MANAGEMENT SCIENCE
Vol. 58, No. 12, December 2012, pp. 2234–2250
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1549
© 2012 INFORMS

On the Efficiency-Fairness Trade-off

Dimitris Bertsimas, Vivek F. Farias
MIT Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 {dbertsim@mit.edu, vivekf@mit.edu}

Nikolaos Trichakis
Harvard Business School, Harvard University, Boston, Massachusetts 02163, ntrichakis@hbs.edu

This paper deals with a basic issue: How does one approach the problem of designing the "right" objective for a given resource allocation problem? The notion of what is right can be fairly nebulous; we consider two issues that we see as key: efficiency and fairness. We approach the problem of designing objectives that account for the natural tension between efficiency and fairness in the context of a framework that captures a number of resource allocation problems of interest to managers. More precisely, we consider a rich family of objectives that have been well studied in the literature for their fairness properties. We deal with the problem of selecting the appropriate objective from this family. We characterize the trade-off achieved between efficiency and fairness as one selects different objectives and develop several concrete managerial prescriptions for the selection problem based on this trade-off. Finally, we demonstrate the value of our framework in a case study that considers air traffic management.

Key words: programming; nonlinear; theory; decision analysis; multiple criteria; games-group decisions; bargaining; fairness
History: Received December 1, 2010; accepted December 21, 2011, by Yossi Aviv, operations management. Published online in *Articles in Advance* September 11, 2012.

1. Introduction

Operations managers are frequently concerned with problems of resource allocation. They must build quantitative decision models for such problems, calibrate these models, and then use a suitable decision support/optimization tool to make implementable decisions or "allocations." There is a vast amount of academic research in operations management and associated fields available to complement each of the steps above. At the risk of belaboring the obvious, the following examples serve to specify this connection with resource allocation:

• *Call center design*. Pools of specialized agents must be utilized to provide service to various classes of customers. Decisions include staffing levels across agent pools and routing protocols to assign customers to agents. If delays experienced by customers are associated with dollar values, a natural objective is minimizing the expected delay costs incurred across customer classes.

• *Healthcare scheduling*. Beds (and the associated resources of doctors, nurses, and equipment) must be allocated over time to patients in need of care. In the case of an operating room, a natural objective might be (and, at least nominally, frequently is) the maximization of throughput. In an urgent care setting, one may care about delay related objectives. For instance, in the case of scheduling a specialized intensive care unit, a natural objective is minimizing the expected waiting time for a bed. In more sophisticated settings, the objectives may be directly related to physiological outcomes—for instance, minimizing mortality.

• *Air traffic control*. In case of inclement weather, the U.S. Federal Aviation Administration (FAA) needs to reallocate landing and takeoff slots among the airlines. Delays on the ground and in the air are associated with dollar values and a natural objective to consider is then reallocating slots in a manner that minimizes the total dollar impact of the resulting delays.

• *Allocation of cadaveric organs*. The United Network for Organ Sharing oversees the allocation of cadaveric organs (e.g., kidneys, livers etc.) to patients in need of them. Medical researchers and statisticians have built sophisticated models that predict the physiological outcome of allocating a specific organ (as measured by a number of attributes) to specific patients. These outcomes are frequently measured in terms of the number of quality-adjusted life years (QALYs) the transplant will add to the patient's life. A natural objective is to assign organs in a manner that maximizes the expected QALYs added across the population over time.

The list above is somewhat idiosyncratic—there are a number of other examples that one could include. What the examples above do share in common, however, is their undoubted relevance from the perspective of the social utility at stake in their solution.

INFORMS holds copyright to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Academic work on these problems frequently tends to focus on decision support related issues. For example, how does one design a routing scheme that minimizes delays in a particular queueing model? Or how does one make organ allocation decisions given the uncertainties in supply, demand, and the acceptance behavior of patients? These are difficult questions to answer.

The present paper focuses on a more basic issue. How does one come up with the right operational objective in each of the scenarios above? Is the "obvious" objective the right one? To return to the examples above, it is hard to argue that minimizing the dollar impact of delays is not a noble objective for the FAA—in fact, a vast number of proposals attempt to do just that. Of course, this noble objective fails to account for the outcome an individual airline might have to endure as part of such an allocation. Similarly, in the case of organ allocation, it is difficult to argue against the value of an allocation scheme that maximizes the number of life years generated via transplantation activities. Unfortunately, this objective fails to account for inequities such a scheme might imply for a particular group of patients (based, for instance, on their age, or peculiar physiological characteristics). Designing the "right" objective is a first-order issue, and the tensions inherent in doing so are frequently complex as the examples we have just noted make clear. This crucial design task is nonetheless frequently executed in an ad hoc fashion. The present paper attempts to provide some structure to guide this task.

• *An abstract framework*. We view resource allocation problems through the lens of welfare economics. In particular, we imagine that any resource allocation problem may be viewed as one where the system designer (or operations manager, in this case) must decide on an allocation of *utilities* to several parties from some set of feasible utilities. How might we select an allocation from among the many efficient allocations possible? A little reflection shows that the criterion implicitly employed in the examples above is a *utilitarian* criterion—one simply seeks to maximize the sum of utilities. In this paper, we adopt the utilitarian criterion as our measure of efficiency. We will return to this notion later, but for now simply note that this criterion can in many situations be unambiguously interpreted as *the* criterion by which to measure efficiency. Put mathematically, the manager's job is selecting an allocation of utilities to n parties, $u \in \mathbf{R}^n_+$, from some set of feasible utilities U. The utilitarian criterion seeks to find an allocation u to maximize $\sum_j u_j$, where u_j is the utility derived by the jth party.

• *Inequity*. The utilitarian criterion is neutral toward inequity. Coupled with the fact that in many of the examples we have encountered above, an auxiliary mechanism for monetary compensation is not implementable, this inequity is the root cause of the tensions in designing an appropriate objective. Fortunately, we have available to us an axiomatic treatment of attitudes toward inequity. This axiomatic treatment has deep roots in early philosophy and has quantitatively culminated over the last 50 years in a family of *social welfare functions* parameterized by a single parameter that measures the attitude of the system designer toward inequities. This family is given by[1]

$$\sum_{j=1}^{n} \frac{u_j^{1-\alpha}}{1-\alpha}.$$

The parameter $\alpha \geq 0$ measures an aversion to inequality. This family of "α-fair" welfare functions subsumes the well-known Nash ($\alpha \to 1$) and Kalai–Smorodinsky ($\alpha \to \infty$) solutions and is the fairness scheme we adopt and analyze in this paper; a more detailed discussion of the α-fair welfare functions and their connection with the Nash and Kalai–Smorodinsky solutions is included in §2.2.

• *The design problem*. The above setting allows us to reduce the problem of designing an appropriate objective to the selection of a single parameter. A natural trade-off implicit in selecting this parameter (at least, as seen from the operational perspective), is the loss in total system utility, or loosely, efficiency, incurred in the pursuit of equity. We seek to quantify this trade-off via a worst-case analysis. In particular, we show that this loss (measured in relative terms) scales in the worst-case like $1 - \Theta(n^{-(\alpha/(\alpha+1))})$, where n is the number of parties and α a design choice that measures the importance of equity. Conversely, another trade-off that arises from the selection of the parameter is the loss in fairness incurred in the pursuit of efficiency. To this end, we show that a natural measure of fairness, namely, the minimum utility that every party is guaranteed to derive, degrades (measured in relative terms) in the worst-case like $1 - \Theta(n^{-1/\alpha})$. The above quantifications are among the principal theoretical contributions in this paper, and to the best of our knowledge are the first general characterizations of the very natural underlying trade-off curves.

Using the aforementioned trade-off curves, we discuss ways in which a manager might choose an appropriate α for his problem so as to balance fairness and equity. Note that there are alternative methods one can devise to carry out the task of choosing α. One such method could be to utilize historical data of instances of the underlying allocation problem so as

[1] It is tempting to confuse this welfare function with the well-known Arrow–Pratt utility function; it is important to not conflate the notions of a utility function and welfare function.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

to computationally construct similar trade-off curves and base the selection of α thereon. Below we outline the merits of our approach that make it relevant and important to managerial practice:

• Our worst-case analysis relies on a minimal number of structural assumptions about the problem at hand. Thus, it can be used by managers to derive general rules of thumb about how to deal with balancing equity and efficiency. In particular, it can be utilized to characterize "sweet spots" in the trade-off between efficiency and fairness that are general and instance-free; see §3.3 for more details.

• The fact that our analysis is problem instance independent is particularly important in settings where managers are faced with complex, multifaceted operational problems. In such settings, choosing α based on a worst-case analysis can be viewed as a long-term strategic decision that can be applied across all operational activities. In particular, this choice of α can then be taken as a given in calibrating the objectives of operational optimization problems that one solves over time. Section 4, which presents a case-study from the airline industry, serves as a useful example in this regard. Allowing the balance between equity and efficiency to depend on problem instance specific data in that setting is unlikely to gain traction given the inherent variability in the specification of these instances and the strongly competing interests of the stakeholders.

We provide a concrete illustration of the value of the framework and our analysis above by implementing it in the context of the air traffic management problem mentioned earlier. In particular, we present a concrete, quantitative statement of the design problem a system manager seeking the "right" operational objective might solve, and then we explore the consequences of various solutions in a study using detailed historical air traffic data.

Finally, despite our focus on problems of particular interest to operations managers, note that the key ideas and results of this paper can be extended and applied to general resource allocation problems, many of which are reviewed in §1.1.

The structure of this work is as follows. In the next subsection, we review relevant applications in the literature where the need for the design of objectives that balance equity and efficiency is apparent. We will also review important developments in the welfare economics and bargaining literature that yield the foundations of our framework. In §2, we introduce our framework rigorously, placing it in the context of welfare economics. Section 3 establishes the trade-off curves that, as we have discussed, can guide the design of an equitable objective. Section 4 considers a concrete design problem in this vain in the context of air traffic management. This case study uses actual air traffic data and illustrates the value of our framework. Concluding remarks are included in §5.

1.1. Literature Review

1.1.1. Economic Theory. A typical setting in welfare economics concerns the scenario where a central planner must make an allocation of goods in an economy to a number of distinct entities. The planner is aware of the preferences of the entities, and one typically assumes these are described via cardinal utilities. The central problem in welfare economics is then concerned with how the central planner should go about making these allocations. Samuelson (1947) provided the first formulation in which the relevant constraint set for the planner was the set of achievable utility allocations, or the *utility possibility set*, an idea which became central in this area. In fact, our framework is based on exactly that notion. The welfare economics problem can then be stated as the problem of picking a point in the utility set (for more details, see §2).

One prominent way of addressing the allocation problem above has been the identification of a real-valued social welfare function of the allocation of utilities, which is used by the central decision maker to rank allocations. The approach in which the welfare function reflects the distributional value judgment of the central planner was first taken by Bergson (1938) and Samuelson (1947). Some of the most important instances of social welfare functions are the utilitarian, maximin, and constant elasticity functions. For the merits of the utilitarian function, see Harsanyi (1955). The maximin function is based on the *Rawlsian justice*, introduced by Rawls (1971). For details regarding the constant elasticity function, see §2.2. We refer the reader to Young (1995) and Sen and Foster (1997) for a thorough overview of the above work. Mas-Colell et al. (1995) provide a nice introduction.

Another approach to dealing with the allocation problem is provided by *bargaining* theory. Here one formulates axioms that any allocation must satisfy and then seeks allocation rules that satisfy these axioms. The standard form of the bargaining problem was first posed by Nash (1950). Nash (1950) provided a set of axioms that an allocation must satisfy, and demonstrated the unique allocation rule satisfying these axioms, all in a two-player setting. An alternative solution (and axiomatic system) for the two-player problem was introduced by Kalai and Smorodinsky (1975). The work by Lensberg (1988) extended these solutions to a setting with multiple players. For other axiomatic formulations, see, Roth (1979). Finally, see Young (1995) and Mas-Colell et al. (1995) for surveys of the literature.

1.1.2. Applications. As is evident from our introductory remarks, the need to design resource allocation objectives that in addition to being "efficient"

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

in an appropriate sense are also equitable is ubiquitous. Below we discuss a biased sample of related applications:

Healthcare. The fundamental question in this area is how to balance equity in health provision and medical utility, which typically corresponds to the aggregate health of the population (see Wagstaff 1991). This natural dichotomy between equity and efficiency is apparent across a wide spectrum of healthcare operations. For instance, in managing operations in a hospital's intensive care unit, one cannot simply maximize throughput without accounting for fairness and medical urgency (see Swenson 1992, Chan et al. 2012b). Furthermore, in their book, *Medicine and the Market: Equity v. Choice*, Callahan and Wasunna (2006) discuss the use of markets and government funding to balance efficiency and equity (respectively), for the purposes of insurance policies and a healthcare reform. See also Pauly (2010) for a related discussion. The efficiency-fairness trade-off is also particularly important in the allocation of research funds by the National Institute of Health (NIH) of the United States over various biomedical research projects. Each of the projects deals with improving the care provided to patients of particular diseases (e.g., cancer, HIV, etc.). A primary goal of the allocation is then to maximize clinical efficiency, that is, to allocate the funds such that the resulting research gains lead to the highest possible anticipated increase in QALYs of the population. Such practice, however, may potentially be unethical and result in age or race discrimination. To ensure an equitable health treatment, the NIH needs to diversify its allocation, trading off clinical efficiency and fairness (see Resnick 2003, Bisias et al. 2012). Finally, similar considerations arise in the allocation of deceased-donor kidneys to patients on a waiting list; see Su and Zenios (2004, 2006) and Bertsimas et al. (2012) for a detailed discussion.

Service Operations. Other settings where the equity-efficiency trade-off is of importance include call center design and other associated queueing problems, supply chain, and service applications. As discussed previously, the maximization of the throughput or the minimization of average waiting time are the typical objectives for a service manager in designing a queueing system. Several studies have acknowledged the importance of accounting for inequity in these settings by employing alternative objectives such as the variability in service times or queue lengths, etc. (see Shreedhar and Varghese 1996, Armony and Ward 2010, Chan et al. 2012a). Within the supply chain literature, Cui et al. (2007) incorporate the concept of fairness into the conventional dyadic channel to investigate how fairness may affect the interactions between the manufacturer and the retailer. Finally, Wu et al. (2008) study the impact of fair processes on the motivation of employees and their performance in execution. They examine the trade-offs involved and study under which circumstances management should use fair processes or not.

Yet another application, revisited in §4 for a case study, is the air traffic control problem, alluded to in the previous discussion. There is an extant body of research devoted to formulating and solving the problem of minimizing the total system delay cost (see Odoni and Bianco 1987, Bertsimas and Stock-Patterson 1998). Although this objective is natural, a somewhat surprising fact is that existing practice (at least within the United States) does not take into account delays in making such reallocation decisions. The emphasis, rather, is on an allocation that may be viewed as equitable or fair to the airlines concerned. Recent research work deals with combining those two objectives (see Vossen et al. 2003, Barnhart et al. 2012, Bertsimas and Gupta 2012).

Networks. The trade-off between efficiency and fairness is hardly specific to just operations management problems. In particular, it is well recognized and studied in many engineering applications as well, ranging from networking and bandwidth allocation, and job scheduling to load balancing. For instance, the network utility maximization problem has been heavily studied in the literature. In that problem, a network administrator needs to assign transmission rates to clients sharing bandwidth over a network, accounting for efficiency (e.g., net throughput of the network) and fairness (e.g., "equal" bandwidth assignment). For more details, see Bertsekas and Gallager (1987), Kelly et al. (1998), and Mo and Walrand (2000).

1.1.3. Worst-Case Analysis. Recent work has focused on studying the worst-case degradation of the utilitarian objective, i.e., the sum of the utilities, under a fair allocation compared to the allocation that maximizes the utilitarian objective. Bertsimas et al. (2011) provide uniform bounds on the worst-case degradation if the system designer chose either the *proportional* or *max-min* fair allocations (each via a separate analysis). They do so for a broad class of resource allocation problems. From a theoretical perspective, the present paper provides a new, unified geometric framework through the lens of which we may reconstruct those earlier results and *simultaneously* develop a host of *novel* guarantees of managerial relevance. First, we may characterize the worst-case degradation of the utilitarian objective or price of fairness for a *continuum* of allocation rules (or equivalently, degrees of fairness) parameterized by the inequality aversion parameter α. We can simultaneously understand an entirely different sort of trade-off, which measures the loss in *fairness* as one seeks more efficient solutions. In particular, we may characterize how the minimum utility allocated to a player varies under

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

varying choices of α. Together, these new trade-off curves allow a manager to make an informed decision on precisely how to balance fairness and inequity in designing an appropriate objective.

In addition to the above work, Butler and Williams (2002) show that the degradation is zero under a max-min fair allocation for a specialized facility location problem. Correa et al. (2007) also analyze the degradation under a max-min fair allocation for network flow problems with congestion. Chakrabarty et al. (2009) show that when the set of achievable utilities is a polymatroid, the worst-case degradation is zero under all Pareto resource allocations. This is a somewhat restrictive condition and a general class of resource allocation problems that satisfy this condition is not known. Relative to the above literature, the present paper provides the first analysis that is simultaneously applicable to *general* resource allocation problems for a *general* family of allocation rules.

2. A General Framework

We describe a general framework that captures the majority of the applications discussed in the introduction. We then review allocation mechanisms that account for the objectives of equity and efficiency alluded to in the introduction.

Consider a resource allocation problem, in which a central decision maker (CDM) needs to decide on the allocation of scarce resources among n players. Each player derives a nonnegative utility, depending on the allocation decided by the CDM (e.g., via means of a utility function). For a given allocation of resources, there is thus a corresponding *utility allocation* $u \in \mathbf{R}_+^n$, with u_j equal to the utility derived by the jth player, $j = 1, \ldots, n$.

A utility allocation $u \in \mathbf{R}_+^n$ is *feasible* if and only if there exists an allocation of resources for which the utilities derived by the players are $u_1, u_2, \ldots, u_n$ accordingly. We define the *utility set* $U \subset \mathbf{R}_+^n$ as the set of all feasible utility allocations. Encapsulated in the notion of the utility set are the preferences of the players and the way they derive utility, as well as individual constraints of the players or the CDM, constraints on the resources, etc. Thus, the utility set provides a condensed way of describing the problem. Given the utility set, the CDM then needs to decide which utility allocation among the players to select or, equivalently, which point from the utility set to select. The notion of the utility set was introduced by Samuelson (1947).

The above setup has been studied within the research areas of fair bargaining and welfare economics (see §1.1). Note that these utilities may not be quasi-linear; that is to say, there is no reason to assume that an allocation to a specific party might be substituted by a cash payment to that party. In fact, allowing for such cash payments greatly simplifies the aforementioned utility allocation problem, as discussed in §2.1; this work is thus relevant in cases where such cash payments are not feasible.

To illustrate the applicability of the setup, we discuss below an example.

Example 1. As a concrete application of the model above, consider the call center design problem alluded to in the introduction. An operations manager (the central decision maker) needs to decide on staffing levels across agent pools (the scarce resources) and routing protocols to serve n different customer classes (the players). Suppose that a specific set of decisions results in the jth customer class experiencing an expected waiting time of w_j, $j = 1, \ldots, n$, during steady-state operation of the center. The vector of steady-state expected waiting times of the customer classes is commonly referred to as the *performance vector*. Suppose also that the utility derived by the jth customer class is $v_j - c_j w_j$, where v_j is the constant nominal utility derived by that particular class for the service and c_j is effectively the value of time to the jth class. Let W be the set of achievable performance vectors, known as the *achievable performance set* or *space*. Note that the description of W might be very complex. The utility set in that case is

$$U = \{v_1 - c_1 w_1, \ldots, v_n - c_n w_n \mid w \in W\}.$$

Note that a lot of work has been devoted to providing tractable descriptions of the underlying achievable performance space, W, or approximations of it, under different settings. Results of that kind are very powerful, as they allow one to maximize concave functions of the waiting times (e.g., utilities) very efficiently. We refer the reader to Gelenbe and Mitrani (1980), Federgruen and Groenevelt (1988), Tsoucas (1991), and Bertsimas et al. (1994) for early results in that field.

In the next section, we review social welfare functions and allocation mechanisms that give rise to efficient and fair allocations.

2.1. Utilitarian Allocations

A natural objective for the central decision maker is to maximize an efficiency metric of the system (defined appropriately). In this work, we adopt the sum of utilities (derived by the players) as our metric of system efficiency, as discussed in the introduction. This is referred to as the utilitarian criterion. Our rationale in doing so is twofold:

1. The utilitarian criterion emerges as the natural efficiency metric employed in practice. The examples alluded to earlier are cases in point, and by themselves are sufficient to justify this benchmark.

2. In a general setting where cash payments or any general monetary transfers are allowed as a

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

mechanism to compensate for inequity, the sum of utilities is the *only* admissible criterion of efficiency. It stands to reason then, that the allocation induced by such a criterion may be viewed as efficient, whether or not monetary transfers are possible.

In mathematical terms, given a utility set U, a utilitarian allocation corresponds to an optimal solution of the problem

$$\begin{aligned} &\text{maximize} && \mathbf{1}^T u \\ &\text{subject to} && u \in U, \end{aligned}$$

with variable $u \in \mathbf{R}^n_+$, and $\mathbf{1}$ is the vector of all ones. We denote the optimal value of this problem with SYSTEM(U), i.e.,

$$\text{SYSTEM}(U) = \sup\{\mathbf{1}^T u \mid u \in U\}.$$

As discussed above, we will regard this value as corresponding to the highest possible level of system efficiency (or social utility) achievable.

The sum of utilities is among the most well-studied social welfare functions, and is known as the Bentham utilitarian function given the philosophical justification of this criterion provided by Jeremy Bentham (see Mas-Colell et al. 1995, Young 1995). The utilitarian principle of maximizing the sum of utilities is neutral toward inequalities among the utilities derived by the players. As a result, it is considered to lack fairness considerations (see Young 1995).

2.2. Fair Allocations

Because of the subjective nature of fairness and different possible interpretations of equity, there is no principle that is universally accepted as "the most fair." In particular, there has been a plethora of proposals in the literature under axiomatic bargaining, welfare economics, as well as in applications ranging from networks, air traffic management, healthcare, and finance. We refer the reader to Young (1995) and Bertsimas et al. (2011) for a more detailed exposition.

A fairness scheme of particular interest, and one on which we will focus our attention in this work, is the *α-fairness* scheme, which was studied early on by Atkinson (1970), building on notions of individual risk aversion introduced by Pratt (1964) and Arrow (1965), and using these instead as notion of aversion to inequity (for more details, see also Mas-Colell et al. 1995, Barr 1987). According to α-fairness, the CDM decides on the allocation by maximizing the constant elasticity social welfare function W_α, parameterized by $\alpha \geq 0$, and defined for $u \in \mathbf{R}^n_+$ as

$$W_\alpha(u) = \begin{cases} \displaystyle\sum_{j=1}^{n} \frac{u_j^{1-\alpha}}{1-\alpha} & \text{for } \alpha \geq 0,\ \alpha \neq 1, \\ \displaystyle\sum_{j=1}^{n} \log(u_j) & \text{for } \alpha = 1. \end{cases}$$

A resulting utility allocation, denoted by $z(\alpha)$, is such that

$$z(\alpha) \in \arg\max_{u \in U} W_\alpha(u), \tag{1}$$

and is referred to as an *α-fair allocation*.

Under the constant elasticity welfare function, the proportional increase in welfare attributed to a given player for a given proportional increase of her utility, is the same at all utility levels. Moreover, because the constant elasticity function is concave and component-wise increasing, it exhibits diminishing marginal welfare increase as utilities increase. In other words, if player A derives a lower utility than player B, then a marginal increase in the utility of player A would yield a higher welfare increase compared to a marginal increase in the utility of player B. As such, the marginal increase in the utility of player A would be more desirable for the CDM. This property of W_α typically leads to more even or fair distributions of utility among players and can thus provide an explanation of why the constant elasticity welfare function yields fair allocations. Furthermore, the rate at which marginal increases diminish is controlled by the parameter α; in the example above, the difference in the welfare increase between the cases of marginally increasing the utility of player A or player B, increases with the parameter α, thus making the scenario of marginally increasing the utility of player A instead of player B even more desirable for the CDM. For that reason, the parameter α is called the *inequality aversion parameter*.

The α-fairness scheme can be useful in practice for a CDM, as it facilitates an understanding of the efficiency-fairness trade-off. In particular, as we discussed above, a higher value of the inequality aversion parameter is thought to correspond to a "fairer" scheme (see also Tang et al. 2006, Barr 1987, Lan et al. 2010). Note that for the smallest value of $\alpha = 0$, we recover the utilitarian principle, which is neutral toward inequalities. Thus, the CDM can adjust attitudes toward inequalities by means of a single parameter.

Furthermore, the α-fairness scheme captures as special cases two important fair bargaining solutions, which have been studied extensively in the literature; for $\alpha = 1$, the scheme corresponds to proportional fairness (introduced by Nash 1950), whereas for $\alpha \to \infty$, the α-fair allocation converges to the utility allocation suggested by max-min fairness (see Kalai and Smorodinsky 1975, Mas-Colell et al. 1995).

Although the α-fairness scheme has been studied both from a theoretical and a practical perspective, most prominently in networks (Mo and Walrand 2000, Bonald and Massoulié 2001) and healthcare (Wagstaff 1991), the underlying efficiency-fairness trade-off is still not well understood. Recent work by Lan et al. (2010) has been devoted to theoretically characterizing

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

what it actually means for a higher value of α to be more fair. The impact of a higher value of α on the system efficiency (i.e., sum of utilities) has also received a lot of attention. Bertsimas et al. (2011) present tight upper bounds on the efficiency loss for the special cases of proportional and max-min fairness.

What is lacking is a precise understanding of the efficiency-fairness trade-off implicit in a selection of the inequality aversion parameter α; an understanding of this trade-off would provide the system manager with a useful design tool. The next section sheds light toward this direction.

3. The Efficiency-Fairness Trade-off

Consider a resource allocation problem, as described in §2, and suppose that the central decision maker wishes to implement α-fairness to trade off efficiency and fairness. As discussed above, such an implementation requires the calibration of the inequality aversion parameter α that controls the trade-off.

Although there is some understanding on how the fairness properties of the α-fairness scheme behave with respect to varying α, there is no theoretical work focusing on the potential efficiency degradation. As such, the selection of the parameter in a practical setting can be very challenging. In particular, to make decisions, a manager needs to understand (a) what the *efficiency loss* might be and (b) what the *fairness loss* might be for a specific value of the parameter α. This section sheds light on exactly those matters, by quantifying what the maximum efficiency and fairness loss can be for a given fixed value of α. We next formally define the notions of the efficiency and fairness loss and discuss the main results.

3.1. Efficiency Loss and the Price of Fairness

As the central decision maker incorporates fairness considerations, the efficiency of the system (measured as the sum of utilities), is likely to decrease, compared to the efficiency under the utilitarian solution.

Suppose the CDM adopts α-fairness, using a fixed value for the inequality aversion parameter α, and the utility set U is such that an α-fair allocation exists (e.g., U is compact). Then, the efficiency of the system under the α-fairness scheme will be the sum of the components of the α-fair utility allocation $z(\alpha)$ (as in (1)), and denoted by

$$\text{FAIR}(U;\alpha) = \mathbf{1}^T z(\alpha).$$

The *efficiency loss* is the difference between the maximum system efficiency, $\text{SYSTEM}(U)$, and the efficiency under the fair scheme, $\text{FAIR}(U;\alpha)$. The efficiency loss relative to the maximum system efficiency is the so-called *price of fairness* (Bertsimas et al. 2011), defined in this case as

$$\text{POF}(U;\alpha) = \frac{\text{SYSTEM}(U) - \text{FAIR}(U;\alpha)}{\text{SYSTEM}(U)}.$$

This price is a number between zero and one, and corresponds to the percentage efficiency loss compared to the maximum system efficiency. It is a key quantity to understanding the efficiency-fairness trade-off.

Note that for $\alpha = 0$, the α-fairness scheme corresponds to the utilitarian principle, because W_0 is the sum of utilities, $W_0(u) = \mathbf{1}^T u$. Hence, for any compact utility set U, the sum of utilities is the same under both schemes, i.e., $\text{SYSTEM}(U) = \text{FAIR}(U;0)$, and

$$\text{POF}(U;0) = 0.$$

For $\alpha > 0$, we have the following result. A useful quantity for the presentation of the result is the *maximum achievable utility* of each player, defined (for the jth player) as

$$u_j^\star = \sup\{u_j \mid u \in U\}, \quad \text{for all } j = 1, \ldots, n.$$

THEOREM 1. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}_+^n$, be compact, convex, and such that the players have equal maximum achievable utilities (greater than zero). For the α-fairness scheme, $\alpha > 0$, the price of fairness is bounded by*

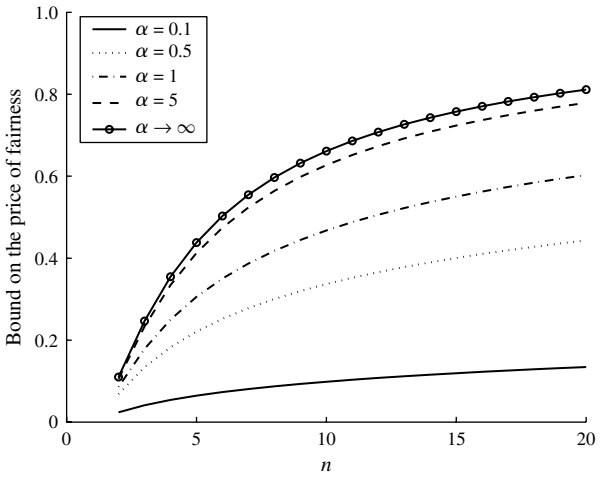
$$\begin{aligned}\text{POF}(U;\alpha) &\leq 1 - \min_{x\in[1,n]} \frac{x^{1+1/\alpha} + n - x}{x^{1+1/\alpha} + (n-x)x} \\ &= 1 - \Theta(n^{-\alpha/(\alpha+1)}).\end{aligned}$$

In Theorem 1, we assume that the utility set is compact and convex. This assumption is standard in the literature of fair bargains and also very frequently satisfied in practice. In particular, compactness of the utility set follows from limited resources and bounded and continuous functions that map resource allocations to utility for each player. Also, in case of nonconvex utility sets, randomization over possible utility allocations results in a convex set (of expected utilities). For more details, we refer the reader to Young (1995) and Bertsimas et al. (2011).

Furthermore, note that the negative of the function that needs to be minimized to compute the exact bound in Theorem 1, is unimodal (see the online appendix, §C, available at http://fileserver.hbs.edu/ntrichakis/onTheEfficiencyFairnessTradeoff_Appendices.pdf). As such, one can efficiently compute the unique minimizer and the associated minimum function value. Figure 1 depicts bounds on the price of fairness implied by Theorem 1, for different values of the inequality aversion parameter α, as functions of the number of players n. The graph illustrates the dependence of the bound on the number of players, for different values of α; in particular, the worst-case

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Figure 1 Bounds on the Price of Fairness of α-Fair Allocations for Different Values of α Implied by Theorem 1



Note. The bounds are plotted as functions of the number of players n.

price is increasing with the number of players and the value of α.

A natural question arising with regard to the results of Theorem 1 is whether the bounds are tight. The surprising fact is that the bounds are very strong, near-tight.

Before moving on, we briefly discuss an extension to the bounds above. In case players have unequal maximum achievable utilities, one can generalize our framework to deal with this case, albeit at the expense of additional technical effort. For instance, if under the same setup of Theorem 1 we also assume that the maximum achievable utilities of the players satisfy

$$L \le \min_{j=1,\dots,n} u_j^\star \le \max_{j=1,\dots,n} u_j^\star = B,$$

for some $0 < L \le B$, we have that the price of fairness is bounded by

$$\mathrm{POF}(U;\alpha) \le 1 - \min_{x\in[1,n]} \frac{(B/L)^{1/\alpha}x^{1+1/\alpha} + n - x}{(B/L)^{1/\alpha}x^{1+1/\alpha} + (n-x)(B/L)x}. \quad (2)$$

Note that the bounds we obtain in this case depend on the ratio of highest to lowest maximum achievable utility B/L; in particular, as the ratio B/L increases, the bounds tend to become worse. The case we focus on (equal maximum achievable utilities), however, is particularly important, because utility levels of different players are commonly normalized, so as the intercomparison of utilities between them becomes meaningful (see Mas-Colell et al. 1995, Harsanyi 1955).

3.1.1. Near Worst-Case Examples for the Price of Fairness. We discuss the construction of near worst-case examples under which the price of fairness is very close to the bounds implied by Theorem 1 for any values of the problem parameters, i.e., the number of players n, and the inequality aversion parameter α. To illustrate the fact that the near worst-case examples are not pathological by any means, but rather have practical significance, we present them in a realistic setup under the context of network management. The setup is relevant to many other applications including traffic management and routing. After discussing the structure of the near worst-case examples, we compare their price of fairness with the established bounds and demonstrate that the bounds are essentially tight. Technical details of the construction of the examples are included in the online appendix, §B.

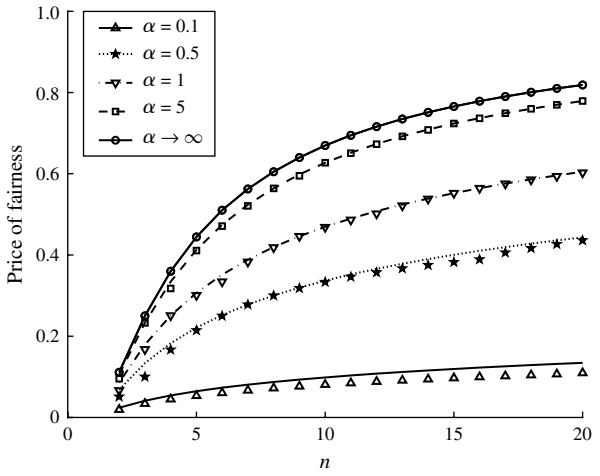
Near Worst-Case Bandwidth Allocation. Consider a network consisting of hubs (nodes) that are connected via capacitated links (edges). Clients, or flows, wish to establish transmission from one hub to another over the network via a prespecified and fixed route. The network administrator needs to decide on the transmission rate assigned to each flow, subject to capacity constraints. The resources to be allocated in this case are the available bandwidth of the links, the players are the flows, and the central decision maker is the network administrator. The utility derived by each player is equal to his assigned transmission rate.

For the purposes of constructing near worst-case examples, we study a line-graph network, which is a specific network topology that has received a lot of attention in the literature and in practice (see Bonald and Massoulié 2001, Tang et al. 2006). Specifically, suppose we have n players or flows. The network consists of y links of unit capacity, where the routes of the first y flows are disjoint and they all occupy a single (distinct) link. The remaining $n-y$ flows have routes that utilize all y links. Each flow derives a utility equal to its assigned nonnegative rate. Note that in this setup, each player has a maximum achievable utility of one, which is trivially achieved if all other flows are assigned zero rates. Thus, Theorem 1 applies.

Suppose we further fix a desired inequality aversion parameter $\alpha > 0$. In that case, one can select y (under some technical conditions) so that the price of fairness is exactly equal to the bound implied by Theorem 1. Technical details about the selection of y are included in the online appendix, §B.

Note that the described worst-case topology pertains to a case of resources shared by n players, who can be of two types; players of the first type (short flows) consume resources at a lower rate, for a unit of utility, compared to players of the second type (long flows). This can be generalized as follows. Consider a knapsack-style problem where a unit of a single resource is shared by n players. Players $1,\dots,l$ consume the resource at a rate of γ_1 for a unit of utility,

Figure 2 Price of Fairness for Constructed Examples (Markers) for Different Values of α in §3.1.1



Notes. The corresponding bounds are also plotted (lines). The values/bounds are plotted as functions of the number of players n.

whereas players $l+1, \ldots, n$ consume the resource at a rate of γ_2. The described utility set is then

$$U = \{u \in \mathbf{R}_+^n \mid \gamma_1 u_1 + \cdots + \gamma_1 u_l + \gamma_2 u_{l+1} + \cdots + \gamma_2 u_n \le 1, \; u \le \mathbf{1} \; \forall j\}.$$

In the online appendix, §B, we present a simple algorithmic procedure of selecting parameters l, γ_1, and γ_2 for a fixed number of players n and α, such that the price of fairness $\mathrm{POF}(U;\alpha)$ for the set U is very close to the price implied by Theorem 1. Figure 2 illustrates the prices achieved by following that procedure for various values of α and n. The average discrepancy between the bound and the values is 0.005, and the largest discrepancy is 0.023.

3.2. Fairness Loss and the Price of Efficiency

Having analyzed the efficiency of different α-fair allocations, we now focus on their fairness properties. To quantify and compare the fairness properties of different allocations, we need to select a fairness metric to adopt. Because of the subjective nature of fairness, note that such a selection can be nebulous, in a similar way to the selection of a fairness scheme (see §2.2).

For the family of α-fair allocations, a natural measure of fairness could be the associated inequality aversion parameter α or the constant elasticity welfare function (see Atkinson 1970, Lan et al. 2010). Those measures, however, are not easy to interpret. Other standard fairness metrics that have been well studied in the literature and are perhaps easier to interpret include the minimum utility, the difference between the maximum and the minimum utility, the standard deviation or the coefficient of variation of the utilities, the Jain index, the Theil index, the mean log deviation of the utilities, the Gini coefficient, etc.

To guide the selection of a fairness metric, we further require that under it the max-min fair allocation (i.e., the α-fair allocation for $\alpha \to \infty$) is optimal among all Pareto allocations for any utility set. Recall that under the premises of α-fairness, the max-min fair allocation is deemed as the "most fair" allocation (see §2.2). Thus, we require that the max-min fair allocation preserves this property under the selected fairness metric as well. To this end, the fairness metric we adopt in this work is the minimum utility. That is, given a utility allocation u, we measure its fairness properties by $\min_j u_j$. This fairness metric was advocated by Rawls (1971) and can be interpreted as a minimum guarantee of utility to all the players. The minimum utility is the only fairness metric from the ones discussed above that is easy to interpret and also satisfies the max-min fair allocation optimality requirement.[2]

For a particular utility set U, our fairness metric attains its highest value for the α-fair allocation corresponding to $\alpha \to \infty$ and is equal to

$$\max_{u \in U} \min_{j=1,\ldots,n} u_j.$$

This utility value is then the highest possible minimum utility guarantee for all players the CDM can set. As the CDM puts more emphasis on efficiency (e.g., by selecting a lower value of α), the minimum utility guarantee is likely to decrease.

Suppose the CDM adopts α-fairness, using a fixed value for the inequality aversion parameter α, and the utility set U is such that an α-fair allocation exists (e.g., U is compact). Under the associated allocation $z(\alpha)$, the fairness metric evaluates to

$$\min_{j=1,\ldots,n} z_j(\alpha).$$

The *fairness loss* is the difference between the fairness metric evaluated at the max-min fair allocation and the α-fair allocation. We then call the fairness loss relative to the maximum value of the fairness metric as the *price of efficiency*, defined as

$$\mathrm{POE}(U;\alpha) = \frac{\max_{u \in U} \min_{j=1,\ldots,n} u_j - \min_{j=1,\ldots,n} z_j(\alpha)}{\max_{u \in U} \min_{j=1,\ldots,n} u_j}.$$

The price of efficiency can be interpreted as the percentage loss in the minimum utility guarantee

[2] Consider the utility set consisting of two points, $U = \{[0.5\ 0.5\ 0.8]^T, [0.45\ 0.55\ 0.72]^T\}$. The first point is the max-min fair allocation of U; however, under the fairness metrics of the difference between the maximum and the minimum utility, the standard deviation, the coefficient of variation of the utilities, the Jain index, the Theil index, the mean log deviation of the utilities, and the Gini coefficient, the second point is preferred. Note that this result remains true for convex sets, e.g., for the convex hull for the points in U, 0, and the unit vectors in $\mathbf{R}_+^3$.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

compared to the maximum minimum utility guarantee. In case the CDM implements max-min fairness ($\alpha \to \infty$), the price of efficiency is zero. As we depart from the max-min fairness doctrine, perhaps to achieve higher system efficiency, the price of efficiency is likely to increase.

We now analyze the worst-case degradation of the minimum utility guarantee among all players. We have the following result. Recall that the maximum achievable utility of the jth player is defined as

$$u_j^\star = \sup\{u_j \mid u \in U\}, \quad \text{for all } j = 1, \ldots, n.$$

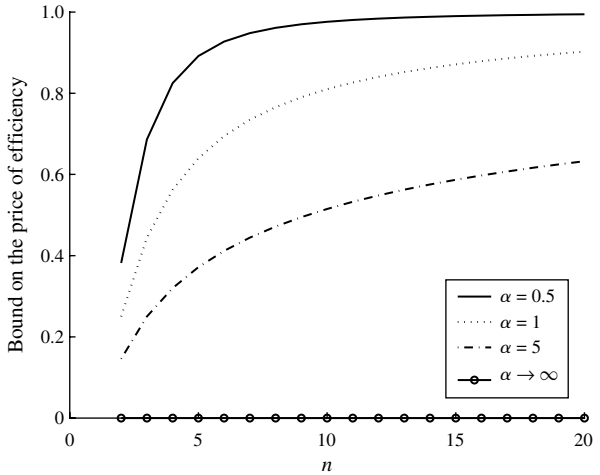
THEOREM 2. *Consider a resource allocation problem with n players, $n \geq 2$. Let the utility set, denoted by $U \subset \mathbf{R}_+^n$, be compact, convex, and such that the players have equal maximum achievable utilities (greater than zero). For the α-fairness scheme, $\alpha > 0$, the price of efficiency is bounded by*

$$\mathrm{POE}(U;\alpha) \leq 1 - \min_{x \in [\rho, 1]} \frac{(n-1)x + x^{1-\alpha}}{n-1+x^{1-\alpha}} = 1 - \Theta(n^{-1/\alpha}),$$

where ρ is the unique root of $n - 1 + x^{-\alpha}(x-1) = 0$ in $(0, 1)$.

Similarly to Theorem 1, one can show that the negative of the function that needs to be minimized to compute the exact bound in Theorem 2 is unimodal; thus, the minimum function value can be efficiently computed. Figure 3 depicts bounds on the price of efficiency implied by Theorem 2, for different values of the inequality aversion parameter α, as functions of the number of players n. The graph illustrates the dependence of the bound on the number of players for different values of α; in particular, the worst-case price is increasing with the number of players and decreasing with the value of α.

Figure 3 **Bounds on the Price of Efficiency of α-Fair Allocations for Different Values of the Inequality Aversion Parameter α Implied by Theorem 2**



Note. The bounds are plotted as functions of the number of players n.

Finally, the bounds on the price of efficiency presented in Theorem 2 are tight.

3.2.1. Worst-Case Examples for the Price of Efficiency. For any values of the problem parameters, i.e., the number of players n and the value of the inequality aversion parameter α, one can construct worst-case examples under which the price of efficiency is equal to the bounds implied by Theorem 2.

The setup of the worst-case examples for the price of efficiency is identical with the setup discussed in §3.1.1 for the price of fairness. In particular, the worst-case topology pertains to a case of a single resource shared by n players, with $n-1$ of them consuming the resource at a rate of γ_1 for a unit of utility, whereas the nth player consumes the resource at a rate of γ_2, for

$$\gamma_1 = \frac{1}{n-1+\xi^{-\alpha}}, \quad \gamma_2 = \frac{\xi^{1-\alpha}}{n-1+\xi^{-\alpha}},$$

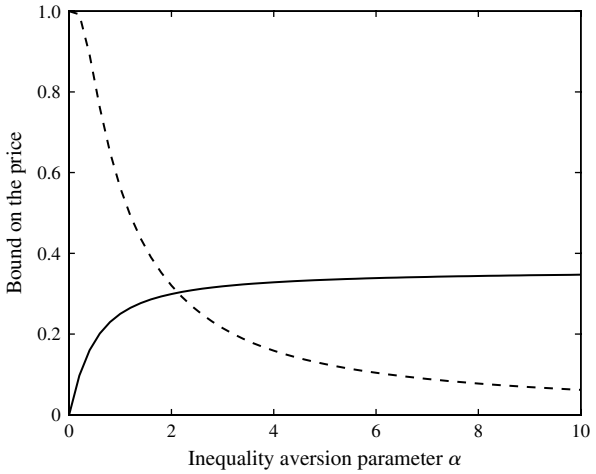
where ξ is the (unique) minimizer from Theorem 2. The proof is similar to the proof of Proposition 1 in the online appendix, §B, and is omitted.

3.3. Potential Managerial Implications

The framework implicit in our characterization of the fairness and efficiency properties of α-fair allocations already provides an interesting and general *quantitative* formalism of balancing equity and fairness in a real-world resource allocation problem. In the introduction, we discussed the relative merits of performing this balancing task, i.e., selecting α, based on the worst-case analysis we provide. We now discuss two potential concrete prescriptions for selecting α, although many others are possible. Consider a setting where the manager must make a choice of objective in a resource allocation problem that impacts four distinct parties. Further, let us imagine that the utilities allocated to a given party are normalized relative to their maximum achievable utility. Using the analysis of the preceding sections, one may construct the curves described in Figures 4 and 5. These curves present two distinct ways of visualizing the price of fairness (i.e., the loss in efficiency due to the requirement of fairness) and the price of efficiency (i.e., the loss in fairness due to the requirement of efficiency) as the manager varies his choice of objective by varying his inequality aversion parameter α. Two easily explained ways of arriving at a choice of α are as follows:

- *Via a tolerance on efficiency and/or unfairness.* The manager might decide that he is willing to be as fair as possible while allowing no more than a certain degradation in efficiency. For instance, the manager might decide to be as fair as possible while guaranteeing no more than a 20% decrease in overall system

Figure 4 **Bounds on the Price of Fairness (Solid) and the Price of Efficiency (Dashed) of α-Fair Allocations for $n = 4$ Players**

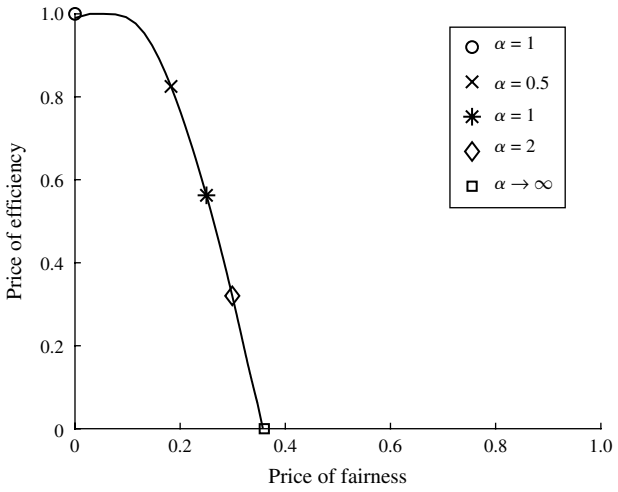


Note. The bounds are plotted as functions of inequality aversion parameter α.

efficiency. Figure 4 then suggests that the appropriate choice of α is about 1. As another example, the manager might want to be as *efficient* as possible while allowing only a limited amount of unfairness. For instance, if the manager were willing to tolerate at most a 20% drop in the utility garnered by the player that is worst off relative to a scheme that maximizes the utility of the worst-off player, Figure 4 would then suggests that the appropriate choice of α is about 3.

• *By balancing equity and fairness*. The manager might choose α so as to appropriately balance the degradation in efficiency and fairness. There are obviously a multitude of ways he may choose to do so and Figure 5 provides a valuable trade-off curve in making such a decision. For instance, the manager might choose to pick α so as to balance the prices of efficiency and fairness. In that case, the trade-off curve described by Figure 5 identifies $\alpha = 2$ as an appropriate choice of α. We will examine both these prescriptions in an empirical study of air traffic flow management in the next section.

Figure 5 **Bounds on the Prices of Fairness and Efficiency of α-Fair Allocations for $n = 4$ Players and Various Values of α**



Finally, we conclude with noting yet another insight derived from Figure 5: In a setting with four players, a manager will likely only want to consider choices of α roughly between 0.5 and 2 and the extreme choices of utilitarianism (i.e., full efficiency with $\alpha = 0$) or, at the other extreme, max-min fairness ($\alpha \to \infty$). In particular, if the manager were to choose an α above 2, he might as well ignore efficiency altogether and select $\alpha = \infty$ (i.e., be as fair as possible) because the change in the price of fairness beyond that point is small (or equivalently, the drop in efficiency in going from $\alpha = 2$ to $\alpha = \infty$ is marginal). Similarly, if the manager were to choose an α below 0.5 he might as well set α to 0, i.e., be utilitarian and ignore fairness altogether because he would get a dramatic increase in efficiency for a relatively small price. To summarize, the manager may choose either one of the extremes or else from a relative small range of inequality aversion parameters, based on the appropriate "prices" of fairness and efficiency.

The case study in the next section illustrates how one can utilize the above ideas in practice, specifically in the context of air traffic management.

4. A Case Study in Air Traffic Flow Management

The tools we have introduced thus far provide a principled (as opposed to ad hoc) approach to the design of appropriate operational objectives. This section is devoted to illustrating this value concretely in the context of the air traffic flow management problem. This problem presents the opportunity to save many billion dollars of unnecessary delay costs every year and is viewed as a key priority for the FAA.

Consider the problem faced by the FAA in allocating landing and take-off slots to airlines, as well as routing them across U.S. airspace, in case of reduced capacity due to unpredictable inclement weather. By this allocation, the FAA is effectively allocating unavoidable delays across airlines, as allocations of unfavorable slots result in delayed flights. Currently, the FAA is allocating slots using a *ration by schedule* (RBS) principle, which prioritizes flights based on the original schedule, and is considered as fair. Proposals in the literature, however, promise to reduce total delay by a significant amount (close to 10%), by using mathematical programming models to minimize total delay (see Bertsimas and Stock-Patterson 1998, Odoni and Bianco 1987). Despite the rising delay costs (Airlines for America 2011), none of these proposals have

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

been implemented. One of the principal reasons for this is that those models do not address the question of whether the gains from optimization will be equitably split among the stakeholders. To this end, recent work deals with minimizing system delay in a fair way to all airlines (see Vossen et al. 2003, Barnhart et al. 2012, Bertsimas and Gupta 2012). The notion of what it means to be fair in these pieces of work is ad hoc.

We now consider a principled approach to solving the above problem, based on the model and analysis presented in §§2 and 3. The relative merits of such an approach, compared with the proposals in the literature, are the following:

- The notions of fairness we consider are eminently defensible.
- It will be possible to present a clear analysis of the trade-off inherent in injecting "equity"; presumably this will provide a meaningful basis for the design of a suitable allocation mechanism.

Our framework will apply to this setting in the following way: The airlines correspond to the players, and the FAA to the central decision maker. Because the current policy debate centers around departures from the RBS policy, a natural choice for the utility of each airline is its delay reduction, compared to the RBS policy, which attempts to follow the original schedule in a first-come, first-served fashion. If for an airline a new allocation results in a delay reduction by x minutes, compared to the RBS policy, then that airline derives x units of utility.[3] We consider two ways of measuring the delay of an airline: either by (a) the net delay of the flights it operates or (b) the net delay experienced by the passengers it serves. With those definitions in place, proposals that minimize total system delay, correspond to the utilitarian principle that maximizes the sum of utilities of the players. Accordingly, the FAA can incorporate fairness considerations by utilizing the α-fairness scheme; that is, the FAA carries out the allocation by maximizing the constant elasticity welfare function of the airlines' utilities. By the choice of α, one can then trade off efficiency for fairness.

Furthermore, in case the maximum achievable utilities (i.e., delay reductions) of the airlines are equal, the bounds on the maximum relative efficiency and fairness losses, established in Theorems 1 and 2, are applicable. Numerical studies indicate that when measuring airline delay by flight delays, then the maximum achievable utilities of similar-sized airlines are for all practical purposes equal (see §4.2). When measuring airline delay by passenger delays, the maximum achievable utilities are not equal; we can then use (2) to bound the price of fairness.

[3] As it turns out, there is also an agreed upon dollar figure associated with this delay.

In the introduction, we discussed the relative merits of choosing α using the worst-case analysis we present in this paper for general resource allocation problems. These merits become very important and relevant in the context of the specific problem and industry we consider in this study. In particular, consider the approach of computationally constructing explicit trade-off curves using historical instances of capacity realizations and choosing α based on those. This approach hardly solves the problem at hand, as the selection of the historical instances influences the shape of the trade-off curves and hence the decision. As such, the selection of historical instances is likely to create tension between the airlines. On the other hand, the worst-case analysis is an instance-free approach, better suited for decisions in such an environment. Moreover, it remains unclear how brittle a decision made using historical information might be to unprecedented capacity shocks. Ex post facto, the FAA can always account for such shocks. On the contrary, decisions made based on a worst-case analysis are more robust to such unpredictable scenarios and allow managers to be proactive (rather than reactive) in hedging against them. Finally, in this study, we only consider the problem of allocating landing, takeoff, and airspace slots. In reality, the FAA also coordinates a range of other airline ground operations, in which equity and efficiency need to be balanced as well. The worst-case analysis we present here, however, is problem independent. This would potentially allow the FAA to strategically choose α in concert with all participating airlines to balance efficiency and fairness, and then apply that decision across the range of airline operations.

4.1. The Model

To characterize the utility set, we use a well-accepted model introduced by Bertsimas and Stock-Patterson (1998). The model is highly detailed and specifies a schedule for each flight. In particular, the model specifies for each flight its scheduled location across the national airspace sectors or airports for every time step. The model accounts for the forecasted capacity of each sector and airport, the maximum and nominal speed of the aircraft used for each flight, as well as potential connectivity of flights (through common usage of aircraft or crew). A self-contained mathematical description of the model is included in the online appendix, §D. We refer the reader to the original paper by Bertsimas and Stock-Patterson (1998) for more details.

We model the utilities as follows. We have a set of flights, $\mathcal{F} = \{1, \ldots, F\}$, operated by a set of airlines, $\mathcal{A} = \{1, \ldots, A\}$ over a discrete time period. Let $\mathcal{F}_a \subset \mathcal{F}$ be the set of flights operated by airline $a \in \mathcal{A}$. The flights utilize a capacitated airspace that is divided

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

into sectors, indexed by j. The decision variables used in the model are defined as

$$w_{ft}^{j} = \begin{cases} 1 & \text{if flight } f \text{ arrives at sector } j \text{ by} \\ & \quad \text{time step } t, \\ 0 & \text{otherwise.} \end{cases}$$

We denote the scheduled departure and arrival time of flight f with d_f and r_f, and the origin and destination airports with o_f and k_f, respectively. Then, the associated ground and airborne delays experienced by flight f are

$$g_f = \sum_t t(w_{ft}^{o_f} - w_{f,t-1}^{o_f}) - d_f,$$

$$b_f = \sum_t t(w_{ft}^{k_f} - w_{f,t-1}^{k_f}) - r_f - g_f.$$

The net delay experienced by flight f is $g_f + b_f$. Note that airborne delay typically incurs a higher cost compared to ground delay because of higher fuel consumption, safety issues, etc. As such, it has been proposed to differentiate the impact of ground and airborne delays (see Bertsimas and Gupta 2012). Accordingly, we measure the net delay experienced by flight f by

$$(\text{delay of flight } f) = g_f + 1.5\, b_f.$$

(a) *Flight delay*. Suppose we measure airline delay by the delay of flights. Then, the utility of the ath airline, that is the reduction of its delay compared to the RBS scheme, is equal to

$$u_a = \sum_{f \in \mathcal{F}_a} \text{RBS}_f - \sum_{f \in \mathcal{F}_a} (g_f + 1.5\, b_f), \tag{3}$$

where RBS_f is the delay of flight f under the RBS scheme.

(b) *Passenger delay*. Suppose we measure airline delay by the delay of passengers. Let $p_{a,f}$ be the number of passengers in flight f operated by airline a. Then, the utility of the ath airline, that is the reduction of its delay compared to the RBS scheme, is equal to

$$u_a = \sum_{f \in \mathcal{F}_a} p_{a,f} \text{RBS}_f - \sum_{f \in \mathcal{F}_a} p_{a,f}(g_f + 1.5\, b_f). \tag{4}$$

Our framework provides a means to account for fairness in this fairly complicated setup. In particular, the framework focuses only on the utilities of the airlines, that is the important outcomes of the allocation.

4.2. Numerical Experiments

We focus on scheduling flights over a course of a day for four airlines (as many as the large airlines currently in the United States), which operate at 54 airports, administering in total around 4,000 flights.[4] We use historical data of scheduled and actual flight departure/arrival times on different days to study the performance of α-fairness. In particular, we use the model described above to implement the solution that minimizes total delay, or equivalently in our setting, maximizes the total delay reduction or sum of utilities (utilitarianism). We then implement the α-fairness scheme for different values of the parameter α. We do so both for the case of measuring airline delay by (a) flight delay and (b) passenger delay.

We record the maximum possible system delay reduction (for $\alpha = 0$), and the system delay reduction under the α-fairness scheme, for various positive values of α, particularly, 0.5, 1 (proportional fairness), and 2. We also implement the max-min fairness scheme ($\alpha \to \infty$) and record the system delay reduction. To evaluate the fairness properties of the different schemes, beyond the interpretation based on the value of α, we also record the individual delay reductions of the airlines and their minimum value, which corresponds to our fairness metric (see §3.2).

Table 1 summarizes the numerical results when we measure airline delay by the delay of flights. The results are for two representative (actual) days, on which inclement weather severely affected operations across the country. For each day and airline, the actual cumulative delay (in minutes) across its flights on that day is reported. We calculate the delay reductions that the utilitarian and the α-fairness schemes would achieve, for different values of α, including the special cases of proportional ($\alpha = 1$) and max-min fairness ($\alpha \to \infty$). The utilitarian scheme achieves roughly a 10% reduction compared to the current RBS policy, which is the largest possible for the schemes we consider. The α-fair allocations yield lower delay reductions, but still the price is relatively small and increasing with α. Note also that the distribution of delay reductions changes rapidly as we are varying α. In particular, note that the utilitarian scheme does not equitably split the gains from optimization, because some airlines incur the same delay as in RBS, and others achieve large reductions. On the contrary, under max-min fairness, all airlines are granted almost the same delay reduction. Figure 6 illustrates the associated prices of fairness and efficiency, as a function of α. We also plot the worst-case bounds implied by Theorems 1 and 2. As expected, increasing the inequality aversion parameter α yields an increase in the efficiency loss and a decrease in the fairness loss.

To support our claim that the maximum achievable utilities of similar-sized airlines are for all practical

[4] There are around 35,000 domestic flights scheduled on a daily basis in the United States.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

Table 1 Numerical Results for §4.2 for Two Days and Four Airlines

		Delay reduction				
	RBS delay (under RBS)	Utilitarian ($\alpha = 0$)	α-fair ($\alpha = 0.5$)	Prop. fair ($\alpha = 1$)	α-fair ($\alpha = 2$)	Max-min fair ($\alpha \to \infty$)
May 13, 2005						
Airline 1	13,985	0	802.5	892.5	930	967.5
Airline 2	7,182	990	862.5	922.5	945	967.5
Airline 3	15,239	990	862.5	922.5	945	967.5
Airline 4	10,415	2,250	1,492.5	1,215	1,087.5	967.5
Total	46,821	4,230	4,020	3,952.5	3,907.5	3,870
July 27, 2006						
Airline 1	12,095	1,537.5	802.5	795	795	787.5
Airline 2	8,933	0	675	735	757.5	787.5
Airline 3	9,531	1,800	960	870	825	787.5
Airline 4	9,551	0	735	765	772.5	787.5
Total	40,110	3,337.5	3,172.5	3,165	3,150	3,150

Note. For each airline, we report the actual delay (in minutes) across its flights on that day (under the RBS policy) and the associated delay reductions that different allocations would achieve.

purposes equal for the experiments above (where we measure flight delays), note that their coefficients of variation are 0.015 and 0.007 for May 13, 2005, and July 27, 2006, respectively.

Similarly, Table 2 summarizes the numerical results when we measure airline delay by the delay of passengers. The results are for the same two days as above. For each day and airline, the actual cumulative delay (in thousand minutes) of its passengers on that day is reported, along with the associated delay reductions of different schemes. Note that in this situation, the α-fair allocations yield again lower delay reductions, but the prices of fairness and efficiency are now larger than in the case of measuring flight delay. Figure 7 depicts the associated prices of fairness and efficiency for that case, as a function of α. Note that for the two days we consider, airlines 3 and 4 serve 1.6 times more passengers per flight (on average) compared to airlines 1 and 2. Recall that when we measured flight delays above, the maximum achievable utilities of the airlines were equal. Hence, when we measure passenger delay, the maximum achievable utilities of the airlines are no longer equal, in particular, the maximum achievable utilities of airlines 3 and 4 are (roughly) 1.6 times larger compared to airlines 1 and 2 (see (3) and (4)). One can then use (2) to obtain a bound for the associated price of fairness (for $B = 1.6$ and $L = 1$), which is also plotted in Figure 7(a).

4.3. Conclusions from Empirical Study

We conclude with a few takeaways from our case-study.

1. *Quality of Decisions*. In our discussion of a managerial prescription, one concrete prescription was to choose an α that balanced the respective prices of

Figure 6 (a) Price of Fairness and (b) Price of Efficiency for the Case of Measuring Flight Delays in the Numerical Experiments in §4.2 for Different Values of the Parameter α

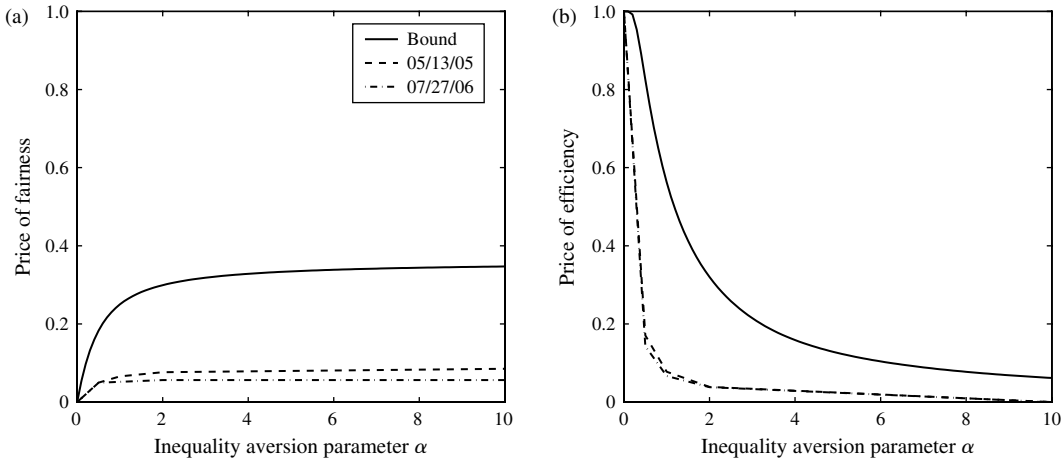


Table 2 Numerical Results for §4.2 for Two Days and Four Airlines

		Delay reduction				
	RBS delay (under RBS)	Utilitarian ($\alpha = 0$)	α-fair ($\alpha = 0.5$)	Prop. fair ($\alpha = 1$)	α-fair ($\alpha = 2$)	Max-min fair ($\alpha \to \infty$)
May 13, 2005						
Airline 1	1,174.7	0	51.6	75.9	88	99.5
Airline 2	603.3	0	55.5	78.4	89.9	99.5
Airline 3	2,057.3	269.7	141.5	125.6	113.5	99.5
Airline 4	1,406	306	245.4	165.1	130.1	99.5
Total	5,241.3	575.7	494.1	445	421.4	397.8
July 27, 2006						
Airline 1	1,028.1	0	51.6	67.6	74.6	81.6
Airline 2	759.3	0	44	61.8	72	81.6
Airline 3	1,286.7	244.8	159.4	117.9	98.8	81.6
Airline 4	1,289.4	200.8	122.4	103.3	92.4	81.6
Total	4,363.5	445.6	377.4	350.6	337.9	326.4

Note. For each airline, we report the actual delay (in thousand minutes) experienced by its passengers on that day (under the RBS policy) and the associated delay reductions that different allocations would achieve.

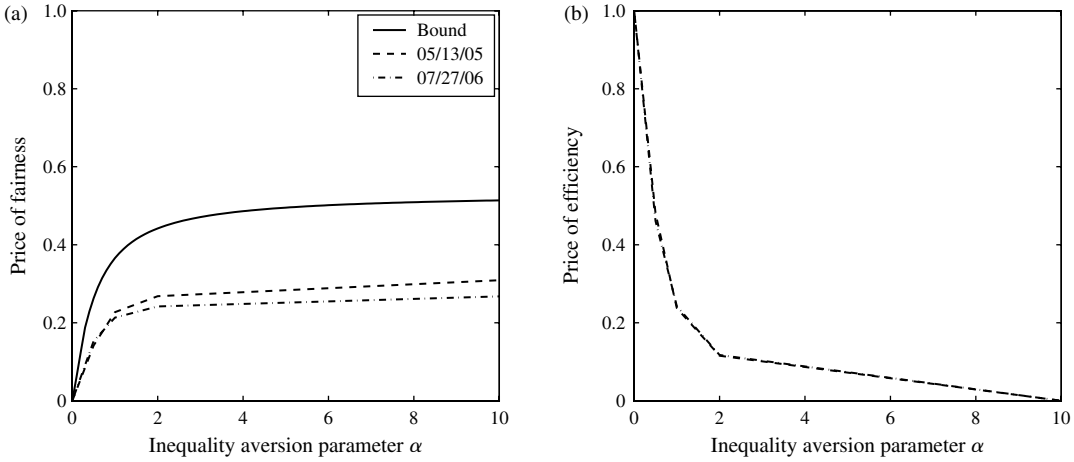
efficiency and fairness. In the setting where utilities were normalized (or equal), we saw that an appropriate choice of α that accomplished this was $\alpha = 2$. For our first set of experiments (where maximum achievable utilities are essentially equal), we see that this decision would incur a price of fairness of about 7.6% and a price of efficiency of about 3.8%. Put another way, this choice of α yields a total utility that is within about 93% of the most efficient solution, whereas the utility of the worst-off player is simultaneously within 97% of that it would have been under a scheme that maximized the utility of the worst-off player. Note that it is not the case that this is simply because the choice of α here "did not matter." For instance, if the manager picked the utilitarian solution ($\alpha = 0$), efficiency would obviously be 100%, but at the price of completely excluding one of the players (i.e., the price of efficiency would be 100%)!

If one were to make this decision using trade-off curves computed *explicitly* for this problem instance, one might be led to choose $\alpha = 1$, which would incur a price of fairness of about 6.5% and a price of efficiency of about 7.7%. Although this choice does take the opportunity to reduce the price of fairness presented by this specific example, the improvement is relatively small (from 7.6% to 6.5%), so that we lost very little in using our robust framework in making a selection of α here.

In summary, our prescription successfully navigated a fairly subtle trade-off.

2. *Robustness Matters.* In looking at the *absolute* prices of fairness and efficiency relative to their worst-case values in our first example (Figure 6), one would deduce that if the manager were to choose α based on some budget on the price of efficiency (or fairness) the resulting might be conservative. In particular, if the

Figure 7 (a) Price of Fairness and (b) Price of Efficiency for the Case of Measuring Passenger Delays in the Numerical Experiments in §4.2 for Different Values of the Parameter α



INFORMS holds copyright to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.

manager were willing to be as fair as possible provided the resulting loss in efficiency was less than 20% (i.e., the price of fairness was no more than 20%), he would pick an α of about 0.5. This would realize an efficiency loss of only 5%; in this case it would have actually been possible for the manager to be fully fair (i.e., pick $\alpha = \infty$) while satisfying the constraint of an efficiency loss of less than 20%. Notice, however, that with the second objective, the conservativeness is substantially less as is evidenced by Figure 7; there a conservative choice would pick $\alpha = 0.5$ for which the *actual* efficiency loss is about 15%, which is not nearly as conservative as in the first instance!

In summary, it is not the case that the worst-case *prices* necessarily arise from pathological examples; here we see a *real-world* problem instance that comes fairly close. That said, if one is aware of additional invariants in the decision problem, this information could be used to further constrain the description of the utility set considered in §2, and one could then hope to computationally construct a trade-off curve, as we did analytically for the case where U is simply required to be convex and compact.

5. Concluding Remarks

We dealt with the problem of designing operational objectives, particularly, balancing efficiency and fairness in the context of resource allocation. We reviewed a plethora of problems in the broad area of operations management for which this dichotomy constitutes a central issue.

Despite the fact that fairness is of a subjective nature, we identify a notion of fairness that is well documented in the welfare economics literature and is of practical interest: the notion of α-fairness. That notion provides a family of welfare functions that is canonical in that it captures the utilitarian allocation, the max-min fair allocation (or Kalai–Smorodinsky) and the proportionally fair (or Nash bargaining) allocation. It also permits the decision maker to trade off efficiency for fairness by means of a single parameter.

For the above notion, we provide near-tight upper bounds on the relative efficiency loss compared to the efficiency-maximizing solution, where we measure efficiency by the sum of player utilities. Similarly, we provide tight upper bounds on the relative fairness loss, where we measure fairness by the minimum utility of players. The bounds are applicable to a broad family of problems; they also suggest when the loss is likely to be small, and illustrate its dependence on the numbers of parties involved and the chosen balance between efficiency and fairness. Such a contribution has been elusive in the literature, to the best of our knowledge, and now provides the means for central decision makers to select their attitudes toward fairness and efficiency using quantitative arguments.

Acknowledgments

The authors thank the anonymous reviewers for providing constructive feedback. The authors thank Bill Moser and Mark Weber of Lincoln Labs and Shubham Gupta for providing the data for their case study. The research was partially supported by the National Science Foundation [Grants DMI-0556106, EFRI-0735905].

References

Airlines for America (2011) Annual and per-minute cost of delays to U.S. airlines. Accessed November 2011, http://www.airlines.org/Pages/Annual-and-Per-Minute-Cost-of-Delays-to-U.S.-Airlines.aspx.

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* 58(3):624–637.

Arrow KJ (1965) *Aspects of the Theory of Risk-Bearing* (Yrjö Jahnssonin Säätiö, Helsinki, Finland).

Atkinson AB (1970) On the measurement of inequality. *J. Econom. Theory* 2(3):244–263.

Barnhart C, Bertsimas D, Caramanis C, Fearing D (2012) Equitable and efficient coordination in traffic flow management. *Transportation Sci.* 46(2):262–280.

Barr N (1987) *The Economics of the Welfare State* (Weidenfeld and Nicolson, London).

Bergson A (1938) A reformulation of certain aspects of welfare economics. *Quart. J. Econom.* 52(2):310–334.

Bertsekas D, Gallager R (1987) *Data Networks* (Prentice-Hall, Upper Saddle River, NJ).

Bertsimas D, Gupta S (2012) On fairness and collaboration in network air traffic flow management: An optimization approach. Working paper, Massachusetts Institute of Technology, Cambridge.

Bertsimas D, Stock-Patterson S (1998) The air traffic flow management problem with enroute capacities. *Oper. Res.* 46(3):406–422.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Oper. Res.* 59(1):17–31.

Bertsimas D, Farias VF, Trichakis N (2012) Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Oper. Res.* Forthcoming.

Bertsimas D, Paschalidis IC, Tsitsiklis JN (1994) Optimization of multiclass queuing networks: Polyhedral and nonlinear characterizations of achievable performance. *Ann. Appl. Probab.* 4(1):43–75.

Bisias D, Lo A, Watkins J (2012) Estimating the NIH efficient frontier. *PLoS ONE* 7(5):e34569.

Bonald T, Massoulié L (2001) Impact of fairness on Internet performance. *SIGMETRICS Perform. Eval. Rev.* 29(1):82–91.

Butler M, Williams HP (2002) Fairness versus efficiency in charging for the use of common facilities. *J. Oper. Res. Soc.* 53(12):1324–1329.

Callahan D, Wasunna AA (2006) *Medicine and the Market: Equity v. Choice* (Johns Hopkins University Press, Baltimore).

Chakrabarty D, Goel G, Vazirani VV, Wang L, Yu C (2009) Some computational and game-theoretic issues in Nash and nonsymmetric bargaining games. Working paper, Georgia Institute of Technology, Atlanta.

Chan CW, Armony M, Bambos N (2012a) Fairness in overloaded parallel queues. Working paper, Columbia University, New York.

Chan CW, Farias VF, Bambos N, Escobar GJ (2012b) Maximizing throughput of hospital intensive care units with patient readmissions. *Oper. Res.* Forthcoming.

Correa JR, Schulz AS, Stier-Moses NE (2007) Fast, fair, and efficient flows in networks. *Oper. Res.* 55(2):215–225.

Cui TH, Raju JS, Zhang ZJ (2007) Fairness and channel coordination. *Management Sci.* 53(8):1303–1314.

Federgruen A, Groenevelt H (1988) $M/G/c$ queueing systems with multiple customer classes: Characterization and control of

achievable performance under nonpreemptive priority rules. *Management Sci.* 34(9):1121–1138.

Gelenbe E, Mitrani L (1980) *Analysis and Synthesis of Computer Systems* (Academic, London).

Harsanyi JC (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *J. Political Econom.* 63(4):309–321.

Kalai E, Smorodinsky M (1975) Other solutions to Nash's bargaining problem. *Econometrica* 43(3):513–518.

Kelly FP, Maulloo A, Tan D (1998) Rate control for communication networks: Shadow prices, proportional fairness, and stability. *J. Oper. Res. Soc.* 49(3):237–252.

Lan T, Kao D, Chiang M, Sabharwal A (2010) An axiomatic theory of fairness in network resource allocation. *INFOCOM'10 Proc. 29th Conf. on Inform. Comm.* (IEEE, Piscataway, NJ), 1343–1351.

Lensberg T (1988) Stability and the Nash solution. *J. Econom. Theory* 45(2):330–341.

Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory* (Oxford University Press, New York).

Mo J, Walrand J (2000) Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Networking* 8(5):556–567.

Nash J (1950) The bargaining problem. *Econometrica* 18(2):155–162.

Odoni AR, Bianco L (1987) *Flow Control of Congested Networks, Chapter the Flow Management Problem in Air Traffic Control* (Springer-Verlag, Berlin).

Pauly MV (2010) Avoiding side effects in implementing health insurance reform. *New England J. Medicine* 362(8):671–673.

Pratt JW (1964) Risk aversion in the small and large. *Econometrica* 32(1/2):122–136.

Rawls J (1971) *A Theory of Justice* (Harvard University Press, Cambridge, MA).

Resnick DB (2003) Setting biomedical research priorities in the 21st century. *Virtual Mentor* 5(7), http://virtualmentor.ama-assn.org/2003/07/msoc1-0307.html.

Roth A (1979) *Axiomatic Models of Bargaining* (Springer-Verlag, Berlin).

Samuelson P (1947) *Foundations of Economic Analysis* (Harvard University Press, Cambridge, MA).

Sen A, Foster JE (1997) *On Economic Inequality* (Oxford University Press, New York).

Shreedhar M, Varghese G (1996) Efficient fair queueing using deficit round-robin. *IEEE/ACM Trans. Networking* 4(3):375–385.

Su X, Zenios SA (2004) Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing Service Oper. Management* 6(4):280–301.

Su X, Zenios SA (2006) Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Sci.* 52(11):1647–1660.

Swenson MD (1992) Scarcity in the intensive care unit: Principles of justice for rationing ICU beds. *Amer. J. Medicine* 92(5):551–555.

Tang A, Wang J, Low SH (2006) Counter-intuitive throughput behaviors in networks under end-to-end control. *IEEE/ACM Trans. Networking* 14(2):355–368.

Tsoucas P (1991) The region of achievable performance in a model of Klimov. Research Report RC16543, IBM T. J. Watson Research Center, Yorktown Heights, NY.

Vossen T, Ball M, Hoffman R (2003) A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. *Air Traffic Control Quart.* 11(4):277–292.

Wagstaff A (1991) QALYs and the equity-efficiency trade-off. *J. Health Econom.* 10(1):21–41.

Wu Y, Loch CH, Van der Heyden L (2008) A model of fair process and its limits. *Manufacturing Service Oper. Management* 10(4):637–653.

Young PH (1995) *Equity: In Theory and Practice* (Princeton University Press, Princeton, NJ).

INFORMS holds **copyright** to this article and distributed this copy as a courtesy to the author(s).
Additional information, including rights and permission policies, is available at http://journals.informs.org/.