Counter-Intuitive Throughput Behaviors in Networks Under End-to-End Control

Ao Tang, *Student Member, IEEE*, Jiantao Wang, *Student Member, IEEE*, and Steven H. Low, *Senior Member, IEEE*

***Abstract*—It has been shown that as long as traffic sources adapt their rates to aggregate congestion measure in their paths, they implicitly maximize certain utility. In this paper we study some counter-intuitive throughput behaviors in such networks, pertaining to whether a fair allocation is always inefficient and whether increasing capacity always raises aggregate throughput. A bandwidth allocation policy can be defined in terms of a class of utility functions parameterized by a scalar α that can be interpreted as a quantitative measure of fairness. An allocation is *fair* if α is large and *efficient* if aggregate throughput is large. All examples in the literature suggest that a fair allocation is necessarily inefficient. We characterize exactly the tradeoff between fairness and throughput in general networks. The characterization allows us both to produce the first counter-example and trivially explain all the previous supporting examples. Surprisingly, our counter-example has the property that a fairer allocation is *always* more efficient. In particular it implies that maxmin fairness can achieve a higher throughput than proportional fairness. Intuitively, we might expect that increasing link capacities always raises aggregate throughput. We show that not only can throughput be reduced when some link increases its capacity, more strikingly, it can also be reduced when *all* links increase their capacities by the same amount. If all links increase their capacities proportionally, however, throughput will indeed increase. These examples demonstrate the intricate interactions among sources in a network setting that are missing in a single-link topology.**

***Index Terms*—Fairness, flow control, optimization, throughput.**

I. Introduction

A. Motivation

A central issue in networking is how to allocate bandwidth to flows *efficiently* and *fairly*, in a decentralized manner. A series of recent work, e.g., [18], [22]–[25], [28], [34], has shown that a bandwidth allocation policy can be expressed in terms of a utility function $U_i(x_i)$ in the sense that the desired bandwidth allocation $x^* = (x_i^*, \text{ all sources } i)$ solves the utility maximization problem formulated in [18]:

$$\max_x \sum_i U_i(x_i) \quad \text{subject to link capacity constraints.} \qquad (1)$$

It is remarkable that as long as traffic sources adapt their rates to the aggregate congestion measure in their paths, they are implicitly maximizing some utility.[1] Hence, the optimization problem (1) is a convenient characterization of many networks under end-to-end control, such as networks controlled by Transmission Control Protocol (TCP). Indeed, one can derive the utility functions that are implicitly solved by the various TCP congestion control algorithms proposed in the literature [23], [25]. In this paper we study some counter-intuitive behaviors of such networks.

A common figure of merit is the aggregate throughput $\sum_i x_i^*$. It measures the total traffic through the network and, in this sense, the efficiency of the bandwidth allocation policy under which the network operates.[2] If users pay a constant charge per unit bandwidth, the aggregate throughput is also proportional to network revenue. How do we balance fairness and aggregate throughput in designing bandwidth allocation policies? Will adding additional link capacities necessarily result in higher aggregate throughput? There are many examples in the literature that point to an inevitable tradeoff between fairness and aggregate throughput (see Section IV-A and IV-B), yet, we cannot find a general theorem clarifying this folklore. In this paper we provide an answer to both problems. The second problem also provides insight on where to invest capacity to best improve network efficiency or revenue.

Our emphasis is on general networks with multiple links and on mathematical analysis. Often, interesting and counter-intuitive behaviors arise only in a network setting where sources interact through multiple shared links in intricate and surprising ways. Such behaviors are absent in a single-link topology. The discovery of some of these behaviors, and their explanation, demonstrate that the formal approach taken here is indeed necessary and rewarding.

We now summarize our main results.

B. Summary

Suppose all policies (utility functions U_i) are parameterized by a common scalar $\alpha \geq 0$. We first derive explicit expressions for the changes in throughputs and prices when the parameter α or the capacities change (Theorem 2). Here, the prices are the Lagrange multipliers of the utility maximization problem [24]. This result applies to general utility functions.

We then specialize to a particular class of utility functions proposed in [28] that turns out to characterize various TCP variants and includes various fairness notions proposed in the networking literature as special cases (see Section IV-A). Following [3] and [27], we interpret the parameter α as a *quantitative* measure of fairness in the sense that a fairer allocation is one with a larger α. All examples we can find in the literature indicate that a fair allocation is necessarily inefficient, i.e.,

Manuscript received October 21, 2004; approved by IEEE/ACM Transactions on Networking Editor N. Shroff.

The authors are with the EAS Division, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: aotang@caltech.edu; jiantao@caltech.edu; slow@caltech.edu).

Digital Object Identifier 10.1109/TNET.2006.872552

[1]Provided that they all react to the same congestion price, possibly in different ways, see [32] otherwise.

[2]This notion of efficiency is different from Pareto optimality commonly used in economics.

an allocation policy with a larger α seems to always lead to a smaller aggregate throughput. We characterize exactly the tradeoff between fairness and throughput in general networks (Theorem 5). The characterization allows us both to produce the first counter-example (Theorem 8) and trivially explain all the previous supporting examples (Corollary 6). Surprisingly, in our counter-example, a fairer allocation is *always* more efficient. In particular it implies that maxmin fairness can achieve a higher aggregate throughput than proportional fairness.

Intuitively, we might expect that the aggregate throughput will always rise as long as some links increase their capacities and no links decrease theirs. This turns out not to be the case, and we characterize exactly the condition under which this is true (Theorem 9). Not only can the aggregate throughput be reduced when some link increases its capacity, more strikingly, it can also be reduced even when *all* links increase their capacities by the same amount (Theorem 11). Moreover, this holds under all bandwidth allocation policies (fairness α). This paradoxical result seems less surprising in retrospect: raising link capacities always increases the aggregate utility, but mathematically there is no *a priori* reason that it should also increase the aggregate throughput. If all links increase their capacities proportionally, however, the aggregate throughput will indeed increase, for the class of utility functions proposed in [28] (Theorem 12).

It is well-known that counter-intuitive behavior can arise in a distributed system where agents optimize their own objectives. The earliest and the most famous example is the Braess paradox in transportation networks, discovered theoretically in 1968 (see, e.g., [4], [11], [29]) and widely verified in the real world years later, e.g., [7]. There the *addition* of a new road segment to a network changes the traffic flow to a new equilibrium in which every car incurs a *longer* travel time. Subsequent paradoxes have been discovered in mechanical and electrical networks [8], in queueing networks [1], [6], [9], [20], [21], and in computer systems [16], [17]. Even though our results have the same flavor, they differ in important ways from the Braess paradox. First, in Braess paradox, the performance degradation is due to misalignment of individual and social optimalities. In our case, even though TCP sources also selfishly optimize their own net benefits (e.g., [24]), congestion prices align their individual optimalities with social optimality. The performance degradation in our case is due to misalignment of two social objectives (utility maximization versus throughput maximization). Second, in Braess paradox, the addition of new capacity (path) leads to degraded performance for *all* flows, and hence the new equilibrium point is not Pareto optimal. In our case, all equilibrium points are Pareto optimal, and hence some flows are worse off and some better off in the new equilibrium point. The conventional wisdom is that the addition of capacity should change the equilibrium to a new point where the aggregate throughput should be at least as high. Our results show that while this is true for simple networks, it does not always hold in general networks. Finally, examples of Braess paradox always involve the addition of new paths and flows that re-route to maximize their own objectives. In our case, only link capacities are changed, while network topology and routing are fixed.

The paper is organized as follows. We describe in Section II our model. We establish in Section III the basic results on the effect of parameter α and link capacities c on equilibrium throughputs and prices. We study in Section IV whether a fair allocation is always inefficient, and in Section V whether increasing capacity always raises aggregate throughput. We conclude in Section VI with limitations of this work. Partial and preliminary results have appeared in [31], which also provides alternative proofs for some of the results in Section IV.

II. MODEL

Consider a network of L links, indexed by l, with finite capacities c_l. It is shared by N sources, indexed by i. Let R be the $L \times N$ routing matrix: $R_{li} = 1$ if source i uses link l and 0 otherwise. Let x_i be the transmission rate of source i, and $U_i(x_i;\alpha)$ be its utility as a function of its rate x_i. Suppose all the utility functions $U_i(x_i;\alpha)$ are parameterized by a scalar $\alpha \geq 0$. Suppose $U_i(x_i;\alpha)$ are concave in x_i for $\alpha \geq 0$ and strictly concave when $\alpha > 0$. When α and c are clear from the context, we may use $U_i(x_i)$ in place of $U_i(x_i;\alpha)$. In general, z denotes the vector $z = (z_1, \ldots, z_n)^T$ when z_i are previously defined, and ^{T} denotes transpose. We use log to denote natural logarithm.

Consider the utility maximization problem defined in [18]:

$$\max_{x\geq 0} \sum_i U_i(x_i;\alpha) \quad \text{subject to} \quad Rx \leq c \tag{2}$$

and its Lagrangian dual [24]:

$$\min_{p\geq 0} \sum_i \max_{x_i \geq 0} \left(U_i(x_i;\alpha) - x_i \sum_l R_{li} p_l \right) + \sum_l c_l p_l. \tag{3}$$

By duality theory, a maximizer $x = x(\alpha, c)$ for (2) and a minimizer (Lagrange multiplier) $p = p(\alpha, c)$ for (3) exist for $\alpha \geq 0$, $c > 0$. Moreover, x is unique if $\alpha > 0$ when the utility functions are strictly concave. We will call a dual optimal solution p the link *prices* (see e.g., [24] for interpretation).

Unless otherwise specified, we will assume that $\alpha > 0$ so that the utility functions are *strictly* concave and the optimal solution denoted by $x = x(\alpha, c)$ is unique. The aggregate throughput $T = T(\alpha, c)$ is defined in terms of the unique solution

$$T(\alpha, c) := \sum_i x_i(\alpha, c). \tag{4}$$

From Lemma 1, $x(\alpha, c)$ is a continuous function of α and c. Moreover, $x(\alpha, c)$ is differentiable except at a finite number of points when the active constraint set at optimal $x(\alpha, c)$ changes as α or c is perturbed. Hence, we can study $\partial T/\partial\alpha$ and $\partial T/\partial c$ in between these points.[3] For the rest of the paper, we will thus focus on the utility maximization with equality constraints that represent only those constraints that are active at optimality:

$$\max_{x_i \geq 0} \sum_i U_i(x_i;\alpha) \quad \text{s.t.} \quad Rx = c. \tag{5}$$

In this case the dual problem (3) should be interpreted as the Lagrangian dual of (2) with a possibly reduced R, as opposed

[3]Hence, all our statements below on $\partial T/\partial\alpha$ and $\partial T/\partial c$ should be interpreted piecewise in between nondifferentiable points of α and c.

to the dual of (5).[4] If every link has a single-link flow, then all constraints are necessarily tight.

Suppose the $L \times N$ routing matrix R has full row rank. Suppose $N \geq L$ and let $M = N - L$ be the difference between the number of sources and the number of links. Then M is the dimension of the null space of R. Let $(z_m, m = 1, \ldots, M)$, $z_m \in \Re^N$, be any basis of the null space of R, and let $Z = [z_1\ z_2\ \ldots\ z_M]$ be the matrix with z_m as its columns. Let $V = V(x;\alpha) := \sum_i U_i(x_i;\alpha)$ be the aggregate utility function. Let $D = D(\alpha, c)$ denote the curvature of the aggregate utility function

$$D := -\frac{\partial^2 V}{\partial x^2} \tag{6}$$

and $b = b(\alpha, c)$ be

$$b := \frac{\partial^2 V}{\partial x \partial \alpha} \tag{7}$$

at the optimal allocation $x = x(\alpha, c)$.

III. Basic Results

We start with an important property of $x(\alpha, c)$, quoted directly from [33].

Lemma 1: For any $\alpha > 0$, $c > 0$, the unique solution $x(\alpha, c)$ of (5) is continuous and differentiable at (α, c).

Here, we have used the assumption that the (α, c) considered throughout this paper is such that the active constraint set is unchanged when α or c is perturbed locally [i.e., we consider problem (5) instead of problem (2)], so that R is independent of α.

The basic results on how throughputs and prices vary as the utility parameter α and capacity c change are given in the next theorem. These results apply to general utility functions. In the next two sections, we will specialize to a particular class of utility functions to study throughput–fairness tradeoff and whether increasing capacity always raises aggregate throughput.

Theorem 2: The optimal rates $x = x(\alpha, c)$ of (5) and optimal prices $p = p(\alpha, c)$ of (3) satisfy the following equations:

$$\begin{aligned}
\frac{\partial x}{\partial \alpha} &= \left(D^{-1} - D^{-1}R^T(RD^{-1}R^T)^{-1}RD^{-1}\right) b \\
\frac{\partial x}{\partial c} &= D^{-1}R^T(RD^{-1}R^T)^{-1} \\
\frac{\partial p}{\partial \alpha} &= (RD^{-1}R^T)^{-1}RD^{-1}b \\
\frac{\partial p}{\partial c} &= -(RD^{-1}R^T)^{-1}
\end{aligned}$$

where matrix D and vector b are defined in (6) and (7), respectively.

Proof: At the optimal point, the Karush–Kuhn–Tucker condition holds. We have

$$Rx = c \quad \text{and} \quad R^T p - \frac{\partial V}{\partial x} = 0. \tag{8}$$

[4]To be precise, the routing matrix R in both (5) and (3) should be denoted $R(\alpha, c)$ to indicate that it consists only of "bottleneck" links at the given (α, c), but we will abuse notation to use R to denote both the full routing matrix and the reduced matrix when there is no danger of confusion.

Define

$$y := \begin{pmatrix} x \\ p \end{pmatrix} \quad w := \begin{pmatrix} c \\ \alpha \end{pmatrix}$$

and

$$G(w, y) = \begin{pmatrix} Rx - c \\ R^T p - \frac{\partial V}{\partial x} \end{pmatrix}.$$

Then (8) can be rewritten as $G(w, y) = 0$. The derivatives of function G are

$$\begin{aligned}
\frac{\partial G}{\partial y} &= \begin{pmatrix} R & 0 \\ -\frac{\partial^2 V}{\partial x^2} & R^T \end{pmatrix} = \begin{pmatrix} R & 0 \\ D & R^T \end{pmatrix} \\
\frac{\partial G}{\partial w} &= \begin{pmatrix} -I & 0 \\ 0 & -\frac{\partial^2 V}{\partial x \partial \alpha} \end{pmatrix} = -\begin{pmatrix} I & 0 \\ 0 & b \end{pmatrix}.
\end{aligned}$$

Since R is full row rank and D is positive definite, $RD^{-1}R^T$ is positive definite. Then $\partial G / \partial y$ is always invertible, and it can be checked that

$$\begin{pmatrix} R & 0 \\ D & R^T \end{pmatrix}^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where

$$\begin{aligned}
A_{11} &:= D^{-1}R^T(RD^{-1}R^T)^{-1} \\
A_{12} &:= D^{-1} - D^{-1}R^T(RD^{-1}R^T)^{-1}RD^{-1} \\
A_{21} &:= -(RD^{-1}R^T)^{-1} \\
A_{22} &:= (RD^{-1}R^T)^{-1}RD^{-1}.
\end{aligned}$$

All the above matrices are well defined because $RD^{-1}R^T$ is invertible. From the implicit function theorem, the vector y can be uniquely solved in terms of w locally. Moreover

$$\begin{aligned}
\frac{dy}{dw} &= -\left(\frac{\partial G}{\partial y}\right)^{-1} \frac{\partial G}{\partial w} \\
&= \begin{pmatrix} R & 0 \\ D & R^T \end{pmatrix}^{-1} \begin{pmatrix} I & 0 \\ 0 & b \end{pmatrix} \\
&= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & b \end{pmatrix}.
\end{aligned}$$

From the definitions of y and w, we have

$$\begin{aligned}
\frac{\partial x}{\partial \alpha} &= A_{12} b \quad & \frac{\partial x}{\partial c} &= A_{11} \\
\frac{\partial p}{\partial \alpha} &= A_{22} b \quad & \frac{\partial p}{\partial c} &= A_{21}.
\end{aligned}$$

Substituting in the definitions of A_{ij} yields the desired results. □

Since the optimal x always satisfies the constraints $Rx = c$, for a fixed c, the change in x should be in the null space of R as α varies. This is captured by the following corollary.

Corollary 3: The derivative $\partial x / \partial \alpha$ can also be expressed as

$$\frac{\partial x}{\partial \alpha} = Z(Z^T D Z)^{-1} Z^T b$$

where the columns of matrix Z form a basis of the null space of R.

Proof: Let

$$\Delta = D^{-1} - D^{-1}R^T(RD^{-1}R^T)^{-1}RD^{-1} - Z(Z^T D Z)^{-1} Z^T.$$

From Theorem 2 and the definition of Δ, we only need to show $\Delta = 0$. By the definition of matrix Z we have

$$RZ = 0 \quad \text{and} \quad Z^T R^T = 0.$$

It is clear that

$$\begin{bmatrix} R \\ Z^T D \end{bmatrix} \Delta = \begin{bmatrix} RD^{-1} - RD^{-1} - 0 \\ Z^T - 0 - Z^T \end{bmatrix} = 0.$$

The next step is to show that the matrix $\left[\begin{smallmatrix} R \\ Z^T D \end{smallmatrix}\right]$ is full rank so that Δ must be the zero matrix. Suppose it is not, then there exists a nonzero vector v such that

$$\begin{bmatrix} R \\ Z^T D \end{bmatrix} v = 0. \tag{9}$$

Hence, $Rv = 0$, i.e., v is in the null space of R. Since the columns of Z form a basis of the null space of R, there exists w such that $v = Zw$ Substituting into (9), we have

$$Z^T D v = Z^T D Z w = 0.$$

Since $Z^T DZ$ is positive definite and invertible, we must have $w = 0$ and $v = Zw = 0$. This contradicts the assumption that $v \neq 0$. Therefore, $\left[\begin{smallmatrix} R \\ Z^T D \end{smallmatrix}\right]$ is full rank and $\Delta = 0$. □

We now apply these results to a particular class of utility functions to study the effect of changes in α and c on aggregate throughput in general networks.

IV. Is Fair Allocation Always Inefficient

In this section, we apply the expression for $\partial x / \partial \alpha$ in Corollary 3 to study the effect of changes in α on aggregate throughput $T(\alpha) = T(\alpha, c)$, for a fixed $c > 0$. It clarifies a folklore about tradeoff between efficiency and fairness of a bandwidth allocation policy. We first define a quantitative measure of fairness and state formally the conjecture.

A. Conjecture

A series of recent work, e.g., [18], [22]–[25], [28] has shown that a bandwidth allocation policy can be expressed in terms of a utility function $U_i(x_i)$ in the sense that the desired bandwidth allocation x solves problem (2), and the associated price vector p solves (3). Indeed, various TCP congestion control algorithms can be interpreted as solving the same utility maximization problem, and its dual [24], with different utility functions [23], [25]. The authors of [18] introduce *proportional fairness*, characterized by $U_i(x_i) = \log x_i$. In [27], an allocation policy called *minimum potential delay* is proposed with $U_i(x_i) = -1/x_i$, which is shown in [22] to approximate the fairness of the TCP Reno deployed on the current Internet. In [28], the following class of utility functions is proposed:

$$U(x_i, \alpha) = \begin{cases} (1-\alpha)^{-1} x_i^{1-\alpha} & \text{if } \alpha \neq 1 \\ \log x_i & \text{if } \alpha = 1 \end{cases} \tag{10}$$

for $\alpha \geq 0$. This includes all the previously considered allocation policies as special cases—maximum throughput ($\alpha = 0$), proportional fairness ($\alpha = 1$), minimum potential delay ($\alpha = 2$), and maxmin fairness ($\alpha = \infty$)—and provides a convenient way to compare the fairness of different allocation policies. Moreover, it also includes the fairness of major TCP congestion control algorithms, Reno ($\alpha = 2$), HSTCP [10] ($\alpha = 1.2$), and Vegas [25], FAST [15], STCP [19] ($\alpha = 1$ for all), as special cases. See [3] for application of this class of utility functions to study the effect of fairness on stability of a dynamic network where sources randomly join and depart.

The above discussion suggests the interpretation of α as a *quantitative* measure of fairness. Note that we are not concerned with fairness *across different flows* under the same allocation policy represented by a given α value, as, e.g., Jain's fairness index is [14]. Rather, we want to compare fairness *across allocation policies*. While there are no generally accepted criteria to compare the fairness of allocation policies, many examples in the networking literature (e.g., [3], [26]–[28], [30]) informally compare specific allocation policies in terms of their α. For instance, the choice of $\alpha = 0$ maximizes throughput $T(\alpha)$ but can be extremely unfair (see Example 1 below). Proportional fairness ($\alpha = 1$) is considered fairer, and maxmin fairness ($\alpha = \infty$) the fairest because it generalizes equal sharing at a single resource to a network of resources in a way that maintains Pareto optimality [2], [12]. Indeed, for the linear network in Example 1 below, [27] explicitly compares the fairness of these policies, and shows that the minimum-potential-delay policy ($\alpha = 2$) "penalizes more (less) severely long routes than maxmin (proportional) fairness." Here, we extrapolate this intuition from these special cases to a continuum of allocation policies, indexed by $\alpha \geq 0$, and interpret α as a quantitative measure of fairness. We will also provide shortly a technical reason for this extrapolation.

For the rest of this paper, we assume that all sources use the common utility function given by (10) with the same $\alpha \geq 0$. This assumption is made both for analytical simplicity and to focus our attention on behaviors that arise from the fairness of allocation policies, not from source heterogeneity.

Is a fairer policy (one with larger α) always less efficient (has a smaller aggregate throughput $T(\alpha)$)? This conjecture is prompted by the various examples in resource allocation in the literature of wired networks [3], [27], [28], wireless networks, [26], [30], economics, [5], etc. (also see the next subsection). These examples seem to illustrate (quoted from [26])

> *"the fundamental conflict between achieving flow fairness and maximizing overall system throughput. ... The basic issue is thus the tradeoff between these two conflicting criteria."*

Formally, we have

Conjecture 4: $T(\alpha)$ is nonincreasing:

$$\frac{\partial T}{\partial \alpha} \leq 0 \quad \text{for} \quad \alpha > 0.$$

B. Special Cases

In this subsection, we review several examples in the literature that have motivated the conjecture. The conjecture is shown to be true in special networks for maxmin fairness, minimum potential delay, proportional fairness, and the

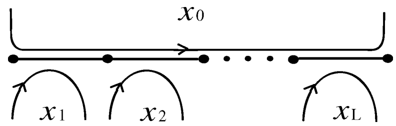


Fig. 1. Linear network.

maximum-throughput policy, by analytically solving (2) or numerically computing $T(\alpha)$. These techniques are not applicable to general networks specified by arbitrary (R, c). By interpolating the space of allocation policies to a continuum indexed by α, we can apply implicit function theorem to compare these allocation policies in general networks.

As we will explain in the next subsection, the underlying network topology in all the examples in this subsection possesses a special structure that is far from apparent in previous analysis but that leads to trivial sufficient conditions for the conjecture to be true; see the discussion preceding Corollary 6.

Example 1: Linear Network With Uniform Capacity [3], [27]: Consider the classical linear network with L links, indexed by $l = 1, \ldots, L$, and $N = L + 1$ sources, indexed by $i = 0, 1, \ldots, L$, shown in Fig. 1. Source 0 goes through all the L links and sources $i \geq 1$ go through links $l = i$. All links have the same capacity of 1 unit.

In [27], the throughput of each source and their aggregate have been calculated for several α values:

$$\begin{aligned}
&\text{maxmin fairness}: \quad T(\infty) = \frac{1}{2}(L+1)\\
&\text{min potential delay}: \quad T(2) = L - \sqrt{L} + 1\\
&\text{proportional fairness}: T(1) = \frac{L^2+1}{L+1}\\
&\text{max throughput}: \quad T(0) = L.
\end{aligned}$$

Hence, the conjecture is true for these specific values of α:

$$T(\infty) \leq T(2) \leq T(1) \leq T(0).$$

After examining this special case, the authors of [27] made a cautious comment: "It is not known whether the same ordering holds for arbitrary network topologies."

Is the conjecture true for other values of α for this topology? In [3], the rates $x_i(\alpha)$ are computed by solving (2), as follows:

$$x_0(\alpha) = \frac{1}{L^{\frac{1}{\alpha}}+1}, \quad x_i(\alpha) = \frac{L^{\frac{1}{\alpha}}}{L^{\frac{1}{\alpha}}+1}, \quad i \geq 1$$

Using this, we can easily check that, for $\alpha > 0$,

$$\frac{\partial T}{\partial \alpha} = \frac{-L^{\frac{1}{\alpha}}(L-1)\log L}{\alpha^2\left(1+L^{\frac{1}{\alpha}}\right)^2} \begin{cases} = 0, & L = 1 \\ < 0, & L \geq 2. \end{cases}$$

Hence, except for the single link case $(L = 1)$, $T(\alpha)$ is strictly decreasing in α for the linear network with uniform link capacity.

Example 2: Linear Network With Nonuniform Capacity [28]: The linear network of Example 1 is considered in [28] with $L = 2$, but with different link capacities $c_1 < c_2$. The authors calculated the source rates under maxmin fairness:

$$x_0(\infty) = x_1(\infty) = \frac{c_1}{2}, \quad x_2(\infty) = c_2 - \frac{c_1}{2}$$

and pointed out that source rate x_0 will be higher under proportional fairness, highlighting the fact that different fairness criteria can produce different throughput in general networks.

Indeed, it is not hard to solve (2) directly to obtain the source rates under proportional fairness for this example:

$$\begin{aligned}
x_0(1) &= \frac{1}{3}\left(c_1 + c_2 - \sqrt{c_1^2 + c_2^2 - c_1c_2}\right)\\
x_1(1) &= \frac{1}{3}\left(2c_2 - c_1 + \sqrt{c_1^2 + c_2^2 - c_1c_2}\right)\\
x_2(1) &= \frac{1}{3}\left(2c_1 - c_2 + \sqrt{c_1^2 + c_2^2 - c_1c_2}\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
T(1) &= \frac{2}{3}c_1 + \frac{1}{3}\sqrt{c_1^2 + c_2^2 - c_1c_2} + \frac{2}{3}c_2\\
&> c_2 + \frac{c_1}{2} = T(\infty).
\end{aligned}$$

The throughputs for proportional and maxmin fairness support the conjecture for $\alpha = 1$ and $\alpha = \infty$.

Example 3: Linear Network With Two Long Flows: Consider a linear network with two long flows, as shown in Fig. 2. We choose $c = (500, 400, 300, 200, 500)^T$ and numerically solve for $T(\alpha)$ for $\alpha > 0$. The result is shown in Fig. 3. It suggests that the conjecture is true for all $\alpha > 0$ for this network. Corollary 6 below implies that, indeed, it is.

We now investigate the conjecture in general networks.

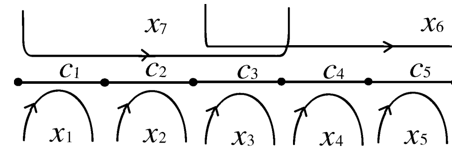


Fig. 2. Linear network with two long flows.

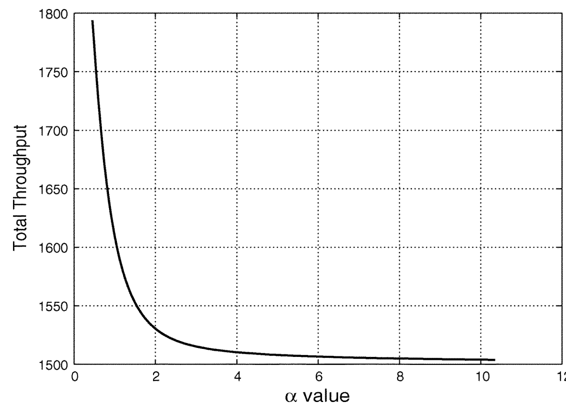


Fig. 3. Fairness–efficiency tradeoff $T(\alpha)$: linear network with two long flows.

C. Necessary and Sufficient Conditions

In this subsection, we will abuse notation and use T to denote both a function of α, as before, and a function of source rates $x(\alpha)$:

$$T\left(x(\alpha)\right) = \mathbf{1}^T x(\alpha) \tag{11}$$

where $\mathbf{1} = (1, \dots, 1)^T$. The meaning should be clear from the context. From Corollary 3, we have that

$$\frac{\partial T}{\partial \alpha} = \mathbf{1}^T Z (Z^T D Z)^{-1} Z^T b. \tag{12}$$

When the utility function $U(x, \alpha)$ is defined as in (10) the matrix D and vector b defined in (6) and (7), respectively, take the forms

$$D = \alpha \operatorname{diag}\left(x_1^{-\alpha-1}, \dots, x_N^{-\alpha-1}\right)$$
$$b = -\left(x_1^{-\alpha}\log x_1, \dots, x_N^{-\alpha}\log x_N\right)^T$$

where $x = x(\alpha) = x(\alpha, c)$ are the optimal rates. Let $\mu = \mu(\alpha, c)$, $\beta = \beta(\alpha, c)$, and $A = A(\alpha, c)$ be defined by

$$\mu_i := z_i^T b, \quad \beta_i := -\mathbf{1}^T z_i, \quad A := Z^T D Z \tag{13}$$

where z_i are the ith columns of Z. Note that A is positive definite and hence invertible. Let $\bar{A}_i(\alpha, c)$ denote the matrix obtained from replacing the ith row of A with row vector $\beta^T = (\beta_1, \beta_2 \dots \beta_M)$. From the above definitions and (12) we have

$$\frac{\partial T}{\partial \alpha} = -\beta^T A^{-1} \mu. \tag{14}$$

Our first main result is a necessary and sufficient condition for the conjecture to hold. Note that the condition is a function of α even though this is not explicit in the notation.

Theorem 5: For any $\alpha > 0$

$$\frac{\partial T}{\partial \alpha} \le 0 \quad \text{if and only if} \quad \sum_{i=1}^{M} \mu_i \det \bar{A}_i \ge 0.$$

Proof: The key observation is the following expression for the row vector:

$$\beta^T A^{-1} = \frac{1}{\det A}\left(\det \bar{A}_1, \det \bar{A}_2, \dots, \det \bar{A}_n\right) \tag{15}$$

which follows from the following formula for matrix inverse [13]:

$$A^{-1} = \frac{1}{\det A} A^*$$

where A^* is the adjoint matrix of A. Combining (14) and (15), we have

$$\frac{\partial T}{\partial \alpha} = -\frac{1}{\det A}\sum_{i=1}^{M} \mu_i \det \bar{A}_i.$$

□

Theorem 5 characterizes exactly the set of networks (R, c) in which Conjecture 4 is true. Though difficult to understand intuitively, this characterization leads directly to two sufficient conditions that explain all the examples in Section IV-B. The first condition implies that the conjecture is true when every link has a single-link flow and there is only one long flow. This condition is satisfied by Examples 1 and 2. The second condition implies that the conjecture is true when there are two long flows but both pass through the same number of links. This condition is satisfied by Example 3. The corollary implies that, while the diversity of capacities c_l in Examples 2 and 3 makes the optimization problem (2) hard to solve and the previous analysis methods complicated, they are not relevant at all to the truth of the conjecture for these examples.

Corollary 6: Suppose every link has a single-link flow.

1) If $dim(Z) = 1$, then $\partial T/\partial \alpha \le 0$ for all $\alpha > 0$.
2) If $dim(Z) = 2$ and the only two long flows pass through the same number of links, then $\partial T/\partial \alpha \le 0$ for all $\alpha > 0$.

To prove the corollary, we now specialize to a particular basis Z of the null space of the routing matrix R, making use of the fact that every link has a single-link flow. Rearrange the column of routing matrix R to express R as

$$R = [I_L \quad R_1]$$

where I_L is the $L \times L$ identity matrix and R_1 is a $L \times M$ matrix, $N = L + M$. We choose a set of basis for the null space of R such that matrix Z can be expressed as

$$Z = \begin{bmatrix} -R_1 \\ I_M \end{bmatrix}. \tag{16}$$

Clearly $\operatorname{rank}(Z) = \dim(Z) = M$.

Lemma 7: Suppose every link has a single-link flow. For Z in the form of (16), we have

1) $\mu_m \ge 0$ for $m = 1, \dots, M$.
2) $a_{mm} \ge a_{mn}$ for all $m, n = 1, \dots, M$.

Proof: The Karush–Kuhn–Tucker condition implies that there are nonnegative $p \in \Re^L$ such that (see e.g., [24])

$$R^T p = \frac{\partial V}{\partial x}.$$

By definition, $Z^T R^T = 0$. Hence, we have

$$Z^T \frac{\partial V}{\partial x} = 0. \tag{17}$$

From (10),

$$\frac{\partial V}{\partial x} = \left(x_1^{-\alpha}, \dots, x_N^{-\alpha}\right)^T.$$

Suppose $z_m = (z_{m1}, \dots, z_{mN})^T$ is the mth row of matrix Z^T. Then (16), (10), and (17) imply that, for $m = 1, \dots, M$

$$x_{L+m}^{-\alpha} + \sum_{j=1}^{L} z_{mj} x_j^{-\alpha} = 0. \tag{18}$$

Since R_1 is a 0–1 matrix, we have $-z_{mj} = 0$ or 1, $j = 1, \dots, L$. Hence, $x_{L+m}^{-\alpha} \ge -z_{mj} x_j^{-\alpha}$ and

$$x_{L+m} \le x_j \quad \text{for } j = 1, \dots, L, \;\; z_{mj} \ne 0. \tag{19}$$

From (7), b is

$$b = \frac{\partial^2 V}{\partial x \partial \alpha} = -\left(x_1^{-\alpha}\log x_1, \dots, x_N^{-\alpha}\log x_N\right)^T.$$

Then, for $m = 1, \ldots, M$,

$$\begin{aligned}\mu_m &= z_m^T b \\ &= -z_m^T \left(x_1^{-\alpha} \log x_1, \ldots, x_N^{-\alpha} \log x_n\right)^T \\ &= -x_{L+m}^{-\alpha} \log x_{L+m} - \sum_{j=1}^{L} z_{mj} x_j^{-\alpha} \log x_j \\ &\geq -\log x_{L+m} \left(x_{L+m}^{-\alpha} + \sum_{j=1}^{L} z_{mj} x_j^{-\alpha} \right) = 0\end{aligned}$$

where the last equality follows from (18) and the inequality follows from (19). This proves the first assertion.

To prove the second assertion, the matrix D is given by

$$D = -\frac{\partial^2 V}{\partial x^2} = \alpha \operatorname{diag}\left(x_1^{-\alpha-1}, \ldots, x_N^{-\alpha-1}\right).$$

Then

$$a_{mm} = z_m^T D z_m = \alpha \sum_{j=1}^{N} z_{mj}^2 x_j^{-\alpha-1}$$
$$a_{mn} = z_m^T D z_n = \alpha \sum_{j=1}^{N} z_{mj} z_{nj} x_j^{-\alpha-1}$$

and hence

$$a_{mm} - a_{mn} = \alpha \sum_{j=1}^{N} \left(z_{mj}^2 - z_{mj} z_{nj}\right) x_j^{-\alpha-1}.$$

Since $z_{mj} = 1, 0$ or -1, $z_{mj}^2 \geq z_{mj} z_{nj}$, and, hence,

$$a_{mm} \geq a_{mn}.$$

□

We are now ready to prove Corollary 6.

Proof of Corollary 6:

1) In this case, $M = 1$ and $Z \in \Re^{N \times 1}$ is a column vector. There are L single-link flows, one at each of the L links, and exactly one other flow which can traverse one or more links. This means $\sum_{j=1}^{L} -z_{1j} \geq 1$ since the long flow at least transverses one link. Hence,

$$\beta_1 = -\mathbf{1}^T z_1 = \sum_{j=1}^{L} -z_{1j} - 1 \geq 0.$$

From Lemma 7, we know that $\mu_1 > 0$. From Theorem 5 we have

$$\frac{\partial T}{\partial \alpha} = -\frac{\mu_1 \beta_1}{\det A} \leq 0$$

since matrix A is positive definite.

2) In addition to the L single-link flows, there are two flows that traverse one or more links. Since they traverse the same number of links, we have

$$\beta_1 = \beta_2 = -\mathbf{1}^T z_1 \geq 0 \qquad (20)$$

as in the first assertion. We also have

$$\mu_1 \det \bar{A}_1 + \mu_2 \det \bar{A}_2 = \beta_1 \left[\mu_1 (a_{22} - a_{21}) + \mu_2 (a_{11} - a_{12})\right].$$

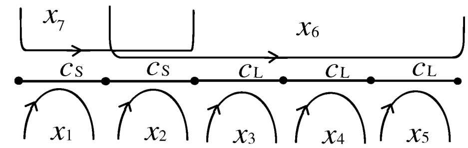


Fig. 4. Network for counter-example in Theorem 8.

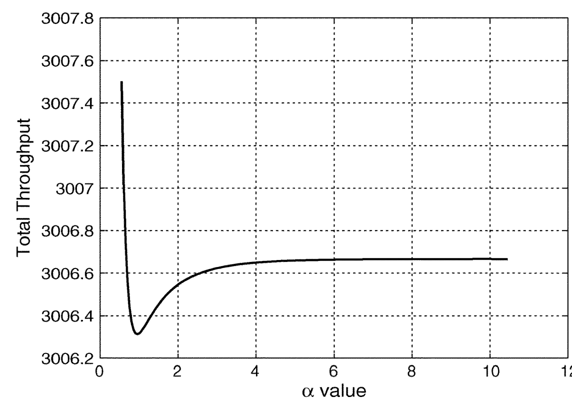


Fig. 5. Counter-example: $c_S = 10$ and $c_L = 1000$.

Lemma 7 and (20) then imply that the above quantity is nonnegative. Hence,

$$\frac{\partial T}{\partial \alpha} = -\frac{\mu_1 \det \bar{A}_1 + \mu_2 \det \bar{A}_2}{\det A} \leq 0.$$

□

D. Counter-Example

The condition in the second part of Corollary 6 that both long flows pass through the same number of (bottleneck) links is important. When that fails, there are networks where the *opposite* of the conjecture is true!

Theorem 8: When $dim(Z) \geq 2$, for any $\alpha_0 > 0$, there exists a network such that

$$\frac{\partial T}{\partial \alpha} > 0 \quad \textit{for all } \alpha > \alpha_0.$$

Proof: See Appendix A. □

Example 4: Counter-Example: Consider the linear network in Fig. 4 with $L = 5$ links and $N = 7$ sources. The null space of R has a dimension $\dim(Z) = N - L = 2$. There are five one-link flows with rates $x_1, \ldots, x_5$ and two long flows with rates x_6, x_7. Links 1 and 2 have a small capacity c_S and links 3, 4 and 5 have a large capacity c_L. We solve the utility maximization (5) numerically to compute $T(\alpha)$, for $\alpha \in [0.5, 10]$.

The aggregate throughput $T(\alpha)$ is plotted in Fig. 5 as a function of αq, for $c_S = 10$ and $c_L = 1,000$. The minimal throughput is achieved around $\alpha_0 = 0.95$ and will be achieved around $\alpha_0 = 0.75$ if we change c_L to be 5,000. $T(\alpha)$ is strictly increasing beyond α_0. In particular,

$$T(\infty) > T(2) > T(1).$$

The example is surprising at first because the conventional wisdom in networking is that increasing α favors long flows

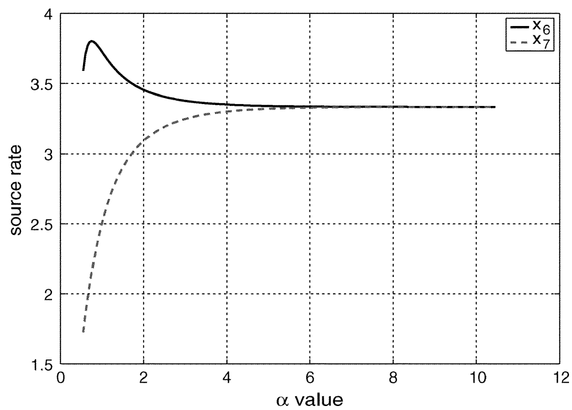


Fig. 6. Counter-example: $x_6(\alpha)$ and $x_7(\alpha)$ as functions of α, for $c_S = 10$ and $c_L = 1000$.

which take up more resources, leading to a drop in aggregate throughput. This is not exactly right. Recall that the price p_l at a link is a precise measure of congestion at that link. A more precise intuition is that increasing α favors "expensive" flows, flows that have the largest sum of link prices in their paths. In Example 4, the link capacity c_S is small and c_L is large, so that prices are high at links 1 and 2, and low at links 3, 4, 5. Even though x_6 traverses more links, it has a lower aggregate price over its path than x_7. Hence, when α increases, x_7 increases, leading to a reduction in x_6 (because of sharing at link 2). This reduction allows increases in flows x_3, x_4, x_5, so that the net change in aggregate throughput $T(\alpha)$ is positive. Hence, the counter-example relies on the design that the longest flow is not the most expensive.

Indeed, one can prove that for the network in Fig. 4, $\partial x_7/\partial\alpha > 0$ and $\partial x_6/\partial\alpha < 0$ for all $\alpha > \alpha_0$, as illustrated in Fig. 6. See the end of Appendix A for the arguments.

In this example, the decrease in x_6 allows increases in just three one-link flows, and yet this is enough to produce a net increase in the aggregate throughput. Our example actually is compact in that our proof shows that x_6 has to pass through at least three links (link 3,4,5) to make $\partial T/\partial\alpha > 0$ (see the remark after the proof of Theorem 8 in Appendix A).

The amount of increment in Fig. 5 is quite small. Indeed, an easy and loose upper-bound for the increment of aggregate throughput is $c_S/2$. We do not know whether the variation is small only because this example is compact, or it is small for general networks (R, c) as well.

V. Does Increasing Capacity Always Raise Throughput

We have seen in the last section how fairness, as measured by α, can affect efficiency, as measured by T, in unexpected ways due to interaction among sources in general networks. In this section, we will use the expression for $\partial x/\partial c$ in Theorem 2 to study how increasing capacity c affects the aggregate throughput T. If users are charged a constant fee to carry each unit of their flows, then T is also proportional to network revenue. The results here can thus be useful in deciding in which links resources should be invested to maximize aggregate throughput or revenue.

Let $\delta \in R^L$ be the vector that represents the direction in which link capacities are changed. For instance, when $\epsilon\delta = \epsilon e_l$, where $\epsilon > 0$ and e_l is an L-vector that has all its entries 0 except the lth entry which is 1, it means only link l increases its capacity from c_l to $c_l + \epsilon$. Similarly, When $\epsilon\delta = \epsilon\mathbf{1}$, then all links increase their capacities by ϵ unit. When $\epsilon\delta = \epsilon c$, then all links increase their capacities by amounts proportional to their current capacities, i.e., link l increases its capacity by ϵc_l.

The change in aggregate throughput per unit of an infinitesimal change along direction δ in capacities is measured by the directional derivative $\mathbf{D}T$ of T in direction δ, defined as

$$\mathbf{D}T(\alpha,\delta) = \mathbf{D}T(\alpha,\delta;c) := \lim_{\epsilon\to 0}\frac{T(\alpha, c+\epsilon\delta) - T(\alpha, c)}{\epsilon}.$$

Then, from (11), we have

$$\mathbf{D}T(\alpha,\delta) = \mathbf{1}^T\frac{\partial x}{\partial c}\delta$$

where $\partial x/\partial c$ is evaluated at the optimal rate $x(\alpha, c)$. In the following, we will take δ to denote the *direction* of increase in capacity (i.e., its magnitude has no significance), with the understanding that $\epsilon\mathbf{D}T(\alpha,\delta)$ provides an estimate of change in aggregate throughput when c is changed to $c + \epsilon\delta$ for small ϵ. This is also important as all our results should be interpreted in the context of small perturbations that do not change the active constraint set in (2).

Define $B = RD^{-1}R^T$, $\eta = \mathbf{1}^T D^{-1}R^T$, $\bar{B}_i$ is the matrix obtained by replacing ith row of B by η. A similar argument to the proof of Theorem 5 yields the following

Theorem 9: For any δ, $\alpha > 0$

$$\boldsymbol{D}T(\alpha,\delta) \ge 0 \quad \text{if and only if} \quad \sum_{i=1}^{L}\delta_i \det\bar{B}_i \ge 0.$$

Theorem 9 characterizes exactly the set of all networks (R, c), and directions δ, in which aggregate throughput will increase, for *all* fairness $\alpha > 0$. An easy consequence is the following

Corollary 10: If R has only two rows, then

$$\boldsymbol{D}T(\alpha,\delta) \ge 0$$

for any $\alpha > 0$ and any $\delta \ge 0$.

Proof: Let B_{ij} denote the (i, j) element of B. A similar argument to the proof of Lemma 7 shows that

1) $\eta_i = B_{ii}$ for $i = 1,2$
2) $B_{ii} \ge B_{ij} = B_{ji}$ for $i, j = 1,2$

Then

$$\begin{aligned}\det(\bar{B}_1) &= \eta_1 B_{22} - \eta_2 B_{21}\\ &= b_{22}(B_{11} - B_{21}) \ge 0.\end{aligned}$$

Similarly we have $\det(\bar{B}_2) \ge 0$. Then from Theorem 9, we have $\mathbf{D}T(\alpha,\delta) \ge 0$. □

Corollary 10 says that increasing link capacity always raises aggregate throughput, provided there are only two bottleneck links. Intuitively, one might expect this to hold more generally. This is however not the case. We provide three interesting examples, with different instantiations of direction δ, as an illustration.

The first result says that not only can the aggregate throughput be reduced when some link increases its capacity, paradoxically, it can also be reduced when *all* links increase their capacities by the same amount. This is true for almost all fairness α.

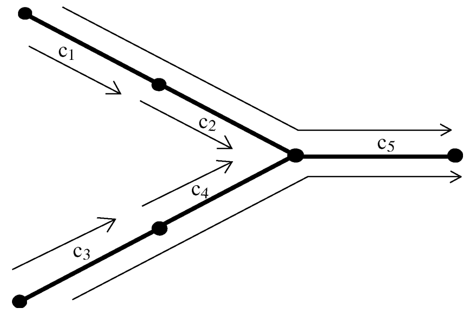


Fig. 7. Counter-example for Theorem 11 (1).

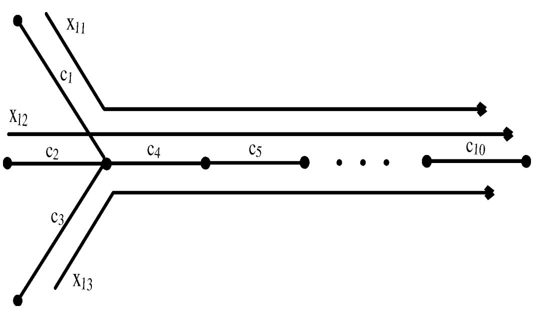


Fig. 8. Counter-example for Theorem 11 (2).

Theorem 11: Given any $\alpha_0 > 0$

1) there exists a network (R, c) such that for all $\alpha > \alpha_0$, $\boldsymbol{D}T(\alpha, e_l) < 0$ for some link l.
2) there exists a network (R, c) such that for all $\alpha > \alpha_0$, $\boldsymbol{D}T(\alpha, \mathbf{1}) < 0$.

Proof: The proof is by construction.

1) Consider the network shown in Fig. 7. There is a single-link flow x_l at each link l, for $l = 1, \ldots 4$. The flow x_5 transverses links 1, 2, 5, and flow x_6 transverses links 3, 4, 5 respectively. The capacities of the links are

$$c_1 = c_2 = c_3 = c_4 = c_L, \quad c_5 = c_S.$$

The corresponding routing matrix is

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

We increase only link 5's capacity by 1 which corresponds to $\delta = e_5$. For any fixed $\alpha_0 > 0$, we can choose c_L/c_S large enough, such that for any $\alpha > \alpha_0$ all links are fully utilized. Using Mathematica, the change in aggregate throughput is evaluated to be

$$\begin{aligned} \mathbf{D}T(\alpha, e_5) &= \mathbf{1}^T \frac{\partial x}{\partial c} e_5 \\ &= \mathbf{1}^T D^{-1} R^T (R D^{-1} R^T)^{-1} e_5 \\ &= -1. \end{aligned}$$

2) The detailed proof of $\mathbf{D}T(\alpha, \mathbf{1}) < 0$ is presented in Appendix B. □

Example 5: $\mathbf{D}T(\infty, \mathbf{1}) < 0$ for Some c in Fig. 8.

To illustrate, we calculate the change in aggregate throughput for the network in Fig. 8 under maxmin policy $\alpha = \infty$. The link capacities are

$$c_S = 2, \quad c_L = 10.$$

We have $x_1 = x_2 = x_3 = 1$, $x_4 = \ldots = x_{10} = 7$, $x_{11} = x_{12} = x_{13} = 1$. The aggregate throughput is 55. When all capacities are increased by 2ϵ with $0 \leq \epsilon \leq 1$. The rates are: $x_1 = x_2 = x_3 = 1 + \epsilon$, $x_4 = \ldots = x_{10} = 7 - \epsilon$, $x_{11} = x_{12} = x_{13} = 1 + \epsilon$. The aggregate throughput now is

$$\begin{aligned} T(\epsilon) &= 6 \times (1 + \epsilon) + 7 \times (7 - \epsilon) \\ &= 55 - \epsilon \end{aligned}$$

i.e., it is a decreasing function of ϵ. Indeed $\mathbf{D}T(\infty, \mathbf{1}) = \lim_{\epsilon \to 0} (T(\epsilon) - T(0))/(2\epsilon) = -1/2 < 0$.

If link capacities are increased proportionally, i.e., if c is increased to $(1 + \epsilon)c$, then aggregate throughput will always rise. Note that even though increasing all link capacities proportionally may be interpreted as changing the unit of capacity, it does *not* imply for general utility functions proportional increase in optimal flow vector x^* or higher aggregate throughput. It does however for the class of utility functions that satisfy the following condition: for any $\lambda > 0$, there exists a constant $\mu(\lambda) > 0$, dependent on λ, such that for all $x \geq 0$,

$$U'(\lambda x) = \mu(\lambda) U'(x).$$

In particular, this is satisfied by the class of utility functions defined in (10), as shown in the following theorem.

Theorem 12: For any network (R, c) and for all $\alpha > 0$, $\boldsymbol{D}T(\alpha, c) > 0$.

Proof: The necessary and sufficient condition for any $x \geq 0$ and $p \geq 0$ to be primal and dual optimal are

$$Rx = c \quad \text{and} \quad R^T p = \frac{\partial V}{\partial x} = \left(x_1^{-\alpha}, \ldots, x_N^{-\alpha}\right)^T.$$

Suppose x and p are optimal with link capacities c. When c is increased to $(1 + \epsilon)c$ for $\epsilon > 0$, we claim that $x(1 + \epsilon)$ and $(1 + \epsilon)^{-\alpha} p$ are the new optimal rate and price vectors respectively. We can check that these vectors satisfy the the optimality condition for capacity $(1 + \epsilon)c$:

$$\begin{aligned} Rx(1 + \epsilon) &= c(1 + \epsilon) \\ R^T p (1 + \epsilon)^{-\alpha} &= (1 + \epsilon)^{-\alpha} \left(x_1^{-\alpha}, \ldots, x_N^{-\alpha}\right)^T = \frac{\partial V}{\partial x}. \end{aligned}$$

Therefore, the aggregate throughput is increased from the original value T to $(1 + \epsilon)T$. Hence, we have

$$\mathbf{D}T(\alpha, c) = \lim_{\epsilon \to 0} \frac{(1 + \epsilon)T - T}{\epsilon} = T > 0.$$

□

VI. Conclusion

A bandwidth allocation policy can be defined in terms of utility functions parameterized by some protocol parameter α. We have studied how throughputs and prices change as link capacities or α changes. We then focus on a specific class of utility functions where α can be interpreted as a quantitative measure of fairness. We say an allocation is fair if α is large and efficient if the aggregate throughput is large. We use this model to investigate whether a fairer allocation is always more inefficient and whether increasing link capacities always raises throughput. We characterize exactly the set of all networks (R, c) in which the answers are "yes". Though these characterizations are difficult to understand intuitively, they have led to simple corollaries that explain all the examples we found in the literature and to the discovery of the first counter example.

There are a number of ways this preliminary work can be extended. First, we have focused on how throughputs x change in response to changes in α and c, which is only half of Theorem 2. The application of the other half of Theorem 2 on how prices change has not been exploited. Second, the necessary and sufficient conditions for the conjectures are hard to understand intuitively and check for large networks. It is not clear whether this condition is likely to hold or fail in practice. It would be useful to derive equivalent but more intuitive characterizations and more general corollaries than ones reported here. Finally, we have assumed every source has the same utility function. It would be interesting to see how the fairness definition and the results here should generalize when sources have the same class of utility functions but with different α_i parameters, or have different utility functions.

APPENDIX

A. Proof of Theorem 8

It is sufficient to prove the assertion with the network shown in Fig. 4 which has $\dim(Z) = 2$. This is because given any $\dim(Z) > 2$, we can always embed this network as a subnetwork within a larger network that has the given $\dim(Z)$ and scale up the capacity of this subnetwork with respect to the capacity in other parts of the network such that the decrease in throughput on this subnetwork dominates the changes in throughput on other parts of the network, as α increases. Hence, consider the network in Fig. 4, with five links $l = 1, 2, \ldots, 5$. Let their capacities be $c_1 = c_2 = c_S \geq 3$ and $c_3 = c_4 = c_5 = c_L$. Fix any $\alpha_0 > 0$.

There are five single-link flows with rates $x_i, i = 1, 2, \ldots, 5$ and two long flows with rates x_6, x_7. The routing matrix R and matrix Z are

$$R = [I_5 \quad R_1], \quad Z = \begin{bmatrix} -R_1 \\ I_2 \end{bmatrix}$$

where

$$R_1^T = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

From (13), we have

$$\beta_1 = -\mathbf{1}^T z_1 = 3, \quad \beta_2 = -\mathbf{1}^T z_2 = 1.$$

We will show that we can choose the link capacity c_L such that

$$\sum_{m=1}^{2} \mu_m \det \bar{A}_m = \mu_1 \det \begin{bmatrix} 3 & 1 \\ a_{21} & a_{22} \end{bmatrix} + \mu_2 \det \begin{bmatrix} a_{11} & a_{12} \\ 3 & 1 \end{bmatrix} = \mu_1(3a_{22} - a_{21}) + \mu_2(a_{11} - 3a_{12}) < 0.$$

Theorem 5 then implies that $\partial T/\partial \alpha > 0$ for all $\alpha > \alpha_0$.

The basic idea of the proof is as follows. From Lemma 7, we know that $\mu_m > 0$, $m = 1, 2$ (strict inequality here), and $a_{mm} \geq a_{mn}$. Hence, the first term is positive. We will show that the second term is negative by making $a_{11} - 3a_{12}$ strictly negative through an appropriate choice of c_L. Moreover, even though both terms go to zero when α goes to infinity, it is possible to choose link capacity c_L such that, for all $\alpha > \alpha_0$, the first term is smaller than the second in magnitude, so that their sum is strictly negative. We will prove the final result after 5 lemmas.

Let p_l, $l = 1, \ldots, 5$, be the link "prices" (Lagrange multipliers) and $q_i = \sum_l R_{li} p_l$ be the end-to-end prices. The following facts are the direct consequences of the optimality condition for the utility maximization problem (see e.g., [24]), which will be used extensively later.

Lemma 13: At optimality, we have
1) $x_1 + x_7 = c_S; x_2 + x_6 + x_7 = c_S; x_i + x_6 = c_L, i = 3, 4, 5$.
2) $q_i = p_i, i = 1, \ldots, 5; q_6 = \sum_{i=2}^{5} p_i; q_7 = p_1 + p_2$.
3) $q_i = x_i^{-\alpha}$ and $x_i = q_i^{-1/\alpha}$.

Define the following positive constants:
1) $K_1 = (3c_S)^{\alpha_0}$;
2) $K_2 = 2^{-1/\alpha_0} c_S/3$;
3) $K_3 = 3K_1 c_S/\alpha_0$;
4) $K_4 = K_3/K_2 + 3K_1 \max(1, 1 - \log K_2)$;
5) $K_5 = \log(3/2)$;
6) $K_6 = 5 + 6 \times 2^{1/\alpha_0}$;
7) $K_7 = 3K_1 c_S((\alpha_0 + 1)/\alpha_0) 2^{\alpha_0 + 2/\alpha_0}$;
8) $K_8 = 1/4c_S$;
9) $K = K_5 K_8/(K_4 K_6 + K_5 K_7)$.

Let $\epsilon = \min(K/2, 1/4, K_8/(2K_7))$. Then we can choose M_1 large enough so that for all $M > M_1$, $(M/c_S - 1)^{-\alpha_0} \log M < \epsilon$, and choose M_2 large enough, so that $M_2/c_S - 1 > 3c_S$. Now choose $c_L = \max(M_1, M_2)$. Then immediately, the following inequalities hold:

$$\frac{c_L}{c_S} - 1 > 3c_S, \quad \left(\frac{c_L}{c_S} - 1\right)^{-\alpha_0} \log c_L < \epsilon. \tag{21}$$

For any $\alpha > \alpha_0$,

$$\left(\frac{c_L}{c_S} - 1\right)^{-\alpha} < \left(\frac{c_L}{c_S} - 1\right)^{-\alpha_0} \log c_L < \epsilon. \tag{22}$$

Using (21), we can have a much tighter bound than (22):

$$\left(\frac{c_L}{c_S} - 1\right)^{-\alpha} < \epsilon \left(\frac{c_L}{c_S} - 1\right)^{-(\alpha - \alpha_0)} < K_1(3c_S)^{-\alpha} \epsilon.$$

The next lemma gives the ranges of all the rates.

Lemma 14:
1) $c_S/2 \leq x_1 \leq c_S; c_S/3 \leq x_2 \leq c_S$;
2) $c_L - c_S \leq x_i \leq c_L, i = 3, 4, 5; K_2 \leq x_7 \leq c_S/3$; $K_2 \leq x_6 \leq c_S/2$.

Proof:
1) $x_1 = c_S - x_7 \leq c_S; q_7 \geq q_1$, so $x_1 \geq x_7, x_1 \geq c_S/2$. $x_2 = c_S - x_6 - x_7 \leq c_S; q_2 \leq q_6, q_2 \leq q_7$, so $x_2 \geq x_6$, $x_2 \geq x_7$. Therefore: $x_2 \geq c_S/3$.
2) For $i = 3, 4, 5$, $x_i = c_L - x_6 \leq c_L$; on the other hand, since $x_6 \leq c_S$, $x_i = c_L - x_6 \geq c_L - c_S$.

3) $q_7 = p_1 + p_2 = x_1^{-\alpha} + x_2^{-\alpha} \le 2(c_S/3)^{-\alpha}$. Then

$$\begin{aligned} x_7 = q_7^{-\frac{1}{\alpha}} &\ge 2^{-\frac{1}{\alpha}}\frac{c_S}{3} \ge 2^{-\frac{1}{\alpha_0}}\frac{c_S}{3} = K_2 \\ q_7 - q_6 &= p_1 - p_3 - p_4 - p_5 \\ &\ge c_S^{-\alpha} - 3(c_L - c_S)^{-\alpha} \\ &= c_S^{-\alpha}\left(1 - 3\left(\frac{c_L}{c_S} - 1\right)^{-\alpha}\right) \\ &\ge c_S^{-\alpha}(1 - 3\epsilon) \ge 0. \end{aligned}$$

Hence, $q_7 \ge q_6$ and $x_7 \le x_6$. Since $c_S = x_2 + x_6 + x_7 \ge 3x_7$, $x_7 \le c_S/3$.

4) $x_6 \ge x_7 \ge K_2$ and $x_6 = c_S - x_2 - x_7 \le c_S - x_2 \le c_S/2$. □

The next step is to upper bound the difference between x_2 and x_6. The intuition is that by choosing c_L large enough, p_l, $l = 3, 4, 5$, will be negligible compared to p_2. Hence, the difference between q_2 and q_6 can be very small, so is the difference between x_2 and x_6.

Lemma 15: $x_2 - x_6 \le K_3(3c_S)^{-\alpha}\epsilon$

Proof: First

$$\begin{aligned} q_6 - q_2 &= p_3 + p_4 + p_5 \\ &\le 3(c_L - c_S)^{-\alpha} \\ &= 3c_S^{-\alpha}\left(\frac{c_L}{c_S} - 1\right)^{-\alpha} \\ &\le 3K_1 c_S^{-\alpha}(3c_S)^{-\alpha}\epsilon. \end{aligned} \tag{23}$$

Since all sources have the same utility function, we can think of equilibrium rates x_i as given by a common function $x(q_i)$ of end-to-end price, evaluated at different prices:

$$x_i = x(q_i) := q_i^{-\frac{1}{\alpha}}.$$

Using the intermediate value theorem and (23), we have

$$\begin{aligned} x_2 - x_6 = &\left.\frac{dx}{dq}\right|_{q=q_\xi}(q_2 - q_6), \quad q_2 \le q_\xi \le q_6 \\ &\le \frac{1}{\alpha} q_2^{-1-\frac{1}{\alpha}} 3K_1 c_S^{-\alpha}(3c_S)^{-\alpha}\epsilon \\ &= \frac{3K_1}{\alpha} x_2^{1+\alpha} c_S^{-\alpha}(3c_S)^{-\alpha}\epsilon \\ &\le \frac{3K_1 c_S}{\alpha_0}(3c_S)^{-\alpha}\epsilon. \end{aligned}$$

□

We now derive an upper bound for μ_1 and a lower bound for μ_2, which will be used in proving the final result.

Lemma 16: $\mu_1 \le K_4 c_S^{-2\alpha}\epsilon$, $\mu_2 \ge K_5 c_S^{-\alpha}$

Proof: First,

$$\begin{aligned} \mu_1 &= \sum_{i=2}^{5} x_i^{-\alpha}\log x_i - x_6^{-\alpha}\log x_6 \\ &\le \sum_{i=2}^{5}(\log x_i - \log x_6)x_i^{-\alpha}. \end{aligned}$$

Using Lemmas 14 and 15 and the intermediate value theorem, we have

$$\begin{aligned} \log x_2 - \log x_6 &= \frac{1}{x_\xi}(x_2 - x_6) \quad \text{for some } x_\xi \in [x_6, x_2] \\ &\le \frac{1}{x_6}(x_2 - x_6) \le \frac{K_3}{K_2}(3c_S)^{-\alpha}\epsilon. \end{aligned}$$

Hence,

$$\begin{aligned} (\log x_2 - \log x_6)x_2^{-\alpha} &\le \frac{K_3}{K_2}(3c_S)^{-\alpha}\epsilon\left(\frac{c_S}{3}\right)^{-\alpha} \\ &= \frac{K_3}{K_2}c_S^{-2\alpha}\epsilon. \end{aligned}$$

For $i = 3, 4, 5$, we have

$$\sum_{i=3}^{5}(\log x_i - \log x_6)x_i^{-\alpha} \le 3(\log x_3 - \log K_2)x_3^{-\alpha}. \tag{24}$$

Now

$$\begin{aligned} 3x_3^{-\alpha}\log x_3 &\le 3\log c_L(c_L - c_S)^{-\alpha} \\ &\le 3c_S^{-\alpha}\log c_L\left(\frac{c_L}{c_S} - 1\right)^{-\alpha} \\ &\le 3^{1-\alpha}K_1 c_S^{-2\alpha}\epsilon \le 3K_1 c_S^{-2\alpha}\epsilon \\ -3\log(K_2)x_3^{-\alpha} &\le 3\max(0, -\log K_2)(c_L - c_S)^{-\alpha} \\ &\le 3K_1\max(0, -\log K_2)c_S^{-2\alpha}\epsilon. \end{aligned}$$

In summary,

$$\begin{aligned} \mu_1 &\le \left(\frac{K_3}{K_2} + 3K_1\max(0, -\log K_2)\right)c_S^{-2\alpha}\epsilon \\ &= K_4 c_S^{-2\alpha}\epsilon \end{aligned}$$

and

$$\begin{aligned} \mu_2 &= x_1^{-\alpha}\log x_1 + x_2^{-\alpha}\log x_2 - x_7^{-\alpha}\log x_7 \\ &\ge x_1^{-\alpha}(\log x_1 - \log x_7) \\ &\ge c_S^{-\alpha}\left(\log\frac{c_S}{2} - \log\frac{c_S}{3}\right) \\ &= c_S^{-\alpha}\log\frac{3}{2} = K_5 c_S^{-\alpha}. \end{aligned}$$

□

With the upper bound for μ_1 and lower bound for μ_2, to prove $\mu_1(3a_{22} - a_{21}) + \mu_2(a_{11} - 3a_{12})$ is negative, we only need an upper bound for the positive term $3a_{22} - a_{21}$ and a lower bound for the magnitude of the negative term $a_{11} - 3a_{12}$. These are provided by the next lemma.

Lemma 17: $0 \le 3a_{22} - a_{21} \le \alpha K_6$

$$-(a_{11} - 3a_{12}) \ge (K_8 - K_7\epsilon)c_S^{-\alpha}\alpha > 0. \tag{25}$$

Proof: From (13),

$$\begin{aligned} a_{11} &= \alpha\left(x_2^{-\alpha-1} + x_3^{-\alpha-1} + x_4^{-\alpha-1} + x_5^{-\alpha-1} + x_6^{-\alpha-1}\right) \\ a_{12} &= a_{21} = \alpha x_2^{-\alpha-1} \\ a_{22} &= \alpha\left(x_2^{-\alpha-1} + x_1^{-\alpha-1} + x_7^{-\alpha-1}\right). \end{aligned}$$

Hence,

$$3a_{22} - a_{21} = \alpha\left(2x_2^{-\alpha-1} + 3x_1^{-\alpha-1} + 3x_7^{-\alpha-1}\right) \ge 0.$$

Also,

$$\begin{aligned}3a_{22}-a_{21} &= \alpha\left(2x_2^{-\alpha-1}+3x_1^{-\alpha-1}+3x_7^{-\alpha-1}\right)\\ &\le \alpha\left(\frac{c_S}{3}\right)^{-\alpha-1}\left(2+3+3\times 2^{1+\frac{1}{\alpha}}\right)\\ &\le \alpha\left(5+6\times 2^{\frac{1}{\alpha_0}}\right)=K_6\alpha.\end{aligned}$$

Now

$$3a_{12}-a_{11}=\alpha\left(2x_2^{-\alpha-1}-3x_3^{-\alpha-1}-x_6^{-\alpha-1}\right).$$

Since

$$\begin{aligned}x_6^{-\alpha-1}-x_2^{-\alpha-1} &= \left.\frac{dx^{-\alpha-1}}{dx}\right|_{x=x_\xi}(x_6-x_2),\ \ x_6\le x_\xi\le x_2\\ &\le(\alpha+1)x_6^{-\alpha-2}(x_2-x_6)\\ &\le(\alpha+1)2^{\frac{\alpha+2}{\alpha}}\left(\frac{c_S}{3}\right)^{-\alpha-2}(x_2-x_6)\\ &\le\frac{\alpha_0+1}{\alpha_0}2^{\frac{\alpha_0+2}{\alpha_0}}3c_SK_1c_S^{-\alpha}\epsilon=K_7c_S^{-\alpha}\epsilon\end{aligned}$$

and

$$\begin{aligned}x_2^{-\alpha-1}-3x_3^{-\alpha-1} &\ge c_S^{-\alpha-1}-3(c_L-c_S)^{-\alpha-1}\\ &= c_S^{-\alpha-1}\left(1-3\left(\frac{c_L}{c_S}-1\right)^{-\alpha-1}\right)\\ &\ge\frac{1}{c_S}\left(1-3\epsilon\left(\frac{c_L}{c_S}-1\right)^{-1}\right)c_S^{-\alpha}\\ &\ge\frac{1}{4c_S}c_S^{-\alpha}=K_8c_S^{-\alpha}\end{aligned}$$

we have

$$\begin{aligned}-(a_{11}-3a_{12}) &\ge \left(K_8c_S^{-\alpha}-K_7c_S^{-\alpha}\epsilon\right)\alpha\\ &=(K_8-K_7\epsilon)c_S^{-\alpha}\alpha.\end{aligned}$$

By the choice of ϵ, we have

$$K_8-K_7\epsilon>0.$$

Now we are ready to evaluate

$$\begin{aligned}\sum_{m=1}^{2}\mu_m\det\bar{A}_m &= \mu_1\det\begin{bmatrix}3 & 1\\ a_{21} & a_{22}\end{bmatrix}+\mu_2\det\begin{bmatrix}a_{11} & a_{12}\\ 3 & 1\end{bmatrix}\\ &\le\alpha\left(K_4c_S^{-2\alpha}\epsilon K_6-K_5c_S^{-\alpha}(K_8-K_7\epsilon)c_S^{-\alpha}\right)\\ &=\alpha c_S^{-2\alpha}(K_4K_6\epsilon-K_5(K_8-K_7\epsilon))\\ &=\alpha c_S^{-2\alpha}((K_4K_6+K_5K_7)\epsilon-K_5K_8)\end{aligned}$$

where the inequality follows from Lemmas 16 and 17. Our choice of ϵ implies

$$\epsilon<\frac{K_5K_8}{K_4K_6+K_5K_7}=K. \tag{26}$$

Then $\sum_{m=1}^{2}\mu_m\det\bar{A}_m<0$ for all $\alpha>\alpha_0$. Finally, by Theorem 5, for all $\alpha>\alpha_0$,

$$\frac{\partial T}{\partial\alpha}>0.$$

□

Remark: If link 5 did not exist, then the coefficient for μ_2 would be

$$a_{11}-2a_{12}=\alpha\left(2x_3^{-\alpha-1}+x_6^{-\alpha-1}-x_2^{-\alpha-1}\right)>0.$$

Then $\sum_{m=1}^{2}\mu_m\det\bar{A}_m>0$ and the conjecture would hold. Hence, our counter-example is compact. □

As we discussed at the end of Section IV-D, when $\alpha>\alpha_0$, x_6 decreases and x_7 increases with increasing α, which is the reason for the increase in aggregate throughput. Here is a short formal proof for $\partial x_7/\partial\alpha>0$ and $\partial x_6/\partial\alpha<0$.

Replace the vector $\mathbf{1}$ with $e_7=(0,0,0,0,0,0,1)^T$ in (11) and in the definition of β in (13). Now $\beta_1=-e_7^Tz_1=0$ and $\beta_2=-e_7^Tz_2=-1$. Then a similar argument to the proof of Theorem 5 yields

$$\frac{\partial x_7}{\partial\alpha}\le 0\quad\text{if and only if}\quad\sum_{m=1}^{2}\mu_m\det\bar{A}_m\ge 0.$$

Using the corresponding $\bar{A}$, we have

$$\begin{aligned}\sum_{m=1}^{2}\mu_m\det\bar{A}_m &= \mu_1\det\begin{bmatrix}0 & -1\\ a_{21} & a_{22}\end{bmatrix}+\mu_2\det\begin{bmatrix}a_{11} & a_{12}\\ 0 & -1\end{bmatrix}\\ &=-\mu_2a_{11}+\mu_1a_{21}.\end{aligned}$$

From Lemma 16, we have $\mu_2\gg\mu_1$. Lemma 7 implies that $a_{11}\ge a_{21}$. Then we have $-\mu_2a_{11}+\mu_1a_{21}<0$. Hence, for $\alpha>\alpha_0$,

$$\frac{\partial x_7}{\partial\alpha}>0.$$

Similarly, we can show that, for $\alpha>\alpha_0$,

$$\frac{\partial x_6}{\partial\alpha}<0.$$

B. *Proof of Theorem 11*

Consider the network shown in Fig. 8. There is a single-link flow x_l at each link l, for $l=1,\ldots,10$, which is omitted from the figure. The link capacities are

$$\begin{aligned}c_l=c_S &\quad\text{for}\quad l=1,2,3\\ c_l=c_L &\quad\text{for}\quad l=4,\ldots,10.\end{aligned}$$

For the network in Fig. 8, the corresponding routing matrix is

$$R=\begin{bmatrix}I_{10} & \begin{matrix}I_3\\ \mathbf{1}_{7\times3}\end{matrix}\end{bmatrix}$$

where $1_{7\times3}$ is a 7×3 matrix with every entry set to 1.

By symmetry of the network, we can define vector v as

$$\begin{aligned}v_1 &:= x_1=x_2=x_3\\ v_2 &:= x_4=\ldots=x_{10}\\ v_3 &:= x_{11}=x_{12}=x_{13}.\end{aligned}$$

We use vector q to denote the corresponding end-to-end prices for rate vector v.

Using Mathematica, we evaluate $\mathbf{D}T(\alpha, \mathbf{1})$ as

$$\mathbf{1}^T \frac{\partial x}{\partial c}\mathbf{1} = \mathbf{1}^T D^{-1} R^T (R D^{-1} R^T)^{-1}\mathbf{1} = \frac{10 r_1 r_2 + 63 r_1 r_3 - 11 r_2 r_3}{r_1 r_2 + 21 r_1 r_3 + r_2 r_3}$$

where $r_i = v_i^{1+\alpha}$.

Given any $\alpha_0 > 0$, $c_S \geq 2$, we show through the following lemmas that we can find c_L such that, for any $\alpha > \alpha_0$, $10 r_1 r_2 + 63 r_1 r_3 - 11 r_2 r_3 < 0$.

Define the following positive constants:

1) $G_1 = c_S^{\alpha_0}$;
2) $G_2 = 7 G_1 c_S ((1+\alpha_0)/\alpha_0)$;
3) $G_3 = 8^{-1+(1/\alpha_0)} (c_S/2)^{1+\alpha_0}$.

Let $\epsilon = G_3 / 2(10 G_2 + 63)$. Choose c_L large enough so that the following two inequalities hold:

$$\frac{c_L}{c_S} - 1 > c_S \quad (27)$$

$$\left(\frac{c_L}{c_S} - 1\right)^{-\alpha_0} < \epsilon. \quad (28)$$

Immediately, we have the following inequality:

$$\left(\frac{c_L}{c_S} - 1\right)^{-\alpha} < \epsilon \left(\frac{c_L}{c_S} - 1\right)^{-\alpha+\alpha_0} < G_1 c_S^{-\alpha} \epsilon. \quad (29)$$

The first Lemma here upper and lower bounds all the rates.

Lemma 19:

1) $c_S/2 \leq v_1 \leq c_S$;
2) $c_L - c_S \leq v_2 \leq c_L$;
3) $8^{-1/\alpha}(c_S/2) \leq v_3 \leq c_S$.

Proof:

1) $v_1 \leq c_S$ is obvious. Since $q_1 \leq q_3$, $v_1 \geq v_3$. By noticing $v_1 + v_3 = c_S$, we have $v_1 \geq c_S/2$.
2) $v_2 \leq c_L$ is obvious. On the other hand, $v_2 = c_L - v_3 \geq c_L - c_S$.
3) $v_3 \leq c_S$ is obvious. On the other hand,

$$\begin{aligned} q_3 &= v_1^{-\alpha} + 7 v_2^{-\alpha} \\ &\leq \left(\frac{c_S}{2}\right)^{-\alpha} + 7(c_L - c_S)^{-\alpha} \\ &= \left(\frac{c_S}{2}\right)^{-\alpha} \left(1 + 7 \times 2^{-\alpha} \left(\frac{c_L}{c_S} - 1\right)^{-\alpha}\right) \\ &\leq 8 \left(\frac{c_S}{2}\right)^{-\alpha}. \end{aligned}$$

Therefore, $v_3 = q_3^{-1/\alpha} \geq 8^{-1/\alpha} c_S/2$.

The next step is to get an upper bound for $|r_1 - r_3|$. The intuition is that by choosing c_L large enough, the difference between q_1 and q_3 can be very small, so is the difference between r_1 and r_3.

Lemma 20: $r_1 - r_3 \leq G_2 \epsilon$

Proof: Using Lemma 19 and (29), we have

$$\begin{aligned} q_3 - q_1 = 7 v_2^{-\alpha} &\leq 7 (c_L - c_S)^{-\alpha} \\ &= 7 c_S^{-\alpha} \left(\frac{c_L}{c_S} - 1\right)^{-\alpha} \\ &< 7 G_1 c_S^{-2\alpha} \epsilon. \end{aligned}$$

Using this inequality and the intermediate value theorem, we have

$$\begin{aligned} v_1 - v_3 &= \left.\frac{dx}{dq}\right|_{q=q_\xi} (q_1 - q_3) \quad \text{for some } q_1 \leq q_\xi \leq q_3 \\ &\leq \frac{1}{\alpha} q_1^{-1-\frac{1}{\alpha}} (q_3 - q_1) \\ &\leq \frac{1}{\alpha} v_1^{1+\alpha} \times 7 G_1 c_S^{-2\alpha} \epsilon \\ &\leq \frac{7 G_1 c_S}{\alpha} c_S^{-\alpha} \epsilon \end{aligned}$$

where the last inequality follows from Lemma 19. Using the last inequality and the intermediate value theorem, we can upper bound $r_1 - r_3 = v_1^{1+\alpha} - v_3^{1+\alpha}$:

$$\begin{aligned} r_1 - r_3 &= (1+\alpha) v_\xi^{\alpha} (v_1 - v_3) \quad \text{for some } v_3 \leq v_\xi \leq v_1 \\ &\leq (1+\alpha) v_1^{\alpha} \times \frac{7 G_1 c_S}{\alpha} c_S^{-\alpha} \epsilon \\ &\leq 7 \frac{1+\alpha}{\alpha} G_1 c_S \epsilon \\ &\leq 7 \frac{1+\alpha_0}{\alpha_0} G_1 c_S \epsilon \\ &= G_2 \epsilon \end{aligned}$$

where the second inequality follows from Lemma 19. □

The next lemma upper bounds r_1 in terms of r_2.

Lemma 21: $r_1 < \epsilon r_2$.

Proof: From Lemma 19, and (27), (28), we have

$$\begin{aligned} \frac{r_1}{r_2} = \frac{v_1^{1+\alpha}}{v_2^{1+\alpha}} &\leq \left(\frac{c_S}{c_L - c_S}\right)^{1+\alpha} \\ &\leq \left(\frac{c_L}{c_S} - 1\right)^{-\alpha} \\ &\leq \left(\frac{c_L}{c_S} - 1\right)^{-\alpha_0} < \epsilon \end{aligned}$$

where we have used the facts $c_L/c_S - 1 > c_S > 1$ and $\alpha > \alpha_0$. □

We can also bound r_3 from below.

Lemma 22: $r_3 \geq G_3$

Proof: From Lemma 19,

$$\begin{aligned} r_3 = v_3^{1+\alpha} &\geq 8^{-1-\frac{1}{\alpha}} \left(\frac{c_S}{2}\right)^{1+\alpha} \\ &\geq 8^{-1-\frac{1}{\alpha_0}} \left(\frac{c_S}{2}\right)^{1+\alpha_0} = G_3. \end{aligned}$$

□

Now we can finish the proof of Theorem 11 by examining the sign of the following quantity:

$$\begin{aligned} &10 r_1 r_2 + 63 r_1 r_3 - 11 r_2 r_3 \\ &\quad \leq 10(G_2 \epsilon + r_3) r_2 + 63 r_1 r_3 - 11 r_2 r_3 \text{ (Lemma 20)} \\ &\quad \leq 10(G_2 \epsilon + r_3) r_2 + 63 r_2 r_3 \epsilon - 11 r_2 r_3 \text{ (Lemma 21)} \\ &\quad = ((10 G_2 + 63)\epsilon - r_3) r_2 \\ &\quad \leq ((10 G_2 + 63)\epsilon - G_3) r_2 \text{ (Lemma 22)} \\ &\quad < 0 \quad \text{(by the choice of } \epsilon). \end{aligned}$$

□

Remark: For the example above, when α is large enough, we have $r_2 \gg r_1 \approx r_3$. Hence,

$$\frac{10r_1r_2 + 63r_1r_3 - 11r_2r_3}{r_1r_2 + 21r_1r_3 + r_2r_3} \approx -\frac{1}{2}.$$

This slope agrees with Example 5 where the aggregate throughput decreases by ϵ when every link's capacity increases by 2ϵ.

ACKNOWLEDGMENT

The authors thank G. Varghese of UCSD for a helpful discussion that motivated the work on throughput–fairness tradeoff, and L. Li and W. Luo of Caltech for helpful discussions.

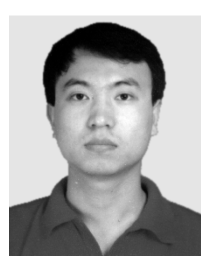
REFERENCES

[1] N. Bean, F. Kelly, and P. Taylor, "Braess's paradox in loss networks," *J. Appl. Prob.*, vol. 34, pp. 155–159, 1997.
[2] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[3] T. Bonald and L. Massoulie, "Impact of fairness on internet performance," in *Proc. ACM SIGMETRICS*, Jun. 2001, pp. 82–91.
[4] D. Braess, "Über ein Paradoxen aus der Verkehrsplanung," *Unternehmensforschung*, vol. 12, pp. 258–268, 1968.
[5] M. Butler and H. Williams, The allocation of shared fixed costs. [Online]. Available: http://www.lse.ac.uk/collections/operationalResearch/pdf/lseor02-52.pdf%, 2002
[6] B. Calvert, W. Solomon, and I. Ziedins, "Braess's paradox in a queuing network with state depending routing," *J. Appl. Prob.*, vol. 34, pp. 134–154, 1997.
[7] C. Fisk and S. Pallottino, "Empirical evidence for equilibrium paradoxes with implications for optimal planning strategies," *Transport. Res.*, vol. 15A, pp. 245–248, 1981.
[8] J. Cohen and P. Horowitz, "Paradoxical behavior of mechanical and electrical networks," *Nature*, vol. 352, pp. 699–701, Aug. 1991.
[9] J. Cohen and F. Kelly, "A paradox of congestion in a queuing network," *J. Appl. Prob.*, vol. 27, pp. 730–734, 1990.
[10] S. Floyd, HighSpeed TCP for large congestion windows. [Online]. Available: http://www.icir.org/floyd/hstcp.html, Feb. 2003, Internet draft draft-floyd-tcp-highspeed-02.txt, work in progress
[11] M. Frank, "The Braess paradox," *Math. Program.*, vol. 20, pp. 283–302, 1981.
[12] E. Gafni and D. Bertsekas, "Dynamic control of session input rates in communication networks," *IEEE Trans. Autom. Control*, vol. AC-29, no. 1, pp. 1009–1016, Jan. 1984.
[13] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[14] R. Jain, W. Hawe, and D. Chiu, A quantitative measure of fairness and discrimination for resource allocation in shared computer systems Digital Equipment Corp., Tech. Rep. DEC-TR-301, Sep. 1984.
[15] C. Jin, D. Wei, and S. H. Low, "TCP FAST: motivation, architecture, algorithms, performance," in *Proc. IEEE INFOCOM*, Mar. 2004.
[16] H. Kameda, E. Altman, T. Kozawa, and Y. Hosokawa, "Braess-like paradoxes in distributed computer systems," *IEEE Trans. Autom. Control*, vol. 45, no. 9, pp. 1687–1690, 2000.
[17] H. Kameda and O. Pourtallier, "Paradoxes in distributed decisions on optimal load balancing for networks of homogeneous computers," *J. ACM*, vol. 49, pp. 407–433, 2002.
[18] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.
[19] T. Kelly, Scalable TCP: improving performance in highspeed wide area networks. [Online]. Available: http://www-lce.eng.cam.ac.uk/ctk21/scalable/, Dec. 2002
[20] Y. Korilis, A. Lazar, and A. Orda, "Capacity allocation under noncooperative routing," *IEEE Trans. Autom. Control*, vol. 42, no. 3, pp. 309–325, 1997.
[21] ——, "Avoiding the Braess paradox in noncooperative networks," *J. Appl. Prob.*, vol. 36, pp. 211–222, 1999.
[22] S. Kunniyur and R. Srikant, "End-to-end congestion control: utility functions, random losses and ECN marks," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 689–702, Oct. 2003.
[23] S. H. Low, "A duality model of TCP and queue management algorithms," *IEEE/ACM Trans. Netw.*, vol. 11, no. 4, pp. 525–536, Aug. 2003.
[24] S. H. Low and D. Lapsley, "Optimization flow control, I: basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.
[25] S. H. Low, L. Peterson, and L. Wang, "Understanding Vegas: a duality model," *J. ACM*, vol. 49, no. 2, pp. 207–235, Mar. 2002.
[26] H. Luo, S. Lu, V. Bharghavan, J. Cheng, and G. Zhong, "A packet scheduling approach to QoS support in multihop wireless networks," *ACM J. Mobile Networks and Applications (MONET), Special Issue on QoS in Heterogeneous Wireless Networks*, vol. 9, no. 3, pp. 193–206, Jun. 2004.
[27] L. Massoulie and J. Roberts, "Bandwidth sharing: objectives and algorithms," *IEEE/ACM Trans. Netw.*, vol. 10, no. 3, pp. 320–328, Jun. 2002.
[28] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.
[29] J. Murchland, "Braess's paradox of traffic flow," *Transpn. Res.*, vol. 4, pp. 391–394, 1970.
[30] R. Srinivasan and A. Somani, "On achieving fairness and efficiency in high-speed shared medium access," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 111–124, Feb. 2003.
[31] A. Tang, J. Wang, and S. H. Low, "Is fair allocation always inefficient," in *Proc. IEEE INFOCOM*, Mar. 2004.
[32] A. Tang, J. Wang, S. H. Low, and M. Chiang, "Network equilibrium of heterogeneous congestion control protocols," in *Proc. IEEE INFOCOM*, 2005, pp. 1338–1349.
[33] H. Varian, *Microeconomic Analysis*, 3 ed. New York: Norton, 1992.
[34] H. Yaiche, R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 667–678, Oct. 2000.

Ao Tang (S'01) received the B.E. and M.E. degrees from Tsinghua University, Beijing, China, in 1999 and 2001, respectively, and the M.S. degree from California Institute of Technology (Caltech), Pasadena, CA, in 2002, all in electrical engineering. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering at Caltech.

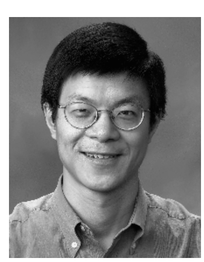
His research interests include resouce allocation and congestion control in networks.

Mr. Tang was the recipient of the 2002 Outstanding Master Thesis Award from Tsinghua University, for his 2001 master thesis "Spatial Characteristics of outdoor wireless channels". At Caltech, he was awarded the Atwood Fellowship and the SISL/IST Fellowship.



Jiantao Wang (S'05) received the B.S. and M.S. degrees in automatic control from Tsinghua University, Beijing, China, in 1997 and 1999, respectively. He is currently pursuing the Ph.D. degree in control and dynamical systems at California Institute of Technology (Caltech), Pasadena, CA.

His research interests include network congestion control, active queue management, and routing.



Steven H. Low (M'92–SM'99) received the B.S. degree from Cornell University, Ithaca, NY, and the M.S. and Ph.D. degrees from the University of California at Berkeley, all in electrical engineering.

He is an Associate Professor at Caltech, where he leads the FAST Project, and a Senior Fellow of the University of Melbourne, Australia. He was with AT&T Bell Laboratories, Murray Hill, NJ, from 1992 to 1996, and with the University of Melbourne from 1996 to 2000. His research interests are in the control and optimization of networks and protocols.

Prof. Low was a co-recipient of the IEEE William R. Bennett Prize Paper Award in 1997 and the 1996 R&D 100 Award. He is on the editorial boards of the IEEE/ACM TRANSACTIONS ON NETWORKING, *ACM Computing Surveys*, *Computer Networks Journal*, *NOW Foundations and Trends in Networking*, and is a Senior Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.