Is Fair Allocation Always Inefficient*

Ao Tang
Dept. of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125
aotang@caltech.edu

Jiantao Wang
Dept. of Control and Dynamical System
California Institute of Technology
Pasadena, CA 91125
jiantao@cds.caltech.edu

Steven H. Low
Dept. of Electrical Engineering and Dept. of Computer Science
California Institute of Technology
Pasadena, CA 91125
slow@caltech.edu

Abstract

A bandwidth allocation policy can be defined by a class of utility functions parameterized by a scalar $\alpha > 0$. An allocation is fair if $\alpha$ 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 trade-off 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 even higher throughput than proportional fairness.

1 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 examples below). On the other extreme, max-min fairness [2] 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 trade-off between fairness and throughput in general networks.

How to compare fairness criteria of different allocation policies? It is shown in [3] that an allocation policy can be expressed in terms of a utility function $U_i(x_i)$, as a function of

---

*This paper is based on work supported by NSF through grants ANI-0113425 and ANI-0230967, Caltech Lee Center for Advanced Networking, AFOSR through grant F49620-03-1-0119, ARO through grant DAAD19-02-1-0283, and Cisco.
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 resource capacity constraints. The authors further advocate proportional fairness, characterized by \( U_i(x_i) = \ln x_i \). In [8], an allocation policy called minimum potential delay is proposed with \( U_i(x_i) = -1/x_i \), which is shown in [4] to approximate the fairness of the Transmission Control Protocol (TCP) on the current Internet. In [9], the following class of utility functions is proposed

\[
U_i(x_i, \alpha) = \begin{cases} 
(1 - \alpha)^{-1} x_i^{1-\alpha} & \text{if } \alpha \neq 1 \\
\ln x_i & \text{if } \alpha = 1 
\end{cases}
\]

for \( \alpha \geq 0 \). 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 \( \alpha \)) 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 [8, 9, 1], in wireless networks, [7, 6], in economics, [5] 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 trade-off between these two conflicting criteria.’’

We review in Section 3 some of the networking examples, after presenting our model and stating formally our conjecture in Section 2.

The truth of the conjecture turns out to depend critically on the network topology in terms of routing and link capacities. The simplicity of the examples in Section 3 allows one to prove the conjecture in the affirmative without exploiting this underlying structure. In Section 4, we clarify this structure and present our main results. 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 arbitrarily small \( \alpha \)) more efficient! We finally conclude in Section 5.

Due to the page limitation, most proofs are omitted.

2 Model

Consider a set of links \( j = 1, \ldots, l \) with finite capacities \( c_j \), shared by a set of sources \( i = 1, \ldots n \). Let \( R \) be the \( l \times n \) routing matrix: \( R_{ji} = 1 \) if source \( i \) uses link \( j \) and 0 otherwise. Suppose all sources have a common utility function given in (1) with the same \( \alpha \). When \( \alpha \) 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.

Consider the utility maximization problem:

\[
\max_{x \geq 0} \quad U(x, \alpha) := \sum_i U_i(x_i, \alpha) 
\]

subject to

\[
Rx \leq c
\]

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 throughput

\[
T(\alpha) := \sum_i x_i(\alpha)
\]  

(4)

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

\[
\frac{dT}{d\alpha} \leq 0 \quad \text{for} \quad \alpha > 0
\]

### 3 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, and prove it for the the simple case that involves only two flows. These special cases motivate our conjecture and illustrate the means by which this question has been studied previously: by analytically solving (2)–(3), numerically computing \( T(\alpha) \), or proving the conjecture for simple cases.

As we will explain in the next section, the underlying network topology in all these examples possess 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 and \( n = l + 1 \) sources, index by \( i = 0, 1, \ldots, n \), shown in Figure 1. Source 0 goes through all the \( n \) links and sources \( i \) go through \( 1 \) link. All links have the same capacity of 1 unit.

In [8], the throughput of each source and their aggregate have been calculated for several \( \alpha \) values:

- **max-min fairness:** \( x_i(\infty) = \frac{1}{2}, \ i \geq 0; \)
  \[ T(\infty) = \frac{1}{2}(l + 1) \]

- **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:** \( x_0(1) = \frac{1}{l+1}, \)

![Figure 1: Linear network.](image-url)
\[ x_i(1) = \frac{l}{l + 1}, \quad i \geq 1 \]
\[ T(1) = l - \frac{l - 1}{l + 1} \]

max throughput: \[ x_0(0) = 0, \quad x_i(0) = 1, \quad i \geq 1 \]
\[ T(0) = l \]

Hence, the conjecture is true for these specific values of \( \alpha \):
\[ T(\infty) \leq T(2) \leq T(1) \leq T(0) \]

The authors of [8] made a cautious comment: “It is not known whether the same ordering holds for arbitrary network topologies”. Numerical examples of aggregate throughput, normalized by the maximum, are shown in Table 1 for these allocation policies. These examples suggest

<table>
<thead>
<tr>
<th># of links</th>
<th>maximum throughput</th>
<th>proportional fairness</th>
<th>minimum potential delay</th>
<th>max-min fairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>l=3</td>
<td>100%</td>
<td>83%</td>
<td>76%</td>
<td>67%</td>
</tr>
<tr>
<td>l=10</td>
<td>100%</td>
<td>92%</td>
<td>78%</td>
<td>55%</td>
</tr>
</tbody>
</table>

that the loss in efficiency of max-min fairness becomes more severe as the number of links increases.

Is the conjecture true for other values of \( \alpha \) for this topology? In [1], the rates \( x_i(\alpha) \) are computed by solving (2)–(3), as follows:
\[ x_0(\alpha) = \frac{1}{l_\alpha + 1}, \quad x_i(\alpha) = \frac{l_\alpha^{1/\alpha}}{l_\alpha^{1/\alpha} + 1}, \quad i \geq 1 \]

Using this, we can easily check that, for \( \alpha > 0 \),
\[
\frac{dT}{d\alpha} = -\frac{l_\alpha^{1/\alpha}(l - 1) \ln l}{\alpha^2(1 + l_\alpha^{1/\alpha})^2} \\
= \begin{cases} 
0, & l = 1 \\
< 0, & l \geq 2
\end{cases}
\]

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

**Example 2: Linear network with nonuniform capacity**
The linear network of Example 1 is considered in [9] with \( l = 2 \), but with different link capacities \( c_1 < c_2 \). The authors calculated the source rates under max-min fairness:
\[ x_0(\infty) = x_1(\infty) = \frac{c_1}{2}, \quad x_2(\infty) = c_2 - \frac{c_1}{2} \]

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)–(3) directly to obtain the source rates under proportional fairness for this example:
\[ x_0(1) = \frac{1}{3} \left( c_1 + c_2 - \sqrt{c_1^2 + c_2^2 - c_1 c_2} \right) \]
\[
x_1(1) = \frac{1}{3} \left( 2c_2 - c_1 + \sqrt{c_1^2 + c_2^2 - c_1 c_2} \right)
\]
\[
x_2(1) = \frac{1}{3} \left( 2c_1 - c_2 + \sqrt{c_1^2 + c_2^2 - c_1 c_2} \right)
\]

The throughputs for proportional and max-min fairness are
\[
T(1) = \frac{2}{3} \left( c_1 + \sqrt{c_1^2 + c_2^2 - c_1 c_2} \right) + \frac{2}{3} c_2 > c_2 + \frac{c_1}{2} = T(\infty)
\]

supporting the conjecture for \(\alpha = 1\) and \(\alpha = \infty\).

**Example 3: A 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 numerically solve for \(T(\alpha)\) for \(\alpha > 0\). The result is shown in Figure 3. It suggests that the conjecture is true for all \(\alpha > 0\) for this network.

![Figure 2: Linear network with two long flows.](image)

![Figure 3: Fairness-efficiency trade-off \(T(\alpha)\): linear network with two long flows.](image)

Corollary 6 below implies that, indeed, it is.

All the above examples use linear network. We next prove the conjecture for max-min fairness, minimum potential delay, proportional fairness, and maximum throughput, for a general network, but with only two sources.
Theorem 2. For a general network with only two sources, we have

\[ T(\infty) \leq T(2) \leq T(1) \leq T(0) \]

Proof. \( T(0) \) upper bounds all \( T(\alpha) \) by definition. Since minimum potential delay minimizes \( \sum_i x_i^{-1} \), we have

\[ \frac{1}{x_1(2)} + \frac{1}{x_2(2)} \leq \frac{1}{x_1(\infty)} + \frac{1}{x_2(\infty)} \]  \hspace{1cm} (5)

Without loss of generality, assume \( x_1(2) \leq x_2(2) \) and \( x_1(\infty) \leq x_2(\infty) \). Since \( x(\infty) \) is max-min fair, \( x_1(\infty) \geq x_1(2) \). Then we have \( x_2(\infty) \leq x_2(2) \) for otherwise (5) will be violated. Therefore we have

\[ x_1(2) \leq x_1(\infty) \leq x_2(\infty) \leq x_2(2) \]

For the sake of contradiction, suppose \( x_1(2) + x_2(2) < x_1(\infty) + x_2(\infty) \). Consider \( x'_2 > x_2(2) \) such that \( x_1(2) + x'_2 = x_1(\infty) + x_2(\infty) \) and

\[ x_1(2) \leq x_1(\infty) \leq x_2(\infty) \leq x_2(2) < x'_2 \]  \hspace{1cm} (6)

Observe that, for \( \gamma_1 > 0, \gamma_2 > 0 \) with a constant sum, \( 1/\gamma_1 + 1/\gamma_2 \) is strictly smaller the closer \( \gamma_1 \) and \( \gamma_2 \) are. Hence (6) implies that

\[ \frac{1}{x_1(2)} + \frac{1}{x_2(2)} > \frac{1}{x_1(\infty)} + \frac{1}{x_2(\infty)} \]

contradicting (5). Hence, \( T(\infty) \leq T(2) \).

To prove that \( T(2) \leq T(1) \), we have again

\[ \frac{1}{x_1(2)} + \frac{1}{x_2(2)} \leq \frac{1}{x_1(1)} + \frac{1}{x_2(1)} \]

or equivalently

\[ \frac{x_1(2) + x_2(2)}{x_1(2)x_2(2)} \leq \frac{x_1(1) + x_2(1)}{x_1(1)x_2(1)} \]

Since \( x(1) \) maximizes \( \sum_i \ln x_i \), we

\[ x_1(2)x_2(2) \leq x_1(1)x_2(1) \]

Hence

\[ \frac{x_1(2) + x_2(2)}{x_1(2)x_2(2)} \leq \frac{x_1(1) + x_2(1)}{x_1(1)x_2(1)} \leq \frac{x_1(1) + x_2(1)}{x_1(2)x_2(2)} \]

and \( T(2) \leq T(1) \). \( \square \)
4 Main results

It turns out that the fairness-efficiency trade-off depends critically on the network topology expressed by the routing matrix \( R \). 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. \( x(\alpha) \) is also a continuous function of \( \alpha \)[11]. Moreover, \( x(\alpha) \) is differentiable except at a finite number of points when the active constraint set at optimal \( x(\alpha) \) changes as \( \alpha \) is perturbed. Hence, we can study \( dT/d\alpha \) in between these points, and, without loss of generality, consider the utility maximization with equality constraints that represent only those constraints that are active at optimality:

\[
\max_x U(x, \alpha) \quad \text{s.t.} \quad Rx = c
\]

If every link has a single-link flow, then all constraints are necessarily tight.

Let the number of links be \( l = n - m \) and suppose the \( l \times n \) routing matrix \( R \) has full row rank. Then \( m \) is the dimension of the null space of \( R \). Let \( (z_i, i = 1, \ldots, m) \) be any basis of the null space of \( R \), and let \( Z = [z_1 \ z_2 \ \ldots \ z_m] \) be the matrix with \( z_i \) as its columns.

As we will see below, the null space represented by \( Z \) and its dimension \( m \), which is the number of sources minus the number of (bottleneck) links, play a critical role in determining whether the conjecture is true.

Let \( D \) denote the curvature of the utility function:

\[
D := -\frac{\partial^2 U}{\partial x^2} = \alpha \ \text{diag}(x_1^{-\alpha - 1}, \ldots, x_n^{-\alpha - 1})
\]

and

\[
b := \frac{\partial^2 U}{\partial x \partial \alpha} = -(x_1^{-\alpha} \ln x_1, \ldots, x_n^{-\alpha} \ln x_n)^T
\]

Let

\[
\mu_i := z_i^T b, \quad \beta_i := -e^T z_i, \quad A := Z^T D Z,
\]

where \( e \) is a column vector. We will set \( e = (1, \ldots, 1)^T \) throughout this paper unless specified otherwise. Let \( \bar{A}_i \) denote the matrix obtained from replacing the \( i \)th row of \( A \) with row vector \((\beta_1, \beta_2, \ldots, \beta_m)\).

Our first result derives \( dx/d\alpha \) explicitly and is the starting point of our analysis.

**Theorem 3.**

\[
\frac{dx}{d\alpha} = Z(Z^T D Z)^{-1} Z^T b
\]

where \( D \) and \( b \) are defined in (8) and (9) respectively.

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 \( x_0 \)). In this situation, intuitively, the moving direction will be “some projection” of Newton direction to this subspace, which is exactly expressed by (11).

Based on Theorem 3 and the relation between rate vector \( x \) and price vector \( p \), the following corollary follows.

\^[1\] Hence, all our statements below on \( dT/d\alpha \) should be interpreted piecewise in between non-differentiable points of \( \alpha \).
Corollary 4. Then the derivative of link price which is also called the dual variable can be expressed as:

\[
\frac{dp}{d\alpha} = (RR^T)^{-1}R(b - D\frac{dx}{d\alpha}) \quad (12)
\]

\[
= (RD^{-1}R^T)^{-1}RD^{-1}b \quad (13)
\]

Our first main result is a necessary and sufficient condition for the conjecture to hold.

Theorem 5.

\[
\frac{dT}{d\alpha} \leq 0 \quad \text{if and only if} \quad \sum_{i=1}^{m} \mu_i \det \bar{A}_i \geq 0
\]

This characterization leads directly to two sufficient conditions that explain all the examples in Section 3. The first condition implies that the conjecture is true when every (bottleneck) link has a short flow and there is only one long flow, or there are two long flows but both pass through the same number of links.

Corollary 6. Suppose every congested 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 \( \frac{dT}{d\alpha} \leq 0 \).

For Examples 1 and 2 in Section 3, there is only one long flow and hence the dimension of the null space of \( R \) is 1. Therefore the first part of Corollary 6 is satisfied. The network in Example 3 has two long flows both passing through 3 links, satisfying the sufficient condition of Corollary 6.

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, the opposite of the conjecture is true!

Theorem 7. When \( \dim(Z) \geq 2 \), for any \( \alpha_0 > 0 \), there exists a network such that

\[
\frac{dT}{d\alpha} > 0 \quad \text{for all} \quad \alpha > \alpha_0
\]

The proof of the theorem 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). The short flows will ensure that all constraints are active at optimality, and hence the opposite of the conjecture is true also for the case \( Rx \leq c \).

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 and two long flows. 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 (7) 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 [10] is satisfied up to an accuracy \( 10^{-6} \).

For capacities \( c_S = 10 \) and \( c_L = 1,000 \), the aggregate throughput \( T(\alpha) \) is plotted in Figure 5 as a function of \( \alpha \). \( T(\alpha) \) attains its minimum around \( \alpha = 0.95 \), and is strictly increasing beyond the minimum. In particular,

\[
T(\infty) > T(2) > T(1)
\]

For capacities \( c_S = 10 \) and \( c_L = 5,000 \), the corresponding plot is shown in Figure 6. The minimal throughput is achieved around \( \alpha = 0.75 \).
Figure 4: Network for counter-example in Theorem 7.

Figure 5: numerical verification with $c_S = 10$ and $c_L = 1000$.

Figure 6: numerical verification with $c_S = 10$ and $c_L = 5000$. 

767
5 Conclusion

In this paper, we study the trade-off between fairness and efficiency for general networks. We provide an exact characterization for the conjecture (that a fairer allocation is necessarily less efficient) to hold. The characterization leads to a simple sufficient condition which explains all the supporting examples in the literature, and allows us to construct a class of simple networks in which a fairer allocation is always more efficient.

References


