Submitted to the Annals of Applied Probability

STRUCTURAL PROPERTIES OF PROPORTIONAL FAIRNESS: STABILITY AND INSENSITIVITY

By Laurent Massoulié

Microsoft Research, Cambridge, U.K.

In this note we provide a novel characterization of the proportionally fair bandwidth allocation of network capacities, in terms of the Legendre-Fenchel transform of the network capacity region. We use this characterization to prove stability of network dynamics under proportionally fair sharing, by exhibiting a suitable Lyapunov function. Our stability result extends previously known results to a more general model including Markovian users routing. In particular, it implies that the stability condition previously known under exponential service time distributions remains valid under so-called phase-type service time distributions.

We then exhibit a modification of proportional fairness, which coincides with it in some asymptotic sense, is reversible (and thus insensitive), and has explicit stationary distribution. Finally we show that the stationary distributions under modified proportional fairness and balanced fairness, a sharing criterion proposed because of its insensitivity properties, satisfy the same large deviations characteristics.

These results give support to the choice of proportional fairness as a default bandwidth allocation criterion, combining the desirable properties of ease of implementation with performance and insensitivity.

1. Introduction. The abstract network bandwidth allocation (NBA) problem can be formulated as follows. A network supports connections of distinct types, indexed by r, the index r spanning the set of types $\mathcal{R}$, assumed finite. Given the number x_r of users of each type $r \in \mathcal{R}$, with $x_r \in \mathbb{N}$, the problem is to determine the total capacity allocated to type r users, denoted be λ_r, with $\lambda_r \in \mathbb{R}_+$. The allocation vector $\lambda := \{\lambda_r\}_{r\in\mathcal{R}}$ is constrained to lie in a set $\mathcal{C} \subset \mathbb{R}_+^{|\mathcal{R}|}$.

The set $\mathcal{C}$ is a suitable abstraction of all the physical capacity constraints of the actual network under consideration. For specific descriptions of the sets $\mathcal{C}$ corresponding to e.g. wired networks with fixed routing –such as the Internet–, or overlay networks with multipath routing, or wireless networks with fixed routing and interferences, we address the reader to references [4] and [14].

In the present note we only require the set $\mathcal{C}$ to be convex and non-increasing, two assumptions that are met in all the applications above mentioned.

Mo and Walrand [16] introduced the following criterion for determining the allocation vector λ. Given weights $w_r > 0$, and a parameter $\alpha \geq 0$, the so-called (w, α)-fair allocation vector is the solution to the optimisation problem

$$\max_{\lambda \in \mathcal{C}} \sum_{r \in \mathcal{R}} x_r U_r(\lambda_r / x_r), \tag{1}$$

where

$$U_r(y) = \begin{cases} \frac{w_r}{1-\alpha} y^{1-\alpha} & \text{if } \alpha \neq 1, \\ w_r \log(y) & \text{if } \alpha = 1. \end{cases} \tag{2}$$

This parametric family of allocation criteria contains the so-called proportional fairness criterion, introduced by Kelly [11], which corresponds to the special case $\alpha = 1$ and $w_r \equiv 1$. In the limit $\alpha \to \infty$, the (w, α)-fair allocation coincides with the so-called max-min fair allocation (see [2] for a definition).

The rationale for proportional fairness, as explained in [11], lies in the following desirable decomposition property. Assume that the ultimate goal of bandwidth allocation is to maximise the sum of utility functions, U_r, of the rates λ_r / x_r allocated to users of class r, exactly as in Equation (1), these utility functions being known to the users but not to the network. Then the decomposition result of [11] states that this can be done by letting on the one hand the network allocate bandwidth according to proportional fairness, with weights w_r specified by the network users, and on the other hand the network users selecting these weights w_r appropriately, given their (privately known) utility functions U_r, and the network allocation in response to distinct weights w_r.

Alternatively, the unweighted proportional fairness allocation can be justified as follows. The results of Stefanescu and Stefanescu [20] (see also Mazumdar et al. [15] for further discussion) imply that it is the only allocation of bandwidth that satisfies four natural axioms introduced by Nash [17][1], assuming that users' utility is a linear function of the rate they receive. It is in fact the natural extension of Nash's bargaining solution, originally derived in the special context of two users, to an arbitrary number of users.

The rationales for candidate NBA solutions we have just reviewed originate from microeconomic theory of utility, and game (bargaining) theory,

[1] namely, invariance with respect to affine utility transformations, Pareto optimality, independence of irrelevant alternatives, and symmetry.

and assume a static set of network users. There is another line of approach to the NBA problem, which is essentially motivated by performance issues in a dynamic setting.

Specifically, assume that network users arrive and leave the system, the arrivals of type r users being at the instants of a Poisson process of rate ν_r. Assume further that users remain in the system until they have transferred a file of a given size, files associated with type r users being exponentially distributed with parameter μ_r. The state variable $x = \{x_r\}_{r\in\mathcal{R}}$ is then a Markov process, with non-zero transition rates

$$\begin{array}{ll} x_r \to x_r + 1 & \text{with rate } \nu_r, \\ x_r \to x_r - 1 & \text{with rate } \mu_r\lambda_r. \end{array} \tag{3}$$

A suitable rationale for selecting a NBA is to guarantee desirable properties of the above Markov process. One such property is ergodicity, as it in turn implies that sojourn times of users are almost surely finite. Ergodicity cannot be guaranteed for all sets of traffic parameters ν_r, μ_r and network capacity sets $\mathcal{C}$. In particular, when the vector $\rho = \{\rho_r\}_{r\in\mathcal{R}}$ does not belong to the capacity set $\mathcal{C}$, the process cannot be ergodic (for a proof, see e.g. [4]). When ρ is on the boundary of $\mathcal{C}$, Kelly and Williams [13] have established that the process cannot be positive recurrent, for sets $\mathcal{C}$ corresponding to wired networks with fixed routes. Their proof extends to the case of general convex non-increasing capacity sets $\mathcal{C}$ with minor modifications.

A reasonable performance requirement is thus that, provided the traffic intensity vector ρ lies in the interior $\mathring{\mathcal{C}}$ of $\mathcal{C}$, then the above Markov process is ergodic. Such a property is in fact satisfied for all (w, α)-fair bandwidth allocation criteria, as follows from the Lyapunov function-based stability proof of Bonald and Massoulié [3] (see also de Veciana et al. [9] who first established the result for the case of max-min fairness, Ye [22] and Key and Massoulié [14] for an extension to more general utility functions U_r in the allocation definition (1)).

Thus, the requirement of achieving ergodicity for the largest possible set of traffic intensity vectors ρ, being met by all (w, α)-fair NBA, does not distinguish one such criterion as superior to the others.

A more stringent requirement has been suggested by Bonald and Proutière [5], namely that not only the stability region (defined to be the set of vectors ρ such that the system is ergodic) be maximal, but also that the corresponding Markov process be *insensitive* to the distribution of sizes of the files transferred by each class of users. Roughly speaking, insensitivity means that the stationary distribution of the numbers of users in the system is unaffected if the service time distributions are modified, provided their mean is left unchanged. For characterizations of insensitive systems, we refer the reader to Schassberger [19] and references therein. If insensitivity holds, the system remains ergodic under the natural stability condition for arbitrary phase-type (that is, mixtures of convolutions of exponential distributions; see e.g. Asmussen [1], p. 80), not necessarily exponential, service time distributions. Note that ergodicity under the natural stability conditions, and for general, non-exponential service time distributions, has so far been established for the max-min fairness NBA in a recent article of Bramson [6], but a similar result has been missing for all other (w, α) NBA's. Although restricted to max-min fairness, the results of Bramson apply under very weak integrability assumptions on the service time distributions, and are not restricted to phase-type distributions.

Bonald and Proutière have identified a new NBA, the so-called *balanced fairness allocation*, which meets the two requirements of maximal stability region and insensitivity, and moreover maximises the fraction of time during which the system is empty, among all allocations meeting these two requirements. They have furthermore identified special network topologies for which balanced fairness coincides with proportional fairness, and have shown that for all other network topologies, balanced fairness is distinct from any utility maximisation NBA.

This leaves several questions open regarding the choice of an NBA. On the one hand, utility maximisation allocations, such as (w, α)-fairness or more specifically proportional fairness, can be implemented in a distributed manner (see for instance the seminal paper by Kelly, Maulloo and Tan [12]), and are motivated by microeconomic theory and game theory arguments in a static setting. In addition, they satisfy the criterion of maximal stability region in the dynamic setting, but do not seem to meet the more stringent requirement of insensitivity. On the other hand, balanced fairness does meet the latter requirement, but no simple distributed technique for realizing this NBA is known, if we except the special network topologies, identified in [5], where it coincides with proportional fairness.

In the present note, we will provide a novel characterization of proportional fairness, and use it to improve upon this unsatisfactory state of affairs. Indeed, relying on this structural property, we will show that the seemingly fortuitous coincidence of balanced fairness and proportional fairness on specific network topologies in fact reflects a deeper relationship between the two NBA's, that holds for any network topology as captured by the set $\mathcal{C}$. More precisely, we will exhibit a third NBA, namely modified proportional fairness, which coincides in some asymptotic sense with proportional fairness. Under modified proportional fairness, the system is reversible, and hence insensitive. Furthermore, the steady state distributions under modified proportional fairness and balanced fairness admit the same large deviations characteristics, described by a simple explicit rate function.

As a by-product, we give a new proof of ergodicity of proportional fairness, which extends to a more general model of network dynamics including Markovian users routing. This in turn implies that the usual stability conditions still hold with service time distributions that are of phase type rather than exponential.

These results give support to the choice of proportional fairness as a default NBA. Indeed, it is motivated by (i) the decomposition property of [11], (ii) axiomatic arguments from bargaining theory [20], and (iii) as an implementable approximation to balanced fairness, meeting the additional criteria of performance and insensitivity.

The structure of the paper is as follows. Section 2 gives the novel characterization of proportional fairness. Stability properties with Markovian user routing are proven in Section 3. Section 4 establishes the relationships between balanced fairness and modified proportional fairness, and in particular the fact that the corresponding equilibrium distributions have the same large deviations characteristics. Conclusions are drawn in Section 5.

2. Characterization of proportional fairness via convex duality.

It is convenient to consider the logarithms of the allocated capacities λ_r, rather than the λ_r themselves. Denote by K the subset of $\mathbb{R}^{|\mathcal{R}|}$ in which these must lie, i.e.

$$\gamma = \{\gamma_r\} \in K \Leftrightarrow \lambda = \{\exp(\gamma_r)\} \in \mathcal{C}.$$

Given γ, γ' in K, and $\epsilon \in (0,1)$, by convexity of the exponential function, for all $r \in \mathcal{R}$, one has

$$\exp(\epsilon\gamma_r + (1-\epsilon)\gamma'_r) \leq \epsilon \exp(\gamma_r) + (1-\epsilon)\exp(\gamma'_r),$$

and thus since $\mathcal{C}$ is convex, non-increasing, then so is K. Denote by $\gamma^{PF}(x)$ the vector of logarithms of proportionally fair allocations, that necessarily belong to K.

Denote by δ_K the function that equals zero on K, and $+\infty$ outside of K. The original characterization of $\lambda^{PF}(x)$ as a maximizer of $\sum_{r\in\mathcal{R}} x_r \log(\lambda_r)$ over $\lambda \in \mathcal{C}$ readily implies that

$$\gamma^{PF}(x) \in \operatorname{argsup}_{\gamma\in\mathbb{R}^{\mathcal{R}}} \left(\langle \gamma, x\rangle - \delta_K(\gamma)\right). \tag{4}$$

Let now δ_K^* denote the Fenchel-Legendre convex conjugate function of δ_K, i.e.

$$\delta_K^*(x) = \sup_{\gamma \in \mathbb{R}^{\mathcal{R}}} \left(\langle \gamma, x \rangle - \delta_K(\gamma) \right).$$

We then have the following compact characterization of the function γ^{PF}.

LEMMA 1. *The function γ^{PF} satisfies for all $x \in \mathbb{R}_+^{\mathcal{R}}$,*

$$\gamma^{PF}(x) \in \partial \delta_K^*(x), \tag{5}$$

where $\partial \delta_K^*(x)$ denotes the subgradient of the convex function δ_K^* at x.

PROOF. It follows from Theorem 23.5, page 218 in Rockafellar [18] that conditions (4) and (5) are equivalent for any *proper* convex function δ_K^*. Recall that a convex function is proper if it nowhere takes the value $-\infty$, and it takes finite values at some points. Both conditions hold for δ_K^*, which establishes the Lemma. □

This simple result allows to use the powerful theory of convex duality in the study of the function $x \to \gamma^{PF}(x)$. For instance, we have the following:

LEMMA 2. *The function δ_K^* is continuously differentiable on $(0,\infty)^{\mathcal{R}}$, and thus on $(0,\infty)^{\mathcal{R}}$, $\gamma^{PF}(x)$ coincides with the ordinary gradient of δ_K^* at x, and depends continuously on x.*

PROOF. By Theorem 25.1, p. 242 in [18], at a point x where a convex function admits a unique subgradient, it is differentiable, and its subgradient reduces to its ordinary gradient. The original allocation vector $\lambda^{PF}(x)$ is uniquely defined at x whenever $x_r > 0$ for all $r \in \mathcal{R}$, by strict concavity of the log function. Thus, $\gamma^{PF}(x)$ is also uniquely defined at $x \in (0,\infty)^{\mathcal{R}}$, and hence it coincides with the ordinary gradient of δ_K^* at x.

Furthermore, by Theorem 25.5, p.246 in [18], the gradient of a proper convex function is continuous on the domain where the function is differentiable. The claimed continuity of the allocation vector $\gamma^{PF}(x)$ on $x \in (0,\infty)^{\mathcal{R}}$ follows. □

Lemma 1 suggests that the Markov process (x_r), under proportionally fair allocations, is close to insensitive. Indeed, consider the alternative NBA, denoted PF' for modified proportional fairness, and defined by

$$\lambda_r^{PF'}(x) = \exp\left(\delta_K^*(x) - \delta_K^*(x - e_r)\right).$$

It is readily verified that, under this allocation strategy, the Markov process is reversible, and admits the stationary measure

$$\pi^{PF'}(x) = \exp(-L(x)), \tag{6}$$

where the function L is defined by

$$L(x) = \delta_K^*(x) - \sum_{r\in\mathcal{R}} \log(\rho_r) x_r. \tag{7}$$

Fix now $y \in \mathbb{R}_+^{\mathcal{R}}$, and let $x = ny$, where n is large. The heuristic calculation

$$\begin{aligned} \lambda_r^{PF'}(x) &= \exp\left(n\delta_K^*(y) - n\delta_K^*(y - n^{-1}e_r)\right) \\ &\approx \exp\left(\partial_r \delta_K^*(y)\right) \\ &= \lambda_r^{PF}(x), \end{aligned}$$

where we have used the homogeneity property of δ_K^*, according to which $\delta_K^*(ny) = n\delta_K^*(y)$, suggests that the behaviour of the systems under PF and PF' are similar, at least far from the origin. We believe that there exists a strong relationship between the behaviour of the system under PF and under PF'. At this stage we content ourselves with making the following conjecture:

CONJECTURE 1. *Let X^{PF} and $X^{PF'}$ denote the number of customers in steady state under PF and PF', respectively. We conjecture that the rescaled vectors $n^{-1}X^{PF}$ and $n^{-1}X^{PF'}$ satisfy, as $n \to \infty$, a Large Deviations principle with the same rate function L as defined in (7).*

3. Stability properties of proportional fairness. We now apply the above characterization to the study of stability properties of the Markov process describing the number of users in the system under proportional fairness. We do this by studying the *fluid dynamics* of the system. We first consider the original Markov process, before moving to the modified process with Markovian users routing.

COROLLARY 1. *The function L defined in (7) decreases along the trajectories $x(t)$ of the fluid dynamics*

$$\frac{d}{dt} x_r(t) = \nu_r - \mu_r \lambda_r^{PF}(x) \tag{8}$$

defined by the drift vector field of the Markov process. Under the stability condition that $\rho \in \mathring{\mathcal{C}}$, it is a Lyapunov function for these fluid dynamics.

Proof. Note that

$$\frac{d}{dt}L(x(t)) = \sum_{r\in\mathcal{R}} \mu_r \left[\partial_r \delta_K^*(x(t)) - \log(\rho_r)\right]\left[\rho_r - e^{\gamma_r^{PF}(x(t))}\right].$$

The two factors in parentheses are of opposite signs, since by Lemma 1 it holds that $\gamma_r^{PF}(x) = \partial_r \delta_K^*(x)$. Thus the function $L(x)$ decreases along the trajectories $t \to x(t)$.

Note that, by the stability condition $\rho \in \mathring{\mathcal{C}}$, there exists some $\epsilon > 0$ such that $(1+\epsilon)\rho \in \mathcal{C}$, and thus

$$\delta_K^*(x) \geq \sum_{r\in\mathcal{R}} x_r \log((1+\epsilon)\rho_r).$$

It follows that

$$L(x) \geq \log(1+\epsilon)\sum_{r\in\mathcal{R}} x_r.$$

Thus the function L is such that the level sets $\{x : L(x) \leq A\}$ are compact for all finite $A \in \mathbb{R}$. In other words it is a valid Lyapunov function. □

With some more care, we can show that, when $x(t) \neq 0$, the time derivative $(d/dt)L(x(t))$ is less than a strictly negative constant. Indeed, since the vector ρ is assumed to be in the interior of the capacity set, there exists $\epsilon' > 0$ such that whenever $x(t) \neq 0$, for some class r, $\lambda_r^{PF}(x(t)) \geq \epsilon' + \rho_r$. Thus the time derivative $(d/dt)L(x(t))$ is at most $-\inf_r \epsilon' \log(1+\epsilon'/\rho_r)$. In view of the results of Dai [8], this implies ergodicity of the Markov process, thereby providing an alternative to the proof of [3].

An important extension of the original Markovian model is amenable to the same stability analysis. We now describe the corresponding model. As before, users are of different types, $r \in \mathcal{R}$. External arrivals of type r users are according to a Poisson process with intensity $\overline{\nu}_r$; the service times of type r users are again exponential with parameter μ_r. However, after completing service, type r users will re-enter the system as type s users with some probability p_{rs}. Thus the non-zero transition rates are now given by

$$\begin{array}{lll} & x \to x + e_r & \text{with rate } \overline{\nu}_r, \\ (9) & x \to x - e_r + e_s & \text{with rate } \mu_r \lambda_r^{PF}(x) p_{rs}, \\ & x \to x - e_r & \text{with rate } \mu_r \lambda_r^{PF}(x)(1 - \sum_{s\in\mathcal{R}} p_{rs}). \end{array}$$

It is assumed that the matrix $P = (p_{rs})_{r,s\in\mathcal{R}}$ is sub-stochastic, and that its spectral radius is strictly less than 1. Thus, there exists a unique vector $\nu = (\nu_r)_{r\in\mathcal{R}}$ solving the traffic equations

$$\nu_r = \overline{\nu}_r + \sum_{s\in\mathcal{R}} p_{sr}\nu_s, \quad r \in \mathcal{R},$$

also written in matrix form

$$\left(I - P^T\right)\nu = \overline{\nu},$$

where P^T is the transposition of the routing probability matrix P. Introduce the notation $\rho_r = \nu_r/\mu_r$, and $\rho = (\rho_r)_{r\in\mathcal{R}}$. Then the natural stability condition for this system with routing is as before

$$\rho \in \mathring{\mathcal{C}}. \tag{10}$$

We now establish the following:

THEOREM 1. *Consider the fluid dynamics of the system with routing:*

$$\frac{d}{dt}x_r(t) = \overline{\nu}_r - \mu_r\lambda_r^{PF}(x) + \sum_{s\in\mathcal{R}} p_{sr}\mu_s\lambda_s^{PF}(x). \tag{11}$$

The function L as defined in (7) is non-increasing along the trajectories of this system. Moreover, under the stability assumption (10), the function L has compact level sets, and there exists $\epsilon > 0$ such that, whenever $x(t) \neq 0$, one has

$$\frac{d}{dt}L(x(t)) \leq -\epsilon.$$

Consequently, the process with Markovian users routing, defined by the transition rates (9), is ergodic under (10).

PROOF. Write

$$\frac{d}{dt}L(x(t)) = \sum_{r\in\mathcal{R}} \left[\log\lambda_r^{PF}(x(t)) - \log\rho_r\right] \left[\overline{\nu}_r - \mu_r\lambda_r^{PF}(x(t)) + \sum_{s\in\mathcal{R}} p_{sr}\mu_s\lambda_s^{PF}(x(t))\right].$$

We introduce the notation $u_r = \log(\lambda_r^{PF}(x(t))/\rho_r)$. The former equation can be rewritten as

$$\frac{d}{dt}L(x(t)) = \sum_{r\in\mathcal{R}} u_r \left[\overline{\nu}_r - \nu_r e^{u_r} + \sum_{s\in\mathcal{R}} p_{sr}\nu_s e^{u_s}\right],$$

or equivalently, in matrix form,

$$\frac{d}{dt}L(x(t)) = \left\langle u, \overline{\nu} - (I - P^T)(\nu e^u)\right\rangle.$$

We use the notation $\lceil \eta \rfloor$ to denote the diagonal matrix with diagonal entries provided by the coordinates of the vector η. Elementary manipulations entail that

$$\begin{aligned} \frac{d}{dt}L(x(t)) &= -\left\langle u\ ,\ (I-P^T)\lceil \nu \rfloor (e^u - 1)\right\rangle \\ &= -\langle \nu\ ,\ \lceil (I-P)u \rfloor (e^u-1)\rangle \\ &= -\langle \overline{\nu}\ ,\ (I-P)^{-1}\lceil (I-P)u \rfloor (e^u-1)\rangle . \end{aligned} \tag{12}$$

In order to show that the previous expression is non-positive, it is enough to show that for each $r \in \mathcal{R}$, the coefficient of $\overline{\nu}_r$ is non-positive, that is:

$$\sum_{n\geq 0}\sum_{s\in\mathcal{R}} p_{rs}^{(n)}\left(e^{u_s}-1\right)\left[u_s - \sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell\right] \geq 0, \quad r\in\mathcal{R}. \tag{13}$$

In order to complete the proof, we appeal to the following lemma:

Lemma 3. *For any substochastic matrix $P = (p_{rs})_{r,s\in\mathcal{R}}$ with spectral radius strictly less than 1, and any real numbers u_r, $r \in \mathcal{R}$, the inequality (13) is satisfied.*

Moreover, there is equality in (13) only if for all states s such that $\sum_{n\geq 0} p_{rs}^{(n)} > 0$, one has $u_s = 0$.

Let us show how this implies the results of the theorem. That the time derivative $(d/dt)L(x(t))$ is non-positive follows from the first half of the Lemma (note that this holds even without the stability assumption (10); however without (10), L is not a proper Lyapunov function).

By (10), there must exist some $\delta > 0$ such that, when $x(t) \neq 0$, $u_s \geq \delta$ for some $s \in \mathcal{R}$. There must also exist some $r \in \mathcal{R}$ such that $\overline{\nu}_r > 0$, and $\sum_{n\geq 0} p_{rs}^{(n)} > 0$. It is also the case that the u_k are bounded from above by some constant, say A, since the allocations λ_k^{PF} are bounded from above.

Denoting by $F_r(u)$ the left-hand side of (13), as will follow from the proof of the Lemma, it holds that $F_r(u) \geq F_r(u^+)$, where $u^+ := (u_s^+)_{s\in\mathcal{R}}$, and $u_s^+ = \max(u_s, 0)$. This yields the following:

$$x(t) \neq 0 \Rightarrow \frac{d}{dt}L(x(t)) \leq -\inf_{r:\overline{\nu}_r>0} \overline{\nu}_r \inf_{u\in S} F_r(u),$$

where the set S is defined as

$$S = \{u \in [0, A]^{\mathcal{R}} : \max_{k\in\mathcal{R}} u_k \geq \delta\}.$$

Since the function F_r is continuous and the set S is compact, the infimum of $F_r(u)$ over S is attained; however it cannot be zero, in view of the second part of the Lemma and the definition of S. □

Proof. (of Lemma 3). Using the notation $x^{\pm} = \max(0, \pm x)$, note that the factor of $p_{rs}^{(n)}$ in (13) reads

$$\begin{aligned}
(e^{u_s}-1)\left[u_s - \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell\right] &= \left[(e^{u_s}-1)^+ - (e^{u_s}-1)^-\right] \times \cdots \\
&\quad \cdots \left[u_s^+ - u_s^- - \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^+ + \sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^-\right] \\
&= (e^{u_s}-1)^+\left[u_s^+ - \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^+\right] \\
&\quad + (e^{u_s}-1)^-\left[u_s^- - \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^-\right] \\
&\quad + (e^{u_s}-1)^+ \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^- \\
&\quad + (e^{u_s}-1)^- \textstyle\sum_{\ell\in\mathcal{R}} p_{s\ell}u_\ell^+ .
\end{aligned}$$

In order to obtain the above expansion, we have used the fact that $(e^{u_s} - 1)^{\pm}u_s^{\mp} = 0$. Note that the last two terms in this expansion are non-negative. Note also that the first two terms both read

$$(e^{v_s}-1)\left[v_s - \sum_{\ell\in\mathcal{R}} p_{s\ell}v_\ell\right]$$

for adequate choices of v_s, namely $v_s = u_s^+$ for the first term, and $v_s = -u_s^-$ for the second term. This shows that, in order to prove the two claims of the lemma, it is sufficent to restrict attention to the case where the u_s all have the same sign, which we now assume.

Introduce the notations

$$\begin{cases}
M(s,\ell) &:= \sum_{n\geq 0} p_{rs}^{(n)} p_{s\ell}, \\
N(s,\ell) &:= M(s,\ell) + \mathcal{I}_{\ell=r}\sum_{n\geq 0} p_{rs}^{(n)}(1-\sum_{k\in\mathcal{R}} p_{sk}) \\
&= \sum_{n\geq 0} p_{rs}^{(n)}\left(p_{s\ell} + \mathcal{I}_{\ell=r}(1-\sum_{k\in\mathcal{R}} p_{sk})\right).
\end{cases}$$

Condition (13) thus reads

$$\sum_{s,\ell\in\mathcal{R}} M(s,\ell)\,(e^{u_s}-1)\,u_\ell \leq \sum_{s,\ell\in\mathcal{R}} N(s,\ell)\,(e^{u_s}-1)\,u_s.$$

Let us now show that this last condition is satisfied. Note first that it is enough to prove the same inequality, with N instead of M in the left-hand side, since

$$\begin{aligned}
&\textstyle\sum_{s,\ell\in\mathcal{R}} N(s,\ell)\,(e^{u_s}-1)\,u_\ell - \sum_{s,\ell\in\mathcal{R}} M(s,\ell)\,(e^{u_s}-1)\,u_\ell = \\
&\textstyle\sum_{s\in\mathcal{R}}\sum_{n\geq 0} p_{rs}^{(n)}(1-\sum_{k\in\mathcal{R}} p_{sk})\,(e^{u_s}-1)\,u_r,
\end{aligned} \tag{14}$$

and this difference is indeed non-negative under the current assumption that the u_s all have the same sign.

We thus need to show that

$$\sum_{s,\ell\in\mathcal{R}} N(s,\ell)\,(e^{u_s}-1)\,u_\ell \le \sum_{s,\ell\in\mathcal{R}} N(s,\ell)\,(e^{u_s}-1)\,u_s. \tag{15}$$

Note now that the marginals of the measure $N(\cdot,\cdot)$ coincide. Indeed,

$$\begin{array}{rcl} \sum_{\ell\in\mathcal{R}} N(s,\ell) & = & \sum_{n\ge 0} p_{rs}^{(n)}, \\ \sum_{s\in\mathcal{R}} N(s,\ell) & = & \sum_{n\ge 0} p_{r\ell}^{(n+1)} + \mathcal{I}_{\ell=r} \sum_{n\ge 0}\sum_{s\in\mathcal{R}} \left(p_{rs}^{(n)} - p_{rs}^{(n+1)}\right) \\ & = & \sum_{n\ge 0} p_{r\ell}^{(n)}. \end{array}$$

Thus, after renormalisation of both sides of (15) by the total mass of the measure N, it equivalently reads

$$\mathbf{E}\left[(e^U-1)V\right] \le \mathbf{E}\left[(e^U-1)U\right], \tag{16}$$

where the random variables U, V have the same distributions. An inequality due to Hoeffding [10] (see also [21] and [7] for more easily accessible references) states that, given two random variables U, V with identical distributions, for any two non-decreasing functions $f,g:\mathbb{R}\to\mathbb{R}$ such that $f(U)$ and $g(U)$ have finite variances, one has:

$$\mathbf{E}[f(U)g(V)] \le \mathbf{E}[f(U)g(U)].$$

Note that the inequality we need to prove is of that form, with as non-decreasing functions $f(U)=U$ and $g(U)=e^U-1$. Finiteness of variances is trivially satisfied as the random variables U take only finitely many values. This concludes the proof of the first claim of the Lemma.

Let us now show that, in order to have equality in (13), all u_s such that $\sum_{n\ge 0} p_{rs}^{(n)} > 0$ must be zero. Equality in (13) implies equality in (16). However, the latter holds if and only the distributions of $(f(U),g(V))$ and $(f(U),g(U))$ coincide; see e.g. [21]. As the functions f, g are strictly increasing, this in turn holds if the distributions of (U,V) and (U,U) coincide. This means that we can partition the set $\mathcal{R}$ such that on each subset of the partition, the u_s are constant, and for s,ℓ in two different subsets of the partition, $N(s,\ell)=0$. Thus, for all s such that $\sum_{n\ge 0} p_{rs}^{(n)} > 0$, one must have $u_r = u_s$. This is needed to ensure equality in (15) . However, in order to ensure equality in (13), the right-hand side of Formula (14) must also be zero, which, using the fact that all u_s coincide, also reads

$$u_r\,(e^{u_r}-1)\sum_{s\in\mathcal{R}} p_{rs}^{(n)}\left(1-\sum_{k\in\mathcal{R}} p_{sk}\right) = 0.$$

Equivalently, one must have $u_r(\exp(u_r) - 1) = 0$, that is, $u_r = 0$, which concludes the proof of the Lemma. □

REMARK 1. *Note that the statement of Lemma 3 remains true if we replace the terms* $[\exp(u_s) - 1]$ *by* $f(u_s)$ *in (13), where* f *is any strictly increasing function such that* $f(0) = 0$.

We now apply Theorem 1 to systems with general phase-type distributions rather than exponential service time distributions. More precisely, we consider the same setting as before, with user classes $r \in \mathcal{R}$, and capacity set $\mathcal{C} \subset \mathbb{R}_+^{\mathcal{R}}$. New type r users arrive as usual according to a Poisson process with intensity ν_r.

The service time distribution of type r customers is now defined as follows. A finite set I_r, referred to as the set of service phases, is given. The total service time is characterized as the aggregation of service times required in subsequent visits to phases. At each visit to phase i, a corresponding service time that is exponentially distributed, with parameter $\mu_{r,i}$, is required. A visit to phase i is followed by a visit to phase j with probability $p_{r;ij}$. A probability distribution $\{\alpha_i\}_{i \in I_r}$ on I_r specifies the phase in which service starts. The transition matrix $P_r := (p_{r;ij})_{i,j \in I_r}$ is assumed to be sub-stochastic, with spectral radius strictly less than 1. It is easily checked that the above description is equivalent to the definition of phase-type distributions given in [1], p.83.

Denote by $\widehat{\mathcal{R}}$ the set of pairs (r, i) with $r \in \mathcal{R}$ and $i \in I_r$. For all $(r, i) \in \widehat{\mathcal{R}}$, let $x_{r,i}$ denote the number of class r users who are currently in phase i of their service.

The process $(x_{r,i})_{(r,i) \in \widehat{\mathcal{R}}}$ is then a Markov process of the kind covered by Theorem 1. More precisely, it corresponds to the following parameters. For the class $s = (r, i) \in \widehat{\mathcal{R}}$, the external arrival rate $\overline{\nu}_s$ is given by $\nu_r \alpha_{r,i}$ and the corresponding service time parameter is $\mu_s = \mu_{r,i}$. For two classes $s = (r, i)$, $s' = (r', i')$, the corresponding routing probability $p_{ss'}$ is zero if $r \neq r'$, and otherwise equals $p_{r;ii'}$. Finally, the capacity set $\widehat{\mathcal{C}}$ is determined from the original capacity set $\mathcal{C}$ as follows. The allocation vector $(\lambda_s)_{s \in \widehat{\mathcal{R}}}$ belongs to $\widehat{\mathcal{C}}$ if and only if the allocation vector $(\lambda_r)_{r \in \mathcal{R}}$ belongs to $\mathcal{C}$, where λ_r is given by $\sum_{i \in I_r} \lambda_{(r,i)}$[2].

We then have the following

[2]This last statement can be rigorously verified from the definition of proportional fairness.

THEOREM 2. *The process tracking the numbers x_r of users of class r, under proportionally fair allocation of resources characterized by the set $\mathcal{C}$, assuming Poisson arrivals and phase type distributions as just described, is ergodic under the usual condition (10), where $\rho_r = \nu_r \sigma_r$, and σ_r is the mean service time for class r users.*

PROOF. By Theorem 1, ergodicity holds provided the vector $\left(\rho_{(r,i)}\right)_{(r,i)\in\widehat{\mathcal{R}}}$ belongs to the interior of $\hat{\mathcal{C}}$. Equivalently, it holds if the vector with r-th coordinate $\sum_{i\in I_r} \rho_{(r,i)}$ belongs to $\mathring{\mathcal{C}}$.

With the specific routing probability matrix P obtained from the characteristics of the phase type service distributions, one has

$$\begin{aligned} \rho_{(r,i)} &= \tfrac{1}{\mu_{r,i}} \textstyle\sum_{j\in I_r} \sum_{n\geq 0} p^{(n)}_{r;ji} \overline{\nu}_{r,j} \\ &= \tfrac{1}{\mu_{r,i}} \textstyle\sum_{j\in I_r} \sum_{n\geq 0} p^{(n)}_{r;ji} \alpha_{r,j} \nu_r. \end{aligned}$$

This in turn implies that

$$\sum_{i\in I_r} \rho_{(r,i)} = \nu_r \sum_{i\in I_r} \frac{1}{\mu_{r,i}} \sum_{j\in I_r, n\geq 0} \alpha_{r,j}\, p^{(n)}_{r;ji}.$$

Noting that in the above expression, the last sum over $j \in I_r$ and $n \geq 0$ gives the average number of visits to phase i in a class r service time, it readily follows that this last expression coincides with $\nu_r \sigma_r$, which completes the proof. □

4. Relationships between balanced fairness and proportional fairness. In this section we define the balanced fairness NBA, give an equivalent characterization and then use it to relate the stationary distributions under balanced fairness and modified proportional fairness.

The balanced fairness NBA, introduced in [5], is best defined in terms of the *balance function*. The balance function, denoted ψ, is defined by induction on $\mathbb{Z}_+^{\mathcal{R}}$, starting from $\psi(0) = 1$, $\psi(x) = 0$ for any x not in $\mathbb{R}_+^{\mathcal{R}}$, and

$$\psi(x) = \inf\left\{a > 0 : \{a^{-1}\psi(x - e_r)\}_{r\in\mathcal{R}} \in C\right\},$$

where e_r is the r-th unit vector in $\mathbb{R}^{\mathcal{R}}$. The balanced fairness rate allocation vector λ^{BF} is then defined as

$$\lambda_r^{BF}(x) = \frac{\psi(x - e_r)}{\psi(x)},\ x \in \mathbb{Z}_+^{\mathcal{R}},\ r \in \mathcal{R}.$$

As for proportional fairness, it is convenient to consider the logarithms γ_r of the allocated capacities λ_r, rather than the λ_r themselves. Denote by $\gamma^{BF}(x)$ the vector of logarithms of balanced fair allocations, that must lie in the convex non-increasing set K. Introduce the notation $\phi(x) = -\log \psi(x)$. Thus one has:

$$\gamma_r^{BF}(x) = \phi(x) - \phi(x - e_r).$$

In other words, the vector $\gamma^{BF}(x)$ is given by the increments of the function ϕ at x, and can be seen as an approximate gradient of ϕ at x. We introduce the notation $\nabla_d f(x) = \{f(x) - f(x - e_r)\}_{r\in\mathcal{R}}$. Note that a stationary measure for the Markov process counting users of all types is given in terms of the function ϕ by:

$$\pi^{BF}(x) = \frac{1}{Z} \exp\left(-\phi(x) + \sum_{r\in\mathcal{R}} x_r \log(\rho_r)\right) \tag{17}$$

for some normalisation constant Z. A consequence of the reversibility property of the Markov process is that this measure is also stationary for the modified Markov process with Markovian routing [5].

We now give an alternative definition of ϕ.

Lemma 4. *The function ϕ admits the following characterization:*

$$\phi(x) = \sup_{f\in\mathcal{F}}\{f(x)\}$$

where $\mathcal{F}$ is the set of functions defined on $\mathbb{Z}^{\mathcal{R}}$ such that $f(0) = 0$, $f(y) = +\infty$ for $y \notin \mathbb{Z}_+^{\mathcal{R}}$, and $\nabla_d f(y)$ belongs to K for all $y \in \mathbb{Z}_+^{\mathcal{R}}$.

Proof. Denote by $\phi'(x)$ the result of the optimisation problem in the right-hand side of the above expression. Proceed by induction on $x \in \mathbb{Z}_+^{\mathcal{R}}$ to show that $\phi(x) = \phi'(x)$. Clearly, $\phi(0) = \phi'(0) = 0$. Also, as the function ϕ satisfies the conditions over which the optimisation is performed, necessarily one has that $\phi(x) \le \phi'(x)$, for all $x \in \mathbb{Z}_+^{\mathcal{R}}$. Assume thus that $\phi'(y) = \phi(y)$ for all $y \le x$, $y \ne x$. The definition by induction of ψ implies that

$$\phi(x) = \sup\{a : \{a - \phi(x - e_r)\}_{r\in\mathcal{R}} \in K\}.$$

On the other hand, for any f satisfying the assumptions,

$$f(x) \le \sup\{a : \{a - f(x - e_r)\}_{r\in\mathcal{R}} \in K\} \le \sup\{a : \{a - \phi(x - e_r)\}_{r\in\mathcal{R}} \in K\} = \phi(x).$$

We have used for the first inequality the definition of the constraints satisfied by f, for the second we have used the induction hypothesis that $f(x - e_r) \le$

$\phi(x - e_r)$ together with monotonicity of the set K, and the last equality is just the inductive definition of ϕ. □

We shall also rely on the following property of the Fenchel-Legendre conjugate δ_K^*.

LEMMA 5. *The function δ_K^* is such that, for all $x \in \mathbb{R}_+^{\mathcal{R}}$, and all $\epsilon_r > 0$, $r \in \mathcal{R}$,*

$$\left\{ \frac{\delta_K^*(x) - \delta_K^*(x - \epsilon_r e_r)}{\epsilon_r} \right\} \in K. \tag{18}$$

It is understood in this expression that a vector u with coordinates in $\{-\infty\} \cup \mathbb{R}$ belongs to K when the vector e^u with coordinates e^{u_r} belongs to the original convex set C, and $e^{-\infty} = 0$.

PROOF. Let $x \in \mathbb{R}_+^{\mathcal{R}}$. Assume first that all the coordinates x_i are strictly positive. It follows that the vector u achieving the supremum in the original definition of $\delta_K^*(x)$ is uniquely defined. By Lemma 1 above, and Theorem 25.1 p. 242 in [18], it follows that δ_K^* is differentiable at x, its (ordinary) gradient being the vector u achieving that supremum. Also, the function $\epsilon \to \epsilon^{-1}[\delta_K^*(x) - \delta_K^*(x - \epsilon e_r)]$ is non-increasing in $\epsilon > 0$, and achieves its maximum as $\epsilon \searrow 0$, where it equals the coordinate u_r of the gradient (see Theorem 23.1, p. 213-214 in [18]). By monotonicity of the set K, it follows that δ_K^* satisfies the condition (18) at x.

We now show that the same is true when some coordinates of x equal zero. Let $\mathcal{I} \subset \mathcal{R}$ denote the set of indices i for which the coordinate x_i equals zero. We say that x belongs to the *face* $\mathcal{I}$ when $x \in \mathbb{R}_+^{\mathcal{R}}$ and $x_i = 0$ if and only if $i \in \mathcal{I}$. We also denote by $K_{\mathcal{I}}$ the subset of K consisting of these vectors u with $u_i = -\infty$ if and only if $i \in \mathcal{I}$. In the definition of $\delta_K^*(x)$, we may actually replace the optimisation domain by $K_{\mathcal{I}}$ rather than K. There is then a single vector u of $K_{\mathcal{I}}$ which achieves the corresponding supremum. We may conclude as in the previous case that (18) holds in the present case as well. □

We are now ready to establish the following

THEOREM 3. *For any $x \in \mathbb{Z}_+^{\mathcal{R}}$, the following inequalities hold:*

$$\delta_K^*(x) \le \phi(x) \le \delta_K^*(x) + r(x) \tag{19}$$

where

$$r(x) := \sum_{r\in\mathcal{R}:x_r>0} \sum_{m=1}^{x_r} \frac{1}{m}.$$

Proof. The two inequalities shall be established by induction on $\sum_{r\in\mathcal{R}} x_r$ for $x \in \mathbb{Z}_+^{\mathcal{R}}$. They obviously hold true for $x = 0$, as $\phi(0) = \delta_K^*(0) = 0$. Assume thus that they hold for all $y \in \mathbb{Z}_+^{\mathcal{R}}$ such that $\sum_{r\in\mathcal{R}} y_r \leq n$, for some integer $n \geq 0$, and let $x \in \mathbb{Z}_+^{\mathcal{R}}$ be given, $\sum_{r\in\mathcal{R}} x_r = n+1$. By the induction hypothesis and the result of Lemma 4, it holds that

$$\phi(x) = \sup\{a :\ \{a-\phi(x-e_r)\}_{r\in\mathcal{R}} \in K\} \geq \sup\{a :\ \{a-\delta_K^*(x-e_r)\}_{r\in\mathcal{R}} \in K\}.$$

Now, in view of Lemma 5, it holds that

$$\{\delta_K^*(x) - \delta_K^*(x-e_r)\}_{r\in\mathcal{R}} \in K.$$

Therefore,

$$\phi(x) \geq \delta_K^*(x),$$

and the first inequality in (19) is established.

By the induction hypothesis again, we have that

$$\phi(x) \leq \sup\{a :\ \{a - \delta_K^*(x-e_s) - r(x-e_s)\}_{s\in\mathcal{R}} \in K\}. \tag{20}$$

Consider first the case where $x_s > 0$ for all $s \in \mathcal{R}$. We shall rely on the following lemma, the proof of which will be given after the end of the current proof.

Lemma 6. *For all $x, h \in \mathbb{R}^{\mathcal{R}}$, such that x has strictly positive coordinates, and $x+h$ has non-negative coordinates, it holds that*

$$\delta_K^*(x+h) \leq \delta_K^*(x) + \left\langle h, \gamma^{PF}(x)\right\rangle + \sum_{s\in\mathcal{R}} \frac{h_s^2}{x_s}. \tag{21}$$

Thus, in view of the previous equation, we have that

$$\delta_K^*(x-e_s) \leq \delta_K^*(x) - \gamma_s^{PF}(x) + \frac{1}{x_s}.$$

Combining this upper bound with (20), as the vector $(\gamma_s^{PF}(x))_{s\in\mathcal{R}}$ is in K, we have that

$$\phi(x) \leq \delta_K^*(x) + sup_{s\in\mathcal{R}} \left\{ r(x-e_s) + \frac{1}{x_s} \right\}.$$

In view of the definition of $r(y)$, the second term in the right-hand side is clearly upper bounded by $r(x)$, which establishes the desired inequality for x.

To conclude the proof, it remains to deal with the case where some coordinates x_s equal zero. This case is in fact similar to the previous one: if x belongs to face I (i.e., $x_s = 0$ if and only if $s \in I$), the previous argument carries over in $\mathbb{Z}_+^{\mathcal{R}\setminus I}$ by considering the convex set K_I instead of K. □

Proof. (of Lemma 6). Let $x, h \in \mathbb{R}^{\mathcal{R}}$, be fixed, such that x has strictly positive coordinates, and $x+h$ has non-negative coordinates. Let $\gamma \in \mathbb{R}^{\mathcal{R}}$ be such that

$$\delta_K^*(x) = \langle x, \gamma\rangle - \delta_K(\gamma).$$

The pair (x, γ) verifies the relations $x \in \partial\delta_K(\gamma)$, $\gamma \in \partial\delta_K^*(x)$. In addition, the following one-to-one correspondence between sub-gradients of $\delta_{\mathcal{C}}$ and δ_K can be established:

$$x \in \partial\delta_K(\gamma) \Leftrightarrow \{x_s e^{-\gamma_s}\}_{s\in\mathcal{R}} \in \partial\delta_{\mathcal{C}}\left(\{e^{\gamma_s}\}_{s\in\mathcal{R}}\right).$$

Let $h \in \mathbb{R}^{\mathcal{R}}$ be fixed. We have that:

$$\begin{array}{rcl}\delta_K^*(x+h) & = & \sup_{g\in\mathbb{R}^{\mathcal{R}}}\{\langle g, x+h\rangle - \delta_K(g)\} \\ & = & \sup_{u\in\mathbb{R}^{\mathcal{R}}}\{\langle u+\gamma, x+h\rangle - \delta_K(u+\gamma)\} \\ & = & \delta_K^*(x) + \langle h,\gamma\rangle + \sup_{u\in\mathbb{R}^{\mathcal{R}}}\{\langle u, x+h\rangle + \delta_K(\gamma) - \delta_K(\gamma+u)\}.\end{array}$$

However, by convexity of $\delta_{\mathcal{C}}$, and recalling that $e^{-\gamma}x \in \partial\delta_{\mathcal{C}}(e^{\gamma})$, we have the following inequality:

$$\begin{array}{rcl}\delta_K(\gamma+u) & = & \delta_{\mathcal{C}}(e^{\gamma+u}) \\ & \geq & \delta_{\mathcal{C}}(e^{\gamma}) + \langle e^{\gamma+u} - e^{\gamma}, e^{-\gamma}x\rangle \\ & = & \delta_K(\gamma) + \sum_{s\in\mathcal{R}}(e^{u_s}-1)\,x_s.\end{array}$$

Combined with the previous expression for $\delta_K^*(z+h)$, this yields

$$\begin{array}{rcl}\delta_K^*(x+h) & \leq & \delta_K^*(x) + \langle h,\gamma\rangle + \sup_{u\in\mathbb{R}^{\mathcal{R}}}\{\sum_{s\in\mathcal{R}} u_s(x_s+h_s) - (e^{u_s}-1)\,x_s\} \\ & = & \delta_K^*(x) + \langle h,\gamma\rangle + \sum_{s\in\mathcal{R}:x_s+h_s>0}(x_s+h_s)\log(1+h_s/x_s) - h_s \\ & & + \sum_{s\in\mathcal{R}:x_s+h_s=0} x_s.\end{array}$$

The claimed inequality (21) now follows by noting that (i) $\log(1+h_s/x_s) \leq h_s/x_s$, and (ii) $\gamma = \gamma^{PF}(x)$. □

A simple consequence of the Theorem is the following.

COROLLARY 2. *For any $x \in \mathbb{R}^+$, it holds that*

$$\lim_{n\to\infty} \frac{1}{n}\phi(nx) = \delta_K^*(x).$$

PROOF. This follows trivially since the function δ_K^* is positively homogeneous, i.e. $\delta_K^*(nx) = n\delta_K^*(x)$, and since the remainder term $r(nx)$ in (19) is of order $\log(n)$, and a fortiori $o(n)$. □

Equivalently, one has:

COROLLARY 3. *The stationary distribution π^{BF} as in (17) admits the following large deviations asymptotics:*

$$\lim_{n\to\infty} \frac{1}{n}\log \pi^{BF}(nx) = -L(x), \quad x \in \mathbb{R}_+^{\mathcal{R}}, \tag{22}$$

where L is the Lyapunov function (7) used in the study of stability properties of proportional fairness. It thus admits the same large deviations characteristics as the stationary distribution (6) of the system under PF' sharing.

REMARK 2. *The result of Theorem 3 also implies that, if for all $x \in \mathbb{R}_+^{\mathcal{R}}$, there exists a limit $\lim_{n\to\infty} \lambda^{BF}(nx)$ of the allocation vector under balanced fairness, then it must coincide with $\lambda^{PF}(x)$. So far we have not been able to establish the existence of such a limit, except in the special case where $|\mathcal{R}| = 2$, although it seems plausible that the limit exists more generally.*

5. Conclusion. In this note, we have given a simple characterization of the proportionally fair allocation based on convex duality. We have used this to provide a new proof of ergodicity of the Markov process keeping track of the number of users in the system, in a general model with Markovian users routing. As a corollary, we have established that the usual stability conditions known under the assumption of exponentially distributed service times remain valid under phase-type distributions. We have conjectured that the equilibrium distributions under proportional fairness and modified proportional fairness share the same large deviations characteristics. Finally, we have established that the steady state distributions under modified proportional fairness and balanced fairness do share the same large deviations characteristics, with a rate function taking a particularly simple form, separating in one contribution from the network topology and one from the traffic loads.

REFERENCES

[1] S. Asmussen, Applied Probability and Queues, Springer, 2003.

[2] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1992.

[3] T. Bonald and L. Massoulié, Impact of Fairness on Internet Performance, in Proceedings of ACM Sigmetrics, 2001.

[4] T. Bonald, L. Massoulié, A. Proutière and J. Virtamo, A Queueing Analysis of Max-Min Fairness, Proportional Fairness and Balanced Fairness. Submitted to the special issue of Queueing Systems: Queueing Models for fair resource sharing.

[5] T. Bonald and A. Proutière, Insensitive Bandwidth Sharing in Data Networks, Queueing Systems 44(1), 69-100, 2003.

[6] M. Bramson, Communication at the INFORMS Applied Probability Conference, Ottawa, July 2005.

[7] P. Bratley, B.L. Fox and L.E. Schrage, *A guide to simulation*, Springer-Verlag, New York, 1986.

[8] J.G. Dai, On positive Harris recurrence of multiclass queuing networks: a unified approach via fluid limit models, Ann. of Applied Probability 5, 49-77, 1995.

[9] G. De Veciana, T.J. Lee and T. Konstantopoulos, Stability and performance analysis of networks supporting services with rate control - could the Internet be unstable?, in: Proc. IEEE Infocom'99, 1999.

[10] W. Hoeffding, Masstabinvariante korrelationstheorie. Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universitat Berlin 5:179-233, 1940.

[11] F. Kelly, Charging and rate control for elastic traffic, Eur. Trans. Telecommun., vol. 8, pp. 33-37, 1997.

[12] F. Kelly, A. Maulloo and D. Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998) 237-252.

[13] F.P. Kelly and R. J. Williams, Fluid model for a network operating under a fair bandwidth-sharing policy. Annals of Applied Probability **14** 1055–1083, 2004.

[14] P. Key and L. Massoulié, Fluid Limits and Diffusion Approximations for fair network bandwidth sharing. submitted to the special issue of Queueing Systems: Queueing Models for fair resource sharing.

[15] R. Mazumdar, L. Mason and C. Douligeris, Fairness in network optimal flow control: optimality of product forms. IEEE Trans. on Commun. 39: 775-782, 1991.

[16] J. Mo and J. Walrand (2000), Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking* **8**, 556–567.

[17] J. Nash, The bargaining problem, Econometrica, vol. 18, pp. 155-162, 1950.

[18] R. T. Rockafellar, *Convex Analysis*, Princeton University Press 1970 (reedited 1997).

[19] R. Schassberger, Two remarks on insensitive stochastic models, Advances in Applied Probability, **18** 791-814, 1986.

[20] A. Stefanescu and M.W. Stefanescu, The arbitrated solution for multiobjective convex programming, Rev. Roum. Math. Pure Appl., vol. 29, pp. 593-598, 1984.

[21] W. Whitt, Bivariate distributions with given marginals. Annals of Statistics **4** 1280-1289, 1976.

[22] H.Q. Ye, Stability of Data Networks Under an Optimization-Based Bandwidth Allocation. IEEE Transactions on Automatic Control, 48, No. 7, 1238-1242.

E-MAIL: lmassoul@microsoft.com