each user i controls a certain amount of flow $D_i > 0$, and her strategy consists of choosing a feasible fractional (i.e., splittable) flow from s to t of volume D_i. The cost associated with flow f_i is the total delay incurred in routing along f_i, which is the sum over all edges e of the flow $f_{i,e}$ times the delay of edge e.

The system cost of a flow f is the total latency or delay experienced by all the users, defined by $C(f) = \sum_{e \in E} f_e \ell_e(f_e)$. The price of anarchy of a game is the worst possible ratio of $C(f^{NE})/C(o)$, where f^{NE} if a flow at a Nash equilibrium, and o is a system-optimal flow, i.e., it minimizes $C(f)$. It is known that with arbitrary (convex) latency functions, the price of anarchy may be *unbounded* even if G consists of only two parallel s-t links. Thus it becomes important to consider ways by which one can reduce the performance degradation due to selfish behavior and bound the price of anarchy. We investigate two primary methods for reducing the price of anarchy.

(a) Stackelberg routing strategies. A central authority, such as the network manager or router, controls a fixed fraction of the flow and routes this flow in any desired way, then the remaining traffic routes itself selfishly. The goal is to find a Stackelberg strategy, that is, a routing of the centrally controlled traffic, that minimizes the cost of the resulting flow (which includes the Stackelberg flow).

As mentioned in [24], an appealing aspect of Stackelberg strategies is their simplicity: unlike other prominent approaches such as, for example, algorithmic mechanism design or pricing policies, no communication is required between the system and the selfish users, and no notion of money is needed. Moreover, despite this simplicity, various situations that arise in the design of communication networks can be modeled by the above setting. For example, Korilis et al. [17] motivate this problem from its applicability in *virtual private network* design, where the system must allocate bandwidth on preassigned virtual paths so as to handle ongoing and future traffic; the bandwidth assigned on the virtual paths can be viewed as centrally controlled traffic, while the individual users of the network (who may come and go) may be treated as selfish traffic.

(b) Network tolls. A central authority imposes tolls (or taxes) on the network edges. The net latency experienced by user i on an edge e with toll τ_e is given by $\tilde{\ell}_{i,e}(x) = \gamma_i \ell_e(x) + \tau_e$, where the quantity $\gamma_i \geq 0$ indicates the sensitivity of user i to delay. (In the nonatomic game, one can consider the setting where the γ_is form a continuum.) The goal is to compute tolls on the edges such that the resulting Nash flow (with respect to the modified latency functions $\tilde{\ell}_{e,i}(x)$) has low cost (compared to the social optimum).

Network tolls are a classical means of congestion control. They were proposed way back by Pigou [22] in 1920, and various results have demonstrated the effectiveness of tolls for nonatomic routing games [2, 10, 6, 11, 15, 29] by showing that one can always induce an optimal flow via tolls (thus reducing the price of anarchy to 1). Given these positive results, it is natural to investigate whether tolls can help reduce the price of anarchy in more general settings, such as atomic splittable games. Moreover, tolls for atomic traffic can be important in areas such as routing of Internet traffic, where there could be relatively few domains (Google, Yahoo, etc.) that generate a large fraction of the traffic.

Our results. We obtain results demonstrating the effectiveness of both Stackelberg strategies and tolls in controlling the price of anarchy. For Stackelberg strategies, we obtain the first results for nonatomic routing in graphs that are more general than (s,t) parallel-link graphs (Section 3), and we strengthen existing results for parallel-link graphs. We show that in series-parallel graphs, for any $\alpha \in [0,1]$, and *arbitrary latency functions*, one can efficiently compute a Stackelberg strategy controlling α-fraction of flow that reduces the price of anarchy to $\frac{1}{\alpha} + 1$. This underlines the power of Stackelberg strategies in such settings. (Note that without any central control, the PoA with arbitrary latency functions is unbounded even for parallel-link graphs.) Previously, such results were known only for parallel-link graphs, due to Roughgarden [24].

For general graphs, we obtain latency-class specific bounds on the price of anarchy with Stackelberg routing. Our bounds give a continuous trade-off between the fraction of flow controlled and the price of anarchy. The trade-off function is a decreasing function of α and varies between the worst-case PoA when $\alpha = 0$, and 1 when $\alpha = 1$. For linear latencies, we describe an explicit trade-off curve showing how the PoA varies with α.

We obtain stronger results for parallel-link graphs. We show that for *any* given class of latency functions, the price of anarchy (of that class) can always be improved by Stackelberg routing. More precisely, let $\rho(\mathcal{L})$ denote the PoA of a class $\mathcal{L}$ of latency functions (without controlling any flow). We show that by controlling an α-fraction of flow one can reduce the price of anarchy to $\alpha + (1-\alpha) \cdot \rho(\mathcal{L})$. This yields a nice, smooth trade-off, and shows that Stackelberg strategies are especially effective in parallel-link graphs.

In Section 4, we consider the use of tolls to reduce the PoA. As mentioned earlier, various results have

affirmed the usefulness of tolls in nonatomic routing games, and the power of tolls in this setting is relatively well understood. Motivated by these positive results, we therefore investigate the more general setting of atomic splittable games. We show that *there always exist tolls that induce an optimal flow* as a Nash flow in the routing game with tolls. This is true even for *asymmetric* atomic splittable routing games, where different users may have different source-sink pairs. We call such tolls *optimal tolls* and show that they can be computed efficiently by solving a *convex program*. Furthermore, we give a complete characterization of flows that can be induced (as some Nash equilibrium) via tolls (Nash flows in atomic splittable instances are not known to be unique). These results also extend to the setting of general atomic splittable congestion games. To the best of our knowledge, these are the *first* results on the effectiveness of tolls for atomic splittable traffic. Previously, except in some restricted cases, even the *existence* of optimal tolls was not known.

Related work. The nonatomic network routing game was originally proposed in the seminal paper of Wardrop [28] as a way of thinking about road traffic. The notion of equilibrium introduced in [28] coincides with the Nash equilibrium concept in the nonatomic routing game.

Koutsoupias and Papadimitriou [18] introduced the idea of evaluating the performance degradation of a selfish game by considering its worst-case possible Nash Equilibrium. Using the terminology defined above, they considered atomic unsplittable routing games on parallel-link graphs with linear latency functions under the *maximum* user cost objective (as opposed to the sum of user costs). Papadimitriou [21] later coined the term "price of anarchy" to describe the ratio between the system-performance of the worst-case Nash equilibrium equilibrium, and the system-performance of the best centrally-enforced solution. Price of anarchy results for nonatomic routing games were first obtained by Roughgarden and Tardos [26], who obtained tight bounds for linear latency functions. Subsequently, Roughgarden [25] (see also [23]) gave tight bounds for many classes of latency functions.

Much less is known about atomic routing games. For atomic splittable routing, very recently, Cominetti, Correa, and Stier-Moses [7] presented various bounds on the price of anarchy. They also showed that in a symmetric atomic game, that is, when all users control the same amount of flow and share the same source and sink, the cost of a Nash equilibrium is at most the cost of the Nash equilibrium in the corresponding nonatomic game. Hayrapetyan, Tardos and Wexler [13] proved such a result for parallel-link graphs, even when the users control different amounts of flow. Thus, in these settings, the PoA of the atomic splittable game is at most the PoA of the corresponding nonatomic game. Awerbuch, Azar and Epstein [1], and Christodoulou and Koutsoupias [4] provide PoA bounds for atomic, *unsplittable* routing games, and show that these can be worse than the PoA bounds in the nonatomic and atomic splittable setting.

Korilis, Lazar and Orda [17] first considered the use of Stackelberg strategies as a means of improving system performance. The considered atomic, *unsplittable* routing games in graphs with s-t parallel links and latency functions of the form $\ell_e(x) = \frac{1}{u_e - x}$, and obtained necessary and sufficient conditions (e.g., on the amount of flow that is centrally controlled) for the existence of a Stackelberg strategy that induces the optimal flow. Subsequently, Roughgarden [24] considered the question of improving price of anarchy bounds via Stackelberg strategies. The work of [24] also considered parallel-link graphs, and showed that for arbitrary latency functions, and any $\alpha \in [0, 1]$, one can compute a Stackelberg strategy that reduces the price of anarchy to at most $\frac{1}{\alpha}$. Given that the price of anarchy is unbounded in such graphs, the benefit of Stackelberg strategies is particularly striking in this setting. Kumar and Marathe [19] considered the algorithmic problem of computing the best Stackelberg strategy and gave a PTAS for this problem on parallel-link graphs.

Recently, there has been a flurry of work related to Stackelberg routing. Kaporis and Spirakis [14] show that one can efficiently compute the smallest α such that an optimal flow can be induced by controlling α-fraction of flow. Sharma and Williamson [27] consider the related question of finding the smallest α for which the cost of the Nash flow can be improved, for parallel-link instances with linear latencies. Independent of our work, Karakostas and Kolliopoulos [16] have obtained PoA bounds for multicommodity nonatomic games on general graphs with linear latency functions, using the LLF and Scale Stackelberg strategies proposed in [24] (which we also use and analyze). Correa and Stier-Moses [8] have also independently obtained some results on Stackelberg routing.

The use of tolls as a means for congestion control was proposed way back by Pigou [22] in 1920. For nonatomic users with identical trade-offs for delay versus toll, Pigou [22], and more formally Beckman et al. [2], showed that marginal cost tolls induce the optimal flow. Much recent work has been directed toward the setting of *heterogeneous* users, where different users may have distinct trade-offs for delay versus toll, In this setting, it turns out that one can exploit linear-programming duality to obtain tolls that

induce an optimal flow, even for multicommodity traffic [10, 6, 5, 11, 15, 29].

Questions regarding the efficacy of tolls for atomic users have only been considered very recently. Caragiannis et al. [3] considered the atomic *unsplittable* case in parallel-link graphs with linear latency functions. In contrast to our results about the existence of optimal tolls, they showed that optimal tolls need not always exist in the unsplittable case, and gave various lower and upper bounds on the PoA achievable through tolls. We are not aware of any previous work on tolls for atomic splittable traffic. The only known results are those that follow from the above-mentioned results of Hayrapetyan et al. [13] and Cominetti et al. [7], which show that cost of an atomic Nash flow is bounded by that of the nonatomic Nash flow; their results imply that optimal tolls exist for parallel-link graphs, and symmetric users on general graphs.

2 Preliminaries

Let $G = (V, E)$ be the underlying directed graph endowed with nonnegative, monotone nondecreasing latency functions $\{\ell_e(x)\}$ on the edges, and a source and sink $s, t \in V$. We will assume that each $\ell_e(x)$ is continuous. Without loss of generality, we can scale the problem so that the total volume of traffic to be routed from s to t is 1 (i.e., if D is the total flow volume, we can equivalently work with the latency functions $\tilde{\ell}_e(x) = \ell_e(Dx)$). We use $C(f)$ to denote the cost of a flow f, which is the total latency experienced by the flow, defined by $C(f) = \sum_e f_e \ell_e(f_e)$. Let $o = \{o_e\}$ denote an optimal flow, and $OPT = C(o)$ be the cost of this flow. We use (G, ℓ) to denote an instance (with total flow volume 1) in the corresponding nonatomic routing game, and $(G, k, \{D_i\}, \ell)$ to denote an atomic (splittable) routing game with k users, where user i controls $D_i \geq 0$ units of flow (so $\sum_i D_i = 1$). Let $\mathcal{P} = \mathcal{P}_{st}$ be the set of all simple s-t paths. The terms "latency", and "cost" of an edge e refer to the quantities $\ell_e(f_e)$, and $f_e \ell_e(f_e)$ respectively. For a path P, we will sometimes use $\ell_P(f)$ and $c_P(f)$ to denote respectively the total latency and cost of the edges on path P under the flow f. Given a flow f and path P, we use f_P as a shorthand for $\min_{e \in P} f_e$.

A Nash equilibrium is a combination of users' strategies, such that no single user can profit by deviating from his strategy (when the other users' strategies stay fixed). A flow f^{NE} is a *Nash flow* for the *nonatomic routing game* (G, ℓ), if it is feasible, and for every $P, P' \in \mathcal{P}$ with $f^{NE}_P > 0$, $\ell_P(f^{NE}) \leq \ell_{P'}(f^{NE})$. Thus, in a Nash flow f^{NE}, all flow paths have the same common latency L, that is, if $f^{NE}_P > 0$, then $\ell_P(f^{NE}) = L = \min_{P \in \mathcal{P}} \ell_P(f^{NE})$. A flow profile $(f_1, \ldots, f_k)$ is a Nash flow for the *atomic splittable routing game* $(G, k, \{D_i\}, \ell)$, if for each user $i = 1, \ldots, k$, f_i is a feasible flow for i, and f_i minimizes the quantity $\sum_e f_{i,e} \ell_e(f_e)$ where f is the flow $\sum_j f_j$. Let $\ell'(x)$ denote the derivative of $\ell(.)$ at x. The following basic facts about optimal and Nash flows will be useful (details and proofs may be found, for example, in [23]).

(a) If the function $x\ell_e(x)$ is convex for each $e \in E$, then an optimal flow can be computed in polynomial time (up to an arbitrarily small additive error).

(b) Every nonatomic and atomic routing game admits an acyclic Nash flow.

(c) Optimal flows, and Nash flows for nonatomic instances, are essentially unique: if o and $\tilde{o}$ are two optimal flows, and f^{NE}, $\tilde{f}^{NE}$ are two nonatomic Nash flows, then $\ell_e(o_e) = \ell_e(\tilde{o}_e)$, and $\ell_e(f^{NE}_e) = \ell_e(\tilde{f}^{NE}_e)$ for every edge e.

(d) A flow $(f_1, \ldots, f_k)$ is a Nash equilibrium for an atomic instance $(G, k, \{D_i\}, \ell)$ iff for each user i, and any paths $P, P' \in \mathcal{P}$ with $f_{i,P} > 0$, we have $\sum_{e \in P}(\ell_e(f_e) + f_{i,e}\ell'_e(f_e)) \leq \sum_{e \in P'}(\ell_e(f_e) + f_{i,e}\ell'_e(f_e))$, where $f = \sum_j f_j$.

For a latency function $\ell(x)$, define $\rho(\ell) = \sup_{0 \leq y \leq x} \frac{x\ell(x)}{y\ell(y)+(x-y)\ell(x)}$. Let $\rho(\mathcal{L})$ denote the worst-case price of anarchy over all nonatomic instances (G, ℓ), where each function $\ell_e(x)$ lies in class $\mathcal{L}$. Roughgarden [25] showed that $\rho(\mathcal{L})$ is precisely $\sup_{\ell \in L} \rho(\ell)$. Correa, Schulz and Stier-Moses [9] gave an alternative proof of this result, using the quantity $\beta(\ell) = \sup_{0 \leq y \leq x} \frac{y(\ell(x)-\ell(y))}{x\ell(x)}$. Note that $\beta(\ell) = 1 - \frac{1}{\rho(\ell)}$, and hence $\beta(\mathcal{L}) = \sup_{\ell \in \mathcal{L}} \beta(\ell) = 1 - \frac{1}{\rho(\mathcal{L})}$.

3 Stackelberg strategies for nonatomic routing games

We now consider Stackelberg routing for nonatomic routing games and obtain bounds on the price of anarchy. Throughout, we use α to denote the fraction of traffic controlled by the central authority. We first consider series-parallel graphs with arbitrary latency functions, in Section 3.1, and prove that the cost of the flow induced by our Stackelberg strategy is at most a constant (depending on α) times the optimum. In Section 3.2, we focus on obtaining PoA bounds that depend on the latency functions in the instance. We first show that for any "reasonable" Stackelberg strategy, the price of anarchy never increases (as compared to the PoA without any flow-control) due to Stackelberg routing. Although this is a highly intuitive result, one should note that there are various examples, most notably the Braess paradox, where measures taken in the

interest of improving overall performance actually end up hurting the system performance, due to the presence of selfish users. We use this to obtain bounds for any given latency class that quantify the trade-off between the fraction of flow controlled and the PoA. For linear latency functions, we obtain an explicit function describing the variation of the PoA with the fraction of flow controlled. In Section 3.3, we consider parallel-link networks and present various improved results. We show that for *any* class of latency functions $\mathcal{L}$ with PoA $\rho(\mathcal{L})$, by controlling α-fraction of flow, one can reduce the PoA to $r(\rho(\mathcal{L}), \alpha) = \alpha + (1-\alpha)\rho(\mathcal{L})$, thus obtaining a nice, smooth trade-off curve. In Section 3.4, we consider some extensions. We argue that some of the latency-class specific bounds of Section 3.2 (those obtained for the Scale strategy) also extend to multicommodity nonatomic games. Finally, we show that the combination of Stackelberg strategies and tolls can yield benefits that are not achievable by either in isolation. We observe that for any instance (G, ℓ), by using our Stackelberg strategy *in conjunction* with tolls, we can enforce an optimal flow where the maximum toll paid by a user along a flow path is bounded (in terms of OPT).

Throughout, we use g to denote the Stackelberg flow, and h to denote the induced Nash flow, that is, h is the Nash flow with respect to the modified latency functions $\tilde{\ell}_e(x) = \ell_e(g_e + x)$. Since the total flow volume is 1 (due to scaling), g and h are s-t flows of volume α and $1-\alpha$ respectively. We use the terms Stackelberg flow and Stackelberg strategy interchangeably. We will assume that $x_e\ell_e(x)$ is convex for each e, so that an optimal flow can be computed in polynomial time.

We use the following two Stackelberg strategies to obtain our results:

(a) **The Largest-Latency-First (LLF) strategy.** This consists of the following: compute an optimal flow o, and repeatedly saturate the paths used by o starting from the largest-latency path until we have routed α units of flow. More precisely, we set $g_e = 0$ for all edges e initially; while α is positive, we repeatedly find a path P such that $\ell_P(o) = \max_{P:(o-g)_P>0} \ell_P(o)$, and set $g_e \leftarrow g_e + \min(\alpha, (o-g)_P)$ for all $e \in P$, and $\alpha = \max(0, \alpha - (o-g)_P)$. Since o is an acyclic flow, the flow g can be computed in polynomial time from the flow o. Clearly we have $g_e \le o_e$ for all edges e.

This is a generalization of the Largest-Latency-First strategy given in [24] for parallel-link graphs, therefore we adopt the same terminology.

(b) **The Scale strategy.** In this strategy, we set $g_e = \alpha \cdot o_e$ on all edges, that is, we simply scale the optimal flow. This strategy was also mentioned in [24] in the context of parallel-link graphs, but no PoA bounds were obtained using this strategy. We use this strategy in Section 3.2, to obtain latency-class specific PoA bounds for general graphs.

3.1 Series-parallel graphs We show that for directed series-parallel (sepa) graphs with end points (or terminals) s and t, the PoA under the LLF strategy is at most $\frac{1}{\alpha} + 1$.

DEFINITION 3.1. *Directed* series-parallel graphs *with end points s and t are defined inductively as follows. A basic sepa graph is a directed edge (s,t). Given two sepa graphs G_1 and G_2 with end points s_1, t_1, and s_2, t_2 respectively, one can create a new sepa graph G as follows. In a* series combination*, terminal s_2 is identified with terminal t_1 to create graph G with terminals $s = s_1$, $t = t_2$; in a* parallel combination*, terminal s_1 is identified with terminal s_2, and terminal t_1 is identified with t_2, to obtain the new terminals $s = s_1 = s_2$ and $t = t_1 = t_2$. In both cases, $E_G = E_{G_1} \cup E_{G_2}$.*

We will use the following properties of series-parallel graphs, which are easily proved by induction on the series-parallel structure of the graph.

CLAIM 3.1. *Consider a sepa graph with end points s and t.*

1. *Let f, f' be two s-t flows routing D, D' units of flow respectively, with $D \ge D'$, $D > 0$. Then, there is some s-t path P such that $f_P > 0$ and $f'_e \le f_e$ for every $e \in P$.*

2. *Let P be an s-t path, f be an s-t flow, and $e_1, \ldots, e_k$ be the subset of edges of P for which $f_e > 0$. Then there is a path P' containing $e_1, \ldots, e_k$ with $f_{P'} > 0$.*

We will use part (i) of Claim 3.1 (taking $f = o - g$ and $f' = h$) to obtain a sharp bound on the latency of the Nash flow h induced by LLF. Part (ii) of the claim will allow us to bound the increase in latency of a Stackelberg flow-path by at most the Nash latency. This will yield a constant PoA bound for sepa graphs.

THEOREM 3.1. *In series-parallel graphs, the LLF strategy induces a flow of cost at most $\left(\frac{1}{\alpha} + 1\right) \cdot OPT$.*

Proof. We first bound the cost of the Nash flow h. Applying part (i) of Claim 3.1 with $f = o - g$ and $f' = h$, yields a path Q with $(o-g)_Q > 0$, and $g_e + h_e \le o_e$ for all $e \in Q$. The Nash latency is at most the latency of path Q, which is at most L_g.

Hence, $\sum_e h_e\ell_e(g_e+h_e) \le (1-\alpha)\cdot L_g$. Now we bound $\sum_e g_e\ell_e(g_e+h_e)$. Consider any path P with $g_P > 0$. Applying part (ii) of Claim 3.1 with path P and the flow h, shows that there exists a path P' with $h_{P'} > 0$ that contains all the edges of P with $h_e > 0$. The combined latency of all these edges is at most $\ell_{P'}(h+g)$, which is at most L_g by the above bound on the Nash latency. Thus, the total latency of path P is at most $\sum_{e\in P:h_e=0} \ell_e(g_e) + L_g$, implying that the cost of g is at most $\alpha\cdot L_g + OPT$. So the total cost of $g+h$ is at most $L_g + OPT \le \left(\frac{1}{\alpha}+1\right)\cdot OPT$. ∎

3.2 PoA bounds depending on the class of latency functions We now obtain bounds on the PoA that depend on the latency functions in the instance. Let (G,ℓ) be a nonatomic instance, where each $\ell_e(x)$ lies in some class $\mathcal{L}$ of latency functions. Recall that the PoA of class $\mathcal{L}$ is given by $\rho(\mathcal{L}) = \sup_{\ell\in\mathcal{L}} \rho(\ell)$, where $\rho(\ell) = \sup_{0\le y\le x} \frac{x\ell(x)}{y\ell(y)+(x-y)\ell(x)}$. Recall also that $\beta(\ell) = \sup_{0\le y\le x} \frac{y(\ell(x)-\ell(y))}{x\ell(x)} = 1 - \frac{1}{\rho(\ell)}$, and $\beta(\mathcal{L}) = \sup_{\ell\in\mathcal{L}} \beta(\ell) = 1 - \frac{1}{\rho(\mathcal{L})}$. We drop the argument $\mathcal{L}$ below and use β and ρ to denote $\beta(\mathcal{L})$ and $\rho(\mathcal{L})$ respectively. In this section, we will also consider the Stackelberg strategy Scale defined earlier.

We first prove in Lemma 3.1 that any Stackelberg flow with $g_e \le o_e$ on all edges induces a flow of cost at most $\rho\cdot OPT$, where the cost of the Nash flow is at most $\rho\cdot\sum_e(o_e-g_e)\ell_e(o_e)$. We then use these bounds to prove bounds $r_{\mathsf{LLF}}(\rho;\alpha)$ and $r_{\mathsf{Sc}}(\rho;\alpha)$ on the price of anarchy under the LLF and Scale strategies respectively, which vary between ρ and 1 as α goes from 0 to 1. We simplify the expression $r_{\mathsf{Sc}}(\rho;\alpha)$ (which yields a stronger bound) for polynomial latency functions, and obtain a closed-form expression for linear latencies. Independently, Correa and Stier-Moses [8] have also obtained Lemma 3.1.

LEMMA 3.1. *Consider a Stackelberg flow satisfying $g_e \le o_e$ for all e. Then, (i) $C(g+h) = \sum_e(g_e+h_e)\ell_e(g_e+h_e) \le \rho\cdot OPT$, and (ii) $\sum_e h_e\ell_e(g_e+h_e) \le \rho\sum_e(o_e-g_e)\ell_e(o_e)$.*

Proof. We use the proof approach of Correa et al. [9]. The Nash flow h satisfies the variational inequality $\sum_e(h_e-f_e)\ell_e(g_e+h_e) \le 0$ for any s-t flow f of volume $1-\alpha$. Taking $f = o-g$, which is a valid flow, we get that $\sum_e(g_e+h_e)\ell_e(g_e+h_e) \le \sum_e o_e\ell_e(g_e+h_e)$. We write $o_e\ell_e(g_e+h_e) = o_e\ell_e(o_e) + o_e\big(\ell_e(g_e+h_e)-\ell_e(o_e)\big)$. Using the definition of β, the last term is at most $\beta(g_e+h_e)\ell_e(g_e+h_e)$ (this holds even if $o_e > g_e+h_e$). So $(1-\beta)\sum_e(g_e+h_e)\ell_e(g_e+h_e) \le OPT$. This proves part (i), since $\rho = (1-\beta)^{-1}$.

Let $\tilde{\ell}_e(x) = \ell_e(x+g_e)$. As in (i), we start with the inequality $\sum_e h_e\tilde{\ell}_e(h_e) \le \sum_e(o_e-g_e)\tilde{\ell}_e(h_e)$. We have $(o_e-g_e)\tilde{\ell}_e(h_e) = (o_e-g_e)\tilde{\ell}_e(o_e-g_e) + (o_e-g_e)\big(\tilde{\ell}_e(h_e)-\tilde{\ell}_e(o_e-g_e)\big)$, and the last term is at most $\beta(\tilde{\ell})\cdot h_e\tilde{\ell}_e(h_e)$. Finally, note that $\beta(\tilde{\ell}) \le \beta$, so we obtain that $(1-\beta)\cdot\sum_e h_e\ell_e(g_e+h_e) \le \sum_e(o_e-g_e)\ell_e(o_e)$, proving part (ii). ∎

Lemma 3.1 shows that if $g_e \le o_e$, then the quantities $\sum_e g_e\ell_e(o_e)$, and $\sum_e h_e\ell_e(g_e+h_e)$ are bounded in terms of OPT. Our goal will be to bound $g_e\ell_e(g_e+h_e)$ in terms of $g_e\ell_e(o_e)$ and $h_e\ell_e(g_e+h_e)$, and thereby bound $C(g+h)$. For example, for linear latencies, we always have $g_e\ell_e(g_e+h_e) \le g_e\ell_e(g_e) + h_e\ell_e(g_e+h_e)$, so if g is the LLF strategy this gives the (weak) bound $C(g+h) \le \big(\alpha+2\rho(1-\alpha)\big)\cdot OPT$. Let g^{LLF} denote the flow due to the LLF strategy, and $g^{\mathsf{Sc}} = \alpha\cdot o$ denote the flow due to the Scale strategy. In general, for either strategy, there could be various combinations (λ,κ) such that $g_e\ell_e(g_e+h_e) \le \lambda\cdot g_e\ell_e(g_e) + \kappa\cdot h_e\ell_e(g_e+h_e)$, and we would like to obtain the tightest bound. To this end, for any $b \ge 0$, we define $c_{\mathsf{LLF}}(\mathcal{L};b) = \sup_{\ell\in\mathcal{L}}\sup_{x,y\ge 0} \frac{(x-b\cdot y)\ell(x+y)}{x\ell(x)}$. This definition is tailored so that with $x = g_e,\ y = h_e$, we get $g_e\ell_e(g_e+h_e) \le c_{\mathsf{LLF}}(\mathcal{L};b)\cdot g_e\ell_e(g_e) + b\cdot h_e\ell_e(g_e+h_e)$. Define $c_{\mathsf{Sc}}(\mathcal{L};\alpha,b) = \sup_{\ell\in\mathcal{L}}\sup_{x,y\ge 0} \frac{(\alpha x-b\cdot y)\ell(\alpha x+y)}{\alpha x\ell(x)}$. For any $b \ge 0$, we will obtain a bound on the PoA under g^{LLF} and g^{Sc} in terms of b, $\rho(\mathcal{L})$, and c_{LLF} (for g^{LLF}) or c_{Sc} (for g^{Sc}); we will then select the value of b that minimizes this ratio. We drop the $\mathcal{L}$ in c_{LLF} and c_{Sc} in the sequel. Note that both c_{LLF} and c_{Sc} are decreasing functions of b. Also $c_{\mathsf{Sc}}(\alpha,b) \le c_{\mathsf{LLF}}(b)$ for every α (since $0 \le \alpha \le 1$), so our PoA bound for the Scale strategy will be better than that for LLF. But, we include an analysis of LLF since it may be of independent interest.

THEOREM 3.2. *(i) The PoA under the Scale strategy is at most*

$$r_{\mathsf{Sc}}(\rho;\alpha) = \min\big\{\rho,\ \inf_{b\ge 0}\big(\alpha\cdot c_{\mathsf{Sc}}(\alpha,b) + \rho(1-\alpha)(1+b)\big)\big\}.$$

(ii) The PoA under the LLF strategy is at most

$$r_{\mathsf{LLF}}(\rho;\alpha) = \min\big\{\rho,\ \inf_{b\ge 0}\big(\alpha\cdot c_{\mathsf{LLF}}(b) + \rho(1-\alpha)(1+b)\big)\big\},$$

which is a decreasing function of α. The functions r_{LLF} and r_{Sc} take values ρ and 1 at $\alpha = 0$ and $\alpha = 1$ respectively.

Proof. Lemma 3.1 shows that the PoA is at most ρ. The cost of $g+h$ is $C(g+h) = \sum_e g_e\ell_e(g_e+h_e) + \sum_e h_e\ell_e(g_e+h_e)$. Taking $g = g^{\mathsf{Sc}} = \alpha\cdot o$, and using the definition of $c_{\mathsf{Sc}}(\alpha,b)$, we get that for any $b \ge 0$,

$$\sum_e g_e^{\mathsf{Sc}}\ell_e(g_e^{\mathsf{Sc}}+h_e) \le c_{\mathsf{Sc}}(\alpha,b)\sum_e g_e^{\mathsf{Sc}}\ell_e(o_e)+ b\sum_e h_e\ell_e(g_e^{\mathsf{Sc}}+h_e).$$

Let $C_1 := C(g^{\mathsf{Sc}}+h)$. So $C_1 \le \alpha\cdot c_{\mathsf{Sc}}(\alpha,b)\cdot OPT + (b+1)\sum_e h_e\ell_e(g_e^{\mathsf{Sc}}+h_e)$. Using part (ii) of Lemma 3.1 to bound $\sum_e h_e\ell_e(g_e^{\mathsf{Sc}}+h_e)$, we get that $C_1 \le \big(\alpha\cdot c_{\mathsf{Sc}}(\alpha,b)+(1-\alpha)(b+1)\rho\big)\cdot OPT$ for every $b\ge 0$. Thus the PoA is at most $r_{\mathsf{Sc}}(\rho;\alpha) = \min\{\rho, \inf_{b\ge0}\big(\alpha\cdot c(\alpha,b)+(1-\alpha)(b+1)\rho\big)\}$. Clearly $r_{\mathsf{Sc}}(\rho;0)=\rho$. It is easy to see that $\lim_{b\to\infty} c(1,b)=1$, which gives $r_{\mathsf{Sc}}(\rho;1)=1$.

Similarly taking $g = g^{\mathsf{LLF}}$, we get that

$$\sum_e g_e^{\mathsf{LLF}}\ell_e(g_e^{\mathsf{LLF}}+h_e) \le c_{\mathsf{LLF}}(b)\sum_e g_e^{\mathsf{LLF}}\ell_e(g_e^{\mathsf{LLF}})+ b\sum_e h_e\ell_e(g_e^{\mathsf{LLF}}+h_e).$$

Thus, for every $b\ge0$, $C_2 := C(g^{\mathsf{LLF}}+h) \le c_{\mathsf{LLF}}(b)\cdot A + (b+1)\rho\cdot B$, by Lemma 3.1 part (ii), where $A = \sum_e g_e^{\mathsf{LLF}}\ell_e(o_e)$, and $B = \sum_e(o_e - g_e^{\mathsf{LLF}})\ell_e(o_e)$. Let $A = \theta\cdot OPT$, so we have $\theta \ge \alpha$. Thus, in the worst case we obtain the bound $C_2/OPT \le \min\{\rho, \max_{\theta\in[\alpha,1]}\inf_{b\ge0}(\theta\cdot c_{\mathsf{LLF}}(b)+(1-\theta)(b+1)\rho)\}$. The rest of the proof is devoted to showing that the above expression can be simplified to obtain the bound $r_{\mathsf{Sc}}(\rho;\alpha)$, and proving the properties of r_{Sc} stated in the theorem. This is somewhat technical, and we only sketch the main ideas; the details are relatively straightforward to fill in.

If $\alpha=1$, then $\theta=1$, and if $\theta=1$, the infimum is 1 since $\lim_{b\to\infty} c_{\mathsf{LLF}}(b)=1$, so the bound in the theorem clearly holds. So let $\alpha<1$. One can show that $c_{\mathsf{LLF}}(b)$ is a convex function. So for any $\theta<1$, $\theta\cdot c_{\mathsf{LLF}}(b)+(1-\theta)(b+1)\rho$ is minimized at the unique value $b_\theta\ge0$ satisfying $c'_{\mathsf{LLF}}(b_\theta)=\rho(1-1/\theta)$. So b_θ increases with θ. Let $\varphi(\theta)=\theta\cdot c_{\mathsf{LLF}}(b_\theta)+(1-\theta)(b_\theta+1)\rho$. Let $\theta^*\in[0,1]$ be such that $c_{\mathsf{LLF}}(b_{\theta^*})=(b_{\theta^*}+1)\rho$. One can show that $\varphi'(\theta)\le0$ for $\theta\ge\theta^*$, and $\varphi'(\theta)\ge0$ for $\theta\le\theta^*$. Thus, over the range $[\alpha,1]$, $\varphi(\theta)$ is maximized at $\theta=\max(\alpha,\theta^*)$. So if $\alpha\ge\theta^*$, then we get the bound in the theorem. Observe that for $\alpha<\theta^*$, $b_\alpha<b_{\theta^*}$, so $c_{\mathsf{LLF}}(b_\alpha)>(b_\alpha+1)\rho\ge\rho$ and $\varphi(\alpha)>\rho$, and the bound still applies. Thus, in both cases we can bound the PoA by $\min\{\rho,\varphi(\alpha)\} = r_{\mathsf{Sc}}(\rho;\alpha)$. For $\alpha<\theta^*$, the value of $r_{\mathsf{LLF}}(\rho;\alpha)$ stays fixed at ρ; for $\alpha\ge\theta^*$, we have $r_{\mathsf{LLF}}(\rho;\alpha)=\min\{\rho,\varphi(\alpha)\}$, which decreases as α increases. Hence, r_{LLF} decreases with α. Finally, since $\lim_{b\to\infty} c_{\mathsf{LLF}}(b)=1$, we have $r_{\mathsf{LLF}}(\rho;1)=1$. ■

Let $\mathcal{L}_k$ be the class of polynomial latency functions with degree (at most) k. For the class $\mathcal{L}_k$, we can simplify the expression above and obtain an upper bound on $r_{\mathsf{Sc}}(\alpha,b)$. For linear latencies, we can further simplify this bound and obtain a closed-form expression. Our bound for linear latencies is weaker than the bound obtained by Karakostas and Kolliopoulos [16], but it is obtained through a general technique that can be used to compute PoA bounds for other latency classes as well.

THEOREM 3.3. *The PoA for the class $\mathcal{L}_k$ under the Scale strategy is at most*

$$r(k;\alpha) = \alpha + \rho(\mathcal{L}_k)(1-\alpha)\big(1+b(k,\alpha)\big),$$

where $b(k,\alpha)\ge0$ is the unique value such that $\alpha^k\big(\frac{b(k,\alpha)+1}{k+1}\big)^{k+1} = \big(\frac{b(k,\alpha)}{k}\big)^k$. For linear latency functions, this simplifies to yield the bound $r(1;\alpha) = \alpha + \frac{4}{3}(1-\alpha)\cdot\frac{2-2\sqrt{1-\alpha}}{\alpha}$. The function $r(k;.)$ takes values $\rho(\mathcal{L}_k)$ at $\alpha=0$, and 1 at $\alpha=1$ respectively.

3.3 Parallel-link graphs We now consider parallel-link networks and obtain various improvements. We obtain a simple proof showing that for *any* class of latency functions $\mathcal{L}$ with PoA $\rho(\mathcal{L})$, LLF reduces the PoA to $r(\rho(\mathcal{L}),\alpha) = \alpha+(1-\alpha)\rho(\mathcal{L})$. Thus, in parallel-link graphs, the (worst-case) PoA always improves by centrally controlling flow. We begin by giving a very simple proof of the following result from [24].

THEOREM 3.4. ([24]) *In a parallel-link network, the LLF strategy induces a flow of cost at most $\frac{1}{\alpha}\cdot OPT$.*

Proof. Re-index the links in decreasing order of $\ell_e(o_e)$, their latency in the optimal flow o (breaking ties arbitrarily). Let k be the index of the highest index link used by g, so $L_g = \ell_k(o_k)$. Let L be the Nash latency. We claim that $L\le L_g$. If not, then for all $i\ge k$, we have $\ell_i(g_i+h_i)\ge L > L_g \ge \ell_i(o_i)$ since all links have latency at least L under the flow $g+h$. This implies that $g_i+h_i > o_i$ for every $i\ge k$. For $i<k$, $g_i=o_i$, so this implies that the total $(g+h)$ flow has volume greater than 1, giving a contradiction.

The claim implies the following simple facts: (a) for $i\le k$, $\ell_i(g_i+h_i)\le\ell_i(o_i)$ since $\ell_i(o_i)\ge L_g\ge L$ for all these links; (b) thus, $\sum_i g_e\ell_e(g_e+h_e)\le OPT$; and (c) $OPT\ge\alpha L_g\ge\alpha L$. Since the cost of h is $(1-\alpha)L$, this implies that $C(g+h)\le OPT+(1-\alpha)L\le\frac{1}{\alpha}\cdot OPT$. ■

THEOREM 3.5. *In a parallel-link network, for any class $\mathcal{L}$ of latency functions with PoA $\rho(\mathcal{L})$, the PoA under LLF is at most $\alpha+(1-\alpha)\rho(\mathcal{L})$.*

Proof. Let $\rho=\rho(\mathcal{L})$. Let $A=\sum_e g_e\ell_e(o_e)$ and $B=\sum_e(o_e-g_e)\ell_e(o_e)$. Then, $OPT = A+B$ and $A/B\ge\alpha/(1-\alpha)$. As argued in Theorem 3.4, $\sum_e g_e\ell_e(g_e+h_e)\le A$, and by Lemma 3.1, we have $\sum_e g_e\ell_e(g_e+h_e)\le\rho\cdot B$.

Thus $C(g+h) \le A + \rho \cdot B$. The ratio $C(g+h)/OPT$ is maximized when $A/B = \alpha/(1-\alpha)$, and the maximum value is at most $\alpha + (1-\alpha)\rho$. ∎

The bound $\alpha + (1-\alpha)\rho$ is at most $\frac{1}{\alpha}$, for $\alpha \le \min\big(\frac{1}{\rho(\mathcal{L})-1}, 1\big)$. Thus, when $\rho(\mathcal{L}) \le 2$ (e.g., for linear, quadratic, cubic latency functions), the bound in Theorem 3.5 is always at most $\frac{1}{\alpha}$, and yields a tighter bound on the PoA.

For linear latency functions, Roughgarden [24] obtained a tighter bound on the PoA, but the proof in [24] is tailored specifically for linear latencies and does not extend to other latency classes. The proof above works for *any* class of latency functions, and demonstrates that Stackelberg strategies always improve the price of anarchy in parallel-link networks.

3.4 Extensions

Multicommodity networks. The PoA bounds obtained above for the Scale strategy in Section 3.2 extend to multicommodity (or asymmetric) nonatomic routing games. Here the nonatomic users are divided into different types and the total flow volume of users of type i is D_i, which has to be routed from s_i to t_i. Roughgarden [23] defined two types of Stackelberg strategies for multicommodity networks: *weak Stackelberg strategies*, where the central authority controls αD_i units of flow for each user-type i, and *strong Stackelberg strategies*, where the center controls an α fraction of the total flow, i.e., it can control any amount $\beta_i D_i$ flow of type i provided that $\sum_i \beta_i D_i \le \alpha \sum_i D_i$. The Scale strategy (which is well-defined even in the multicommodity setting) is thus a weak strategy. Notice that nowhere in the analysis of Scale in Section 3.2 do we use the fact that we have a single-commodity network. Thus, the analysis and the PoA bounds derived for Scale hold as is even for multicommodity networks.

Combining Stackelberg strategies and tolls. While Stackelberg strategies and tolls have till now been considered separately, a natural question to consider is whether combining these two measures yields any benefits that are not achievable by either in isolation? There are two issues of interest here: the cost of the resulting flow, and the total toll paid by a user (along a flow path). We observe that for any instance (G, ℓ), by using the LLF strategy in conjunction with tolls, we can enforce an optimal flow where the maximum toll paid by a user along a flow path is bounded (in terms of OPT). In contrast, it is known that for arbitrary instances (G, ℓ), Stackelberg strategies alone cannot reduce the PoA below $\frac{1}{\alpha}$ (see [23]), and without any flow control, the total toll paid along a flow path may be unbounded (compared to OPT) [12] (although the optimal flow can always be enforced [6, 29, 11, 15]).

A Stackelberg strategy g used in conjunction with tolls $\tau = \{\tau_e\}$ induces the flow $g + h$, where h is the Nash flow with respect to the latencies $\hat{\ell}_e(x) = \ell_e(x + g_e) + \tau_e$. The above result follows from two facts. (a) For any feasible flow f', there exists a flow $f \le f'$ that can be enforced via tolls [11]. (b) For any flow f enforceable by tolls, there exist tolls that enforce f such that the maximum toll paid by a user is at most $\max_{P \in \mathcal{P}} \ell_P(f)$ [12]; a slight modification of the proof shows that the maximum toll paid is at most $\max_{P: f_P > 0} \ell_P(f)$. Thus, if g is the flow due to the LLF strategy, then considering the instance $(G, \tilde{\ell})$, where $\tilde{\ell}_e(x) = \ell_e(x + g_e)$, one can enforce a flow $f \le o - g$ as the Nash flow, charging a toll of at most $\max_{P:(o-g)_P > 0} \tilde{\ell}_P(o-g) \le \frac{1}{\alpha} \cdot OPT$ to any user. Note that the induced flow $f + g \le o$ is also optimal.

4 Tolls for atomic splittable routing games

In this section, we consider the use of tolls for atomic splittable routing games. We show that tolls that induce an optimal flow always exist and can be efficiently computed, and this is true even in general *asymmetric* atomic routing games, where different atomic users may have different source-sink pairs. We also give a complete characterization of flows that can be induced via tolls. Since Nash flows are not known to be unique for atomic splittable routing games (even in general graphs with a single source-sink pair), by inducing a flow F, we mean that F can be realized as *some* atomic Nash equilibrium via tolls. To our knowledge, this is the first result on tolls for atomic splittable traffic. As mentioned in the Introduction, except in some special cases implied by the results of Cominetti et al. [7] and Hayrapetyan et al. [13], even the existence of tolls inducing an optimal flow was not known. (We remark that in these settings, it is also the case that atomic Nash equilibria are unique; Orda et al. [20] show this for parallel-link graphs, and [7] show this for single-commodity, symmetric instances.)

We consider an *asymmetric* atomic splittable routing game $(G, k, \{D_i\}, \ell)$, where user i controls $D_i \ge 0$ units of flow which has to be routed from s_i to t_i. The users are *heterogeneous*, which means that the net latency experienced by user i on edge e due to a toll τ_e, is given by $\tilde{\ell}_{i,e}(x) = \gamma_i \ell_e(x) + \tau_e$, where $\gamma_i \ge 0$ indicates the sensitivity of user i to delay. Let $\mathcal{P}_i = \mathcal{P}_{s_i t_i}$ denote the set of all simple s_i-t_i paths. Given a flow $f = \sum_i f_i$, where each f_i is a feasible flow for user i, define $\ell^*_{i,e}(x) = \ell_e(x) + f_{i,e}\ell'_e(x)$; this represents the *marginal cost* of increasing flow on edge e for user i. Let $\tilde{\ell}^*_{i,e}(x) = \tilde{\ell}_{i,e}(x) + f_{i,e}\tilde{\ell}'_{i,e}(x) = \gamma_i \ell^*_{i,e}(x) + \tau_e$. The flows $f_1, \ldots, f_k$ will be clear from the context. As usual, let $\ell^*_{i,P}(f)$ and $\tilde{\ell}^*_{i,P}(f)$ denote $\sum_{e \in P} \ell^*_{i,e}(f_e)$ and

$\sum_{e\in P} \tilde{\ell}^*_{i,e}(f_e)$. A flow is at Nash equilibrium if no user can reduce the cost of its flow by rerouting any part of it. The following characterization will be useful: a flow profile $(f_1,\ldots,f_k)$, with $f=\sum_i f_i$, is a Nash equilibrium (or Nash flow) under tolls $\{\tau_e\}$, iff for each user i, (a) f_i is feasible, i.e., it routes D_i units from s_i to t_i, and (b) for any two paths $P,P'\in\mathcal{P}_i$ with $f_{i,P}>0$, we have $\tilde{\ell}^*_{i,P}(f)\le\tilde{\ell}^*_{i,P'}(f)$.

We say that a flow H is *enforceable*, if there exist tolls $\{\tau_e\}$ such that H is realized as *some* Nash equilibrium given these tolls. The proof approach is similar to that in [11, 15]. We show that tolls (if they exist) are obtained as the optimal Lagrangian multipliers (or dual variables) to an appropriate *convex program* (instead of a linear program). Consider the following convex program:

$$\min \quad L(f) := \sum_i \gamma_i \sum_{P\in\mathcal{P}_i} \ell_P(H) f_{i,P} + \sum_i \gamma_i \sum_e \ell'_e(H_e)\cdot\frac{\left(\sum_{P\in\mathcal{P}_i:e\in P} f_{i,P}\right)^2}{2} \qquad (\mathrm{CP}_H)$$

$$\text{s.t.} \quad \sum_i \sum_{P\in\mathcal{P}_i:e\in P} f_{i,P} \le H_e \quad \text{for all } e\in E, \qquad (4.1)$$

$$\sum_{P\in\mathcal{P}_i} f_{i,P} = D_i \quad \text{for all } i, \qquad (4.2)$$

$$f_{i,P} \ge 0 \quad \text{for all } i, P\in\mathcal{P}_i. \qquad (4.3)$$

The objective function of (CP_H) is clearly convex, so (CP_H) is a convex program. Although (CP_H) has exponentially many variables, we can express it compactly using flow-edge variables. Thus, (CP_H) can be solved efficiently up to an arbitrarily small error. Observe that (CP_H) is very similar to the linear program used in [11, 15] to prove the existence of tolls that enforce flow H in the nonatomic setting. The only difference is in the objective function, which now contains a non-linear term that appears due to the fact that atomic users take into account the impact their routing strategy on the flow that they control. Applying the *Kuhn-Karusch-Tucker* (KKT) conditions for convex optimization to (CP_H), we get that $(f_1,\ldots,f_k)$ is an optimal solution to (CP_H) if it is feasible, and there exist Lagrangian multipliers $\tau_e\ge 0$, z_i, $y_{i,P}\ge 0$ corresponding to constraints (4.1), (4.2), and (4.3) respectively, such that the following hold:

1. (Complementary Slackness) for every edge e, $\tau_e\big(\sum_i\sum_{P\in\mathcal{P}_i:e\in P} f_{i,P} - H_e\big) = 0$; for every i, $z_i\big(D_i - \sum_{P\in\mathcal{P}_i} f_{i,P}\big) = 0$; for every i and path $P\in\mathcal{P}_i$, $-y_{i,P}f_{i,P}=0$.
2. (Zero Gradient)

$$\forall i, P\in\mathcal{P}_i,\ \ \frac{\partial}{\partial f_{i,P}}\Big[L+\tau_e\Big(\sum_i\sum_{P\in\mathcal{P}_i:e\in P} f_{i,P}-H_e\Big)+\sum_i z_i\Big(D_i-\sum_{P\in\mathcal{P}_i} f_{i,P}\Big)-\sum_{i,P\in\mathcal{P}_i} y_{i,P}f_{i,P}\Big]=0$$

Simplifying, we get $\sum_{e\in P}\big(\gamma_i(\ell_e(H_e)+f_{i,e}\ell'_e(H_e))+\tau_e\big) - z_i - y_{i,P} = 0$ for every $i,P\in\mathcal{P}_i$, where $f_{i,e}=\sum_{P\in\mathcal{P}_i:e\in P} f_{i,P}$.

THEOREM 4.1. *A feasible flow H is enforceable if and only if there is an optimal solution $(f_1,\ldots,f_k)$ to (CP_H) that satisfies all constraints* (4.1) *with equality.*

Proof. First suppose that H is enforceable via tolls $\{\tau_e\}$. So $H=f=\sum_i f_i$, where $(f_1,\ldots,f_k)$ is an atomic Nash equilibrium induced by these tolls. So for any path $P\in\mathcal{P}_i$ with $f_{i,P}>0$, the value $\sum_{e\in P}\big(\gamma_i(\ell_e(f_e)+f_{i,e}\ell'_e(f_e))+\tau_e\big)$ must be the same, and must be minimum among all paths $P'\in\mathcal{P}_i$; set z_i equal to this value. Set $y_{i,P}=\sum_{e\in P}\big(\gamma_i(\ell_e(f_e)+f_{i,e}\ell'_e(f_e))+\tau_e\big)-z_i$. It is easy to verify that f, and (τ,z,y) satisfy the KKT conditions (note that $f_e=H_e$). Since $(f_1,\ldots,f_k)$ is feasible for (CP_H), it must be an optimal solution, and it satisfies all constraints (4.1) with equality.

Now suppose that there exists an optimal solution $(f_1,\ldots,f_k)$ to (CP_H) such that $f_e:=\sum_i f_{i,e}=H_e$ for every edge e. Then there must exist Lagrangian multipliers (τ,z,y) satisfying the KKT conditions. For any path $P\in\mathcal{P}_i$ with $f_{i,P}>0$, we must have $\gamma_{i,P}=0$ by condition (iii). So by condition (iv), we have $z_i=\sum_{e\in P}\big(\gamma_i(\ell_e(H_e)+f_{i,e}\ell'_e(H_e))+\tau_e\big)$. Also for any path P' we have $z_i\le\sum_{e\in P}\big(\gamma_i(\ell_e(H_e)+f_{i,e}\ell'_e(H_e))+\tau_e\big)$ since $y_{i,P}\ge 0$. Thus $(f_1,\ldots,f_k)$ is an atomic Nash equilibrium induced by the tolls $\{\tau_e\}$, which implies that H is enforceable. ∎

Note that given an optimal solution $(f_1,\ldots,f_k)$ to (CP_H) that satisfies constraints(4.1) with equality, one can obtain the corresponding tolls τ_e by solving the following system of linear inequalities:

$$\begin{aligned} z_i &\le \sum_{e\in P}\big(\gamma_i(\ell_e(H_e)+f_{i,e}\ell'_e(H_e))+\tau_e\big) && \forall i,P\in\mathcal{P}_i,\\ z_i &= \sum_{e\in P}\big(\gamma_i(\ell_e(H_e)+f_{i,e}\ell'_e(H_e))+\tau_e\big) && \forall i,P: f_P>0,\\ \tau_e &\ge 0 && \forall e. \end{aligned}$$

We can even optimize a linear (or convex) objective function of the tolls. As in [11], one can show that for any feasible flow H, one can obtain a flow $H'\le H$

such that every feasible solution, and hence any optimal solution, to ($\mathrm{CP}_{H'}$) must satisfy (4.1) with equality, and hence, H' is enforceable. If H is an optimal solution then so is H'; thus, one can compute tolls enforcing an optimal flow.

These results extend to the setting of *mixed* routing games, where some users are atomic splittable users, and some are nonatomic. The only change is that in the objective function of (CP_H), the nonlinear term appears only for the atomic users. Our results also extend to the setting of generalized atomic splittable congestion games, where there are m resources with associated latency functions, and user i's strategy consists of choosing a fractional assignment $(f_{i,1}, \ldots, f_{i,m})$ of his volume D_i to the m resources, incurring cost $\sum_j f_{i,j}\ell_j(\sum_{i'} f_{i',j})$.

Acknowledgment

I thank Lisa Fleischer for several stimulating discussions, and useful comments on earlier drafts of this paper. In particular, the results in Section 3.1 were obtained in collaboration with her, and I thank her for allowing me to include these results.

References

[1] B. Awerbuch, Y. Azar, and A. Epstein. The price of routing unsplittable flow. *Proc. 37th STOC*, pages 57–66, 2005.

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

[3] I. Caragiannis, C. Kaklamanis, and P. Kanellopoulos. Taxes for linear atomic congestion games. *Proc. 14th ESA*, pages 184–195, 2006.

[4] G. Christodoulou and E. Koutsoupias. The price of anarchy of finite congestion games. *Proc. 37th STOC*, pages 67–73, 2005.

[5] R. Cole, Y. Dodis, and T. Roughgarden. How much can taxes help selfish routing? *Journal of Computer and System Sciences*, 72:444–467, 2006.

[6] R. Cole, Y. Dodis, and T. Roughgarden. Pricing network edges for heterogeneous selfish users. *Proc. 35th STOC*, pages 521–530, 2003.

[7] R. Cominetti, J. R. Correa, and N. Stier-Moses. Network games with atomic players. *Proc. 33rd ICALP*, pages 525–536. Springer, 2006.

[8] J. Correa and N. Stier-Moses. A note on Stackelberg routing. Unpublished manuscript, September 2006.

[9] J. R. Correa, A. S. Schulz, and N. Stier-Moses. Fast, fair, and efficient flows in networks. *Proc. 10th IPCO*, pages 59–73. Springer, 2004.

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

[11] L. Fleischer, K. Jain, and M. Mahdian. Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games. *Proc. 45th FOCS*, pages 277–285, Rome, Italy, 2004.

[12] Lisa Fleischer. Linear tolls suffice: New bounds and algorithms for tolls in single source networks. *Theoretical Computer Science*, 348:217–225, 2005.

[13] Ara Hayrapetyan, É. Tardos, and T. Wexler. The effect of collusion in congestion games. *Proc. 38th STOC*, pages 89–98, 2006.

[14] A. Kaporis and P. Spirakis. The price of optimum in Stackelberg games on arbitrary single commodity networks and latency functions. Proc. 18th SPAA, pages 19–28, 2006.

[15] G. Karakostas and S. Kolliopoulos. Edge pricing of multicommodity networks for heterogeneous users. *Proc. 45th FOCS*, pages 268–276, 2004.

[16] G. Karakostas and S. Kolliopoulos. Stackelberg strategies for selfish routing in general multicommodity networks. Technical report, McMaster University, June 2006.

[17] Y. A. Korilis, A. A. Lazar, and A. Orda. Achieving network optima using Stackelberg routing strategies. *IEEE/ACM Transactions on Networking*, 5(1):161 – 173, 1997.

[18] Elias Koutsoupias and C. Papadimitriou. Worst-case equilibria. *Proc. 16th STACS*, pages 404–413, 1999.

[19] V. S. A. Kumar and A. Marathe. Improved results for Stackelberg scheduling strategies. *Proc. 29th ICALP*, 2002.

[20] A. Orda, R. Rom, and N. Shimkin. Competitive routing in multiuser communication networks. *IEEE/ACM Transactions on Networking*, 1:510–521, 1993.

[21] C. H. Papadimitriou. Algorithms, games, and the internet.

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

[23] T. Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, 2005.

[24] T. Roughgarden. Stackelberg scheduling strategies. *SIAM J. Computing*, 33(2):332–350, 2004.

[25] T. Roughgarden. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67(2):341–364, 2003.

[26] T. Roughgarden and É. Tardos. How bad is selfish routing? *Journal of the ACM*, 49:236–259, 2002.

[27] Y. Sharma and D. P. Williamson. Stackelberg thresholds in network routing games or the value of altruism. Unpublished manuscript, August 2006.

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

[29] H. Yang and H. J. Huang. The multi-class, multi-criteria traffic network equilibria and system optimum problem. *Transportation Research B*, 28:1–15, 2004.