Is Fair Allocation Always Inefficient

Ao Tang Jiantao Wang Steven H. Low
California Institute of Technology, Pasadena, CA 91125, USA
{aotang, jiantao, slow}@caltech.edu

***Abstract*— No.**

I. Introduction

A central issue in networking is how to allocate bandwidth to flows *efficiently* and *fairly*, in a decentralized manner. Here, the efficiency of an allocation policy is measured by the aggregate throughput of all the flows under the policy. Maximizing aggregate throughput however can be extremely unfair (see Example 1 below). On the other extreme, max-min fairness [1] is generally viewed as the ideal fairness criteria that generalizes equal sharing at a single resource to a network of resources in a way that maintains Pareto optimality. It is often believed however that this is achieved at the cost of reduced aggregate throughput compared with other fairness criteria. In this paper, we study the tradeoff between fairness and throughput in a general network.

How should we compare fairness criteria of different allocation policies? It is shown in [5] that an allocation policy can be expressed in terms of a utility function $U_i(x_i)$, as a function of source rate x_i, in the sense that the desired bandwidth allocation $x^* = (x_i^*$, all sources $i)$ maximizes aggregate utility $\sum_i U_i(x_i)$ subject to capacity constraints $Rx \le c$. Indeed, various congestion control algorithms in Transmission Control Protocol (TCP) can be interpreted as solving the same utility maximization problem with different utility functions [9]. The authors of [5] advocate *proportional fairness*, characterized by $U_i(x_i) = \log x_i$. In [11], an allocation policy called *minimum potential delay* is proposed with $U_i(x_i) = -1/x_i$, which is shown in [6] to approximate the fairness of the TCP on the current Internet. In [13], the following class of utility functions is proposed

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

for $\alpha \ge 0$. This includes all the previously considered allocation policies – maximum throughput ($\alpha = 0$), proportional fairness ($\alpha = 1$), minimum potential delay ($\alpha = 2$), and max-min fairness ($\alpha = \infty$) – and provides a convenient way to compare different fairness criteria.

Is a fairer policy (one with larger α) always less efficient (has a smaller aggregate throughput)? This conjecture is prompted by the various examples in resource allocation in the literature in wired networks [11], [13], [2], in wireless networks, [7], [15], in economics, [3], etc. These examples seem to illustrate (quoted from [7])

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

In Section II, we present our model and state this conjecture formally. We review in Section III some special cases considered in the networking literature that confirm the conjecture.

The truth of the conjecture for the general case turns out to depend critically on the network topology in terms of routing and link capacities. The simplicity of the examples in Section III allows one to prove the conjecture in the affirmative without exploiting this underlying structure. In Section IV, we clarify this structure and present a necessary and sufficient condition for the conjecture to hold. This characterization leads to a trivial sufficient condition, which turns out to be satisfied by all the examples in the literature we examined. It also leads us to the first counter-example to the conjecture. Surprisingly, we are able to construct a class of simple networks in which a fairer allocation is *always* (for all $\alpha > 0$) more efficient!

We prove these results in Section V and conclude in Section VI with limitations of this work.

II. Model

Consider a set of L links, indexed by l, with finite capacities C_l. It is shared by a set of 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. Suppose all sources have a common utility function given in (1) with the same α. When α is clear from the context, we will use $U_i(x_i)$ in place of $U_i(x_i, \alpha)$. In general, ^{T} denotes transpose and z denotes the vector $z = (z_1, \ldots, z_n)^T$ when z_i are previously defined. We use $\log$ to deonte natural logarithm.

Consider the utility maximization problem:

$$\max_{x \ge 0} \quad U(x, \alpha) := \sum_i U_i(x_i, \alpha) \tag{2}$$

$$\text{subject to} \quad Rx \le c \tag{3}$$

A maximizer for (2)–(3) always exists since the utility functions are concave, and hence continuous, and the feasible set is compact. It is unique if $\alpha > 0$ when the utility function is strictly concave. Denote by $x(\alpha)$ the unique maximizer when $\alpha > 0$ and a maximizer when $\alpha = 0$. Consider the aggregate

0-7803-8356-7/04/$20.00 (C) 2004 IEEE

throughput

$$T(\alpha) \quad := \quad \sum_i x_i(\alpha) \tag{4}$$

Conjecture 1. *$T(\alpha)$ is nondecreasing:*

$$\frac{dT}{d\alpha} \quad \leq \quad 0 \qquad \textit{for } \alpha > 0$$

III. Special cases

In this section, we review several examples in the literature in which the conjecture is true for max-min fairness, minimum potential delay, and proportional fairness. These special cases motivate the conjecture and illustrate the means by which this question has been studied previously: by analytically solving (2)–(3) or numerically computing $T(\alpha)$.

As we will explain in the next section, the underlying network topology in all these examples possesses a special structure that is far from apparent in previous analysis but that leads to a trivial sufficient condition for the conjecture to be true.

Example 1: Linear network with uniform capacity
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 Figure 1. Source 0 goes through all the

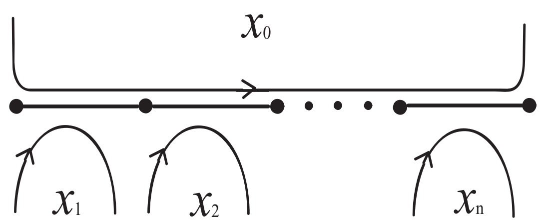

Fig. 1. Linear network.

L links and sources $i \geq 1$ go through links $l = i$. All links have the same capacity of 1 unit.

In [11], the throughput of each source and their aggregate have been calculated for several α values:

max-min fairness: $x_i(\infty) = \frac{1}{2}, \quad i \geq 0$

$T(\infty) = \frac{1}{2}(L+1)$

min potential delay: $x_0(2) = \frac{1}{\sqrt{L}+1},$

$x_i(2) = \frac{\sqrt{L}}{\sqrt{L}+1}, \; i \geq 1$

$T(2) = L - \sqrt{L} + 1$

proportional fairness: $x_0(1) = \frac{1}{L+1},$

$x_i(1) = \frac{L}{L+1}, \; i \geq 1$

$T(1) = L - \frac{L-1}{L+1}$

max throughput: $x_0(0) = 0, \quad x_i(0) = 1, \; i \geq 1$

$T(0) = L$

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

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

The above calculation also shows that the loss in efficiency of max-min fairness ($T(\alpha)/T(0)$) becomes more severe, but that of minimum potential delay and proportional fairness becomes less severe, as the number L of links increases. Numerical examples of aggregate throughput, normalized by the maximum, are shown in Table I for these allocation policies. The authors of [11] make a cautious comment: "It is not known whether the same ordering holds for arbitrary network topologies".

TABLE I
Throughput Comparison

# of links	maximum throughput	proportional fairness	minimum potential delay	max-min fairness
L=3	100%	83%	76%	67%
L=10	100%	92%	78%	55%

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

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

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

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

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
The linear network of Example 1 is considered in [13] with $L = 2$, but with different link capacities $c_1 < c_2$. The authors calculated the source rates under max-min fairness:

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

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)–(3) directly to obtain the source rates under proportional and max-min fairness for this example:

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

The throughputs for proportional and max-min fairness are supporting 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 Figure 2. We choose $c = (500, 400, 300, 200, 500)^T$ and

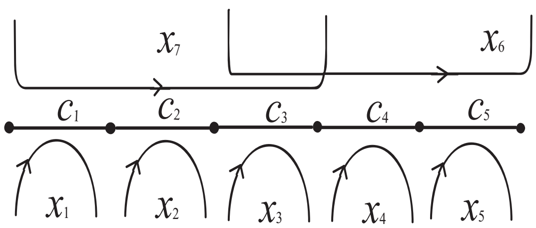

Fig. 2. Linear network with two long flows.

numerically solve for $T(\alpha)$ for $\alpha > 0$. The result is shown in Figure 3. It suggests that the conjecture is true for all $\alpha > 0$

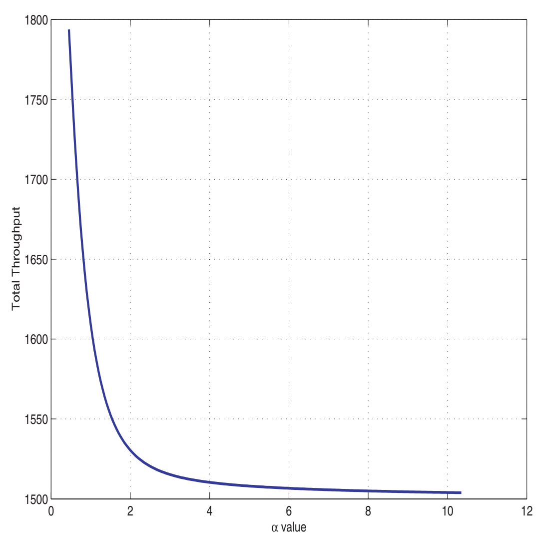

Fig. 3. Fairness-efficiency tradeoff $T(\alpha)$: linear network with two long flows.

for this network. Corollary 3 below implies that, indeed, it is.

IV. MAIN RESULTS

It turns out that the behavior of the fairness-efficiency tradeoff $T(\alpha)$ depends critically on the network topology expressed by the routing matrix R and the capacity vector c. The simplicity of the examples in the previous section (except Example 3) allows one to prove the conjecture without exploiting this underlying structure. It is this structure, however, that leads us to the first counter-example to the conjecture.

For the rest of the paper, consider $\alpha > 0$ so that the utility functions in (1) are *strictly* concave and the solution $x(\alpha)$ of (2)–(3) is unique. From Lemma 5 below, $x(\alpha)$ is a continuous function of α. Moreover, $x(\alpha)$ is differentiable except at a finite number of points when the active constraint set at optimal $x(\alpha)$ changes as α is perturbed. Hence, we can study $dT/d\alpha$ in between these points[1]. Consider the utility maximization with equality constraints that represent only those constraints that are active at optimality:

$$\max_x U(x, \alpha) \quad \text{s.t.} \quad Rx = c \tag{5}$$

The solution $x(\alpha)$ of (2)–(3) is also a local solution of (5)[2]. Since (5) is a convex optimization problem, Local solution $x(\alpha)$ is also the global unique solution of (5). We will stick to problem (5) for the rest of the paper.

Finally, it is worthwhile to notice that 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 $N \ge L$ 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 $(z_m, m = 1, \dots, M)$, $z_m \in \Re^N$, be any basis of the null space of R, and let $Z = [z_1\ z_2\ \dots\ z_M]$ be the matrix with z_m as its columns. As we will see below, the null space represented by Z and its dimension M play a critical role in determining whether the conjecture is true.

Fix an $\alpha > 0$ and recall $x(\alpha)$ is the unique solution of (5). Let $D = D(x(\alpha), \alpha)$ denote the curvature of the utility function at optimal allocation $x = x(\alpha)$:

$$D := -\frac{\partial^2 U}{\partial x^2} = \alpha\, \text{diag}(x_1^{-\alpha-1}, \dots, x_N^{-\alpha-1}) \tag{6}$$

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

$$b := \frac{\partial^2 U}{\partial x \partial \alpha} = -(x_1^{-\alpha}\log x_1, \dots, x_N^{-\alpha}\log x_N)^T \tag{7}$$

Let $\mu = \mu(x(\alpha), \alpha)$, $\beta = \beta(x(\alpha), \alpha)$, and $A = A(x(\alpha), \alpha)$ be defined by:

$$\mu_i := z_i^T b, \quad \beta_i := -e^T z_i, \quad A := Z^T D Z \tag{8}$$

where $e = (1, \dots, 1)^T$. Then $T(x(\alpha)) = e^T x(\alpha)$. Let $\bar{A}_i(x(\alpha), \alpha)$ denote the matrix obtained from replacing the i^{th} row of A with row vector $\beta = (\beta_1, \beta_2 \dots \beta_M)$.

[1] Hence, all our statements below on $dT/d\alpha$ should be interpreted piecewise in between non-differentiable points of α.

[2] We abuse the notation to use R to denote both the full routing matrix in (2)–(3), and the sub-matrix in (5) that consists of only links with positive equilibrium prices.

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 2. *For any* $\alpha > 0$

$$\frac{dT}{d\alpha} \le 0 \quad \textit{if and only if} \quad \sum_{i=1}^{M} \mu_i \det \bar{A}_i \ge 0$$

This characterization leads directly to two sufficient conditions that explain all the examples in Section III. The first condition implies that the conjecture is true when every (bottleneck) link has a single-link flow and there is only one long flow. The second condition implies that the conjecture is true when there are two long flows but both pass through the same number of links.

Corollary 3. *Suppose every link has a single-link flow.*

1) *If* $dim(Z) = 1$*, then* $\frac{dT}{d\alpha} \le 0$ *for all* $\alpha > 0$*.*
2) *If* $dim(Z) = 2$ *and the only two long flows pass through the same number of links, then* $\frac{dT}{d\alpha} \le 0$ *for all* $\alpha > 0$*.*

For Examples 1 and 2 in Section III, there is only one long flow and hence the dimension $\dim(Z)$ of the null space of R is 1. Therefore the first part of Corollary 3 is satisfied. The network in Example 3 has two long flows both passing through 3 links, satisfying the second sufficient condition of Corollary 3.

The condition in the second part of Corollary 3 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 4. *When* $dim(Z) \ge 2$*, for any* $\alpha_0 > 0$*, there exists a network such that*

$$\frac{dT}{d\alpha} \; > \; 0 \qquad \textit{for all } \alpha > \alpha_0$$

Since in reality, there are many more flows than bottleneck links and therefore $\dim(Z)$ is typically large, it is conceivable that the conjecture is wrong more often than right in practice.

The proof of the theorem, in Section V, is by constructing a linear network with a one-link flow at every link and two flows that pass through different number of links (the difference is just 2 links).

Example 4: counter-example

Consider the linear network in Figure 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 $\alpha \in [0.5, 10]$. As the change in total throughput $T(\alpha)$ is small for the parameters we have chosen, accuracy is important in the numerical solution. In our solution, the Karush-Kuhn-Tucker condition [14] is satisfied with an accuracy of 10^{-6}.

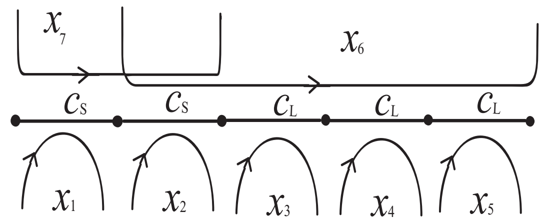

Fig. 4. Network for counter-example in Theorem 4.

The aggregate throughput $T(\alpha)$ is plotted in Figure 5 and Figure 6 as a function of α, for capacities $C_S = 10$ and $C_L = 1,000$ and for capacities $C_S = 10$ and $C_L = 5,000$, respectively. The minimal throughput is achieved around $\alpha_0 = 0.95$ in the former case, and around $\alpha_0 = 0.75$ in the latter case. $T(\alpha)$ is strictly increasing beyond the α_0. In particular,

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

□

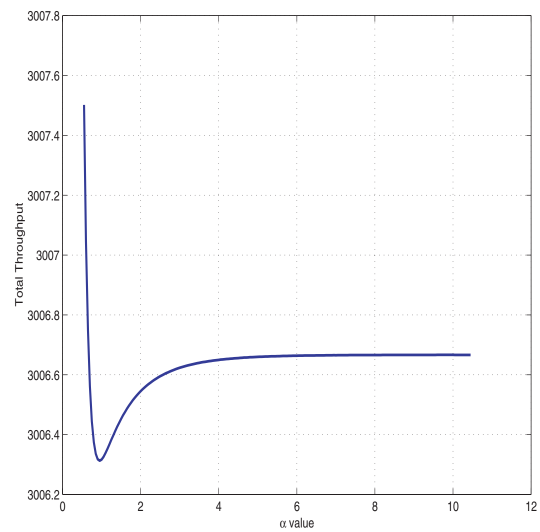

Fig. 5. Counter-example: $C_S = 10$ and $C_L = 1000$.

The example is surprising at first because the conventional wisdom in networking is that increasing α favors long flows which take up more resources, leading to a drop in aggregate throughput. This is not exactly right. Associate with each link the Lagrange multiplier (called *price*) of the utility maximization problem (2)–(3); see e.g. [8]. The price at a link is a precise measure of congestion at that link. Then 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 leads to increase in flows x_3, x_4, x_5 so that the

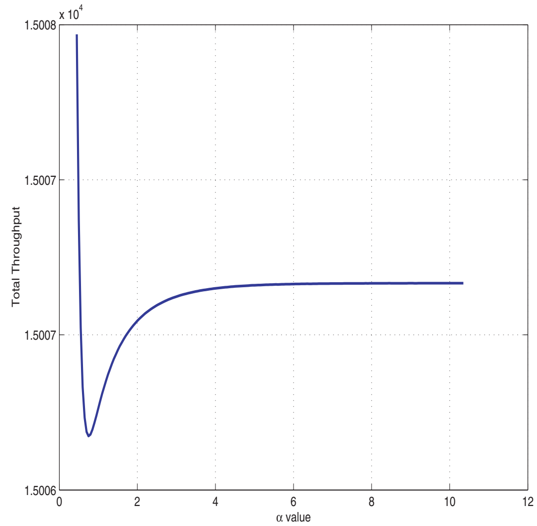

Fig. 6. Counter-example: $C_S = 10$ and $C_L = 5000$.

net change in aggregate throughput $T(\alpha)$ is positive. Hence the counter-example relies on the design that the longest flows are not the most expensive ones.

Indeed, one can prove that for the network in Figure 4, $dx_7/d\alpha > 0$ and $dx_6/d\alpha < 0$ for all $\alpha > \alpha_0$, as illustrated in Figure 7. In this example, as α increases beyond α_0, the decrease in x_6 produces an increase in three one-link flows, leading to a net increase in the aggregate throughput. If flow 6 traverses more than three links, the increase in aggregate throughput should be larger. Our example however is compact in that our proof shows that x_6 has to pass through at least three links (link 3,4,5) to make $dT/d\alpha > 0$ (see the remark after the proof of Theorem 4 in Section V).

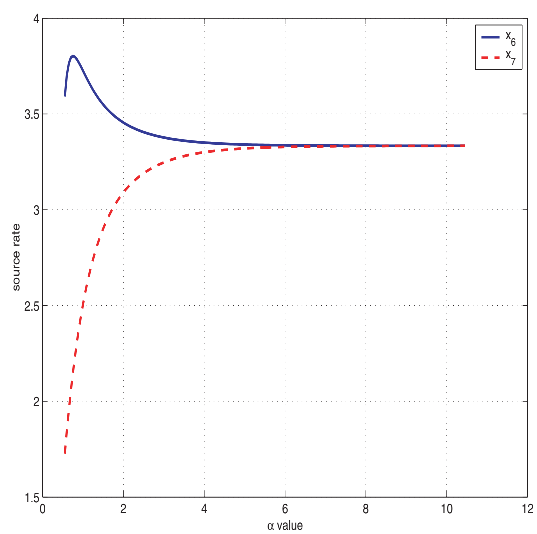

Fig. 7. Counter-example: $x_6(\alpha)$ and $x_7(\alpha)$, as a function of α, with $c_S = 10$ and $c_L = 1000$.

V. Proofs and intuitions

We start with an important property of $x(\alpha)$, quoted directly from [16].

Lemma 5. *For $\alpha > 0$, the unique solution $x(\alpha)$ of (5) is continuous and differentiable.*

The aggregate throughput $T(\alpha)$ is the sum of $x_i(\alpha)$. The next result derives $dx/d\alpha$ explicitly and is the starting point of our analysis.

Theorem 6. *For $\alpha > 0$, the optimal solution $x(\alpha)$ satisfies*

$$\frac{dx(\alpha)}{d\alpha} \quad = \quad Z(Z^T D Z)^{-1} Z^T b \tag{9}$$

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

Proof: Suppose $\overline{x}$ is an arbitrary vector that satisfies $R\overline{x} = c$. Because the range space of Z is the null space of R, every vector x that satisfies the constraint can be uniquely expressed as:

$$x = \overline{x} + Zy, \quad y \in \Re^M \tag{10}$$

Define

$$y(\alpha) \quad = \quad \arg\max_{y \in \Re^M} U(\overline{x} + Zy, \alpha) \tag{11}$$

Note that we can ignore the nonnegativity constraint $x \geq 0$ in the utility maximization problem (2)–(3), because the utility function U defined in (1) is such that the constraint $x_i = 0$ will not be active at optimality. This means that unconstrained maximization in (11) is equivalent to the constrained maximization (2)–(3)). Hence from (10), we have:

$$x(\alpha) = \overline{x} + Zy(\alpha) \quad \text{ and } \quad \frac{dx}{dy} = Z$$

Define, for any $\alpha > 0$,

$$G(y, \alpha) \quad := \quad \frac{\partial U(\overline{x} + Zy, \alpha)}{\partial y}$$

Then[3]

$$G(y, \alpha) \quad = \quad \left(\frac{dx}{dy}\right)^T \frac{\partial U}{\partial x} \quad = \quad Z^T \frac{\partial U}{\partial x} \tag{12}$$

Since $y(\alpha)$ is the unique unconstrained maximizer of $U(\overline{x} + Zy, \alpha)$ for a given α, the optimality condition for unconstrained problem gives

$$\left.\frac{\partial U(\overline{x} + Zy, \alpha)}{\partial y}\right|_{y=y(\alpha)} \quad = \quad 0 \tag{13}$$

or equivalently, $G(y(\alpha), \alpha) \equiv 0$ for all $\alpha > 0$. Moreover

$$\frac{\partial G}{\partial y} \quad = \quad Z^T \frac{\partial^2 U}{\partial x^2} \frac{dx}{dy} \quad = \quad -Z^T D Z \tag{14}$$

where D is defined in (6). Since $U(x, \alpha)$ is strictly concave in x, $D = -\frac{\partial^2 U}{\partial x^2}$ is positive definite for any x. Then for any nonzero vector $v \in \Re^M$,

$$v^T Z^T D Z v = (Zv)^T D (Zv) > 0 \tag{15}$$

[3] In the rest of this proof, $\partial U/\partial\alpha$, $\partial x/\partial y$, etc. are evaluated at $x(\alpha) = \overline{x} + y(\alpha)$.

Hence the matrix $Z^T DZ$ is positive definite, and $\frac{\partial G}{\partial y}$ is an invertible matrix for any positive α.

We can then use the implicit function theorem to conclude that there exists a unique continuous function $y(\alpha) : \Re \to \Re^M$ such that $G(y(\alpha), \alpha) = 0$.[4] Moreover,

$$\begin{aligned} \frac{dy}{d\alpha} &= -\left(\frac{\partial G}{\partial y}\right)^{-1} \frac{\partial G}{\partial \alpha} \\ &= \left(Z^T DZ\right)^{-1} \frac{\partial G}{\partial \alpha} \qquad \text{(from (14))} \\ &= \left(Z^T DZ\right)^{-1} Z^T \frac{\partial^2 U}{\partial x \partial \alpha} \qquad \text{(from (12))} \\ &= \left(Z^T DZ\right)^{-1} Z^T b \end{aligned}$$

where b is defined in (7). Finally,

$$\frac{dx}{d\alpha} = Z \frac{dy}{d\alpha} = Z \left(Z^T DZ\right)^{-1} Z^T b$$

Here, we have used the assumption that the active constraint set is unchanged when α is perturbed locally (i.e., we consider problem (5) instead of problem (2)–(3)), so that Z is independent of α. □

The vector $D^{-1}b$ is usually called the Newton direction, which is the moving direction $dx/d\alpha$ for unconstrained optimization problem. When the constraint is given by $Rx = c$, the feasible set is in the null space of R (with a shift by $\overline{x}$). In this situation, intuitively, the moving direction will be "some projection" of the Newton direction to this subspace, which is exactly expressed by (9).

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

$$T(x(\alpha)) = e^T x(\alpha)$$

The meaning should be clear from the context. Define

$$\lambda = \lambda(\alpha) := \frac{dT}{d\alpha}$$

The next result shows that λ is the nonzero solution of a polynomial.

Lemma 7. *λ is the nonzero root of the equation:*

$$\det\left(Z^T \left(\lambda D - be^T\right) Z\right) = 0$$

Proof: From Theorem 6,

$$\lambda = \frac{dT}{d\alpha} = e^T \frac{dx}{d\alpha} = e^T Z \left(Z^T DZ\right)^{-1} Z^T b$$

We can express it in terms of the matrix trace:

$$\begin{aligned} \lambda &= tr(e^T Z (Z^T DZ)^{-1} Z^T b) \\ &= tr((Z^T DZ)^{-1} Z^T be^T Z) \end{aligned}$$

where the last equality follows from the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrices A, B of the proper dimensions.

Since $\left(Z^T DZ\right)^{-1} Z^T be^T Z$ is a rank one matrix, it can have at most one nonzero eigenvalue. Hence its trace λ (the sum of all the eigenvalues) is exactly this nonzero eigenvalue. Suppose the eigenvector corresponding to λ is v. Then

$$\left(Z^T DZ\right)^{-1} Z^T be^T Zv = \lambda v$$

Multiplying both sides by $Z^T DZ$, we have

$$Z^T be^T Zv = \lambda Z^T DZv$$

and hence

$$Z^T(\lambda D - be^T)Zv = 0$$

Therefore λ is the nonzero solution of the equation:

$$\det\left(Z^T \left(\lambda D - be^T\right) Z\right) = 0$$

Since any solution of the above equation is an eigenvalue of the rank one matrix $\left(Z^T DZ\right)^{-1} Z^T be^T Z$, there can only be at most one nonzero solution. □

Denote the polynomial in Lemma 7 by $F(\lambda)$:

$$F(\lambda) := \det(Z^T(\lambda D - be^T)Z)$$

Since the matrix has at most one nonzero eigenvalue, $F(\lambda)$ must be of the form $F(\lambda) = \lambda^{M-1}(\lambda - \lambda_0)$. The next theorem provides an explicit expression for the nonzero root of $F(\lambda)$. Since the matrix A defined in (8) is positive definite, $\det A > 0$. Hence, the next theorem implies our main result, Theorem 2, of the previous section.

Theorem 8. *For all $\alpha > 0$*

$$\lambda = -\frac{1}{\det A} \sum_{i=1}^{M} \mu_i \det \bar{A}_i$$

Proof: Suppose

$$F(\lambda) = \sum_{m=0}^{M} f_m \lambda^m \tag{16}$$

We will prove directly that $f_m = 0$ for $m = 0, \ldots, m-2$.

Let $\Delta := Z^T \left(\lambda D - be^T\right) Z$. Then

$$\Delta = \lambda A + W \quad \text{and} \quad F(\lambda) = \det \Delta$$

where $A = Z^T DZ$ is positive definite, and W is the rank one matrix $W := \mu\beta$, with μ and β defined in (8).

From (16), we have that

$$f_m = \frac{1}{m!} \left. \frac{d^m F(\lambda)}{d\lambda^m} \right|_{\lambda=0} \tag{17}$$

For $m = 0$,

$$f_0 = \det(W) = 0$$

since $\mathrm{rank}(W) = 1$. The derivatives of the determinant of Δ, whose entries are differentiable, can be calculated as (see [10]):

$$\frac{d(\det \Delta)}{d\lambda} = \det(P_1) + \ldots + \det(P_M) \tag{18}$$

[4]The existence and uniqueness of $y(\alpha)$ also follows from the existence and uniqueness of the solution for (2)–(3).

where P_m is identical to Δ except the entries in the m^{th} row are replaced by their derivatives.

For $m = 0, 1, \ldots, M$, let $S(\Delta, m) :=$ { all matrices which are generated by replacing m rows of Δ with their derivatives }. Then

$$\frac{dF(\lambda)}{d\lambda} = \sum_{P \in S(\Delta,1)} \det P$$

Moreover, since each entry of Δ is a linear function of λ, the second derivative of any row of Δ is zero. When we repeatedly use (18) to calculate the k^{th} derivative of $\det \Delta$, all of the nonzero determinants are generated by matrices with k *different* rows replaced by their corresponding derivatives. Hence, we have, for $m = 1, \ldots, M$,

$$\frac{d^m F(\lambda)}{d\lambda^m} = m! \sum_{P \in S(\Delta,m)} \det P$$

Consider a matrix $P \in S(\Delta, m)$ where $M - m$ rows are the corresponding rows of Δ, and the other m rows are the derivatives of the corresponding rows of Δ. Since $\Delta = \lambda A + W$, these m rows of P are just the corresponding rows of A. The remaining $M - m$ rows of P will be the corresponding rows of W when we set $\lambda = 0$.

For $m = 0, 1, \ldots, M$, let $T(A, m) :=$ { all matrices which are generated by replacing $M - m$ rows of A with the corresponding rows of W }. Then

$$\begin{aligned} \left.\frac{d^m F(\lambda)}{d\lambda^m}\right|_{\lambda=0} &= m! \sum_{P \in S(\Delta,m)} \det P|_{\lambda=0} \\ &= m! \sum_{E \in T(A,m)} \det E \end{aligned}$$

and therefore,

$$f_m = \frac{1}{m!} \left.\frac{d^m F(\lambda)}{d\lambda^m}\right|_{\lambda=0} = \sum_{E \in T(A,k)} \det E$$

Since rank$(W) = 1$, any two rows of W are linearly dependent. This means that any $E \in T(A, m)$ with $m \leq M - 2$ has at least two linearly dependent rows. Hence

$$\det E = 0 \quad \text{for any } E \in T(A, m),\ m \leq M - 2$$

Therefore we have

$$\begin{aligned} f_m &= 0, \quad m = 0, 1, \ldots M - 2 \\ f_{M-1} &= \sum_{E \in T(A,M-1)} \det E \\ f_M &= \sum_{E \in T(A,M)} \det E = \det A \end{aligned}$$

Since $T(A, M - 1)$ contains all matrices that are obtained from A by replacing one of its rows, say the i^{th} row, with $(\mu_i \beta_1, \ldots, \mu_i \beta_M)$, we have

$$f_{M-1} = \sum_{i=1}^{M} \mu_i \det \bar{A}_i$$

Hence

$$F(\lambda) = f_M \lambda^M + f_{M-1} \lambda^{M-1}$$

Then nonzero root of $F(\lambda)$ is given by:

$$\lambda = -\frac{f_{M-1}}{f_M} = -\frac{\sum_{i=1}^M \mu_i \det \bar{A}_i}{\det A}$$

□

We now specialize to a particular basis Z of the null space of the routing matrix R to simplify the proof of the corollaries. Suppose every link has a single-link flow. We can then rearrange the column of routing matrix R to express R as

$$R = \begin{bmatrix} I_L & R_1 \end{bmatrix}$$

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

$$Z = \begin{bmatrix} -R_1 \\ I_M \end{bmatrix} \tag{19}$$

Clearly rank(Z) = dim$(Z) = M$.

Lemma 9. *Suppose every link has a single-link flow. For Z in the form of (19), we have*

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

Proof: The Karush-Kuhn-Tucker condition implies that there are nonnegative $p \in \Re^L$ such that (see e.g. [8])

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

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

$$Z^T \frac{\partial U}{\partial x} = 0 \tag{20}$$

From (1),

$$\frac{\partial U}{\partial x} = (x_1^{-\alpha}, \ldots, x_N^{-\alpha})^T$$

Suppose $z_m = (z_{m1}, \ldots, z_{mN})^T$ is the m^{th} row of matrix Z^T. Then (19), (1), and (20) imply that, for $m = 1, \ldots, M$,

$$x_{L+m}^{-\alpha} = \sum_{j=1}^{L} -z_{mj} x_j^{-\alpha} \tag{21}$$

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

$$x_{L+m} \leq x_j \quad \text{for } j = 1, \ldots, L,\ z_{mj} \neq 0 \tag{22}$$

From (7), b is

$$b = \frac{\partial^2 U}{\partial x \partial \alpha} = -(x_1^{-\alpha} \log x_1, \ldots, x_N^{-\alpha} \log x_N)^T$$

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 (21) and the inequality follows from (22). This proves the first assertion.

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

$$D = -\frac{\partial^2 U}{\partial x^2} = \alpha \ \text{diag}(x_1^{-\alpha-1}, \ldots, x_N^{-\alpha-1})$$

Then

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

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 3 of the previous section, reproduced here.

Corollary 3. *Suppose every link has a single-link flow.*

1) *If $dim(Z) = 1$, then $\frac{dT}{d\alpha} \leq 0$.*
2) *If $dim(Z) = 2$ and the only two long flows pass through the same number of links, then $\frac{dT}{d\alpha} \leq 0$.*

Proof: 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 $\sum_{j=1}^{L} -z_{1j} \leq 1$. Hence

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

From Lemma 9, we know that $\mu_1 > 0$. From Theorem 8 we have

$$\lambda = \frac{dT}{d\alpha} = -\frac{\mu_1 \beta_1}{\det A} \leq 0$$

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

$$\beta_1 = \beta_2 = -e z_1 \geq 0 \tag{23}$$

as in the first assertion. From theorem 8 we have:

$$\begin{aligned}
f_{M-1} &= \mu_1 \det \bar{A}_1 + \mu_2 \det \bar{A}_2 \\
&= \mu_1 \det \begin{bmatrix} \beta_1 & \beta_2 \\ a_{21} & a_{22} \end{bmatrix} + \mu_2 \det \begin{bmatrix} a_{11} & a_{12} \\ \beta_1 & \beta_2 \end{bmatrix} \\
&= \beta_1 \left[\mu_1 (a_{22} - a_{21}) + \mu_2 (a_{11} - a_{12})\right]
\end{aligned}$$

Lemma 9 and (23) then implies $f_{M-1} \geq 0$. Hence

$$\frac{dT}{d\alpha} = -\frac{f_{M-1}}{\det A} \leq 0$$

□

The final result proves our counter-example in the previous section.

Theorem 4. *When $dim(Z) \geq 2$, for any $\alpha_0 > 0$, there exists a network such that*

$$\frac{dT}{d\alpha} > 0 \qquad \textit{for all } \alpha > \alpha_0$$

Proof: It is sufficient to prove the assertion with the network shown in Figure 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 Figure 4, with five links $l = 1, 2, \ldots, 5$. Let their capacities be $c_1 = c_2 = C_S \geq 3$ and $c_3 = c_4 = c_5 = C_L$. Fixed 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:

$$R = \begin{bmatrix} & 0 & 1 \\ & 1 & 1 \\ I_5 & 1 & 0 \\ & 1 & 0 \\ & 1 & 0 \end{bmatrix} \qquad Z = \begin{bmatrix} 0 & -1 \\ -1 & -1 \\ -1 & 0 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{24}$$

From (8), we have

$$\beta_1 = -e^T z_1 = 3 \qquad \beta_2 = -e^T z_2 = 1$$

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

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

Theorem 2 then implies that $dT/d\alpha > 0$ for all $\alpha > \alpha_0$.

The basic idea of the proof is as follows. From Lemma 9, we know that $\mu_m \geq 0$, $m = 1, 2$ (strict inequality here), and

$a_{mm} \geq a_{mn}$. Hence the first term is positive. We can also show that the second term is negative since $a_{11} - 3a_{12}$ can be made strictly negative by appropriate choice of C_L. We will however show that although both terms go to zero when α goes to infinity, it is possible to choose link capacity C_L such that the first term is arbitrarily small compare to the absolute value of the second term for all $\alpha > \alpha_0$, so that the sum of the first and second term 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 [8] for the utility maximization problem, which will be used extensively later.

Lemma 10. *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) $q_i = p_i, i = 1, 2, 3, 4, 5$; $q_6 = \sum_{i=2}^{5} p_i$; $q_7 = p_1 + p_2$
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) $K_2 = 2^{-\frac{1}{\alpha_0}} \frac{C_S}{3}$;
3) $K_3 = \frac{3K_1 C_S}{\alpha_0}$;
4) $K_4 = \frac{K_3}{K_2} + 3K_1 \max(1, 1 - \log K_2)$;
5) $K_5 = \log(\frac{3}{2})$;
6) $K_6 = 5 + 6 \times 2^{\frac{1}{\alpha_0}}$;
7) $K_7 = 3K_1 C_S \frac{\alpha_0+1}{\alpha_0} 2^{\frac{\alpha_0+2}{\alpha_0}}$;
8) $K_8 = \frac{1}{4C_S}$;
9) $K = \frac{K_5 K_8}{K_4 K_6 + K_5 K_7}$.

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

$$\frac{C_L}{C_S} - 1 > 3C_S \quad , \quad (\frac{C_L}{C_S} - 1)^{-\alpha_0} \log C_L < \epsilon \tag{25}$$

For any $\alpha > \alpha_0$,

$$(\frac{C_L}{C_S} - 1)^{-\alpha} < (\frac{C_L}{C_S} - 1)^{-\alpha_0} \log C_L < \epsilon \tag{26}$$

Noting (25), we can also have a much tighter bound than (26):

$$(\frac{C_L}{C_S} - 1)^{-\alpha} < \epsilon(\frac{C_L}{C_S} - 1)^{-(\alpha-\alpha_0)} < K_1(3C_S)^{-\alpha}\epsilon \tag{27}$$

The next Lemma upper and lower bounds all the rates.

Lemma 11. 1) $\frac{C_S}{2} \leq x_1 \leq C_S$; $\frac{C_S}{3} \leq x_2 \leq C_S$;
2) $C_L - C_S \leq x_i \leq C_S, i = 3, 4, 5$;
3) $K_2 \leq x_7 \leq \frac{C_S}{3}$; $K_2 \leq x_6 \leq \frac{C_S}{2}$;

Proof:

1) $x_1 = C_S - x_7 \leq C_S$; $q_7 \geq q_1$, so $x_1 \geq x_7$, $x_1 \geq \frac{C_S}{2}$. $x_2 = C_S - x_6 - x_7 \leq C_S$; $q_2 \leq q_6$, $q_2 \leq q_7$, so $x_2 \geq x_6$, $x_2 \geq x_7$. Therefore, $x_2 \geq \frac{C_S}{3}$.
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) $q_7 = p_1 + p_2 = x_1^{-\alpha} + x_2^{-\alpha} \leq 2(\frac{C_S}{3})^{-\alpha}$, then

$$x_7 = q_7^{-\frac{1}{\alpha}} \geq 2^{-\frac{1}{\alpha}} \frac{C_S}{3} \tag{28}$$

$$\geq 2^{-\frac{1}{\alpha_0}} \frac{C_S}{3} = K_2 \tag{29}$$

Hence

$$\begin{aligned} q_7 - q_6 &= p_1 - \sum_{i=3}^{5} p_i \\ &\geq C_S^{-\alpha} - 3(C_L - C_S)^{-\alpha} \\ &= C_S^{-\alpha}(1 - 3(\frac{C_L}{C_S} - 1)^{-\alpha}) \\ &\geq C_S^{-\alpha}(1 - 3\epsilon) \geq 0 \end{aligned}$$

Hence $q_7 \geq q_6$, $x_7 \leq x_6$. Since $C_S = x_2 + x_6 + x_7 \geq 3x_7$, $x_7 \leq \frac{C_S}{3}$.
4) $x_6 \geq x_7 \geq K_2$, $x_6 = C_S - x_2 - x_7 \leq C_S - x_2 \leq \frac{C_S}{2}$.

□

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, $p_i, i = 3, 4, 5$, 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 12.

$$x_2 - x_6 \leq K_3(3C_S)^{-\alpha}\epsilon \tag{30}$$

Proof.

$$\begin{aligned} q_6 - q_2 &= p_3 + p_4 + p_5 \\ &\leq 3(C_L - C_S)^{-\alpha} \\ &= 3C_S^{-\alpha}(\frac{C_L}{C_S} - 1)^{-\alpha} \\ &\leq 3K_1 C_S^{-\alpha}(3C_S)^{-\alpha}\epsilon \end{aligned}$$

By using the intermediate value theorem then:

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

□

Now we are ready to derive a upper bound for μ_1 and a lower bound for μ_2, which will be directly used in proving the final result.

Lemma 13.

$$\mu_1 \leq K_4 C_S^{-2\alpha}\epsilon \quad , \quad \mu_2 \geq K_5 C_S^{-\alpha}$$

Proof. First,

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

Now

$$\begin{aligned}\log x_2 - \log x_6 &= \frac{1}{x_\xi}(x_2 - x_6) \quad (x_6 \le x_\xi \le x_2) \\ &\le \frac{1}{x_6}(x_2 - x_6) \le \frac{K_3}{K_2}(3C_S)^{-\alpha}\epsilon\end{aligned}$$

Hence

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

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

$$\sum_{i=3}^{5}(\log x_i - \log x_6)x_i^{-\alpha} \le 3(\log x_3 - \log K_2)x_3^{-\alpha} \qquad (31)$$

Now

$$\begin{aligned}3x_3^{-\alpha}\log x_3 &\le 3\log C_L(C_L - C_S)^{-\alpha} \\ &\le 3C_S^{-\alpha}\log C_L(\frac{C_L}{C_S} - 1)^{-\alpha} \\ &\le 3^{1-\alpha}K_1C_S^{-2\alpha}\epsilon \le 3K_1C_S^{-2\alpha}\epsilon\end{aligned}$$

and

$$\begin{aligned}-3\log(K_2)x_3^{-\alpha} &\le 3\max(0, -\log K_2)(C_L - C_S)^{-\alpha} \\ &\le 3K_1\max(0, -\log K_2)C_S^{-2\alpha}\epsilon\end{aligned}$$

In summary,

$$\begin{aligned}\mu_1 &\le (\frac{K_3}{K_2} + 3K_1\max(0, -\log K_2))C_S^{-2\alpha}\epsilon \\ &= K_4C_S^{-2\alpha}\epsilon\end{aligned}$$

and

$$\begin{aligned}\mu_2 &= x_1^{-\alpha}\log x_1 + x_2^{-\alpha}\log x_2 - x_7^{-\alpha}\log x_7 \\ &\ge x_1^{-\alpha}(\log x_1 - \log x_7) \\ &\ge C_S^{-\alpha}(\log\frac{C_S}{2} - \log\frac{C_S}{3}) \\ &= C_S^{-\alpha}\log\frac{3}{2} = K_5C_S^{-\alpha}\end{aligned}$$

□

With the upper bound for μ_1 and lower bound 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 an lower bound for the absolute value of the negative term $a_{11} - 3a_{12}$, which is exactly characterized by the next Lemma.

Lemma 14.

$$0 \le \det\begin{bmatrix} 3 & 1 \\ a_{21} & a_{22}\end{bmatrix} \le \alpha K_6 \qquad (32)$$

$$-\det\begin{bmatrix} a_{11} & a_{12} \\ 3 & 1\end{bmatrix} \ge (K_8 - K_7\epsilon)C_S^{-\alpha}\alpha > 0 \qquad (33)$$

Proof. From (8),

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

Hence

$$\begin{aligned}\det\begin{bmatrix} 3 & 1 \\ a_{21} & a_{22}\end{bmatrix} &= 3a_{22} - a_{21} \\ &= \alpha(2x_2^{-\alpha-1} + 3x_1^{-\alpha-1} + 3x_7^{-\alpha-1}) \ge 0\end{aligned}$$

Also

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

Now

$$\begin{aligned}-\det\begin{bmatrix} a_{11} & a_{12} \\ 3 & 1\end{bmatrix} &= 3a_{12} - a_{11} \\ &= \alpha(2x_2^{-\alpha-1} - 3x_3^{-\alpha-1} - x_6^{-\alpha-1})\end{aligned}$$

Since

$$\begin{aligned}x_6^{-\alpha-1} - x_2^{-\alpha-1} &= \frac{dx^{-\alpha-1}}{dx}|_{x=x_\xi}(x_6 - x_2) \; x_6 \le x_\xi \le x_2 \\ &\le (\alpha+1)x_6^{-\alpha-2}(x_2 - x_6) \\ &\le (\alpha+1)2^{\frac{\alpha+2}{\alpha}}(\frac{C_S}{3})^{-\alpha-2}(x_6 - x_2) \\ &\le \frac{\alpha_0+1}{\alpha_0}2^{\frac{\alpha_0+2}{\alpha_0}}3C_SK_1C_S^{-\alpha}\epsilon = K_7C_S^{-\alpha}\epsilon\end{aligned}$$

and

$$\begin{aligned}x_2^{-\alpha-1} - 3x_3^{-\alpha-1} &\ge C_S^{-\alpha-1} - 3(C_L - C_S)^{-\alpha-1} \\ &= C_S^{-\alpha-1}(1 - 3(\frac{C_L}{C_S} - 1)^{-\alpha-1}) \\ &\ge \frac{1}{C_S}(1 - 3\epsilon(\frac{C_L}{C_S} - 1)^{-1}))C_S^{-\alpha} \\ &\ge \frac{1}{4C_S}C_S^{-\alpha} = K_8C_S^{-\alpha}\end{aligned}$$

we have

$$\begin{aligned}-\det\begin{bmatrix} a_{11} & a_{12} \\ 3 & 1\end{bmatrix} &\ge (K_8C_S^{-\alpha} - K_7C_S^{-\alpha}\epsilon)\alpha \\ &= (K_8 - K_7\epsilon)C_S^{-\alpha}\alpha\end{aligned}$$

By the choice of ϵ, we have

$$K_8 - K_7\epsilon > 0$$

□

Now we are ready to evaluate f_{M-1},

$$
\begin{aligned}
f_{M-1} &= \mu_1 \det \begin{bmatrix} 3 & 1 \\ a_{21} & a_{22} \end{bmatrix} + \mu_2 \det \begin{bmatrix} a_{11} & a_{12} \\ 3 & 1 \end{bmatrix} \\
&\leq \alpha(K_4 C_S^{-2\alpha} \epsilon K_6 - K_5 C_S^{-\alpha}(K_8 - K_7\epsilon) C_S^{-\alpha}) \\
&= \alpha C_S^{-\alpha}(K_4 K_6 \epsilon - K_5(K_8 - K_7\epsilon)) \\
&= \alpha C_S^{-\alpha}((K_4K_6 + K_5K_7)\epsilon - K_5K_8)
\end{aligned}
$$

By the way we choose ϵ,

$$
\epsilon < \frac{K_5K_8}{K_4K_6 + K_5K_7} = K \tag{34}
$$

Then $f_{M-1} < 0$ for all $\alpha > \alpha_0$. This ends the whole proof. □

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

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

Then $f_{M-1} > 0$. Hence our counter-example is compact. □

As we discussed at the end of IV by numerical illustration, when $\alpha > \alpha_0$, x_6 decreases and x_7 increases with increasing α, which is the reason for the increase of aggregate throughput. Here is a short formal proof for that $\frac{dx_7}{d\alpha} > 0$ and $\frac{dx_6}{d\alpha} < 0$.

By setting the $e = (0,0,0,0,0,0,1)^T$, we can check the derivative of flow x_7, based on Theorem 8. The corresponding f_{M-1} is:

$$
\begin{aligned}
f_{M-1} &= \mu_1 \det \begin{bmatrix} 0 & -1 \\ a_{21} & a_{22} \end{bmatrix} + \mu_2 \det \begin{bmatrix} a_{11} & a_{12} \\ 0 & -1 \end{bmatrix} \\
&= -\mu_2 a_{11} + \mu_1 a_{21} \\
&\leq -K_5 C_S^{-\alpha} C_S^{-\alpha-1} + K_4 C_S^{-2\alpha}\epsilon \\
&= C_S^{-2\alpha}\left(K_4\epsilon - \frac{K_5}{C_S}\right)
\end{aligned}
$$

Since ϵ can be arbitrarily small, we can have that

$$
\frac{dx_7}{d\alpha} > 0
$$

Similar method will work for flow x_6's case and yield

$$
\frac{dx_6}{d\alpha} < 0
$$

VI. Conclusion

A bandwidth allocation policy can be defined by a class of utility functions parameterized by a scalar $\alpha > 0$. An allocation is *fair* if α is large and *efficient* if the aggregate source rate is large. All examples in the literature suggest that a fair allocation is necessarily inefficient. In this paper, 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, the class of networks in our counter-example is such that a fairer allocation is *always* more efficient. In particular it implies that max-min fairness may achieve higher throughput than proportional fairness.

There are a number of ways this preliminary work can be extended. First, the necessary and sufficient condition for the conjecture is hard to understand intuitively and check for large networks. It is not clear whether this condition is likely to hold or fail in practice. Second, it would be useful to clarify the relationship of fairness defined in terms of α, as in this paper, with other notions of fairness, such as Jain's index [4]. Finally, we have assumed every source has the same utility function. It would be interesting to see how the fairness definition and tradeoff results should generalize when sources have the same class of utility functions but with different α_i parameters, or have different utility functions.

Acknowledgments: We thank George Varghese of UCSD for helpful discussion that motivated this work, and John Doyle of Caltech for the inspiration for our Abstract. This work is performed as part of the FAST Project supported by NSF, ARO, AFOSR and Cisco.

References

[1] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall,1992.
[2] T. Bonald and L. Massoulie. Impact of fairness on Internet performance. In *Proceedings of ACM Sigmetrics*, pages 82–91, June 2001.
[3] M. Butler and H. Williams. The allocation of shared fixed costs. `http://www.lse.ac.uk/collections/operationalResearch/pdf/lseor02-52.pdf`.
[4] R. Jain, W. Hawe and D. Chiu. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. DEC-TR-301, September 26, 1984
[5] F. Kelly, A. Maulloo, and D. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of Operations Research Society*, 49(3):237–252, March 1998.
[6] S. Kunniyur and R. Srikant. End-to-end congestion control: utility functions, random losses and ECN marks. *IEEE/ACM Transactions on Networking*, 11(5):689-702, Oct 2003
[7] H. Luo, S. Lu, V. Bharghavan, J. Cheng, and G. Zhong. A packet scheduling approach to QoS support in multihop wireless networks. *ACM Journal of Mobile Networks and Applications (MONET), Special Issue on QoS in Heterogeneous Wireless Networks*, 2003.
[8] S. Low and D. Lapsley. Optimization flow control, I: basic algorithm and convergence. *IEEE/ACM Transactions on Networking*, 7(6):861–874, December 1999. `http://netlab.caltech.edu`.
[9] S. Low A duality model of TCP and queue management Algorithms. *IEEE/ACM Trans. on Networking*, 11(4):525-536, August 2003
[10] C. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.
[11] L. Massoulie and J. Roberts. Bandwidth sharing: objectives and algorithms. *IEEE/ACM Transactions on Networking*, 10(3):320–328, June 2002.
[12] P. Milgrom and C. Shannon. Monotone Comparative Statics. *Econometrica*,62(1):157–180,1994
[13] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, October 2000.
[14] J. Nocedal and S. Wrignt. *Numerical optimization*. Springer, 1999.
[15] R. Srinivasan and A. Somani. On achieving fairness and efficiency in high-speed shared medium access. *IEEE/ACM Transactions on Networking*, 11(1):111-124,February 2003.
[16] H. Varian. *Microeconomic analysis*. W. W. Norton & Company, 3 edition, 1992.