The Impact of Stackelberg Routing in General Networks

Vincenzo Bonifaci Tobias Harks Guido Schäfer

Technische Universität Berlin, Institut für Mathematik
Straße des 17. Juni 136, 10623 Berlin, Germany
{bonifaci,harks,schaefer}@math.tu-berlin.small

July 6, 2007

Abstract

We investigate the impact of *Stackelberg routing* in network routing games. In this setting, a fraction α of the entire demand is first routed by a central authority, called the *Stackelberg leader*, while the remaining demand is then routed by selfish (nonatomic) players. The aim is to devise *Stackelberg strategies*, i.e., strategies to route the centrally controlled demand, so as to minimize the price of anarchy of the resulting flow.

Although several advances have been made recently in proving that Stackelberg routing may in fact significantly reduce the price of anarchy for certain network topologies, it is still an open question whether this holds true in general. We answer this question negatively. We prove that the price of anarchy achievable via Stackelberg routing can be unbounded even for single-commodity networks.

In light of this negative result, we consider bicriteria bounds. We develop an efficiently computable Stackelberg strategy that induces a flow whose cost is at most the cost of an optimal flow with respect to demands scaled by a factor of $1+\sqrt{1-\alpha}$. Thus, we obtain a smooth trade-off curve that scales between the absence of centralized control (doubling the demands is sufficient) and completely centralized control (no scaling is necessary).

Finally, we analyze the effectiveness of a simple Stackelberg strategy, called SCALE, for polynomial latency functions. Our analysis is based on a general technique which is simple, yet powerful enough to obtain (almost) tight bounds for SCALE in general networks. For linear latency functions, we derive an upper bound that matches the current best one and show that this bound is tight. For general polynomial latency functions, we obtain upper bounds that improve all previously known ones.

1 Introduction

Over the past years, the impact of the behavior of selfish, uncoordinated users in congested networks has been investigated intensively in the theoretical computer science literature. In this context, *network routing games* have proved to be a reasonable means of modeling selfish behavior in networks. The basic idea is to model the interaction between the selfish network users as a *noncooperative game*. We are given a directed graph with latency functions on the arcs and a set of origin-destination pairs, called *commodities*. Every commodity has a *demand* associated with it, which specifies the amount of flow that needs to be sent from the respective origin to the destination. We assume that every demand represents a very large population of players, each controlling an infinitesimal amount of flow of the entire demand (such players are also called *nonatomic*). The latency that a player experiences to traverse an arc is given by a (non-decreasing) function of the total flow on that arc. We assume that every player acts selfishly and routes his flow along a minimum-latency path from its origin to the destination; this corresponds to a common solution concept for noncooperative games, that of a *Nash equilibrium* (here *Nash flow*). In a Nash flow no player can improve his own latency by unilaterally switching to another path.

It is well known that Nash equilibria can be *inefficient* in the sense that they need not achieve socially desirable objectives [1, 5]. That is, in the context of network routing games, a Nash flow in general does not minimize the total cost; or said differently, selfish behavior may cause a performance degradation in the network. Koutsoupias and Papadimitriou [10] initiated the investigation of the efficiency loss caused by selfish behavior. They introduced a measure to quantify the inefficiency of Nash equilibria which they termed the *price of anarchy*. The price of anarchy is defined as the worst-case ratio of the cost of a Nash equilibrium over the cost of a system optimum.

In recent years, considerable progress has been made in quantifying the degradation in network performance caused by the selfish behavior of noncooperative network users. In a seminal work, Roughgarden and Tardos [17] showed that the price of anarchy for network routing games with nonatomic players and linear latency functions is $4/3$; in particular, this bound holds independently of the underlying network topology. The case of more general families of latency functions has been studied by Roughgarden [13] and Correa, Schulz, and Stier-Moses [2]. (For an overview of these results, we refer to the book by Roughgarden [16].) Despite these bounds for specific classes of latency functions, it is known that the price of anarchy for general latency functions is unbounded even on simple parallel-arc networks [17].

Due to this large efficiency loss, researchers have proposed different approaches to reduce the price of anarchy in network routing games. One of the most promising approaches is the use of *Stackelberg routing* [9, 15]. In this setting, it is assumed that a fraction $\alpha \in [0,1]$ of the entire demand is controlled by a central authority, termed *Stackelberg leader*, while the remaining demand is controlled by the selfish nonatomic players, also called the *followers*. In a *Stackelberg game*, the Stackelberg leader first routes the centrally controlled flow according to a predetermined policy, called the *Stackelberg strategy*, and then the remaining demand is routed by the selfish followers. The aim is to devise Stackelberg strategies so as to minimize the price of anarchy of the resulting combined flow.

Although Roughgarden [15] showed that computing the *best* Stackelberg strategy, i.e., one that minimizes the price of anarchy of the induced flow, is NP-hard even for parallel-arc networks and linear latency functions, several advances have been made recently in proving that Stackelberg routing can indeed significantly reduce the price of anarchy in network routing games. As an example, Roughgarden [15] showed that for parallel-arc networks Stackelberg strategies exist that reduce the price of anarchy to $1/\alpha$, *independently* of the latency functions. That is, even if the Stackelberg leader controls only a small constant fraction of the overall demand, the price of anarchy reduces to a constant (while it is unbounded in the absence of any centralized control). Very recently, Swamy [19] obtained a similar result for single-commodity, series-parallel networks. Besides these efforts, researchers have also tried to characterize the effectiveness of different Stackelberg strategies for specific classes of latency functions.

Our Results. In this paper, we investigate the impact of Stackelberg routing to reduce the price of anarchy in network routing games with nonatomic players. Our contribution is threefold:

1. Albeit the above mentioned advances in the context of Stackelberg routing, a central question is still open: Can we always devise a Stackelberg strategy such that the price of anarchy is bounded? A partial answer to this question was given by Roughgarden [15]. He showed that for certain types of Stackelberg strategies, which he termed *weak* strategies (see Section 2 for a definition), the price of anarchy for multi-commodity networks can be unbounded. However, this does not rule out the existence of such Stackelberg strategies in general.

   We answer this question negatively. We prove that the price of anarchy achievable via Stackelberg routing can be unbounded even for single-commodity networks. Our result holds for arbitrary Stackelberg strategies and independently of the fraction $\alpha \in (0,1)$ controlled by the Stackelberg leader. This settles the open question explicitly posed by Roughgarden [14, Open Problem 4].

2. In light of this negative result, we investigate the effectiveness of Stackelberg routing strategies compared to an optimum flow for a larger demand; i.e., we consider bicriteria bounds.

   We develop an efficiently computable Stackelberg strategy inducing a flow whose cost is at most the cost of an optimal flow with respect to demands increased by a factor of $1+\sqrt{1-\alpha}$. Thus, we obtain a smooth trade-off curve that scales between the absence of centralized control (doubling the demands is sufficient) and completely centralized control (no scaling is necessary). We also prove that this characterization is tight. Our bound is a natural generalization of the bicriteria bound by Roughgarden and Tardos [17]. We demonstrate that this result has a particular nice interpretation for the class of (practical relevant) M/M/1-latency functions that model arc-capacities: In order to beat the cost of an optimal flow, it is sufficient to scale all arc capacities by $1+\sqrt{1-\alpha}$.

3. One of the simplest Stackelberg strategies to implement is SCALE (see also [15]). SCALE simply computes an optimal flow for the entire demand and then scales this flow down by α. The currently best known bound for the price of anarchy induced by SCALE on multi-commodity networks and linear latency functions is due to Karakostas and Kolliopoulos [8]. Their analysis is based on a (rather involved) machinery presented in [12]. Very recently, Swamy [19] derived the first general bounds for polynomial latency functions.

   We introduce a general approach, which we term the λ-approach, to bound the price of anarchy of Stackelberg strategies. This approach is simple, yet powerful enough to obtain better bounds for SCALE in general networks. For linear latency functions, we derive an upper bound that coincides with the bound in [8]. However, our analysis is much simpler; in particular, we do not rely on the machinery in [12]. For general polynomial latency functions, our approach yields upper bounds that significantly improve the bounds by Swamy [19]. We also derive lower bounds for SCALE. We present a generalized Braess instance that shows that for the linear case our bound is tight; a similar instance can be used to show that for higher degree polynomials our bounds are almost tight (though there remains a gap for small values of α). We believe that our λ-approach may also prove useful to derive improved bounds for other Stackelberg strategies.

Related Work. The idea of using Stackelberg strategies to improve the performance of a system was first proposed by Korilis, Lazar, and Orda [9]. The authors identified necessary and sufficient conditions for the existence of Stackelberg strategies that induce a system optimum; their model differs from the one discussed here.

Roughgarden [15] first formulated the problem and model considered here. He also proposed some natural Stackelberg strategies such as SCALE and Largest-Latency-First (LLF). For parallel-arc networks he showed that the price of anarchy for LLF is bounded by $4/(3+\alpha)$ and $1/\alpha$ for linear and arbitrary latency functions, respectively. Both bounds are tight. Moreover, he also proved that it is NP-hard to compute the best Stackelberg strategy. Kumar and Marathe [11] investigated approximation schemes to compute the best Stackelberg strategy. The authors gave a PTAS for the case of parallel-arc networks.

Karakostas and Kolliopoulos [8] proved upper bounds on the price of anarchy for SCALE and LLF. Their bounds hold for arbitrary multi-commodity networks and linear latency functions. Their analysis is based on a result obtained by Perakis [12] to bound the price of anarchy for network routing games with asymmetric and non-separable latency functions. Furthermore, Karakostas and Kolliopoulos [8] showed that their analysis for SCALE is almost tight.

Very recently, Swamy [19] obtained upper bounds on the price of anarchy for SCALE and LLF for polynomial latency functions. In the case of linear latency functions, his bound is inferior to the one given by Karakostas and Kolliopoulos [8]. Swamy also proved a bound of $1+1/\alpha$ for single-commodity, series-parallel networks with arbitrary latency functions.

Correa and Stier-Moses [3] proved, besides some other results, that the use of *opt-restricted strategies*, i.e., strategies in which the Stackelberg leader sends no more flow on every edge than the system optimum, does not increase the price of anarchy. Sharma and Williamson [18] considered the problem of determining the smallest value of α such that the price of anarchy can be improved. They obtained results for parallel-arc networks and linear latency functions. Kaporis and Spirakis [7] studied a related question of finding the minimum demand that the Stackelberg leader needs to control in order to enforce an optimal flow.

2 The Model

In a network routing game we are given a directed network $G=(V,A)$ and k origin-destination pairs $(s_1,t_1),\ldots,(s_k,t_k)$ called *commodities*. For every commodity $i\in[k]$, a demand $r_i>0$ is given that specifies the amount of flow with origin s_i and destination t_i. Let $\mathcal{P}_i$ be the set of all paths from s_i to t_i in G and let $\mathcal{P}=\cup_i\mathcal{P}_i$. A *flow* is a function $f:\mathcal{P}\to\mathbb{R}_+$. The flow f is *feasible* (with respect to r) if for all i, $\sum_{P\in\mathcal{P}_i} f_P=r_i$. For a given flow f, we define the flow on an arc $a\in A$ as $f_a=\sum_{P\ni a} f_P$.

Moreover, each arc $a\in A$ has an associated variable *latency* denoted by $\ell_a(\cdot)$. For each $a\in A$ the latency function ℓ_a is assumed to be nonnegative, nondecreasing and differentiable. If not indicated otherwise, we also assume that ℓ_a is defined on $[0,\infty)$ and that $x\ell_a(x)$ is a convex function of x. Such functions are called *standard* [13]. The latency of a path P with respect to a flow f is defined as the sum of the latencies of the arcs in the path, denoted by $\ell_P(f)=\sum_{a\in A}\ell_a(f_a)$. The triple (G,r,ℓ) is called an *instance*.

The *cost* of a flow f is $C(f)=\sum_{P\in\mathcal{P}} f_P\ell_P(f)$. Equivalently, $C(f)=\sum_{a\in A} f_a\ell_a(f_a)$. The feasible flow of minimum cost is called *optimal* and denoted by o. A feasible flow f is a *Nash flow*, or *selfish flow*, if for every $i\in[k]$ and $P,P'\in\mathcal{P}_i$ with $f_P>0$, $\ell_P(f)\le\ell_{P'}(f)$. In particular, if f is a Nash flow, all s_i-t_i paths to which f assigns a positive amount of flow have equal latency. It is well-known that if f_1 and f_2 are Nash flows for the same instance, then $C(f_1)=C(f_2)$, see e.g. [17].

In a Stackelberg network game we are given, in addition to G, r and ℓ, a parameter $\alpha\in(0,1)$. A *(strong) Stackelberg strategy* is a flow g feasible with respect to $r'=(\alpha_1r_1,\ldots,\alpha_kr_k)$, for some $\alpha_1,\ldots,\alpha_k\in[0,1]$ such that $\sum_{i=1}^k\alpha_ir_i=\alpha\sum_{i=1}^k r_i$. If $\alpha_i=\alpha$ for all i, g is called a *weak Stackelberg strategy*. Thus, both strong and weak strategies route a fraction α of the overall traffic, but a strong strategy can choose how much flow of each commodity is centrally controlled. For single-commodity networks the two definitions coincide. A Stackelberg strategy g is called *opt-restricted* if $g_a\le o_a$ for all $a\in A$.

Given a Stackelberg strategy g, let $\tilde{\ell}_a(x)=\ell_a(g_a+x)$ for all $a\in A$ and let $\tilde{r}=r-r'$. Then a flow h is *induced by* g if it is a Nash flow for the instance $(G,\tilde{r},\tilde{\ell})$. The Nash flow h can be characterized by the following *variational inequality* [4]: h is a Nash flow induced by g if and only if for all flows x feasible with

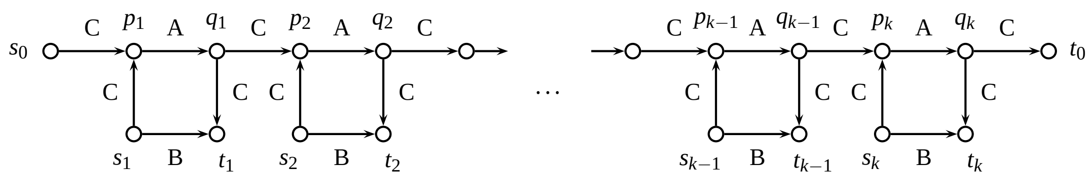

Figure 1: The graph G_k, used in the proof of Theorem 3.1. Arcs are labeled with their type.

respect to $\tilde{r}$,

$$\sum_{a \in A} h_a \ell_a(g_a + h_a) \leq \sum_{a \in A} x_a \ell_a(g_a + h_a). \tag{1}$$

We will mainly be concerned with the cost of the combined induced flow $g+h$, given by $C(g+h) = \sum_{a \in A}(g_a + h_a)\ell_a(g_a + h_a)$. In particular, we are interested in bounding the ratio $C(g+h)/C(o)$, called the *price of anarchy*.

3 The Limits of Stackelberg Routing

In this section, we prove that there does not exist a Stackelberg strategy that induces a price of anarchy bounded by a function of α only. More precisely, we show that for any fixed $\alpha \in (0,1)$, the ratio between the cost of the flow induced by any Stackelberg strategy and the optimum can be arbitrarily large, even in single-commodity networks.

We first show this claim for multi-commodity networks. In this case, such a result was already known to hold for weak Stackelberg strategies [16]; here we prove that it also holds for strong Stackelberg strategies.

Multi-Commodity Networks. We prove the following theorem.

Theorem 3.1. *Let $M > 0$ and $\alpha \in (0,1)$. There is a multi-commodity instance $I = (G, r, \ell, \alpha)$ such that, if g is any strong Stackelberg strategy for I inducing a Nash flow h, and o is an optimal flow for the instance (G, r, ℓ), then $C(g+h) \geq M \cdot C(o)$.*

To prove the theorem we will use an instance based on the graph depicted in Figure 1. For a positive integer k, the graph G_k has $4k+2$ nodes $V_k = \{s_0, t_0, s_1, t_1, p_1, q_1, \ldots, s_k, t_k, p_k, q_k\}$. The arc set A_k is the union of three sets, $\{(p_i, q_i) : i \in [k]\}$, $\{(s_i, t_i) : i \in [k]\}$, and $\{(s_i, p_i), (q_i, t_i), (q_i, p_{i+1}) : i \in [k]\} \cup \{(s_0, p_1), (q_k, t_0)\}$. We call the arcs in these sets of type A, B, and C respectively (see Figure 1). There are $k+1$ commodities $0, 1, \ldots, k$. Commodity i has origin s_i and destination t_i. The demand is $r_0 := (1-\alpha)/2$ for commodity 0, and $r_1 := (1+\alpha)/2k$ for all other commodities; thus, the total demand is $r_0 + kr_1 = 1$.

The latency of an arc is determined by its type. Type B arcs have constant latency 1, and type C arcs have constant latency 0. Type A arcs have latency $\ell_\varepsilon(x)$, where the function $\ell_\varepsilon(x)$ is defined as follows:

$$\ell_\varepsilon(x) = \begin{cases} 0, & \text{if } x \leq r_0 \\ 1 - \frac{r_0 + r_1 - x}{(1-\varepsilon) r_1}, & \text{if } x \geq r_0 + 2\varepsilon r_1 \end{cases}$$

Here ε is any positive constant such that $\varepsilon < \frac{1-\alpha}{1+\alpha}$. In the interval $(r_0, r_0 + 2\varepsilon r_1)$ the function ℓ_ε is defined arbitrarily so that overall it is a standard and convex function (see also Figure 3 in the appendix). In particular, $\ell_\varepsilon(x) \geq 1 - \frac{r_0 + r_1 - x}{(1-\varepsilon) r_1}$ for all x.

Let us first bound the cost of the optimal flow.

Lemma 3.2. $C(o) \leq 1$.

Proof. Consider the flow $\bar{f}$ where each commodity is routed along the shortest path (in terms of number of arcs) from origin to destination. The latency on the s_0-t_0 path is zero, since the load on each arc of the path is r_0 and $\ell_\varepsilon(r_0) = 0$. The latency of each other s_i-t_i path is 1. Then $C(o) \leq C(\bar{f}) = k \cdot r_1 = (1+\alpha)/2 \leq 1$. $\square$

Proof of Theorem 3.1. For $i \in [k]$, let g_i be the amount of flow sent by the Stackelberg strategy over the arc (s_i, t_i). Since the total value of the flow controlled by any Stackelberg strategy is α, we have $\sum_{i \in [k]} g_i \leq \alpha$.

The crucial point is that without loss of generality, all the selfish flow induced by g on an s_i-t_i path, $i \neq 0$, will be sent along the path (s_i, p_i, q_i, t_i). Indeed, if the arc (s_i, t_i) contained some selfish flow $h_i > 0$, the latency of the path (s_i, p_i, q_i, t_i) would be $\ell_\varepsilon(r_0 + r_1 - g_i - h_i) < 1 = \ell_{(s_i,t_i)}(g_i + h_i)$. But this contradicts the definition of Nash flows. Thus the combined flow on each (p_i, q_i) arc is exactly $r_0 + r_1 - g_i$. Now let P_0 be the unique s_0-t_0 path. We have

$$\ell_{P_0}(g+h) \geq \sum_{i \in [k]} \ell_\varepsilon(r_0 + r_1 - g_i) \geq \sum_{i \in [k]} \left(1 - \frac{g_i}{(1-\varepsilon)r_1}\right) \geq k - \frac{\alpha}{(1-\varepsilon)r_1} = \frac{1}{1-\varepsilon} \cdot \left(\frac{1-\alpha}{1+\alpha} - \varepsilon\right) \cdot k.$$

The last inequality follows from $\sum_i g_i \leq \alpha$, and the last equality from $r_1 = (1+\alpha)/2k$. Since $\varepsilon < \frac{1-\alpha}{1+\alpha}$, we conclude that $\ell_{P_0}(g+h) = \Omega(k)$. Together with Lemma 3.2, we obtain

$$C(g+h) \geq r_0 \cdot \ell_{P_0}(g+h) = \tfrac{1}{2} \cdot (1-\alpha) \cdot \Omega(k) = \Omega(k) \cdot C(o).$$

Thus the ratio of $C(g+h)/C(o)$ can be made arbitrarily large by picking a sufficiently large k. $\square$

Single-Commodity Networks. We use the insights gained in the previous section to prove the following, stronger result:

Theorem 3.3. *Let $M > 0$ and $\alpha \in (0,1)$. There is a single-commodity instance $I = (G, r, \ell, \alpha)$ such that, if g is any Stackelberg strategy for I inducing a Nash flow h, and o is an optimal flow for the instance (G, r, ℓ), then $C(g+h) \geq M \cdot C(o)$.*

Theorem 3.3 extends Theorem 3.1 to single-commodity networks. The main idea behind the proof is to simulate the instance used in Theorem 3.1 by creating a supersource s and a supersink t and connecting them to the sources and sinks of the original network (see also Figure 4 in the appendix). If somehow we were able to enforce the s-t flow to split according to the demand vector of the multi-commodity instance, the result would easily follow as in the proof of Theorem 3.1. In order to achieve this, we use latency functions that simulate capacities on the arcs connecting the supersource to the sources and the sinks to the supersink. Although these "capacities" might be exceeded, we will make sure that if the excess flow is too large, the price of anarchy will already be large enough for our purposes.

For a positive integer k, consider the graph $G'_k = (V'_k, A'_k)$ obtained by the graph G_k defined in the proof of Theorem 3.1 by letting $V'_k = V_k \cup \{s, t\}$ and $A'_k = A_k \cup \{(s, s_i), (t_i, t) : i = 0, 1, \ldots, k\}$. There is a single commodity (s, t), with unit demand. We call the two arcs (s, s_0) and (t_0, t) of *type D*, and all arcs (s, s_i), (t_i, t) with $i \in [k]$ of *type E* (see also Figure 4 in the appendix).

The latencies of type B and C arcs are exactly the same as in the proof of Theorem 3.1. For arcs of type A we use, for the sake of simplicity, an idealized version of the latency function used in Theorem 3.1; specifically we use

$$\ell_0(x) = \begin{cases} 0, & \text{if } x \leq r_0 \\ 1 - \frac{r_0 + r_1 - x}{r_1}, & \text{if } x > r_0. \end{cases}$$

Although $\ell_0(x)$ is not differentiable in r_0, it can be approximated with arbitrarily small error by standard functions (indeed, the function $\ell_\varepsilon(x)$ used in the proof of Theorem 3.1 is such an approximation).

For fixed L and τ, let $u_{L,\tau}(x)$ be any standard function satisfying $u_{L,\tau}(L) = 0$ and $u_{L,\tau}(L+\tau) = M/\tau$. Type D arcs have latency $u_{r_0,\delta/3k^3}(x)$, and type E arcs have latency $u_{r_1,\delta/3k^3}(x)$. Here (as in Theorem 3.1) $r_0 = (1-\alpha)/2$ and $r_1 = (1+\alpha)/2k$. We will fix the constant δ later in the proof.

The proof of the following lemmas are deferred to the appendix.

Lemma 3.4. $C(o) \leq 1$.

The following lemma will allow us to focus on the case where the "excess flow" on "capacitated" arcs is less than $\delta/3k^3$.

Lemma 3.5. *For any Stackelberg strategy g inducing a Nash flow h, the following holds:*

(i) If a is a type D arc and $g_a + h_a \geq r_0 + \delta/3k^3$, then $C(g+h) \geq M \cdot C(o)$.

(ii) If a is a type E arc and $g_a + h_a \geq r_1 + \delta/3k^3$, then $C(g+h) \geq M \cdot C(o)$.

For the remainder of the proof we assume that there is no arc satisfying the conditions of Lemma 3.5; otherwise the theorem follows immediately.

Lemma 3.6. *For any Stackelberg strategy g inducing a Nash flow h, the following hold:*

(i) For any arc $a = (q_{i-1}, p_i)$, $i \in [k]$, $g_a + h_a \geq r_0 - \delta/k$.

(ii) For any arc $a = (s, s_i)$, $i \in [k]$, $g_a + h_a \geq r_1 - \delta/k$.

We are now ready to conclude the proof of Theorem 3.3.

Proof of Theorem 3.3. For any $i \in [k]$, consider the ith block in the graph (Figure 2). Let g_i, h_i be the Stackelberg and selfish flow on the arc (s_i, t_i), respectively. We have two cases:

1. $h_i = 0$: in this case, using Lemma 3.6, the flow on arc (p_i, q_i) is at least $r_0 - \delta/k + r_1 - \delta/k - g_i$. The latency on that same arc is thus at least $\ell_0(r_0 + r_1 - 2\delta/k - g_i)$.

2. $h_i > 0$: in this case, the Nash flow on path $P_i' = (s, s_i, t_i, t)$ is strictly positive. Consider the path $P_i'' = (s, s_i, p_i, q_i, t_i, t)$. By definition of Nash flow, $\ell_{P_i''}(g+h) \geq \ell_{P_i'}(g+h)$. Notice that the two paths P_i', P_i'' share all their nonzero-latency arcs except for (s_i, t_i) (only present in P_i') and (p_i, q_i) (only present in P_i''). Thus $\ell_{P_i''}(g+h) \geq \ell_{P_i'}(g+h)$ implies $\ell_{(p_i,q_i)}(g+h) \geq \ell_{(s_i,t_i)}(g+h) = 1$. As a consequence, $\ell_{(p_i,q_i)}(g+h) \geq 1 = \ell_0(r_0 + r_1) \geq \ell_0(r_0 + r_1 - 2\delta/k - g_i)$ since g_i and δ/k are nonnegative.

In both cases, $\ell_{(p_i,q_i)}(g+h) \geq \ell_0(r_0 + r_1 - 2\delta/k - g_i) \geq 1 - \frac{g_i + 2\delta/k}{r_1}$.

The latency on the path $P_0' = (s, s_0, p_1, q_1, \ldots, p_k, q_k, t_0, t)$ is at least

$$\ell_{P_0'}(g+h) \geq \sum_{i=1}^{k} \ell_{(p_i,q_i)}(g+h) \geq \sum_{i=1}^{k} \left(1 - \frac{g_i + 2\delta/k}{r_1}\right) \geq k - \frac{\alpha}{r_1} - \frac{2\delta}{r_1} = \left(\frac{1-\alpha-4\delta}{1+\alpha}\right) k.$$

The last inequality is a consequence of the fact that the total Stackelberg flow is α, so $\sum_i g_i \leq \alpha$.

Choosing $\delta < (1-\alpha)/4$, we can conclude that $\ell_{P_0'}(g+h) = \Omega(k)$. Together with Lemma 3.4 and Lemma 3.6, this gives

$$C(g+h) \geq (r_0 - \delta/k) \cdot \ell_{P_0'}(g+h) \geq (\tfrac{1}{2} \cdot (1-\alpha) - \delta) \cdot \Omega(k) = \Omega(k) \cdot C(o).$$

Thus the ratio $C(g+h)/C(o)$ can be made arbitrarily large by picking a sufficiently large k. $\square$

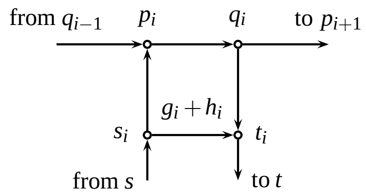

Figure 2: The ith block of the graph $G'(k)$.

4 A General Bicriteria Upper Bound

As we have seen in the previous sections, no Stackelberg strategy controlling a constant fraction of the traffic can reduce the price of anarchy to a constant, even if we consider single-commodity networks. In light of this negative result, we therefore compare the cost of a Stackelberg strategy on an instance $I = (G, r, \ell, \alpha)$ to the cost of an optimal flow for the instance $I^\beta = (G, \beta r, \ell)$ in which the demand vector has been scaled up by a factor $\beta > 1$.

We propose the following simple Stackelberg strategy, which we term *Augmented SCALE (ASCALE)*:

1. Compute an optimal flow o^β for the instance I^β.
2. Define the Stackelberg flow by $g := \frac{\alpha}{\beta} o^\beta$.

We prove that the resulting flow induced by the Stackelberg strategy ASCALE satisfies $C(g+h) \leq C(o^\beta)$ if we choose $\beta = 1 + \sqrt{1-\alpha}$. This result can be seen as a generalization of Roughgarden and Tardos' result that the cost of a Nash flow is always not larger than the cost of the optimal flow for an instance in which demands have been doubled [17]. Our bound gives a smooth transition from absence of centralized control (where doubling the demands is sufficient) to completely centralized control (where no augmentation is necessary). Due to the lack of space, most of the proofs in this section are deferred to the appendix.

Lemma 4.1. *If g is the ASCALE strategy,* $\beta \cdot C(g+h) \leq \sum_{a \in A} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right)$.

Theorem 4.2. *If g is the ASCALE strategy,* $C(g+h) \leq \frac{1}{\beta - 1} \cdot \left(1 - \frac{\alpha}{\beta}\right) \cdot C(o^\beta)$. *Furthermore, this bound is tight.*

Corollary 4.3. *Let* $\beta = 1 + \sqrt{1-\alpha}$. *Then if g is the ASCALE strategy,* $C(g+h) \leq C(o^\beta)$.

The next theorem shows that our result for ASCALE has a consequence for the SCALE strategy as well.

Theorem 4.4. *Let $I = (G, r, \ell, \alpha)$ be an instance and define the modified latency function $\hat{\ell}_a$ by $\hat{\ell}_a(x) = \ell_a(x/\beta)/\beta$ for each arc a, where $\beta = 1 + \sqrt{1-\alpha}$. Let $\hat{g}$ be the SCALE strategy for I, $\hat{h}$ be the Nash flow induced by $\hat{g}$ in $\hat{I} = (G, r, \hat{\ell}, \alpha)$, and o be the optimal flow for (G, r, ℓ). Then, if $\hat{C}(x)$ is the cost of a flow x with respect to the modified latency functions $\hat{\ell}$, we have $\hat{C}(\hat{g} + \hat{h}) \leq C(o)$.*

In the case of M/M/1 latency functions, which are of the form $\ell_a(x) = 1/(u_a - x)$, where u_a intuitively represents the capacity of arc a, the bicriteria bound has a particularly nice interpretation, since $\hat{\ell}_a(x) = \ell_a(x/\beta)/\beta = 1/(\beta(u_a - x/\beta)) = 1/(\beta u_a - x)$. In a purely selfish scenario, this implies that to beat optimal routing it is sufficient to double the capacity of every edge [17]. In the Stackelberg case, Theorem 4.4 shows that it is sufficient to increase the capacities by a factor of $1 + \sqrt{1-\alpha}$ if the SCALE strategy is used.

5 Bounds for Specific Classes of Latency Functions

In the following, we describe a general approach that may be used to analyze the price of anarchy of an opt-restricted Stackelberg strategy. We use this approach to derive upper bounds on the price of anarchy for the SCALE strategy in the case of linear latency functions and polynomial latency functions with nonnegative coefficients.

Lemma 5.1. *For any opt-restricted strategy g, $\sum_{a\in A}(g_a+h_a)\ell_a(g_a+h_a) \le \sum_{a\in A} o_a \ell_a(g_a+h_a)$.*

Proof. Use the variational inequality (1) with $x = o - g$. □

In order to bound the price of anarchy, we use the variational inequality (Lemma 5.1) and bound the cost of the induced flow on every arc by a λ-fraction of the optimal cost plus some ω-fraction of the cost of the induced flow itself:

$$C(g+h) = \sum_{a\in A}(g_a+h_a)\ell_a(g_a+h_a) \le \sum_{a\in A} \lambda \cdot o_a \ell_a(o_a) + \omega(\ell_a; g_a, \lambda)\cdot (g_a+h_a)\ell_a(g_a+h_a). \tag{2}$$

Now, the idea is to determine a λ that provides the tightest bound possible. Choosing $\lambda = 1$, the above approach resembles the one that was previously used by Correa, Schulz, and Stier-Moses [2] to bound the price of anarchy of network routing games; however, optimizing over the parameter λ provides an additional means to obtain better bounds. The idea of introducing the scaling parameter λ was first introduced in the context of bounding the price of anarchy in atomic congestion games (see Harks [6]).

For any latency function ℓ_a and nonnegative numbers g_a, λ, we define the following nonnegative value:

$$\omega(\ell_a; g_a, \lambda) := \sup_{o_a, h_a \ge 0} \frac{o_a}{g_a+h_a} \cdot \frac{\ell_a(g_a+h_a) - \lambda \ell_a(o_a)}{\ell_a(g_a+h_a)}. \tag{3}$$

We assume by convention $0/0 = 0$.

For a given opt-restricted strategy g we further define $\omega(g,\lambda) = \max_{a\in A}\omega(\ell_a; g_a, \lambda)$. Before we state the main theorem, we need one additional definition.

Definition 5.2. Given an opt-restricted strategy g, the *feasible λ-region* is $\Lambda(g) := \{\lambda \in \mathbb{R}_+ \mid \omega(g,\lambda) < 1\}$.

Notice that every $\lambda \in \Lambda(g)$ induces a bound on the price of anarchy.

Theorem 5.3. *Let $\lambda \in \Lambda(g)$. Then $C(g+h) \le \frac{\lambda}{1-\omega(g,\lambda)} C(o)$.*

Proof. The proof follows immediately from (2), Lemma 5.1 and the definition of $\omega(g,\lambda)$. □

SCALE: Linear Latency Functions. In this and the following section, we will analyze the SCALE strategy defined by $g = \alpha o$.

Here, we consider the class of linear latency functions $\mathcal{L}_1 = \{c_1 x + c_0 : c_0, c_1 \ge 0\}$. We first derive a bound on the value $\omega(g,\lambda) = \omega(\alpha o, \lambda)$.

Lemma 5.4. *If $\lambda \in [0,1]$, then $\omega(\alpha o, \lambda) \le \max\left\{\frac{1}{\alpha}(1-\lambda), \frac{1}{4\lambda}\right\}$.*

Proof. Without loss of generality, we can assume that each latency function is a monomial; otherwise, subdivide each arc a into two arcs a_0, a_1 where $\ell_{a_0}(x) = c_0$ and $\ell_{a_1}(x) = c_1 x$. Thus, $\omega(\alpha o, \lambda) = \max_{a\in A}\omega(\ell_a; \alpha o_a, \lambda) \le \sup_{c_0, c_1 \ge 0} \max\{\omega(c_0; \alpha o_a, \lambda), \omega(c_1 x; \alpha o_a, \lambda)\}$.

We start with constant latency functions $\ell_a(x) = c_0$. By definition of ω we get

$$\omega(c_0; \alpha o_a, \lambda) = \sup_{o_a, h_a \ge 0} \frac{o_a}{\alpha o_a + h_a} \cdot \frac{(1-\lambda)c_0}{c_0} = \frac{1}{\alpha}(1-\lambda).$$

For latency functions $\ell_a(x) = c_1 x$, we get

$$\omega(c_1 x; \alpha o_a, \lambda) = \sup_{o_a, h_a \geq 0} \frac{o_a}{\alpha o_a + h_a} \cdot \frac{c_1(\alpha o_a + h_a - \lambda o_a)}{c_1(\alpha o_a + h_a)}.$$

Define $\mu := \frac{h_a}{o_a}$ if $o_a > 0$ and zero otherwise. Then

$$\omega(c_1 x; \alpha o_a, \lambda) = \sup_{\mu \geq 0} \frac{1}{\alpha + \mu} \cdot \frac{\alpha + \mu - \lambda}{\alpha + \mu} = \sup_{0 < t \leq 1/\alpha} t(1 - \lambda t) \leq \frac{1}{4\lambda}.$$

□

We are now prepared to derive an upper bound on the price of anarchy.

Theorem 5.5 (Karakostas and Kolliopoulos [8])**.** *The price of anarchy of the SCALE strategy in the case of linear latency functions is at most* $\frac{(1+\sqrt{1-\alpha})^2}{2(1+\sqrt{1-\alpha})-1}$.

Proof. Let $\lambda = \frac{1}{2}(1 + \sqrt{1-\alpha})$. Then, by Lemma 5.4, $\omega(\alpha o, \lambda) \leq \frac{1}{2(1+\sqrt{1-\alpha})} < 1$. Thus, $\lambda \in \Lambda(\alpha o)$, and by Theorem 5.3,

$$\frac{C(g+h)}{C(o)} \leq \frac{\lambda}{1 - \omega(\alpha o, \lambda)} = \frac{1 + \sqrt{1-\alpha}}{2} \cdot \frac{2(1 + \sqrt{1-\alpha})}{2(1 + \sqrt{1-\alpha}) - 1}.$$

Rewriting proves the claim. □

We next present a family of instances that pointwise match the upper bound of Theorem 5.5 for infinitely many values of α. More precisely, the lower bound is matched for all values of α such that $1/\sqrt{1-\alpha}$ is an integer. To the best of our knowledge, this is the first tight bound for values of $\alpha \neq 0, 1$.

Theorem 5.6. *Let $n \geq 2$ be an integer and let $c = 1 - (n-1)\alpha/n$. Then, the price of anarchy of the SCALE strategy for linear latency functions is at least $\frac{nc^2 + (n-1)\alpha c}{(n-1)c + 1/n}$. Moreover, for all $\alpha = 1 - 1/k^2$, with k a positive integer, there exists an n such that the corresponding bound matches the upper bound of Theorem 5.5.*

The proof of Theorem 5.6 can be found in the appendix.

SCALE: Polynomial Latency Functions. In this section, we consider the class $\mathcal{L}_d$ of polynomials with nonnegative coefficients and degree at most $d \in \mathbb{N}$: $\mathcal{L}_d := \{c_d x^d + \cdots + c_1 x + c_0 \,:\, c_s \geq 0, s = 0, \ldots, d\}$. Similarly to the previous section, we start by bounding $\omega(\ell_a; \alpha o_a, \lambda)$ when $\ell_a(x)$ is a monomial in x.

Lemma 5.7. *For all $s \in \{0, \ldots, d\}$, $\omega(c_s x^s; \alpha o_a, \lambda) = \omega_s(\lambda)$, where*

$$\omega_s(\lambda) := \begin{cases} \frac{1}{\alpha}\left(1 - \frac{\lambda}{\alpha^s}\right), & \text{if } \lambda \leq \frac{\alpha^s}{s+1} \\ \frac{s}{s+1} \cdot \frac{1}{((s+1)\lambda)^{1/s}}, & \text{if } \lambda \geq \frac{\alpha^s}{s+1}. \end{cases}$$

Proof. By definition of ω (Eq. (3)),

$$\begin{aligned} \omega(c_s x^s; \alpha o_a, \lambda) &= \sup_{o_a, h_a \geq 0} \frac{o_a}{\alpha o_a + h_a} \cdot \frac{c_s(\alpha o_a + h_a)^s - c_s \lambda o_a^s}{c_s(\alpha o_a + h_a)^s} \\ &= \sup_{\mu \geq 0} \frac{1}{\alpha + \mu} \cdot \left(1 - \frac{\lambda}{(\alpha + \mu)^s}\right) = \sup_{0 < t \leq 1/\alpha} t(1 - \lambda t^s). \end{aligned}$$

The supremum in the last expression can be attained either at an extreme point ($t = 0$ or $t = 1/\alpha$) or at the local maximum point $t^* = ((s+1)\lambda)^{-1/s}$. If $1/\alpha \leq t^*$, the global maximum will be at $1/\alpha$ and will have a value of $\frac{1}{\alpha}(1 - \frac{\lambda}{\alpha^s})$; otherwise, the global maximum is at t^* and its value is $\frac{s}{s+1} \cdot ((s+1)\lambda)^{-1/s}$. The condition $1/\alpha \leq t^*$ can be rewritten as $\lambda \leq \alpha^s/(s+1)$ through straightforward manipulations. □

Corollary 5.8. $\omega(\alpha o, \lambda) \le \max_{0 \le s \le d} \omega_s(\lambda)$.

Proof. As in the proof of Lemma 5.4, we use the fact that

$$\omega(\alpha o, \lambda) = \max_{a \in A} \omega(\ell_a; \alpha o_a, \lambda) \le \sup_{c_0, c_1, \ldots, c_d \ge 0} \max\{\omega(c_0; \alpha o_a, \lambda), \ldots, \omega(c_d x^d; \alpha o_a, \lambda)\}.$$

The claim then follows by Lemma 5.7. □

The proof of the following lemma can be found in the appendix.

Lemma 5.9. *Let $s \in [d]$. There is a unique $\lambda \in (0,1)$ such that $\omega_s(\lambda) = \omega_0(\lambda)$; call it λ_s. Then:*

(i) $\lambda_s = z_s^s/(s+1)$, *where $z_s > \alpha$ is the unique solution to the equation*

$$z^{s+1} - (s+1)z + \alpha s = 0; \tag{4}$$

(ii) $\lambda_s > 1/(s+1)$;

(iii) $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$;

(iv) $\omega_s(\lambda_d) \le \omega_d(\lambda_d) = \omega_0(\lambda_d)$.

Theorem 5.10. *The price of anarchy of the SCALE strategy in the case of latency functions in the class $\mathcal{L}_d$ is at most*

$$\frac{(d+1)z_d - \alpha d}{(d+1)z_d - d}$$

where $z_d \ge 1$ is the unique solution of the equation $z^{d+1} - (d+1)z + \alpha d = 0$.

Proof. We will use Theorem 5.3 with $\lambda = \lambda_d$. However, in order to apply the theorem, we first need to upper bound $\omega(\alpha o, \lambda_d)$.

Using Corollary 5.8, we know that $\omega(\alpha o, \lambda_d) \le \max_{0 \le s \le d} \omega_s(\lambda_d)$. Notice that by Lemma 5.9 (i,iv) and by definition of ω_d, $\max_{0 \le s \le d} \omega_s(\lambda_d) = \omega_d(\lambda_d) = \frac{d}{d+1} \cdot ((d+1)\lambda_d)^{-1/d} = \frac{d}{d+1} \cdot z_d^{-1} < 1$. This implies $\lambda_d \in \Lambda(\alpha o)$ and we can invoke Theorem 5.3 to obtain a bound on the price of anarchy given by

$$\frac{\lambda_d}{1 - \omega(\alpha o, \lambda_d)} = \frac{z_d^d/(d+1)}{1 - \frac{d}{d+1} z_d^{-1}} = \frac{z_d^{d+1}}{(d+1)z_d - d} = \frac{(d+1)z_d - \alpha d}{(d+1)z_d - d}.$$

□

A lower bound for polynomial latency functions of degree d can be obtained by generalizing the construction used in Theorem 5.6. We use again the network of Figure 5 (see appendix), except that we replace everywhere the latency function x by x^d and the constant c by $(1-(n-1)\alpha/n)^d$. The optimal flow is still split evenly on the direct paths, so that with similar arguments we obtain the following lower bound.

Theorem 5.11. *Let $n \ge 2$ be an integer and let $c = (1-(n-1)\alpha/n)^d$. Then, the price of anarchy of the SCALE strategy for latency functions in the class $\mathcal{L}_d$ is at least $\left(nc^{1+1/d} + (n-1)\alpha c\right) / \left((n-1)c + n^{-d}\right)$.*

Notice that the theorem does not fix n, so it is possible to optimize n based on α as in Theorem 5.6. For latency functions in $\mathcal{L}_2$ and $\mathcal{L}_3$, we compare in Figure 6 (in the appendix) the lower bound thus obtained with the upper bound of Theorem 5.10.

References

[1] D. Braess. Über ein Paradoxon der Verkehrsplanung. *Unternehmenforschung*, 11:258–268, 1968.

[2] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses. Selfish routing in capacitated networks. *Mathematics of Operations Research*, 29:961–976, 2004.

[3] J. R. Correa and N. E. Stier-Moses. Stackelberg routing in atomic network games. Technical report, Columbia Business School, February 2007.

[4] S. Dafermos and F. Sparrow. The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Series B*, 73:91–118, 1969.

[5] P. Dubey. Inefficiency of Nash Equilibria. *Mathematics of Operations Research*, 11:1–8, 1986.

[6] T. Harks. On the price of anarchy of network games with nonatomic and atomic players. Technical report, avalaible at Optimization Online, January 2007.

[7] A. Kaporis and P. Spirakis. The price of optimum in Stackelberg games on arbitrary single commodity networks and latency functions. In *Proc. of the 18th annual ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pages 19–28, New York, NY, USA, 2006. ACM Press.

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

[9] 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.

[10] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In C. Meines and S. Tison, editors, *Proc. of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 1563 of *LNCS*, pages 404–413. Springer, 1999.

[11] V. S. A. Kumar and M. V. Marathe. Improved results for Stackelberg scheduling strategies. In *Proc. of the 33rd International Colloquium of Automata, Languages and Programming (ICALP)*, pages 776–787, 2002.

[12] G. Perakis. The price of anarchy when costs are non-separable and asymmetric. In D. Bienstock and G. Nemhauser, editors, *Proc. of the 10th International Conference on Integer Programming and Combinatorial Optimization (IPCO)*, volume 3064 of *Lecture Notes in Computer Science*, pages 46–58. Springer, Berlin Heidelberg, 2004.

[13] T. Roughgarden. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67:341–364, 2002.

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

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

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

[17] T. Roughgarden and E. Tardos. How bad is selfish routing? *Journal of the ACM*, 49(2):236–259, 2002.

[18] Y. Sharma and D. Williamson. Stackelberg thresholds in network routing games or the value of altruism. In *Proc. of the 8th ACM conference on Electronic Commerce*, pages 93–102, 2007.

[19] C. Swamy. The effectiveness of Stackelberg strategies and tolls for network congestion games. In *Proc. of the 18th annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1133–1142, 2007.

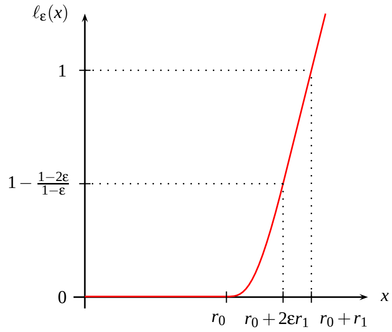

Figure 3: The latency function $\ell_\varepsilon(x)$ used in the proof of Theorem 3.1.

A Proofs of Section 3

Proof of Lemma 3.4. Let P'_0 be the path $(s, s_0, p_1, q_1, p_2, \ldots, p_k, q_k, t_0, t)$, and for $i \in [k]$, let P'_i be the path (s, s_i, t_i, t). Consider the feasible flow $\bar{f}$ such that $\bar{f}_{P'_0} = r_0$ and $\bar{f}_{P'_i} = r_1$ for $i \in [k]$. The latency induced by $\bar{f}$ is 0 on arcs of type A, C, D, E and 1 on arcs of type B. So $C(o) \le C(\bar{f}) = k \cdot r_1 = (1+\alpha)/2 \le 1$. □

Proof of Lemma 3.5. Consider for example (i). We have

$$\begin{aligned} C(g+h) &\ge (g_a + h_a) \cdot \ell_a(g_a + h_a) = (g_a + h_a) \cdot u_{r_0, \delta/3k^3}(g_a + h_a) \\ &\ge (r_0 + \delta/3k^3) \cdot M/(\delta/3k^3) \ge M \ge M \cdot C(o), \end{aligned}$$

where the last inequality follows from Lemma 3.4. Part (ii) is proved similarly. □

Proof of Lemma 3.6. Regarding (i), we will prove by induction on i the stronger claim

$$g_a + h_a \ge r_0 - (2i+1)\delta/3k^2.$$

For $i = 1$, notice that by Lemma 3.5 the flow along each of $(s, s_1), \ldots, (s, s_k)$ is at most $r_1 + \delta/3k^3$, so the flow on (s, s_0) must be at least $1 - \sum_{i=1}^{k} \left(r_1 + \delta/3k^3\right) = 1 - kr_1 - \delta/3k^2 = r_0 - \delta/3k^2$. But the flow on (s, s_0) is the same as that on arc $(s_0, p_1) = (q_0, p_1)$. Notice that a similar argument allows also to conclude that the flow on each (s, s_i) arc ($i \in [k]$) is at least $r_1 - \delta/3k^2$. This implies (ii) for all $i \in [k]$.

To prove (i) for $i > 1$, consider the ith block in the graph (Figure 2) and let $f = g + h$.

By flow conservation, $f_{(q_i, p_{i+1})} = f_{(q_{i-1}, p_i)} + f_{(s, s_i)} - f_{(t_i, t)}$. Using induction and Lemma 3.5,

$$\begin{aligned} f_{(q_i, p_{i+1})} &= f_{(q_{i-1}, p_i)} + f_{(s, s_i)} - f_{(t_i, t)} \\ &\ge (r_0 - (2i-1)\delta/3k^2) + (r_1 - \delta/3k^2) - (r_1 + \delta/3k^3) = r_0 - (2i+1)\delta/3k^2. \end{aligned}$$

□

B Proofs of Section 4

Proof of Lemma 4.1. Consider the flow $(1-\alpha)g/\alpha$; it is a flow feasible with respect to $(1-\alpha)r$. Using the variational inequality (1), we get

$$\sum_{a \in A} h_a \ell_a(g_a + h_a) \le \frac{1-\alpha}{\alpha} \sum_{a \in A} g_a \ell_a(g_a + h_a).$$

Adding $\sum_a g_a \ell_a(g_a+h_a)$ to both sides and using $g = \frac{\alpha}{\beta} o^\beta$, we obtain

$$C(g+h) \le \frac{1}{\alpha} \sum_{a\in A} \frac{\alpha}{\beta} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) = \frac{1}{\beta} \sum_{a\in A} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right).$$

□

Proof of Theorem 4.2. We first show that for *every* arc $a \in A$,

$$o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) \le \left(\frac{\alpha}{\beta} o_a^\beta + h_a\right)\ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) + \left(1-\frac{\alpha}{\beta}\right) o_a^\beta \ell_a(o_a^\beta). \tag{5}$$

There are two cases. When $\frac{\alpha}{\beta} o_a^\beta + h_a \ge o_a^\beta$, the inequality holds simply because its left hand side is upper bounded by the first summand of the right hand side. Otherwise, if $o_a^\beta > \frac{\alpha}{\beta} o_a^\beta + h_a$,

$$\begin{aligned} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) &\le \left(\frac{\alpha}{\beta} o_a^\beta + h_a + o_a^\beta - \frac{\alpha}{\beta} o_a^\beta\right)\ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) \\ &\le \left(\frac{\alpha}{\beta} o_a^\beta + h_a\right)\ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) + \left(1-\frac{\alpha}{\beta}\right) o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) \\ &\le \left(\frac{\alpha}{\beta} o_a^\beta + h_a\right)\ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) + \left(1-\frac{\alpha}{\beta}\right) o_a^\beta \ell_a(o_a^\beta). \end{aligned}$$

Summing (5) over all $a \in A$, we obtain $\sum_{a\in A} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) \le C(g+h) + \left(1-\frac{\alpha}{\beta}\right) C(o^\beta)$. Invoking Lemma 4.1 we get

$$\beta \cdot C(g+h) \le \sum_{a\in A} o_a^\beta \ell_a\left(\frac{\alpha}{\beta} o_a^\beta + h_a\right) \le C(g+h) + \left(1-\frac{\alpha}{\beta}\right) C(o^\beta).$$

Solving for $C(g+h)$ now gives the bound as claimed. The bound is also tight, as can be seen by considering a slightly modified Pigou instance (we omit the details). □

Proof of Theorem 4.4. Observe that the SCALE strategy for I can be obtained by computing the ASCALE strategy for $I^{1/\beta} := (G, r/\beta, \ell, \alpha)$ and scaling it up by a factor of β; that is, $\hat{g} = \beta g$, where g is the ASCALE strategy for $I^{1/\beta}$. Let h be the Nash flow induced by g in $I^{1/\beta}$. By the variational inequality (1),

$$\sum_{a\in A} h_a \ell_a(g_a+h_a) \le \sum_{a\in A} y_a \ell_a(g_a+h_a) \tag{6}$$

for any flow y feasible for $(1-\alpha)r/\beta$. Since $\ell_a(x/\beta)/\beta = \hat{\ell}_a(x)$, we can rewrite (6) as $\sum_a (\beta h_a)\hat{\ell}_a(\hat{g}_a + \beta h_a) \le \sum_a (\beta y_a)\hat{\ell}_a(\hat{g}_a + \beta h_a)$. This implies that βh is a Nash flow induced by $\hat{g}$ in $\hat{I}$. Since the cost of Nash flows is unique, $\hat{C}(\hat{g}+\beta h) = \hat{C}(\hat{g}+\hat{h})$. Finally, since $\hat{C}(\beta x) = C(x)$ for any flow x, we can conclude $\hat{C}(\hat{g}+\hat{h}) = \hat{C}(\beta(g+h)) = C(g+h) \le C(o)$ where the inequality follows from Corollary 4.3. □

C Proofs of Section 5

Proof of Theorem 5.6. We use the network depicted in Figure 5. There is a single commodity (s,t) with unit demand. In the optimal flow the demand is split evenly among the paths (s, p_i, q_i, t), $i \in [n]$. The resulting cost is $C(o) = (n-1)c + 1/n$.

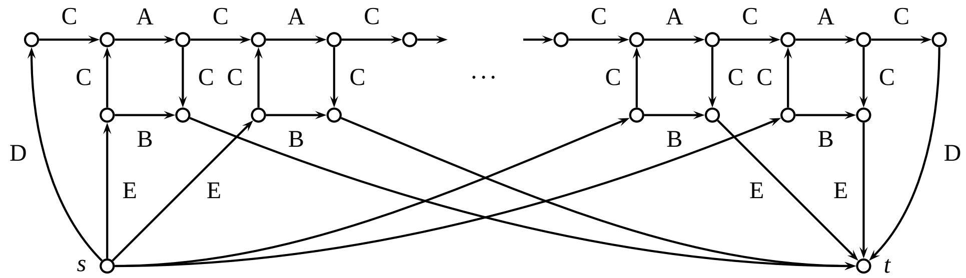

Figure 4: The graph G'_k, used in the proof of Theorem 3.3. Arcs are labeled with their type.

The SCALE strategy sends a flow of value α/n along each direct path (s, p_i, q_i, t), $i \in [n]$. Due to the condition $c = 1-(n-1)\alpha/n$, the Nash flow is sent along the zigzag path $(s, p_1, q_1, p_2, \dots, p_n, q_n, t)$. Thus, the cost of the combined flow $g+h$ is given by

$$C(g+h) = n\Big(1 - \frac{n-1}{n}\alpha\Big)^2 + (n-1)\alpha c = nc^2 + (n-1)\alpha c$$

and the bound follows.

To see that the bound is tight when $\alpha = 1-1/k^2$, pick $n = k+1 = 1+1/\sqrt{1-\alpha}$. After substituting the expressions for n and c into the bound and appropriate rewriting we obtain the same expression as in Theorem 5.5. $\square$

Proof of Lemma 5.9. Let λ be such that $\omega_s(\lambda) = \omega_0(\lambda)$. Assuming that $\lambda \le \alpha^s/(s+1)$, $\omega_s(\lambda) = \omega_0(\lambda)$ is equivalent to $\frac{1}{\alpha}(1-\frac{\lambda}{\alpha^s}) = \frac{1}{\alpha}(1-\lambda)$, which is impossible for $\alpha \in (0,1)$. Thus $\lambda > \alpha^s/(s+1)$ and $\frac{s}{s+1}((s+1)\lambda)^{-1/s} = \frac{1}{\alpha}(1-\lambda)$. If we substitute $\lambda = z^s/(s+1)$, the last relation becomes equivalent to (4). On the other hand, λ_s as defined in (i) satisfies $\omega_s(\lambda_s) = \omega_0(\lambda_s)$. To check that equation (4) has indeed exactly one solution larger than α, use for example Descartes' rule of signs.

Part (ii) follows from (i) and the fact that equation (4) has exactly one solution larger than 1 (again by Descartes' rule of signs), implying $z_s > 1$.

To prove (iii), we now view λ_s, as given in (i), as a real function of a real variable, i.e. $\lambda_s = \lambda(s)$. To show (iii) it is then enough to argue that $\lambda'(s) \le 0$ for $s \in [1,d]$. In the following we omit the dependency of λ and λ' from s to improve readability.

As seen earlier in the proof, the equation (4) is equivalent to $\frac{s}{s+1} \cdot \frac{1}{((s+1)\lambda)^{1/s}} = \frac{1}{\alpha}(1-\lambda)$, which is in turn equivalent to the identity

$$\lambda^{1/s}(1-\lambda) = \frac{\alpha s}{(s+1)^{1+1/s}}. \tag{7}$$

If we differentiate both sides of (7), we obtain

$$\lambda^{1/s}(1-\lambda)\left(-\frac{1}{s^2}\ln\lambda + \frac{1}{s\lambda}\lambda'\right) - \lambda^{1/s}\lambda' = \frac{\ln(s+1)}{s^2}\frac{\alpha s}{(s+1)^{1+1/s}}.$$

Rearranging terms and substituting $\frac{\alpha s}{(s+1)^{1+1/s}}$ with $\lambda^{1/s}(1-\lambda)$, (which is valid by (7)),

$$\lambda^{1/s}\lambda' \cdot \frac{1-(s+1)\lambda}{s\lambda} = \frac{1}{s^2}\lambda^{1/s}(1-\lambda)\ln\lambda + \frac{1}{s^2}\lambda^{1/s}(1-\lambda)\ln(s+1)$$

and finally after further simplifications

$$\lambda' \cdot \frac{1-(s+1)\lambda}{s\lambda} = \frac{1}{s^2}(1-\lambda)\ln((s+1)\lambda).$$

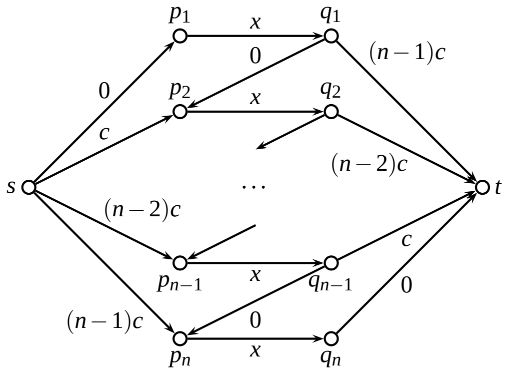

Figure 5: The network used in the proof of Theorem 5.6. Arcs are labeled with their latency function.

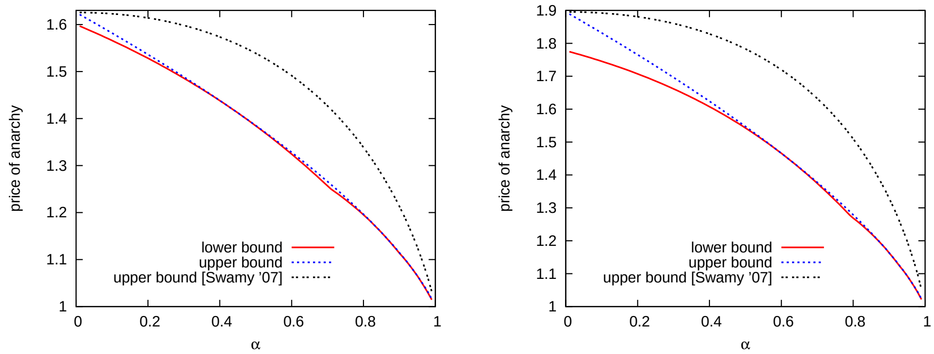

Figure 6: Upper vs. lower bounds for SCALE for latency functions in $\mathcal{L}_2$ (left) and $\mathcal{L}_3$ (right). The plots also show the previously best upper bound by Swamy [19].

Using part (ii) of the lemma, the right hand side has to be positive, while the term multiplying λ' in the left hand side has to be negative, again by part (ii). Thus, $\lambda' < 0$.

To prove part (iv), first notice that the functions ω_s and ω_0 are both continuous and monotonically decreasing, and in the interval $(0,1)$, they intersect exactly once (in λ_s). Since $\omega_s(\varepsilon) < \omega_0(\varepsilon)$ for all sufficiently small $\varepsilon > 0$, we have $\omega_s(\lambda) \leq \omega_0(\lambda)$ for all $\lambda \leq \lambda_s$. But by (iii), $\lambda_d \leq \lambda_s$, so $\omega_s(\lambda_d) \leq \omega_0(\lambda_d)$. Finally, $\omega_0(\lambda_d) = \omega_d(\lambda_d)$ by definition of λ_d. □