View metadata, citation and similar papers at core.ac.uk

Probability in the Engineering and Informational Sciences, **22**, 2008, 495–517. Printed in the U.S.A.
doi:10.1017/S0269964808000302

QUANTIFYING FAIRNESS IN QUEUING SYSTEMS

PRINCIPLES, APPROACHES, AND APPLICABILITY

BENJAMIN AVI-ITZHAK
RUTCOR
Rutgers University
New Brunswick, NJ, USA
E-mail: aviitzha@rutcor.rutgers.edu

HANOCH LEVY*
Computer Engineering and Networks Lab
ETH Zurich, Switzerland
E-mail: hanoch@tik.ee.ethz.ch

DAVID RAZ†
Holon Institute of Technology
Holon, Israel
E-mail: davidra@hit.ac.il

In this article we discuss fairness in queues, view it in the context of social justice at large, and survey the recently published research work and publications dealing with the issue of *measuring* fairness of queues. The emphasis is placed on the underlying principles of the different measurement approaches, on reviewing their methodology, and on examining their applicability and intuitive appeal. Some quantitative results are also presented.

The article has three major parts (sections) and a short concluding discussion. In the first part we discuss fairness in queues and its importance in the broader context of the prevailing conception of social justice at large, and the distinction between *fairness of the queue* and *fairness at large* is illuminated. The second part is dedicated to explaining and discussing three main properties expected of a fairness measure: conformity to the general concept of social justice, granularity, and intuitive

*On leave of absence from the School of Computer Science, Tel-Aviv University.
†A large part of this work was done while at the school of Computer Science, Tel-Aviv University.

Downloaded from Cambridge University Press core.ac.uk at University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

appeal and rationality. The third part reviews the fairness of the queue evaluating and measuring approaches proposed and studied in recent years. We describe the underlying principles of the different approaches, present some of their results, and review them in context of the three main properties expected from a measure. The short discussion that follows centers on future research issues.

1. INTRODUCTION

1.1. Preface

Why are we using ordered queues? Why do they serve in many real-life applications, such as banks, supermarkets, airports, computer systems, communications systems, Web services, call centers, and numerous other systems?

Although the major reason for the formation of queues is economic (i.e., scarcity of resources), the dominant reason for using ordered (disciplined) queues is often the strive to maintain some level of social justice, or in other words, fairness in treatment of everyone involved. (Queues also increase efficiency because who is served next does not have to be negotiated constantly.)

In this sense, a system serving a queue of people is a microcosm social construct. Emotions and resentment might flare if unfairness is practiced or is perceived as being practiced, whereas courtesy, and even camaraderie due to same experience-sharing, might result when fairness in treatment is perceived (see Rafaeli, Barron, and Haber [19]). Notwithstanding its fundamental role, the fairness factor was virtually neglected, or even disregarded, in the published queuing literature until quite recently. Aspects of fairness in queues were recognized and discussed, or mentioned in passing, quite early by a considerable number of authors: Palm [16] dealt with judging the annoyance caused by congestion, Mann [12] discussed the queue as a social system, and Whitt [28] addressed overtaking in queues, to mention just three.

Although almost every child, if asked, can tell you what is fair and what is not, it is not an undemanding undertaking to have a group of people agree on a common definition of fairness, much more so when it comes to defining a quantitative measure of the level of fairness and when the group is large. It is not surprising, then, that extensive research aimed at developing fairness measures for queues, in contrast to the traditional "efficiency" measures of sojourn and waiting times, has been slow in coming.

Traditionally, a first come–first served (FCFS), or a first in–first out (FIFO), queue discipline is considered most fair. This probably derives from experience in queues where the total amount of service the system is able, or willing, to dispose is limited by a maximal number served or by a length of time the system is open for service (i.e., *exhaustible-servers* systems). In such systems (e.g., a line at a gas pump at a time of energy crisis, a line for basic foods in a refugee camp, or, less dramatic, a line for tickets for a show or a sports event), which were very prevalent in the human experience; if you are not early enough in the queue, chances are you will never get the service or product or you might have to come again at a future time (viz., the early bird gets the worm). Placing ahead of you a person who arrived after you is or can be regarded as

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

grossly unfair, particularly if that person is not needier than you. Many present-day queuing systems, however, are not of this type; rather, all birds get their worms, not only the early ones, and, thus, FIFO might not be as crucial in these systems. Fairness of exhaustible-servers queues is an important issue, deserving attention on its own, and is outside the scope of this article, which focuses on nonexhaustive servers.

Larson [11] in his discussion paper on the psychology of waiting recognized the central role played by "Social Justice" (which is another name for fairness). In the first part of his work, dedicated to social justice in queues, he brought several anecdotal actual situations, experienced by him and others, that strongly support the traditional claim of FIFO being the most socially just queue discipline. In fact he practically defined social injustice as violation of FIFO when stating "…customers may become infuriated if they experience social injustice, defined as violation of FIFO."

What would be a fair service order in a supermarket queue or in an airport waiting line? Many people would instinctively embrace Larson's view, responding that FIFO is the fairest order (i.e., serving the most *senior* customer first, where *seniority* is measured in the time the customer has already spent in the line). Already Kingman [9] pronounced this same view by calling FIFO "the fairest queue discipline." The underlying principle, or rationale, of this view can be expressed in one sentence: *The one who has been waiting longest earned the right to be served first.* However, recalling that the server is nonexhaustible (viz. it can serve forever), *is* FIFO undeniably the most fair discipline?

To answer this question, consider a common situation at a supermarket counter, which some readers can associate with their own personal experience (see Fig. 1): Mr. Short arrives at the supermarket counter holding only one item. In the line ahead of him he finds Mrs. Long carrying two fully loaded carts of items. Short says to Long "Excuse me, I only have one item. Would you mind if I go ahead of you?" Would it be fair to have Mrs. Long served ahead of Short and Short waiting for the full processing of Mrs. Long's loaded carts? Or, would it be fairer to advance Short in the queue and serve him ahead of Long? This dilemma might cause some to "relax" their strong belief

FIGURE 1. The supermarket of Mrs. Long and Mr. Short.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

in the absolute fairness of FIFO. In fact, the dilemma brings to the discussion a new factor: that of *service requirement.* The basic intuition thus suggests that prioritizing short jobs over long jobs might also be fair, based on the underlying principle that *the one who demands the least of the server's time should be served first*. It is the trade-off between these two factors, *seniority* (prioritize Mrs. Long) and *service requirement* (prioritize Mr. Short), that creates the dilemma in this case. To demonstrate the conflict, we continue our scenario in two directions: (1) Long responds "Why don't you go ahead of me. I have arrived just a few seconds ago and it is not fair that you will wait that long while your short service will delay me very little." This is one possibility. Alternately, Long might be negative, saying (2) "Look, I have been waiting in this line forever. If not for this lengthy wait I would have been out of here long before your arrival. You can patiently wait too." Clearly, Long weighs their seniority difference against their service requirement difference in deciding what is the fair thing to do. This trade-off, illustrated by the "Long vs. Short" scenario, will accompany us in this article in attempting to understand fairness in queues. (It should be noted that many supermarkets handle this conflict by allocating some of the counters exclusively to Shorts, who also retain the option to select a "regular" counter.)

1.2. What Is "Fairness of the Queue"?

Evidently, there is a need to agree upon the definition of fairness, or at least the underlying principles, or a rationale that forms its foundation. As mentioned earlier, a queuing system is a microcosm social construct and its fairness should conform to the general cultural perception of social justice in the particular society. Social justice has always been, and still is, a cardinal issue in all cultures. It is the cement holding the society together. As such, it has been subject to debate by philosophers, prophets, and spiritual leaders since the beginning of recorded history. In modern time, many economists and social scientists joined the ongoing debate. Because the perception of social justice is culture and time dependent, we are interested in the modern Western societies' perspective. As is to be expected, there is a vast ocean of modern research and publications on this issue, mostly by philosophers, economists, and social and behavioral scientists. Reviewing and interpreting this literature is much beyond the scope of this article and probably also beyond our ability. To readers who would like to dive into this ocean, or just wet their feet at its shores, we recommend visiting the *Stanford Encyclopedia of Philosophy* [27]. A most, some would say *the* most, prominent and comprehensive publication on this issue is Rawls', '*A Theory of Justice*' [21].[1] The book does not make for an easy reading, but, in essence, Rawls' general conception of social justice, as summarized in a nutshell by Piccard [17] is the following:

> All social primary goods—liberty and opportunity, income and wealth, and the bases for self-respect—are to be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

We are back to the traditional economists' approach of achieving social justice by appropriately dividing the "pie," except that the pie here is made of a mix of tangibles and intangibles, whereas the traditional economists' pie is wholly tangible. By Rawls' conception, if all persons involved are equally nonfavored (equally needy), the pie should be equally divided. Obviously, Rawls's conception, although widely recognized, has its dissenters, as is true for practically any social issue. We will use it here as a guideline.

1.2.1. Fairness of the Queue Versus Fairness in a Queuing System. If we accept the above conception, it must also apply to queuing environments as well. We therefore need to differentiate between *fairness in a queuing system* and *fairness of a queue*, which is, roughly put, the fairness component that is attributable to the queue discipline or structure. For illustration, imagine a waiting room packed with patients. The door to the doctor's office opens and a nurse appears and asks: "Who is the sickest?" This order of service is near to longest-job-first (LJF). Still, the many, if not most, will say it is fair by the principle that those most at risk, or those suffering the most, should be attended to first. The fairness issue is cast here in a queuing situation. Alas, it has little to do with fairness of the queue. Very few will categorize a LJF ordering as fair, given that all customers are equally needy.

In this discussion article we define the *fairness/unfairness of the queue* as the fairness/unfairness that can be related to the discipline or configuration of the queue when all customers are equally needy. Customers will be assumed to be equally needy if they are discernable only by their arrival time and service requirement and are identical in all other respects. The doctor's waiting room scenario is very realistic when looking at hospitals' ERs. Arriving patients are categorized into several classes of neediness (urgency, or critical level, of condition) and the classes are prioritized according to their level of neediness. The fairness related to the class prioritizing is determined by the neediness of the customers. The "fairness of the queue" in this situation is related to the order of service within each class, under the assumption that same class patients are practically equally needy.

Note that most customers would not differentiate between fairness of the queue and the fairness of the system, unless specifically guided to so do. Rafaeli, Kedmi, Vashdi, and Barron [20, Study III] conducted an experiment comparing perceived fairness by customers in a multiserver/multiqueue system (each server has its own queue, served in a FIFO order) to that of customers served in the same system that has, in addition to the regular queues, VIP queues (e.g., business class check-in counters in an airport). Only responses of those served in non-VIP queues were considered. Average fairness in the VIP structure was found to be significantly lower than in the structure without VIP queues, unless people knew that those in the VIP queue had paid a special fee in order to join it; that is, the queue was perceived as unfair by participants who thought others are getting a preferential treatment with no justification. However, once they learned that the preference was bought for a special fee, they perceived the same system as being fair. In the first situation we are dealing with the perceived

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

fairness of the queue. In the second situation we are dealing with the perception of the system's fairness at large. Participants perceive buying preferential treatment as fair.

In what follows, fairness is meant to stand for fairness of the queue, unless otherwise specified.

1.2.2. What Is the Pie and How Can It Be Equally Divided? Recall Rawls' conception of appropriately dividing the pie. Assume all customers are equally needy, what then is the "pie" and how can it be "equally divided"? Clearly, the scarce resource, or the pie, is the service the servers are capable of rendering. As for dividing it equally, consider, for example, an $M/D/1$ system where the service requirement is the same constant for all customers. In such a system, all customers *seemingly* receive an equal share of the server's attention; hence, FIFO and LIFO should be equally fair! Not so, in an ongoing serving system: the time of receiving the service might be essential; hence, the pie must be continually divided over time. Therein lies the key to the just division.

One can take either an intuitive-pragmatic approach or a formal-direct one to attaining a just division of the resources over time. In the former, for example, it can be assumed that the customer gets the utility of the service plus the disutility of the wait. Therefore, in the $M/D/1$ case, where equal service time is given to all customers, the waiting times must also be identical to attain *absolute* fairness. Unfairness in this case is produced by deviations from equal wait, and a "natural" measure for it is the waiting time variance. The fairest discipline must produce the smallest waiting time variance. For the $M/D/1$ class of disciplines that are nonpreemptive (Processor Sharing is considered to be preemptive) and work conserving, the smallest variance is produced by FIFO. This is also true for $M/G/1$ [1,9]. Extending this approach to the $M/G/1$ system we note that customers receive unequal shares of the server's attention, giving rise to the Long-versus-Short dilemma. One way to solve this conflict is to assume that absolute fairness is achieved if the waiting disutility of each customer is proportional to his/her service utility (assuming, for simplicity, linearity of both utility functions). The unfairness measure can be derived from the variance of the deviations from this proportionality. In both cases, the particular "equal and timely pie division" results from a pragmatic *perceived* fairness of the queue, instead of vice versa. The conformity to the conception of general social justice is an after-the-fact rationalizing of the pragmatic fairness-of-the-queue principle used.

One formal-direct approach to equally dividing the resource is to assume that the community entitled to a slice of it at any moment is made of the customers present in the system at that time. If there are N customers present, each is entitled to receive $(1/N)$-th of the server's attention (service rate) to achieve absolute fairness. Thus, the pie is equally divided at all points in time. Deviations from this division of the server's rate are unfair, and a summary measure of their variability can serve as an unfairness measure. In this approach, the absolute fairness results from the formal just division of the pie, in contrast to the former approach. Still, it remains to agree on how to measure deviations from the defined just division of the servers' rate. Note

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

that dividing the resources according to this just approach (called processor sharing) is not Pareto optimal because in many situations all customers get slowed down and everyone loses; see a short discussion on this issue at the end of Section 4.

1.3. Importance and Applicability of Fairness of the Queue

As already mentioned, the fairness factor has long been recognized in queuing literature. Nevertheless, queuing theory has been mostly occupied with the performance metrics of waiting time, which is frequently being looked at via its expected value. Under this quantity, the customers' objective is to minimize delay. The fairness factor, although playing an important role in the design and operation of actual waiting systems, has only recently become a topic of interest also to queuing theorists. Rothkopf and Rech [8], in their article discussing perceptions in queues, bring an impressive list of quantifiable considerations showing that combining queues might not be economically advantageous, contra to the "common" belief. At the end, they concede, however, that all of these considerations might not have sufficient weight to overcome the unfairness perceived by customers served in a separate queues structure.

Experimental evidence of the importance of fairness in queues was recently provided in Rafaeli et al. [19]. The experimental studies revealed that for humans waiting in queues, the issue of fairness is highly important, perhaps sometimes more important than the duration of the wait. For the case of common queue versus a separate one at each server, they found that the common queue was perceived as fairer. Probably for this reason we find separate queues mostly in systems where a common queue is physically not practical (e.g., traffic toll boots and supermarkets).

In most situations of limited resources there is a need to utilize, or share, the resources in an efficient and fair way. Thus, an ordered queue is a fairness and efficiency management facility and is perceived as such by most service systems operators. Supermarkets, where common queues are not always practical, try to increase both fairness and efficiency by assigning some of the counters to Shorts only. The same practice is common to toll booths as well, where "Easy Pass" is used. An alternate solution is to make a common queue feasible by allocating the necessary additional resources. For example, if you arrive to Newark airport on an international flight, you find that the passport control queue is common and an extra attendant is assigned to direct people to the next available server and reduce overtaking.

One intriguing situation of fairness applicability is the *blind queue*. In the course of our study of fairness in queues we were asked more than once "Is fairness relevant at all in a blind queue?" There are many situations in which customers cannot see each other and are not informed of the state of the system and the discipline used. Telephone systems operators know from experience that some customers are impatient and are likely to renege after a relatively short wait. Customers who are more patient will hang on for quite a while before hanging up (pun not intended). Therefore, a waiting customer is more likely to be a patient one, as compared to a new arrival. Using a LIFO waiting line discipline might result in retaining more customers and increased profit (see Pla, Casares-Giner, and Martinez [18] for a list of references).

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

However, most customers would consider LIFO as unfair, even if informed of it ahead of time, and outrageous if it is concealed and then revealed to them somehow. In fact, in today's information age, it is hard to expect such practice to remain concealed for long. Suppose, nonetheless, that such LIFO practice can indeed be hidden. Does it make the practice fair? No. Is fairness in this case relevant? This is a question of ethics. Is cheating right if it never gets disclosed and the cheater can get away with it unscathed? The answer depends on your ethical values.

In fact, making the queue less blind might be quite important to customers. Many call centers will inform you of your place in the line and sometimes provide you with an estimate of the wait involved. This allows you to be aware that the order of service is FIFO and enables you to renege now, instead of wasting so much of your time before reneging anyway. Both are fairness considerations. Along these lines, surveys of 911 callers who were classified by the police as "low priority" and kept waiting a long time for police arrival found that callers were not dissatisfied with the service, provided they were told that the police are busy with higher-priority calls and tasks and were also told to expect a long delay (see Larson [11], Chan and Tien [6], and McEwen, Connors, and Cohen [9]). In this case, although we are dealing with fairness based on need rather than fairness of the queue, the knowledge that the system is fair strongly influences the callers' degree of satisfaction and prevents repeated calls and complaints.

1.4. Job Versus Flow-Related Fairness

Queuing model applications can be classified into (1) *job-based systems* and (2) *flow-based systems*. In the former, the ith customer, say C_i, is associated with a single job J_i arriving at epoch a_i and requiring service time $s_i, i = 1, 2, \ldots$. Of interest is therefore the performance experienced by that individual job, which is synonymous with the customer in this article. In the latter, customer C_i is associated with a stream (or flow) of jobs $J_i^1, J_i^2, \ldots$ arriving at epochs $a_i^1, a_i^2, \ldots$, respectively. Of interest is the performance experienced by the whole flow. The applications associated with this latter model are communications networks applications where a customer (sometimes called source) is associated with a stream of packets that are sent through a communications device (e.g., a router). For a brief overview of *flow-based systems* publications, see Avi-Itzhak, Levy, and Raz [2].

Our focus in this work is on job-based systems. Applications that are associated with this model are as follows:

1. **Banks, supermarkets, public offices, and the like,** in which customers physically enter queues where they wait for service and then get served;
2. **Some computer systems,** in which a customer (or a customer's computer application) submits a job to the system, and the customer is satisfied when the service of the job is completed;
3. **Call centers,** in which customers call to receive service, possibly wait in a virtual queue (while listening to some music) until being answered by "the

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

next available agent". Call center queuing systems are conceptually identical to physical queuing facilities, such as banks or airlines counters, except that the queue can be blind unless the operator decides otherwise.

2. PROPERTIES EXPECTED OF A FAIRNESS MEASURE

When introducing a new queuing performance measure for a seemingly intangible entity like fairness, several questions should be discussed. How does the underlying principle (or conception) used conform to the wider, nonqueuing-related approach to dealing with fairness? What quantities should be measured and at what *level of detail*? How *intuitive* and *appealing* is the measure?

These questions relate to three major properties characterizing the measure: (1) *conformity*, (2) *granularity, and* (3) *intuitive appeal and rationality.* In this section we discuss these properties, to be used later in examining the fairness measures proposed recently in the literature.

2.1. Conformity to the General Concept of Social Justice

For many people, fairness perception is very intuitive, almost instinctive. Thus, approaches toward fairness of the queue are mostly based on pragmatic principles (e.g., seniority must be respected, customers requiring little should get priority, waiting time should be in proportion to the service required) not necessarily directly based on some abstract general formal conception offered by "deep thinkers." Nevertheless, general formal conceptions emerge from the same "natural" pragmatic cultural attitudes of society. Therefore, the underlying principle of a queuing fairness measure should conform to the general cultural perception of social justice prevailing in the particular society, either directly or indirectly. If it does not, its acceptance might be deterred.

2.2. Granularity

At what granularity level should the fairness performance metric conform to the underlying fairness principle? Our conclusion is that conformity is desirable on all three granularity levels of the system:

1. At the *individual customer level,* by a quantity representing the deviation of the treatment given to each customer from the absolutely fair treatment as defined by the underlying fairness principle of the measure (referred to henceforth as *discrimination*);
2. At the *scenario level*, by a summary statistic of the discriminations as experienced by a (finite or infinite) set of customers in a particular scenario (a sample path of the stochastic process);

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

3. At the *system level,* by a summary statistic (e.g., expected value or variance) of the performance as experienced by an arbitrary customer (in a stationary or a transient system).

In the context of fairness, it is essential to make explicit use of the individual and scenario quantities in addition to the system's fairness, since humans can feel them better and relate to them better than to the third quantity. This is important to building confidence in the fairness measure, which is somewhat abstract, intangible, and difficult to feel.

In deriving the *system fairness*, one can take two different approaches for dealing with the *stochastic* nature of the system:

1. *Fairness of the actual performance measure:* This approach accounts for the performance measure of each individual customer and then uses some summary statistics function (e.g., the *max* operation or some type of *expectation*) to compare them to each other yielding the scenario or system fairness measures. Thus, the approach compares the *actual performance measures* (*or discriminations*) experienced by the individuals.
2. *Fairness of the mean*: This approach classifies the customers into classes (where a class can be a customer that repeats visiting the system indefinitely) and computes for each class the expected performance measure (or discrimination). Then these expected values are compared to each other across the various classes (or customers) by some summary statistics function to yield a measure of system fairness. Thus, the entities that are compared to each other are the expected performance measures (or discriminations) rather than their actual performance measures.

To illustrate the difference between these approaches, consider any "pie division" problem e.g., a bonus b divided by an employer among n equally deserving employees. The first approach considers the *actual* bonuses, $\{b_1, b_2, \ldots, b_n : b_1 + b_2 + \cdots + b_n = b\}$ given to the employees, compares them to each other, and then uses a summary statistic to summarize them. Because all employees are equally deserving, the *absolutely fair* slicing of the pie is into equal shares (viz. $b_1 = b_2 = \cdots = b_n = b/n$). Then the discrimination (positive or negative) of employee i is expressible as $(b_i - b/n)$ (viz. the deviation from absolute fairness). We note that this is a zero-sum situation; if one employee gets more, it is taken away from other employees. The unfairness of the scenario can conveniently be represented by $\sum_{i=1}^{n} (b_i - b/n)^2/n$. If the employer decides to use a probabilistic mechanism for slicing the pie, with random variables bonuses $B_1, B_2, \ldots, B_n$ summing to b, the overall system's unfairness is given by $\sum_{i=1}^{n} E[(B_i - b/n)^2]/n$.

In the second approach (fairness of the mean), the slicing of the pie (the bonus distribution) is absolutely fair if $\mathrm{E}(B_i) = b/n$ for $i = 1, 2, \ldots, n$. If, for example, the employer uses an "all-or-none" lottery granting one of the employees all the money b, and all others get nothing, it is considered as fair as deterministically splitting the money evenly among the employees, providing that the lottery gives even odds to all

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

employees. The approach can also yield a measure of unfairness in the mean using the expression $\sum_{i=1}^{n}[E(B_i) - b/n]^2$.

2.3. Intuitive Appeal and Rationality

Producing intuitively acceptable results is a highly important, perhaps the most important, property expected of a fairness measure. Surprising results, whose disagreement with intuition cannot be rationally and convincingly explained, are most likely to be rejected. A measure producing such "surprises" is not likely to achieve wide acceptance and might be viewed, instead, as an interesting curiosity. In this subsection we present two, intuitively based, simple tests for the validity of a measure. These do not suffice to label a measure as valid, rather, not passing them is a red light indicating that the measure is questionable. As our goal is to focus on the fairness of the queue and neutralize other external parameters, a customer is assumed to be distinguishable from others only by its arrival epoch and service time requirement.

For convenience of presentation, we use the terms *seniority*, and *service requirement*. The seniority of J_i at epoch t is given by $t - a_i$ and the service requirement of J_i is s_i. One may recall that *seniority* and *service requirement* were in the heart of the dilemma in the Short-versus-Long scenario.

It is natural to expect that a "fair" scheduling discipline will give preferential service to highly senior jobs and to low service-requirement jobs. We can say that a schedule policy adheres to the *seniority preference principle* if for every two jobs whose service time is identical it always prioritizes the more senior one, and it adheres to the *service requirement preference principle* if for every two jobs whose arrival time is identical, it always prioritizes the one with the smaller service requirement. Examining common scheduling policies, we might observe that (1) FIFO adheres to the seniority preference principle and does not adhere to the service requirement preference principle, (2) LIFO, random order of service (ROS), and LJF adhere to neither of the principles, and (3) shortest job first (SJF) and shortest remaining processing time (SRPT) adhere to the service requirement preference principle and do not to the seniority preference principle. Of all these policies, only the processor sharing (PS) scheduling adheres to both principles. Not surprising then, as we will see in the sequel, PS will play a major role in defining queue fairness.

If one accepts these "principles," one might expect a fairness measure to follow them and to associate higher fairness values with schedules that give such preferential service compared to schedules that do not. This can be stated formally in the following two measure tests:

1. *Strong service-requirement preference test:* Consider jobs J_i and J_j, arriving at $a_i = a_j$ and obeying $s_i < s_j$. Let π be a scheduling policy where the service of J_i is completed before that of J_j and π' be identical to π, except for exchanging the service schedule of J_i and J_j. A fairness measure is said to satisfy the strong service requirement preference test if the fairness value it associates with π is higher than that it associates with π'.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

2. *Strong seniority preference test:* Consider jobs J_i and J_j, obeying $s_i = s_j$ and $a_i < a_j$. Let π and π' be scheduling policies as in test 1. A fairness measure is said to satisfy the strong seniority preference test if the fairness value it associates with π is higher than that it associates with π'.

One might view these two preference tests as two axioms expressing one's basic belief in queue fairness. It should be noted that when $a_i < a_j$ and $s_i > s_j$ (the Short-versus-Long case) the principles of giving preference to more senior jobs and to shorter jobs conflict with each other and, thus, the relative fairness of the possible scheduling of J_i and J_j is likely to depend on the relative values of the parameters.

One should note that if a measure satisfies the strong seniority test then for systems with *deterministic* service times FIFO and LIFO are the most fair and most unfair policies, respectively.

A *weak service requirement preference test* can be defined similarly to the strong one, where the requirement that $a_i = a_j$ is replaced by a requirement that the arrival times of all jobs present are *identical.* In a similar manner, a *weak seniority preference test* can be defined.

3. REVIEW OF PROPOSED FAIRNESS MEASURES AND THEIR PROPERTIES

Analytic treatment and quantification of queue fairness have been quite limited in the literature and have been addressed only very recently. The modeling dilemma of seniority versus service requirement seems to be at the heart of these queue fairness-modeling attempts: The approaches proposed in Gordon [7] and in Avi-Itzhak and Levy [2] center on the *seniority* factor. In contrast, the approach proposed by Wierman and Harchol-Balter [29], focuses on the *service requirement* factor. Sandman [26] proposed considering both *seniority* and *service requirement*, and, finally, Raz, Levy, and Avi-Itzhak [23] focused on neither of them and chose to focus on fair *resource allocation*, directly conforming to the general conception of social justice. In this section we review these publications and examine their properties in light of the discussion of expected properties given in Section 2. Due to space considerations, this review is short. A more thorough review can be found in Avi-Itzhak, Levy, and Raz [22].

3.1. Seniority-Based Fairness: Order Fairness

3.1.1. Skips and Slips: An Approach for Fairness Evaluation. This approach for evaluating fairness based on seniority was proposed by Gordon [7]. It aims at quantifying the violation of social justice due to overtaking in the queue. The underlying rationale is that FIFO is just and customer overtaking causes injustice.

The approach defines two types of overtaking events experienced by a tagged customer: (1) A *skip*, when the tagged customer overtakes another customer (viz. it

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

completes service before a customer that arrived ahead of it), and (2) a *slip*, when the tagged customer is overtaken by another customer. Gordon [7] suggested that counting the number of skips and slips can provide an indication of the amount of injustice and analyzes these counts; nonetheless, it does not deal with how to use these as the basis for a fairness measure.

Several systems are studied, including the following: two $M/M/1$ systems in parallel, two $M/M/1$ systems in parallel, where the tagged customer (and only that customer) uses the "join the shortest queue" strategy, the multiserver system $M/M/m$, and the infinite-server system $M/G/\infty$. For these systems the probability laws of the number of skips and the number of slips experienced by an arbitrary customer (denoted N_{SKIPS} and N_{SLIPS}, respectively) are derived. Interesting results are (1) For every system $E(N_{\text{SLIPS}}) = E(N_{\text{SLIPS}})$, (2) for most systems, the distributions of the two variables differ from each other, (3) only one system is found by the author where the distributions equal each other, the $M/G/\infty$ system where the service time distribution is symmetric around its mean, and (4) using the "join the shortest queue" strategy by the tagged customer, when no one else uses it, reduces the number of slips and increases the number of skips he/she experiences.

Reviewing this measure with respect to the properties discussed in Section 2 yields the following:

1. **Conformity:** The underlying principle is pragmatic; namely seniority merits priority. Nonetheless, the approach is not fully sensitive to the extent of seniority differences between customers, as it assigns equal weight to all skips (and slips), regardless of the relative seniority of the involved customers: If customers C_j and C_{j+1} are interchanged, one skip and one slip will be counted regardless of whether C_{j+1} arrives 1 s or 1 h behind. Also, it does not consider service requirements at all; thus, it might apply mainly in systems where seniority is the most important factor, e.g., identical deterministic service times (and possibly exhaustible-servers systems). In the case of nonequal service times, the conformity of this principle to the conception of social justice at large might be questioned.
2. **Granularity:** Accounting for the number of skips and slips can be done at all three granularity levels; individual, scenario, and system. How fairness at the system level can be measured remains open in the work of Gordon [7]. In light of the fact that $E(N_{\text{SKIPS}} - N_{\text{SKIPS}}) = 0$, a possible approach that comes to mind is to take either $E(N_{\text{SKIPS}})$ or $\text{Var}(N_{\text{SKIPS}} - N_{\text{SKIPS}})$ as system fairness metrics. How such measures behave and how they relate to the measure developed in Section 3.1.2 is an open question.
3. **Intuitive appeal and rationality:** The approach is strongly intuitively appealing, as long as only seniority matters. Because no system measure was proposed or studied, the question of whether it satisfies the basic tests is not meaningful.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

3.1.2. A Seniority-Based Fairness Measure. An order fairness measure, based on seniority, was studied in Avi-Itzhak and Levy [1]. The basic underlying model used in that study assumes that all service times are identical. In that context, the major factor of interest is that of job seniority. The study deals with a specific sample path of the system and examines a realization π of the service order (i.e., a feasible sequence of job indexes reflecting the order of service) and with a fairness measure $F(\pi)$ defined on the service order. The article assumes several elementary axioms on the properties of $F(\pi)$. The major axiom is the following:

> ***Monotonicity of F( ) under neighbor jobs interchange*:** If two neighboring jobs are interchanged to modify π and yield a new service order π' then $F(\,)$ increases if the interchange advances the more senior job ahead of the less senior job and decreases if the interchange advances the less senior job ahead of the more senior job. The increase in $F(\,)$ is monotone in the seniority difference and is zero if the jobs are equally senior.

The additional axioms deal with *Reversibility of the interchange*, *Independence on position and time,* and *Fairness change is unaffected by jobs not interchanged.* The results derived show that for a specific sample path, the quantity $c\sum_i a_i\Delta_i + \alpha$, where Δ_i is the *order displacement* of customer C_i (number of positions C_i is pushed ahead or backward on the schedule, compared to FIFO) and where $c > 0$ and α are arbitrary constants, satisfies the basic axioms. This quantity is the unique form satisfying the axioms applied to any feasible interchange (not necessarily of neighbors). Under steady state, this quantity is equivalent to the *variance* of the *waiting time* (with a negative sign), up to a constant multiplier. The *waiting time variance* can thus serve as a surrogate for the *system's unfairness measure*.

Reviewing this measure with respect to the properties discussed in Section 2 yields the following:

1. **Conformity:** The underlying principle is pragmatic (viz., seniority merits priority). In the case of equal service times, for which it is proposed, it can be considered to conform to the basic conception of social justice. One possible way of extending this concept to the nonconstant service times situation is to require that the waiting disutility be divided in proportion to the slice of the resource received by each customer.
2. **Granularity:** The measure is defined, and is applicable, at all three granularity levels; the individual customer level, the scenario level, and the system level.
3. **Intuitive appeal and rationality:** The measure is intuitively appealing when all service times are the same since the Short-versus-Long conflict does not exist. Under this measure, FIFO is the least unfair and LIFO is the most unfair. The measure satisfies the strong *seniority preference test.* The strong *service requirement test* is not applicable. If used for nonequal service times, it still satisfies the strong *seniority preference test* but does not satisfy the *service requirement preference test.* An extension of this measure's approach to nonequal service times is discussed by the authors.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

3.2. An Expected-Slowdown, Service-Requirement-Based, Fairness Criterion

Slowdown is defined in computer-related queuing publications as the conditional response time divided by the conditioning service length: $S(x) \stackrel{\text{def}}{=} T(x)/x$, where $T(x)$ is the response time experienced by a customer whose required service time is of size x. Wierman and Harchol-Balter [29] proposed an expected slowdown criterion for classifying $M/G/1$ disciplines into three classes (based on earlier work on slowdown presented in Bender, Chakrabarti, and Muthukrishnan [4], Bansal and Harchol-Balter [3], and Harchol-Balter, Sigman, and Wierman (2002)). The reason for looking at the slowdown metric is that it makes intuitive sense that the mean response time experienced by a user should be proportional to the service requirement of the job submitted by the user. The PS policy was observed by Kleinrock [10] as being "fair" since, under $M/G/1$-PS, all jobs experience the same mean slowdown: $E(S(x)) = 1/(1-\rho)$, where $\rho < 1$ is the system's load. It turns out, as proven in Wierman and Harchol-Balter [29], that the criterion of $1/(1-\rho)$ is "tight" in the sense that no scheduling policy can achieve a same constant slowdown value lower than $1/(1-\rho)$. The aim of this approach is to use the PS policy as the standard of fairness and to ask which other scheduling policies achieve this standard as follows:

- A scheduling policy is said to be fair for a given load and service distribution if $E(S(x)) \leq 1/(1-\rho)$ for all values of x.
- A service policy is *always fair* if it is fair under all loads and all service distributions. A service policy is *always unfair* if it is not fair under all loads and all service distributions. Other policies are *sometimes unfair,* meaning fair under some loads and distributions and unfair under others.

Their work analyzed a wide set of scheduling disciplines under the $M/G/1$ setup and classified them into these three classes.

Although this is a criterion, in contrast to a measure that assigns a numerical value to each $M/G/1$ discipline, we include it here because it raises interest in the computer science queuing-related community and it resembles the pragmatic principle of waiting and service proportionality, mentioned in Section 1.2.

Reviewing this criterion with respect to the properties discussed in Section 2 yields the following:

1. **Conformity:** The criterion is based on a pragmatic conceptual principle that fairness is violated whenever the conditional mean response time exceeds the one obtained in PS. This leads to a *fairness-of-the-mean* criterion, which focuses directly on job size and indirectly, to a lesser extent, on seniority. We note that the criterion does not strive for proximity (or equality) of the expected slowdowns for all values of x—only that they all be bounded by $E(S(x)) \leq 1/(1-\rho)$.

   An additional rationale for the use of the expected slowdown as a fairness criterion was offered to us in personal communications. In systems where the

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

customer does not see other customers (such as in many computer systems), the customer can view his response time only relative to his service requirement, not relative to other customers concurrently served with him in the system. Note, however, that to adopt this rationale the customer must, somehow, be able to relate to his *expected* response time and must not care about how the blind queue internally schedules jobs.

2. **Granularity:** The criterion is based on expected values. As such, it yields to *system fairness* analysis for a relatively wide class of disciplines for the $M/G/1$ system. It is limited to steady-state systems and is not applicable to classifying unfairness to individuals or unfairness of a scenario. Also, as a criterion and not a measure, it does not rank the relative fairness rankings of individual disciplines, in case they belong to the same class. For example, consider the class of "always fair" policies, for which Wierman and Harchol-Balter [29] raised the open question of whether it contains more policies in addition to PS and LIFO-PR. As discussed in Avi-Itzhak et al. [2], this class is indefinitely large (containing the class of symmetric queues) and one possible way of distinguishing fairness among its policies is by considering the variance of the conditional waiting time, $\mathrm{Var}(W(x))$, for each policy; further, the value of this variance can vary drastically (up to relative difference of infinity) across the various "always fair" policies.
3. **Intuitive appeal and rationality:** The criterion classifies as *always unfair* all conservative policies that are (1) nonpreemptive non-size-based (e.g., FIFO, LIFO, and ROS), (2) preemptive size-based (e.g., preemptive shortest service first), or (3) age-based (age of a job is defined as service already received). It classifies as *sometimes unfair* all conservative policies that are (1) nonpreemptive size based (e.g., shortest service first), and (2) preemptive shortest remaining processing (all other remaining processing based are either *always unfair* or *sometimes unfair*). As stated earlier, it classifies as *always fair* all symmetric queues (including LIFO-PR and PS). The tests of fairness defined in Section 2.3 do not apply, since the criterion does not have the necessary granularity.

If one focuses primarily on service times, the results are intuitively appealing—for example, that nonpreemptive policies are always unfair (due to large jobs blocking short jobs) and LIFO-PR and PS are always fair. The criterion is likely to have intuitive appeal mainly in situations where seniority is unobservable and customers are frequent repeaters, perhaps in some computer systems where users do not see the other jobs. In many daily life queuing applications where seniority is observable, the results of the criterion (LIFO-PR is *always fair*, whereas FIFO is *always unfair*, even for $M/D/1$) are likely not to appeal to common humans.

3.3. A Service Requirement and Seniority-Based Fairness

A fairness approach based on accounting both for service requirement and seniority (as a combination) is offered in Sandmann [26]. The approach aims at counting events

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

of fairness violation, adding "size" violation events to the "order" violation events of Gordon [7]. Specifically, a tagged customer C can be subject to two types of "discriminating" events: (1) an *overtaking event*, in which C is overtaken by another customer (identical to the slip event of Section 3.1.1) and (2) a *large job event*, occurring if upon the arrival of C to the system, it finds there C' whose residual service is greater than or equal to the service requirement of C and (later) C' departs from the system ahead of or concurrently with C. Let N_i^{over} and N_i^{large} be the numbers of these events, experienced by customer $C_i, i = 1, 2, \ldots$, respectively. The discrimination frequency[2] of C_i is defined to be $\text{DF}_i = N_i^{\text{over}} + N_i^{\text{large}}$, and the discrimination frequency of a sample path π is defined as $\text{DF}(\pi) = \sum_i \text{DF}_i$. Let N^{over} and N^{large} be the number of discrimination events experienced by an arbitrary customer when the system is in steady state. The system unfairness under steady state is defined as $E(N^{\text{over}} + N^{\text{large}})$.

Sandmann [26] showed that this value satisfies both strong preference tests. The question of how to derive the expected discrimination of an arbitrary customer is not addressed. To yield this measure, note that $E(N^{\text{over}})$ can be taken from the analysis in Section 3.1.1; the analysis of $E(N^{\text{large}})$ remains as an open issue for research.

Reviewing this measure with respect to the properties discussed in Section 2 yields the following:

1. **Conformity:** The underlying principle is pragmatic; namely seniority and smaller residual service time merit priority. The "overtaking" events are, nonetheless, not fully sensitive to seniority differences (see the remark regarding skip events in Section 3.1.1, Conformity) and the "large" events are not fully sensitive to differences in the remaining service requirement. Also, how to weight overtaking events versus large events is an open question (Sandmann [26] uses a relative weight of 1).
2. **Granularity:** The measure applies at all levels of granularity. These are given by DF_i for the individual customer, $\text{DF}(\pi)$ for a sample path, and the expected value of the DF measure taken for an arbitrary customer in equilibrium.
3. **Intuitive appeal and rationality:** The measure satisfies *both* the strong *seniority preference test* and the strong *service requirement preference test*. It is worth noting that this is the only measure, of the measures reviewed here, that satisfies both strong tests.

3.4. A Resource Allocation-Based Fairness

A resource allocation queuing fairness measure (RAQFM) was introduced in Raz et al. [23]. The measure is directly based on the general conception of justice requiring equal distribution of resources among all equally needy members. The measure accounts, indirectly, for both seniority and service requirements, thus offering a solution to the Short-versus-Long conflict, based on a basic principle of social justice. The method applies to multiple servers (Raz et al. [22]), but for the sake of presentation, we will focus mostly on the single-server system.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

The basic underlying principle is that at any moment in time all customers present are entitled to equal shares of the resource; namely at every epoch t at which there are $N(t)$ jobs (customers) present in the system, each is entitled to $1/N(t)$ of the server's serving rate. This is called the temporal *warranted service rate* to be given to each customer at that epoch. The overall warranted service of C_i (customer i) is given by integrating this value over the duration that C_i stays in the system. Subtracting this warranted service from the granted service (which is its service time s_i) yields the *discrimination* of C_i, denoted $D_i = s_i - \int_{a_i}^{d_i}(1/N(t))dt$, where d_i is the departure epoch of C_i. Note that the discrimination might be positive or negative. Taking summary statistics over all discriminations experienced by the customers in a scenario yields an unfairness measure for the scenario. Taking expectations over all scenarios yields the system unfairness. One of the basic properties of the discrimination function is that it is a zero-sum function (viz. the total discrimination in the system, at every epoch, is zero). Thus, the expected value of discrimination is meaningless, and the proper summary statistics is the second moment[3] (or variance) or expected absolute value of discrimination. Note that the definition of this measure results in PS being the least unfair policy in the $G/G/1$ case.

The measure yields to exact analysis (via numerical procedures) for the family of multiple-server Markovian (including M/phase-type/m type and, in particular, M/Coxian/m type) queues. It is an open subject for research whether (and how) it yields to mathematical analysis for $M/G/1$-type systems and to systems with arbitrary service times, at large. It does yield, for example, to exact mathematical analysis of the $M/G/1$ LIFO-preemptive queue (Brosh, Levy, and Avi-Itzhak [5]).

Reviewing this measure with respect to the properties discussed in Section 2 yields the following:

1. **Conformity:** The underlying principle of the measure conforms directly to the basic conception of social justice.
2. **Granularity:** The measure is defined, and is applicable, at all three granularity levels; the individual customer level, the scenario level, and the system level.
3. **Intuitive appeal and rationality:** This measure is not based on a pragmatic intuitive principle, but rather on a general conception of social justice. As such, it might not be intuitively appealing at a first glance. Therefore, an extensive examination of its properties under various systems and conditions, and their agreement with intuitive appeal, is required. Next, we review only some of these properties (a more thorough review can be found in Avi-Itzhak et al. [2]; see also the analysis in Raz, Levy, and Avi-Itzhak [23,24] and Brosh et al. [5]). Unless otherwise stated, the properties are phrased for single-server systems. The properties show very good agreement with intuition:
    1. The measure satisfies the *strong seniority preference test* for work conserving and nonpreemptive service policies (Raz et al. [24]). It also satisfies the *weak service requirement preference test* for such policies. Nonetheless, it does not satisfy the strong version of this test.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

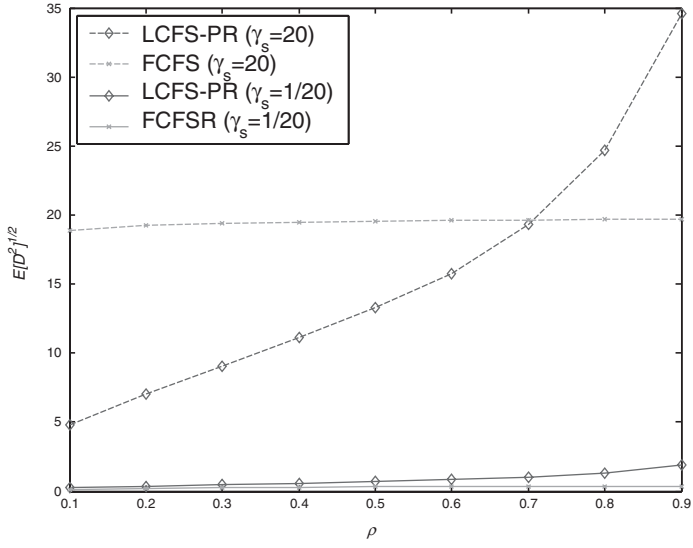

FIGURE 2. RAQFM: unfairness for high ($\gamma_s = 20$) and low ($\gamma_s = 1/20$) service time variability in an $M/G/1$ system.

2. An interesting question is which of the two "extreme policies," LIFO-PR and FIFO, is fairer. Although the seniority-based approaches and the service-requirement-based criterion have opposite views on this issue, the resource allocation approach bridges this gap: When service times are of small variability, the measure ranks FIFO as fairer than LIFO-PR, in agreement with the seniority-based measures; the intuition is that size differences are of small importance due to the small variability, and seniority is what matters. When service times are of high variability (and load is not too small), it ranks LIFO-PR as fairer than FIFO, in agreement with the size-based criterion; the intuition is that size now does matter due to the existence of large and small jobs. This is depicted in Figure 2, in which the square root of the unfairness measure is plotted versus the load.
3. For multiple-server systems, the measure evaluates (in the case of exponential or deterministic service times) the common (single) FIFO queue more fair than the equivalent multiqueue. These are in agreement with the experimental findings of Rafaeli et al. [20] and those mentioned by Larson [11].

4. DISCUSSION: MEASURE APPLICABILITY AND FUTURE RESEARCH

Attempting to measure fairness in queues calls for fresh new approaches, deviating from the traditional ones of queuing theory, focusing—almost exclusively—on

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

efficiency. This area is young, but gradually gaining recognition, and the importance of managing fairness *and* efficiency (as well as the balance between them) becomes more evident. As such, it is a highly challenging, but also promising, area for researchers and practitioners alike. In this article we exposed the existing, mostly very recent, research works on quantifying fairness of queues. In what follows, we indicate several possible directions of future research. We recognize our inherent bias due to personal research involvement in this area and hope that fresh minds will generate new ideas and approaches, far beyond ours.

When trying to compare the measures and approaches suggested thus far, the question of the degree of universal applicability comes to the surface—that is, How wide is the range of systems each measure applies to, either in its original form or via a generalization. Wide applicability is one of the most important requirements of a measure, since if not applicable to many systems, it might not be useful as a scale of reference. The seniority-based fairness measure presented in Section 3.1.2 applies only to the case in which service times are equal or to the case in which the servers are exhaustible and the major performance factor is getting the service. The approach presented in Section 3.1.1 is similar because it focuses on seniority and does not account explicitly for service requirements. The service requirement (expected slowdown)-based criterion of Section 3.2 focuses on service requirement and disregards seniority and, thus, is applicable mainly to systems where seniority is unobservable or not important—most likely computer systems (if disregarding seniority becomes acceptable). In both cases, the degree of universality is quite limited. The service requirement and seniority combination (SSCF) measure (Section 3.3), which accounts for both seniority and service requirements, and the RAQFM (Section 3.4), which reacts to both of these factors, are more universally applicable.

An important question is to what models can the measures be generalized and how. For example, consider a system, proposed by one of the referees, with two servers of different processing rates, μ_1 and μ_2, to be called System A; compare it to System B, with two servers and where the service rate of each server is $\mu = (\mu_1 + \mu_2)/2$. How will the different measures treat these systems? Which of these systems is fairer according to those measures that are applicable? (See Avi-Itzhak et al. [2] for an extended discussion of this problem.)

Extending fairness analysis to more general queuing systems is valuable because practical queuing applications where fair operation is important are often more complex than the single-server system. Evaluating the fairness of such applications or situations might be quite challenging. In fact, queuing operational issues that are relatively simple from performance perspectives (e.g., mean delay) might be more complex (both conceptually and analytically) from fairness perspectives. An example of such a challenging question is how fair are systems where the server might go idle (e.g., consider the supermarket cashier going on a coffee break while you are waiting in the queue)? Other challenging queue settings include (1) multiserver, multiqueue systems (for which initial work was done using the RAQFM measure), (2) queues with reneging and balking, and (3) general queuing networks (e.g., queues in tandem).

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

To provide a practical example of the latter issue, consider airport systems, in which travelers first wait in a luggage security line, then in a (multi-queue) check-in line, and then in another security line. In studying a fairness measure for such settings, it is important that it will be mathematically tractable (to a reasonable extent) to afford exact analysis of the measure or that it will at least be computationally feasible. It is also important to examine whether the results produced fit the basic intuition.

Another worthwhile direction is developing a measure of fairness based on the proportionality principle, namely the *waiting time of a job should be in proportion to the servers' time provided to it*. It is intuitively appealing to require that customers who get more will also wait more. This idea was addressed in Section 1.2.2 and discussed with regard to proportionality of the mean waiting time (the slowdown criterion) in Section 3.2. We propose a general fairness measure based on individual discrimination of C_i, being defined as $W_i - cx_i$, where W_i and x_i are the waiting time and service requirement of C_i, respectively, and c is a constant that might vary from one system to another. This approach is presently under study by the authors.

One more interesting problem relating to the existing measures is how to determine the trade-off between the total weights of the two event types of the SSCF proposed measure (Section 3.3). This issue involves directly resolving the fundamental dilemma of Short versus Long. The RAQFM and the proportionality measure indicated earlier resolve this dilemma indirectly, by trying to adhere to an underlying basic fairness principle. No direct approach has been suggested so far.

The issue of how to account for different values of neediness remains open. Within this context, the waiting line for organ donations, in which fairness might be of utmost importance, might require a completely different approach than those described in this work. Similarly, the issues of how to account for economical factors and of how to combine queue pricing with fairness remain open. The issue of pricing/admission and scheduling received much attention in the queuing literature; a recent book (Hassin and Haviv [8]) is dedicated to this subject.

The question of how to account for both fairness and mean delay when analyzing the performance of a queuing system remains open as well. To this end, we note that these two measures can be in opposition to each other and lead to conflicting operational rules. To demonstrate this, recall our comment from Section 1.2 that the PS service discipline, which is the fairest by some of the measures, is non-Pareto with respect to waiting times, since in many situations all customers get slowed down and everyone loses. For example, in $M/D/1$-PS, everyone, except for the last customer of a busy period, stays longer in the system; thus, in a repeated situation, customers might prefer FCFS because, on average, their time in the system is lower (fairness in this situation is like communism: social justice means everyone is poor). A utility function that accounts for both the mean and the variance of the delay, like the Markowitz mean/variance utility function [13] in portfolio theory, might be useful here.

Finally, and perhaps above all, experimental studies and publication of actual case studies, especially collaborated work with social science researchers, will advance the area of queue fairness significantly.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

Acknowledgments

We thank the anonymous reviewers and the editors of this article for their many insightful and valuable comments. This work was supported in part by the Israeli Ministry of Science and by EURO-NGI.

Notes

1. Nussbaum [15] described Rawls as "the most distinguished moral and political philosopher of our age."
2. The term "discrimination frequency" is used in Sandmann [26]. The term "discrimination count" might be more appropriate in this context.
3. In this case, the units of the measure are time squared (such as delay variance). One can take the square root of it to make the units equivalent to those of mean delay.

References

1. Avi-Itzhak, B. & Levy, H. (2004). On measuring fairness in queues. *Advances of Applied Probability* 36(3): 919–936.
2. Avi-Itzhak, B., Levy, H., & Raz, D. (2005). Quantifying fairness in queueing systems: Principles, approaches and Applicability, Technical report RRR-25-2005, RUTCOR, Rutgers University, New Brunswick, NJ. Available from http://rutcor.rutgers.edu/pub/rrr/reports2005/25_2005.pdf.
3. Bansal, N. & Harchol-Balter, M. (2001). Analysis of SRPT scheduling: Investigating unfairness. In *Proceedings of ACM Sigmetrics 2001 Conference on Measurement and Modeling of Computer Systems*, pp. 279–290.
4. Bender, M., Chakrabarti, S., & Muthukrishnan, S. (1998). Flow and stretch metrics for scheduling continuous job streams. In *Proceedings of the 9th Annual ACMSIAM Symposium on Discrete Algorithms*, pp. 270–279.
5. Brosh, E., Levy, H., & Avi-Itzhak, B. (2005). The effect of service time variability on queue Fairness. Technical Report RRR 24-2005, RUTCOR, Rutgers University, New Brunswick, NJ.
6. Chan, M.F. & Tien, J.M. (1981). An alternative approach to police response. Wilmington Management of Demand Program, National Institute of Justice, Washington DC.
7. Gordon, E.S. (1987). New problems in queues: Social injustice and server production management, Ph. D. dissertation, MIT, Boston, MA.
8. Hassin, R. & Haviv, M. (2002). *To queue or not to queue, equilibrium behavior in queueing systems*. Boston: Kluwer Academic Publishers.
9. Kingman, J.F.C. (1962). The effect of queue discipline on waiting time variance, *Proceedings of the Cambridge Philosophical Society*, 58: 163–164.
10. Kleinrock, L. (1976). *Queueing systems* Vol II: *Computer applications*. New York: Wiley, 1976.
11. Larson, R.C. (1987). Perspective on queues: Social justice and the psychology of queueing, *Operations Research* 35(6): 895–905.
12. Mann, I. (1969). Queue culture: The waiting line as a social system, *American Journal of Sociology* 75: 340–354.
13. Markowitz, H.M. (1991). *Portfolio selection*, 2nd ed. Boston: Blackwell.
14. McEwen, J.T., Connors, E.F., & Cohen, M.I. (1984). *Evaluation of the differential police response field test*. Alexandria, VA: Research Management Associates, Inc.
15. Nussbaum, M. (2001). The enduring significance of John Rawls, the Chronicle of Higher Education. *The Chronicle Review*, July 20, 2001.
16. Palm, C. (1953). Methods of judging the annoyance caused by congestion, *TELE* 4: 189–108.
17. Piccard, D. (2005). Outline of an extended book review, *Stanford Encyclopedia of Philosophy*. Available from http://oak.cats.ohiou.edu/~piccard/entropy/rawls.html (accessed January 2005).

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302

18. Pla, V., Casares-Giner, V., & Martinez, J. (2004). On a multiserver finite buffer queue with impatient customers. In *Proceedings of 16th ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile System*.
19. Rafaeli, A., Barron, G., & Haber, K. (2002). The effects of queue structure on attitudes. *Journal of Service Research* 5(2): 125–139.
20. Rafaeli, A., Kedmi, E., Vashdi, D., & Barron, G. (2005). Queues and fairness: A multiple study investigation. Technical report, Technion—Israel Institute of Technology.
21. Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press; revised edition published 1999.
22. Raz, D., Avi-Itzhak, B., & Levy, H. (2005). Fairness considerations of scheduling in multi-server and multi-queue systems, RUTCOR Technical report RRR-11-2005, Rutgers University, New Brunswick, NJ.
23. Raz, D., Levy, H., & Avi-Itzhak, B. (2004). A resource-allocation queueing fairness measure. In *Proceedings of Sigmetrics 2004; Performance Evaluation Review* 32(1): 130–141.
24. Avi-Itzhak, B., Levy, H., & Raz, D. (2004). A resource allocation queueing fairness measure: Propertius and bounds. *Queueing Systems Theory and Application* 56: 65–71.
25. Rothkopf, M.H. & Rech, P. (1987). Perspectives on queues: Combining queues is not always beneficial. *Operations Research* 35: 6.
26. Sandmann, W. (2005). A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement. In *Proceedings of the 1st Euro NGI Conference on Next Generation Internet Networks Traffic Engineering*.
27. Zalta, E.N. (Ed.) (2005). *Stanford Encyclopedia of Philosophy*. Available from http://plato.stanford.edu/contents.html.
28. Whitt, W. (1984). The amount of overtaking in a network of queues. *Networks* 14(3): 411–426.
29. Wierman, A. & Harchol-Balter, M. (2003). Classifying scheduling policies with respect to unfairness in an $M/GI/1$. In *Proceedings of the ACM Sigmetrics 2003 Conference on Measurement and Modeling of Computer Systems*.

Downloaded from https://www.cambridge.org/core. University of Basel Library, on 30 May 2017 at 14:31:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0269964808000302