

Stability and Benefits of Suboptimal Utility Maximization

Tian Lan¹, Xiaojun Lin², Mung Chiang¹, Ruby Lee¹

¹Department of Electrical Engineering, Princeton University, NJ 08544, USA

²School of Electrical and Computer Engineering, Purdue University, IN 47907, USA

Abstract—Network utility maximization has been widely used to model resource allocation and network architectures. But in practice often it cannot be solved optimally due to complexity reasons. Thus motivated, we address the following two questions in this paper: can suboptimal utility maximization maintain queue stability? Can under-optimization of utility objective function in fact lead to benefits to other network design objectives? We show that a resource allocation that is suboptimal with respect to a utility maximization formulation still maintains the maximum flow-level stability when the utility gap is sufficiently small and the information delay is bounded, and can still provide a guaranteed size of stability region otherwise. Utility-suboptimal rate allocation can also enhance other network performance metrics, e.g., it may increase network throughput and reduce link saturation. Quantifying these intuitions, this paper provides a theoretical support for turning attention from optimal but complex solutions of network optimization to those that are simple even though suboptimal.

I. INTRODUCTION

The framework of Network Utility Maximization (NUM) has been very extensively studied over the last decade since [1]. Formulating many resource allocation problems as maximization of an increasing and concave utility function over a convex constraint set, a large number of publications have developed iterative, distributed algorithms that converge to the optimum.

Achieving optimality is clearly desirable for two reasons. Not only does this attain the benchmark of the highest value of network utility, it also guarantees flow-level stochastic stability. The number of flows varies over time as they are randomly generated by users and served by the network. This system can be viewed as a queuing system where the service rate depends on the resource allocation (e.g., rate control) policy employed by the network. For convex NUM with Markov arrival and zero information delay (i.e. perfect queue-length information), it has been shown that for all rate allocation policies maximizing α -fair utilities with $\alpha > 0$, flow-level stochastic stability can be achieved if and only if the traffic intensity lies within the rate region, see, e.g., [2], [3], [4], [5]. In other words, rate region in the α -fair utility maximization problem is also the maximum stability region under arrival and departure dynamics.

Utility-optimality and flow-level stability are strong benefits of optimizing NUM. However, in practice it is often prohibitive to solve NUM optimally, due to computational complexity and

information delay. There are many situations where only suboptimal solutions to the utility maximization problem are realistically computable given the constraints on protocols and timescale: First, rate allocation algorithms require certain convergence time to compute the optimal rate allocations. As a result, practical rate allocations are subject to a positive random time delay. In addition, if the network configuration (e.g., the number of active users) changes faster than the convergence time, the rate allocations will never be optimal. Second, solving some NUM problems for rate allocation may require solving non-convex problems even if the resulting feasible rate region is convex. For example, when the feasible rate region of a network is obtained by time-sharing among different subsets of users, a non-convex multi-user scheduling problem still needs to be solved in order to find the exact rate region achieved by time-sharing [6]. Due to these reasons, optimal rate allocations might be prohibitive in practice.

The gap between elegant theory and useful practice thus lead us to the following question: between optimality and simplicity, which one should we pick in solving NUM? Driven by the practical need for simple yet suboptimal solutions, we allow suboptimal utility maximization, and then quantify the effects of information delay and utility-gap on flow-level stability, and on other important network performance metrics such as total throughput and link saturation.

In [7], the authors show that for a class of rate allocation algorithms based on the so-called dual solutions, the optimal stability region can be achieved even if the algorithm does not converge to the optimal rate allocation at any time. Similar observations have also been made in switching [8] and scheduling [10] problems. In this paper, we take a different approach. We characterize the capability of a resource allocation algorithm by two features: (1) the gap between its utility; and (2) the optimal utility and the time delay of the queue length information. We study stability as a function of both the utility gap and the information delay. Our results apply to a class of general NUM formulations, in which the flow-level queuing model are not first-order Markov, thus making our proof technique of independent interest to general flow level queueing model. Intuitively, one would think that the maximum stability region may be retained if the utility gap is small and the time delay is bounded, while only a reduced stability region can be achieved when the utility gap becomes large. This is indeed true. In Section III, we show that when information delay is uniformly bounded by a constant and the ratio of the utility gap (caused by a suboptimal rate allocation policy) to the maximum utility approaches zero as queue length tends to infinity, the maximum stability region can be retained.

This work has been in part supported by the National Science Foundation through awards CCF-0448012, CNS-0430487, CNS-0519880, CCF-0635202, CNS-0720570, CNS-0721484, and ONR YIP award N00014-07-1-0864. An earlier version of this paper has appeared in IEEE Infocom 2008.

However, when the utility gap is in proportion to the maximum utility, only a reduced stability region can be achieved. In this case, we can still provide a lower bound for the achievable stability region under rate allocation policies satisfying the information delay and the utility gap conditions. These results characterize the stability of a broad class of suboptimal rate allocation policies.

On the other hand, when information delay is bounded, since suboptimal rate allocations with a small enough utility gap is capable of achieving the maximum stability region, we investigate the potential *benefits* of allowing such a utility gap, i.e., the upside of under-optimizing utility objectives. It is clear that by deliberately under-optimizing a utility, we can achieve network performance improvement in other metrics. What remains unclear is precisely how much improvement we can possibly achieve by under-optimizing the utility with a given allowable gap. We formulate the potential performance improvement as a function of given utility gap, and derive a first-order approximation for these tradeoff curves based on local sensitivity (shadow price) analysis. This formulation generalizes that in [11], which focuses on how network performance can be affected by the choice of α -fair utility and assumes that optimality always holds. Our result not only illustrates the potential benefits of under-optimizing a utility, but also quantitatively characterizes the *tradeoff* between sacrificing utility value and improving other network performance metrics, e.g. the total throughput and link saturation. Our analysis can be easily extended beyond the class of α -fair utility.

The key results of this paper are summarized as follows:

- For general utility functions satisfying certain assumptions, we find a sufficient condition for flow-level stability of networks operating under suboptimal rate-allocation policies. When the utility gap is sufficiently small and the information delay is upper bounded by a constant, maximum stability region can be achieved and is shown to be equal to the feasible rate region, i.e. a network is stable if the average traffic load at each link is less than its capacity.
- When the utility gap of suboptimal rate allocations is in proportion to the maximum utility and the information delay is bounded, we show that the achievable stability region can be strictly smaller than the feasible rate region. We further obtain a lower bound for the achievable stability region in this case.
- We formulate and analyze an expected Lyapunov function (of the queue-length) and its first order derivative, with respect to the utility gap and information delay introduced in this paper. Such a new technique is necessary here because our flow-level queuing model with random information delay does not allow a first-order Markov representation. This approach extends the flow-level stability analysis beyond the first-order Markov queuing model used in our previous work [13].
- We consider the impact of the utility gap on overall network performance. Since each utility function is designated for a particular objective, we show that allowing a utility gap gives us freedom to improve other network performance metrics, such as total throughput and maximum link saturation. Thus we formulate and analysis the tradeoff between the utility gap and the two network performance metrics, by employing a sensitivity analysis of the utility maximization problem. Close-form solutions for the gradient of the tradeoff curves

are obtained.

- The results in this paper give a new perspective to look at suboptimal solutions of the utility maximization problem. We show that suboptimal rate-allocation policies may not always be inferior in performance. More precisely, by under-optimizing a utility and allowing a certain optimization gap, we can still retain the maximum flow-level stability and obtain network performance improvements in other metrics.

The remaining of the paper are organized as follows: In Section II, we introduce the class of utility functions considered in this paper and define the utility gap for suboptimal rate allocations. Two stability results are stated next: in Section III.A, a sufficient condition on utility gap and information delay for achieving the maximum flow-level stability is provided. In Section III.B, when the utility gap of suboptimal rate allocations is proportion to the maximum utility, we show that the achievable stability region can be strictly smaller, and we further obtain a lower bound for all achievable stability regions. In Section IV, we analyze the tradeoff between the utility gap and two network performance metrics: total throughput and link saturation. Results based on sensitivity analysis are derived to measure the benefits of under-optimizing α -fair utility. Simulation results are provided at the end of section III and IV respectively. Proofs of all theorems are collected in the Appendix.

Throughout this paper, we use the following notations: Vectors are denoted in small letter, e.g., x , with their i th component denoted by x_i . Matrices are denoted by capitalized letters, e.g., A , with A_{ij} denoting the $\{i, j\}$ th component. Vector inequalities denoted by $x \succeq y$ are considered component-wise. We use $D(x)$ to denote a diagonal matrix whose diagonal elements are the corresponding components from vector x . Subscripts $(\cdot)^T$ denotes the matrix transpose. $\mathbb{P}(\mathbf{M})$ is the probability of an event \mathbf{M} . We use \mathcal{R} to denote a set of vectors and $\tilde{\mathcal{R}}$ for its interior.

II. UTILITY MAXIMIZATION AND GAP

Consider a communication network shared by a set of data flows, which belong to N distinctive flow classes. We assume that flows of class i arrive to the network according to a Poisson process with rate λ_i , and each flow of class i brings a file for transfer whose size is exponentially distributed with mean $\frac{1}{\mu_i}$. A flow is considered to have left the network when its file transfer is completed. Let $x_i(t)$ denote the number of flows of class i that remain in the system at time t . We refer to the vector $x(t) = [x_1(t), \dots, x_N(t)]$ as the network state. The problem of network rate allocation is to determine the total rate allocated to class- i flows in state $x(t)$, denoted by $\phi_i(t)$. Rate $\phi_i(t)$ is equally shared by all class- i flows, each assigned a rate $\phi_i(t)/x_i(t)$. We refer to the vector $\phi(t) = [\phi_1(t), \dots, \phi_N(t)]$ as the rate allocation in state x . The allocation vector $\phi(t)$ is constrained to lie in a set $\mathcal{R} \subset \mathbb{R}_+^N$. The set \mathcal{R} may represent varying physical, topological, technological, and economic constraints of the network under consideration. A rate allocation $\phi(t)$ is feasible if $\phi(t) \in \mathcal{R}$, which means that the network can use some resource allocation policy to support the rate vector $\phi(t)$. In this paper, we only require the set \mathcal{R} to be convex, closed and bounded, which holds in many settings, e.g., [3], [6].

Various network rate control policies can be derived as solving some utility maximization problem with different utility functions,

i.e.

$$\phi_{\text{opt}}(x(t)) = \arg \max_{\phi \in \mathcal{R}} \sum_{i: x_i(t) \geq 1} x_i(t) U_i \left(\frac{\phi_i}{x_i(t)} \right)$$

where U_i is a utility function for flow class i . In this paper, we assume that the function U_i is continuous and twice differentiable on $(0, +\infty)$. In addition, we assume the utility functions satisfy the following conditions:

- (a) $U(z) \geq 0 \forall z$ and $U(0) = 0$, or $U(z) \leq 0 \forall z$.
- (b) $U(z)$ is concave and monotonically increasing.
- (c) $\lim_{z \rightarrow 0} U'(z) = \infty$.

- (d) There exists a positive constant s , s.t. $\frac{zU''(z)}{U'(z)} \geq -s, \forall z$.

Assumptions (a) and (b) are commonly used in the literature [5]. Assumption (c) can be interpreted as one that prevents starvation, since it implies that slope of the utility function increases to infinity as the rate of the flow class approaches zero. Condition (d) requires that the utility function does not have sharp changes. One example of such utility functions satisfying assumptions (a-d) is a class of so-called α -fair utility functions [12], defined by

$$U_i(z) = \begin{cases} \frac{z^{1-\alpha}}{1-\alpha}, & \alpha > 0 \text{ and } \alpha \neq 1 \\ \log z, & \alpha = 1 \end{cases} \quad (1)$$

where α is a positive constant. It is easy to verify that the assumptions (a-d) are satisfied with $s = \alpha$. Parameter $\alpha \geq 0$ models the level of fairness, which includes several special cases such as proportional fairness and max-min fairness. For example, maximizing the total utility corresponds to maximizing weighted throughput as $\alpha \rightarrow 0$, weighted proportional fairness as $\alpha = 1$, minimum potential delay as $\alpha = 2$ and max-min fairness as $\alpha \rightarrow \infty$.

If the optimal rate allocation policy $\phi_{\text{opt}}(x(t))$ that maximizes problem (1) with an α -fair utility is implemented at each time t , it has been shown in [2], [5], [4] that such rate allocation achieves the maximum stability region (i.e. the interior of the feasible rate region \mathcal{R}). In this paper, we consider a more general scenario where rate allocations are not optimal and thus could possibly reduce network stability and performance. This work is motivated by the following two issues in practical networks: First, all practical rate allocation policies are subject to a positive delay due to the time requirement for gathering network information and for algorithm convergence. In other words, the practical rate allocation vector $\phi(t)$ can at best correspond to the optimal rate allocation $\phi_{\text{opt}}(\hat{x}(t))$ for some vector $\hat{x}(t)$, where each $\hat{x}_i(t)$ is equal to $x_i(t - \tau_i(t))$ for a certain information delay $\tau_i(t)$. In other words, $\hat{x}(t)$ represents the network state ‘‘observed’’ by a practical rate allocation policy at time t . Second, due to computational overhead, even given certain information delay, a practical rate allocation policy may still not be able to compute the exact $\phi_{\text{opt}}(\hat{x}(t))$. In other words, there may exist a utility gap due to the suboptimality of the rate allocation policy. We quantify the suboptimality of a practical rate allocation policies $\phi(t)$ with respect to $\hat{x}(t)$ by a utility gap as follows

$$\Delta(\hat{x}(t)) = \sum_{i: \hat{x}_i \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)} \right) - \sum_{i: \hat{x}_i \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) \quad (2)$$

The gap $\Delta(\hat{x}(t))$ measures the difference between suboptimal rate allocations and the optimal allocation, caused only by the imperfect computation of the rate allocation algorithm. Given certain conditions on the utility gap $\Delta(\hat{x}(t))$ and the information delay process $\tau(t)$, in section III, we will characterize the stability region of networks with an arbitrary suboptimal rate allocation policy. In section IV, we will formulate and analyze the tradeoff between utility gap and two network performance metrics.

III. UTILITY GAP, INFORMATION DELAY AND STABILITY

We first investigate how stability will be affected by utility gap $\Delta(\hat{x}(t))$ and information delay $\tau(t)$. Consider a network where class- i flows arrive as a Poisson process of intensity $\lambda_i \geq 0$ and have i.i.d. exponential file sizes of mean $1/\mu_i$. Let $\rho_i = \lambda_i/\mu_i$ be the traffic intensity of class- i flows. This is the traffic load generated by class- i flows per unit time. Due to the information delay, the rate allocation policy $\phi(t)$ now depends on previous network states at time $t - \tau_i(t)$, for $i = 1, \dots, N$. Thus, the usual method of flow-level stability analysis in [2], [5], which require a first-order Markov model of the queue-length process, are insufficient. In this paper, we consider the queue-length process $x(t)$ and prove stability by evaluating an expected Lyapunov drift. Let $h > 0$ be a small time interval. The evolution of the i th queue is described by the following equation:

$$x_i(t+h) = [x_i(t) + a_i(t, h) - d_i(t, h)]^+, \quad (3)$$

where $a_i(t, h)$ is the number of flows arriving to flow class i during time t to $t+h$ and $d_i(t, h)$ is the number of departure flows. We say the network is stable under a given rate allocation policy $\phi(t)$ if there exists a positive non-decaying function $f(\cdot)$ with $\lim_{z \rightarrow \infty} f(z) = \infty$, such that the queue-length process satisfies

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{E} \left[\sum_{i=1}^N f(x_i(t)) \right] dt < \infty. \quad (4)$$

If the feasible rate region \mathcal{R} is compact and convex, a necessary stability condition has been given in [2], [5], [4]: The traffic intensity vector must belong to the feasible rate region ($\rho \in \tilde{\mathcal{R}}$). Furthermore, it has also been shown that all α -fair rate allocations with arbitrary $\alpha > 0$ maximize the flow-level stability region, i.e. assigning rates $\phi_{\text{opt}}(x)$ achieves the maximum stability region, which is equal to the interior of \mathcal{R} . In other words, $\rho \in \tilde{\mathcal{R}}$ is a sufficient condition for stability with $\phi_{\text{opt}}(x)$ as the optimal rate allocation policy.

In practice, when only suboptimal solutions are computable, a positive utility gap $\Delta(\hat{x}(t))$ exists and a information delay $\tau(t)$ has to be considered. In the next section, we will derive a sufficient condition for achieving maximum stability. When the condition is not satisfied, we prove that the achievable stability region may be strictly smaller than the feasible rate region. The main results on stability are stated in Theorem 1 and 2.

A. A Sufficient Condition for Maximum Stability

Theorem 1: For an arbitrary suboptimal rate allocation policy $\phi(t)$, if the information delay $\tau(t)$ is uniformly bounded by a constant $\Omega > 0$ and the order of the utility gap caused by the imperfectness of rate allocation algorithm is less than the order

of the optimal utility when the number of active flows grows large, i.e.,

$$\limsup_{\max_i \hat{x}_i(t) \rightarrow \infty} \frac{\Delta(\hat{x}(t))}{\left| \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)}\right) \right|} = 0, \quad (5)$$

then the network is stable if the traffic condition $\rho \in \tilde{\mathcal{R}}$ is satisfied, i.e., the maximum stability region can be obtained.

Remark 1: If both the information delay $\tau(t)$ and the utility gap $\Delta(\hat{x}(t))$ are upper bounded by constants for all network states, the maximum stability region can be achieved. This is simply a special case of Theorem 1 and can be easily proved by verifying utility gap condition (5). The statement holds for all fair utility functions, including the α -fair utilities with $\alpha > 0$.

Remark 2: Theorem 1 shows that for achieving the maximum stability region, it is not necessary to solve the optimal solution to the utility optimization problem (1) and to require perfect information on the network states $x(t)$. Thus, in practice, suboptimal rate allocation policies that may only require a much lower computational complexity or operate on a larger time-scale than that of the optimal policies could still stabilize the network, as long as the utility gap and delay conditions (5) are satisfied. The sufficient condition in Theorem 1 thus characterizes a large class of suboptimal rate allocation policies that retain the maximum network stability.

B. A Lower Bound on Achievable Stability Region

When the condition (5) in Theorem 1 is not satisfied and the utility gap $\Delta(\hat{x}(t))$ is on the same order as that of the optimal utility, the achievable stability region could be smaller than the feasible rate region, even if the delay $\tau(t)$ is zero.

Proposition 1: There exists a suboptimal rate allocation policy $\phi(t)$ such that the utility gap is on the same order as the order of the optimal utility and the information delay is zero, i.e., for some constant $\eta \in (0, 1)$,

$$\limsup_{\max_i \hat{x}_i \rightarrow \infty} \frac{\Delta(\hat{x}(t))}{\left| \sum_{i=1}^N \hat{x}_i(t) U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)}\right) \right|} \leq \eta, \quad (6)$$

but the achievable stability region is strictly smaller than \mathcal{R} , even if the rate vector $\phi(t)$ is Pareto-optimal (i.e. $\phi(x)$ lies on the boundary of the feasible rate region).

Proposition 1 implies that if the utility gap is large, there exists a suboptimal rate allocation policy whose achievable stability region is strictly smaller than the interior of the feasible rate region \mathcal{R} , regardless of the information delay. Raised from this example, a challenge is to answer the question: what is the minimum stability region that a suboptimal rate allocation policy can achieve given that condition (6) is satisfied?

In the next theorem, we show that $(1 - \eta)^{\frac{1}{|1-s|}} \tilde{\mathcal{R}}$ is a lower bound of all achievable stability regions, if the ratio of the utility gap and the optimal utility is asymptotically bounded by a constant $\eta < 1$ as the number of active flows grows large. This lower bound is tight in the sense that there exists a suboptimal rate allocation policy whose stability region is exactly $(1 - \eta)^{\frac{1}{|1-s|}} \tilde{\mathcal{R}}$.

Theorem 2: For an arbitrary suboptimal rate allocation policy $\phi(t)$, if the information delay is uniformly bounded by a constant

and the order of the utility gap $\Delta(\hat{x}(t))$ is the same as that of the optimal utility, i.e.,

$$\limsup_{\max_i \hat{x}_i(t) \rightarrow \infty} \frac{\Delta(x(t))}{\left| \sum_{i=1}^N \hat{x}_i(t) U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)}\right) \right|} \leq \eta, \quad (7)$$

then the achievable stability region is lower bounded by $(1 - \eta)^{\frac{1}{|1-s|}} \tilde{\mathcal{R}}$. There also exists a suboptimal rate allocation policy satisfying (7) whose stability region is exactly $(1 - \eta)^{\frac{1}{|1-s|}} \tilde{\mathcal{R}}$, i.e., the lower bound is tight.

Remark 3: Theorem 2 provides a lower bound for achievable stability regions. Of course, under condition (7), there might still exist certain suboptimal rate allocation policies that are capable of achieving the maximum stability. However, the lower bound in Theorem 2 is tight in the sense that there exists a suboptimal rate allocation policy with zero information delay and its stability region is exactly $(1 - \eta)^{\frac{1}{|1-s|}} \tilde{\mathcal{R}}$. Proposition 1 and Theorem 2 together characterize the stability of a broad class of suboptimal rate allocation policies.

C. Numerical Examples

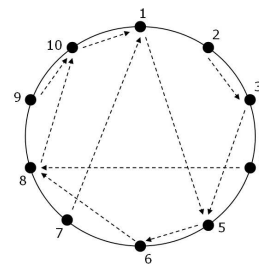


Fig. 1. A ring network with ten users and ten flow classes.

Consider a ring network with $N = 10$ flow classes and $L = 10$ unit-capacity links as shown in Fig.1. Flow class i is initiated by user i and contains $x_i(t) \geq 0$ active flows at time t . Let R be the shortest-distance routing matrix for this ring network. For an α -fair utility function with $\alpha = 1$ (i.e. a logarithmic utility), we compute the optimal rate allocation policy $\phi_{\text{opt}}(x(t)) \forall t$ and then perturb it randomly to construct a set of suboptimal rate allocation $\phi(t)$, such that the resulting information delay and utility gap are constants:

$$\tau_0 = \tau_i(t) \text{ and } \Delta_0 = \Delta(\hat{x}(t)), \quad \forall t. \quad (8)$$

According to Remark 1, since both utility gap and information delay are constants, the suboptimal rate allocation policy $\phi(t)$ constructed above will achieve the maximum stability region that equals to the interior of the feasible rate region $\mathcal{R} = \{\phi \in \mathbb{R}_+^N : R\phi \leq \mathbf{1}, \phi \geq 0\}$.

Figure 2 illustrates flow-level stability of the the network under different suboptimal rate allocation policies for $(\Delta_0, \tau_0) = \{(0, 0), (5, 0), (0, 2), (5, 2)\}$ respectively, by plotting the average total queue length vs. traffic load. In this simulation, we assume that the flow arrival rates for all flow classes are equal, i.e. $\rho_i = \rho_0$ for $i = 1, \dots, 10$. For $\rho_0 \in [0, \frac{1}{3})$, we have $\rho = \rho_0 \cdot \mathbf{1} \in \mathcal{R}$,

which implies that the expected queue-length should remain finite. Figure 2 also shows that the average queue-length of a suboptimal policy $\phi(t)$ approaches that of the optimal rate allocation policy, when utility gap and information delay decrease.

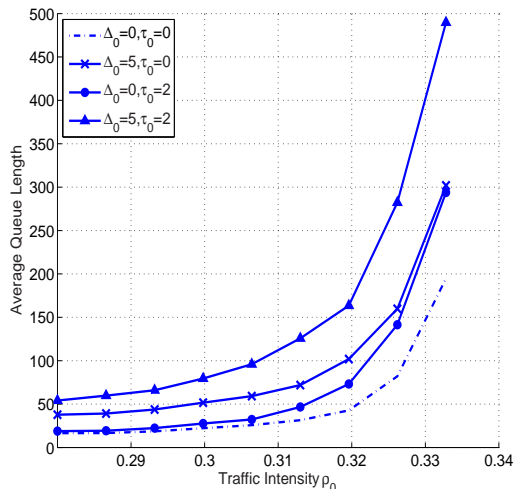


Fig. 2. This figure plots the average total queue-length of the ring network for four different rate allocation policies. It is shown that the three suboptimal rate allocation policies with constant utility gap and information delay still stabilize the queue for traffic intensity $\rho_0 < \frac{1}{3}$, although their delay performances measured by the average queue length are worse compared to that of the optimal rate allocation policy with $(\Delta_0, \tau_0) = (0, 0)$.

IV. UTILITY GAP AND NETWORK PERFORMANCE

Section III showed that when the utility gap is small enough, flow-level stability remains unaffected for uniformly bounded delays. Therefore, a suboptimal rate allocation policy that under-optimizes a certain utility still achieves the maximum stability region. On the other hand, since each utility function is designated to capture one particular network objective, allowing a non-zero utility gap (or, equivalently, under-optimizing the utility) gives us freedom to potentially improve other network performance objectives, such as total throughput and maximum link saturation discussed in this section. More precisely, if we let the information delay to be zero, there exists a tradeoff between utility gap and the maximum network performance improvement we can potentially achieve. In this section we first provide a formulation of this tradeoff. Then we develop a quantitative approximation of the tradeoff curve based on local sensitivity analysis, assuming the utility gap is small. To obtain close-form solutions, we focus on α -fair utility functions in this section, although our result can be extended to all concave utility function. Our approach is different from [11], which is restricted to a throughput-fairness tradeoff for *optimal* solutions only. Instead, in this section, we addresses the following question pertaining to *suboptimality*: by under-optimizing an utility with gap Δ , what is the maximum performance improvement we can possibly achieve?

We focus on the following model for wireline networks, which is an important special case of the model described in section III. Consider a network of L links, indexed by l , each with a finite link capacity c_l . It is shared by N flow classes. Again we use ϕ_i

to denote the total data rate of class- i flows. Then the feasible rate regions are defined by $\mathcal{R} = \{\phi : R\phi \preceq c, \phi \succeq \mathbf{0}\}$, where c is the vector of link capacities and R is the $L \times N$ routing matrix: $R_{li} = 1$ if class- i flows uses link l and 0 otherwise. At each state x , the optimal rate allocation is obtained by solving problem (1) with α -fair utility, i.e.,

$$\begin{aligned} \max_{\phi} \quad & \sum_{i=1}^N x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha} \\ \text{s.t.} \quad & R\phi \preceq c, \phi \succeq \mathbf{0} \end{aligned} \quad (9)$$

Let ϕ_{opt} be the optimal rate allocation that solves the maximization problem (9). Any suboptimal rate allocation $\phi \neq \phi_{opt}$ only achieves a utility value less than the maximum utility. We say a rate allocation ϕ under-optimizes the α -fair utility by a gap Δ if

$$\Delta = U_{opt} - \sum_{i=1}^N x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha} \quad (10)$$

where $U_{opt} = \sum_{i=1}^N x_i^{\alpha} \phi_{opt,i}^{1-\alpha} / (1-\alpha)$ is the optimal utility achieved by rate allocation ϕ_{opt} . Since the α -fair utility is designated for achieving fairness, under-optimizing the α -fair utility with a gap Δ relaxes the maximization problem (9). Thus it gives freedom to system designers to potentially improve other network performance objectives, such as total throughput and maximum link saturation. However, it is unclear how much performance improvement we can achieve by under-optimizing the α -fair utility with a given allowable gap. For example, if we prepare to sacrifice 5% of the utility, how much is the throughput improvement we could expect in return? In the next section, we formulate this type of tradeoff functions and provide a local sensitivity analysis based on examining the Karush-Kuhn-Tucker (KKT) conditions at the optimum allocation.

A. Utility Gap and Total Throughput

First we consider the tradeoff between utility gap and total throughput. For a rate allocation policy ϕ , total throughput T is simply defined as the sum-rate of all flow classes:

$$T = \sum_{i=1}^N \phi_i. \quad (11)$$

We are interested in characterizing the tradeoff between the utility gap and the maximum total throughput. More precisely, we compute the maximum total throughput that can be achieved by under-optimizing the utility with a designated gap. Thus maximum total throughput T is formulated as a function of the utility gap Δ . With some abuse of notation, we refer to this tradeoff function as $T(\Delta)$, which is defined as the maximized objective value of the following optimization problem:

$$\begin{aligned} \max_{\phi} \quad & \sum_{i=1}^N \phi_i \\ \text{s.t.} \quad & R\phi \preceq c, \phi \succeq \mathbf{0} \\ & \sum_{i=1}^N x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha} \geq U_{opt} - \Delta \end{aligned} \quad (12)$$

Thus, the utility gap Δ is the input and the function $T(\Delta)$ gives the maximum possible total throughput under the utility gap constraint.

Remark 4: For the tradeoff function defined by (12), it is easy to see that increasing utility gap relaxes the constraint set of the optimization problem, and leads to a higher optimal objective value. Maximum total throughput $T(\Delta)$, defined by the optimization in (12), is a monotonically increasing function of the utility gap Δ .

Since the α -fair utility functions are concave and the total throughput is linear, we conclude that the maximization problem (12) is convex. Thus we can numerically solve it and compute the maximum-throughput-versus-utility tradeoff curve using any convex optimization solvers. Furthermore, according to the results in section III, a suboptimal rate allocation policy with small enough utility gap can still retain the maximum stability region. When the utility gap is small, we can quantitatively approximate the maximum-throughput-versus-utility tradeoff function using its first order expansion:

$$T - T_0 = \left[\frac{dT}{d\Delta} \Big|_{\Delta=0} \right] \Delta + o(\Delta) \quad (13)$$

where $T_0 = \sum_{i=1}^N \phi_{\text{opt},i}$ is the total throughput at $\Delta = 0$.

In the context of convex optimization, the first order derivative $dT/d\Delta$, also known as shadow price, can be obtained by a local sensitivity analysis, if we make the assumption that the active constraint set in the problem (12) is unchanged when the gap Δ is perturbed locally. Further, we assume that the routing matrix R consists only of ‘bottleneck’ links. These two conditions guarantee that the tradeoff function $T(\Delta)$ is continuous and differentiable at a given Δ . These local sensitive analysis can provide a good approximation of the maximum-throughput-versus-utility tradeoff, when the α -fair utility is slightly under-optimized.

Theorem 3: The maximum-throughput-versus-utility tradeoff function $T(\Delta)$ has the following first order gradient (shadow price) at $\Delta = 0$:

$$\frac{dT}{d\Delta} \Big|_{\Delta=0} = - \frac{\mathbf{1}^T \cdot A \cdot \left(\frac{x^\alpha}{\phi_{\text{opt}}^\alpha} \right)}{\left(\frac{x^\alpha}{\phi_{\text{opt}}^\alpha} \right)^T \cdot A \cdot \left(\frac{x^\alpha}{\phi_{\text{opt}}^\alpha} \right)}, \quad (14)$$

where $A = D^{-1} - D^{-1}R^T(RD^{-1}R^T)^{-1}RD^{-1}$ and $D = \alpha \cdot \text{diag} \left\{ [\phi_{\text{opt},1}^{-\alpha-1}; \dots; \phi_{\text{opt},N}^{-\alpha-1}] \right\}$ is a diagonal matrix. The vector division and power $x^\alpha/\phi_{\text{opt}}^\alpha$ are component-wise.

B. Utility Gap and Maximum Link Saturation

In this section, we consider the maximal link saturation as a network performance metric, defined by

$$Z = \max_{l \in L} \frac{\sum_i R_{il} \phi_i}{c_l}. \quad (15)$$

By under-optimizing the α -fair utility, it is possible to reduce the maximal link saturation and then balance the network traffic over all links. Moreover, reducing Z could potentially minimize the occurrence of ‘bottleneck’ links in the network, and also make the network more robust to link capacity fluctuation and traffic bursts.

We characterize the optimal tradeoff between the utility gap and the maximum link saturation, i.e. we compute the minimum Z that can be achieved by under-optimizing the α -fair utility with a designated gap. This tradeoff function $Z(\Delta)$ can be formulated as follows

$$Z(\Delta) = \min_{\phi} \max_{l \in L} \frac{\sum_s R_{il} \phi_i}{c_l} \quad (16)$$

subject to $R\phi \preceq c, \phi \succeq 0$

$$\sum_{i=1}^N x_i^\alpha \frac{\phi_i^{1-\alpha}}{1-\alpha} \geq U_{\text{opt}} - \Delta$$

Remark 5: For the tradeoff function defined by (16), it is easy to see that increasing utility gap relaxes the constraint set of the optimization problem, and leads to a smaller optimal objective value. Thus maximum link saturation Z is a monotonically decreasing function of the utility gap Δ . Furthermore, it is easy to verify that the optimization problem (16) is convex. The saturation-gap tradeoff can be numerically computed.

Now we conduct a local sensitivity analysis for the saturation-gap tradeoff defined by optimization problem (16). Again, we make the assumption that the active constraint set in the problem (12) is unchanged when the gap δ is perturbed locally. The main result is summarized in the next theorem. Its proof is very similar to that of Theorem 3. We denote Z_0 as the link saturation for $\Delta = 0$.

Theorem 4: When the utility gap is small, the saturation-utility tradeoff function can be approximated using its first order expansion:

$$Z - Z_0 = \left[\frac{dZ}{d\Delta} \Big|_{\Delta=0} \right] \Delta + o(\Delta). \quad (17)$$

The first order derivative (shadow price) of the saturation-utility tradeoff function is given by

$$\frac{dZ}{d\Delta} \Big|_{\Delta=0} = - \frac{1}{c^T (RD^{-1}R^T)^{-1} c}, \quad (18)$$

where $D = \alpha \cdot \text{diag} \left\{ [x_1^\alpha \phi_{\text{opt},1}^{-\alpha-1}, \dots, x_N^\alpha \phi_{\text{opt},N}^{-\alpha-1}] \right\}$ is a diagonal matrix.

C. Numerical Examples

In this section, we plot the two tradeoff curves and their first-order approximations obtained in Section IV.A and Section IV.B, for the ring network described in Section III.C. Since all links have unit capacity, the feasible rate region is given by $\mathcal{R} = \{\phi : R\phi \leq \mathbf{1}, \phi \succeq \mathbf{0}\}$, where R is the routing matrix for the ring network. Let x_i denote the number of active flows for source i . We can solve the two convex optimization problems (12) and (16) for an arbitrary utility gap Δ to obtain the exact tradeoff curves $T(\Delta)$ and $Z(\Delta)$, which are plotted in Figure 3 and Figure 4 using solid lines. In both figures, we assume that the number of active flows are $x_i = 10 \forall i$. A proportional fairness utility function corresponding to $\alpha = 1$ is considered.

When the utility gap Δ is small, the maximum-throughput-versus-utility and the saturation-gap tradeoff curves can be approximated by their first order expansions given by (13) and (17), respectively. Using the close form solutions in Theorem 3

and Theorem 4, we compute the first order gradients as follows $\frac{dT}{d\Delta}\Big|_{\Delta=0} = 0.414$ and $\frac{dZ}{d\Delta}\Big|_{\Delta=0} = -0.010$. Thus the two tradeoff curves can be approximated by

$$T(\Delta) \approx T(0) + 0.414\Delta \quad (19)$$

$$Z(\Delta) \approx Z(0) - 0.01\Delta \quad (20)$$

In Figure 3 and Figure 4, we also plot the corresponding linear approximations for the maximum-throughput-versus-gap and the saturation-gap tradeoff curves in dashed line.

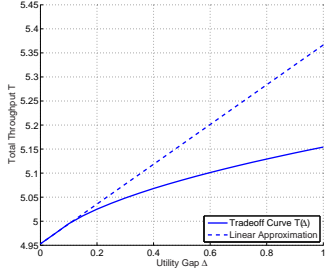


Fig. 3. Throughput-Gap tradeoff curve and its first order approximation.

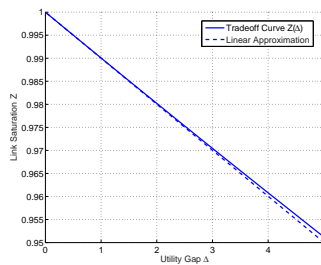


Fig. 4. Saturation-Gap tradeoff curve and its first order approximation.

Figure 4 shows that the saturation-gap tradeoff defined in Section IV.B can be well approximated by its first order expansion, given by the closed-form expressions in Theorem 4, while such an approximation is accurate for the throughput-gap tradeoff only when the utility gap is small. The two tradeoff curves allow us to predict how much performance improvement we can possibly achieve by under-optimizing the utility with a designated small utility gap. For example, if we under-optimize the utility by 1%, i.e. $\Delta = 1\% |U_{\text{opt}}| = 0.312$, it is clear from equation (19) that a maximum total throughput increase of $T - T_0 = +0.129$ (equivalently $+2.52\%T_0$) could be expected in return at the optimum. This result not only illustrates the potential benefits of under-optimizing an α -fair utility, but also quantitatively characterizes the tradeoff between sacrificing utility value and achieving network performance improvement. Suppose the utility of the ring network is under-optimized by 1%. The local sensitive analysis results are summarized as follows: throughput is enhanced by 2.52% ($T - T_0 = +0.129$) and link saturation reduced by 0.31% ($Z - Z_0 = -0.0031$). Whether this particular tradeoff is worth making or not depends on operator's preference, but it is important to provide the choices of tradeoff through the results like those in this section.

V. CONCLUDING REMARKS

Suboptimal resource allocation with a utility gap is simply an inevitable phenomenon in real networking. Fortunately, it may still be able to maintain stability region and even enhance other network performance metrics. Intuition on stability and utility-versus-throughput and utility-versus-saturation tradeoff are quantified with closed-form expressions in this paper. There are still open questions to the study of suboptimal solutions to network optimization, e.g., degradation on fairness due to utility gap and global sensitivity analysis, before we fully understand "how bad is suboptimal rate allocation".

APPENDIX: PROOFS

A. Properties of the Utility Function Satisfying Definition 1.

Lemma 1: If $U(\cdot)$ is a utility function satisfying Assumptions (a-d), then $U'(a) \geq (\frac{a}{b})^{-s} U'(b)$ for all $a \geq b > 0$. Further, if the utility function is negative, then $U(a) \leq (\frac{a}{b})^{-|1-s|} U(b)$. Otherwise, if the utility function is positive, $U(a) \geq (\frac{a}{b})^{|1-s|} U(b)$.

Proof: From Assumption (c), we have $\frac{U''(z)}{U'(z)} \geq -\frac{s}{z}$. Choose $a \geq b > 0$ and integrate both side of the inequality from b to a . We obtain $U'(a) \geq (\frac{a}{b})^{-s} U'(b)$. If the utility function is negative (i.e. case 2 in Assumption (a)), we fix b in this inequality and integrate it from $a = b$ to $a = +\infty$, i.e.

$$b^s U'(b) \int_b^{+\infty} \frac{1}{y^s} dy \leq U(\infty) - U(b) \leq -U(b) \quad (21)$$

where $U(\infty)$ exists because the utility function is monotonically increasing and upper bounded by zero as in Assumption (a). This implies that the integration on the left hand side also exists and thus $s > 1$. We can derive $-\frac{U'(b)}{U(b)} \leq \frac{s-1}{b}$. Integrating it again, we obtain $U(a) \leq (\frac{a}{b})^{1-s} U(b)$, which is the desired result. Similarly, when the utility function is positive, we consider the integral of $U'(a) \geq (\frac{a}{b})^s U'(b)$ from $b = 0$ to $b = a$ and derive the result in Lemma 1. ■

B. Proof of Theorem 1.

Proof: To prove stability under the rate allocation policy $\phi(t)$, we consider the following Lyapunov function

$$V(x(t)) = \sum_{i=1}^N \sum_{n=1}^{x_i(t)} U'_i \left(\frac{c\rho_i}{n} \right). \quad (22)$$

where $c > 0$ is a proper constant defined later in the proof. Since the rate-allocation depends on the history of past states, previous methods in [2], [5] for analyzing flow-level stability are insufficient. In the following proof, we will analyze the expected Lyapunov function $W(t) = \mathbb{E}[V(x(t))]$ and obtain an expression for the drift of the function $W(t)^*$. Then, the flow-level stability (4) can be proven by deriving an upper bound for the drift

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{W(t+h) - W(t)}{h} \\ &= \sum_{i=1}^N \lim_{h \rightarrow 0} \frac{\mathbb{E} \left[\sum_{n=1}^{x_i(t+h)} U'_i \left(\frac{c\rho_i}{n} \right) - \sum_{n=1}^{x_i(t)} U'_i \left(\frac{c\rho_i}{n} \right) \right]}{h} \end{aligned} \quad (23)$$

In order to move the limit (as $h \rightarrow 0$) inside the expectation above, we make the use of the Dominated Convergence Theorem. Recall that $x(t+h) = [x(t) + a(t,h) - d(t,h)]^+$. Because both arrival rate λ_i and departure rate $\mu_i \phi_i(t)$ can be upper bounded by a constant $\omega > 0$, using a simple sample-path argument [7], we can easily show that, for any $h > 0$,

$$-\mathcal{Y}_1 \leq x_i(t+h) - x_i(t) \leq \mathcal{Y}_2, \quad (24)$$

*A similar problem with feedback delay is treated in [10] although the model there does not involve any flow-level dynamics.

where \mathcal{Y}_1 and \mathcal{Y}_2 are some Poisson random variables with mean ωh and independent of $x(t)$ and $\phi(t)$. Then, for each i and $h < 1$, we can upper bound the expectation in (23) by

$$\begin{aligned} & \frac{1}{h} \mathbb{E} \left[\sum_{n=1}^{x_i(t+h)} U'_i \left(\frac{c\rho_i}{n} \right) - \sum_{n=1}^{x_i(t)} U'_i \left(\frac{c\rho_i}{n} \right) \right] \\ & \leq \frac{1}{h} \mathbb{E} \left[\sum_{n=1}^{\mathcal{Y}_2} U'_i \left(\frac{c\rho_i}{x_i(t) + n} \right) \right] \\ & \leq \frac{1}{h} \mathbb{E} \left[\sum_{n=1}^{\mathcal{Y}_2} U'_i \left(\frac{c\rho_i}{x_i(t) + \mathcal{Y}_2} \right) \right] \\ & \leq \frac{1}{h} \mathbb{E} \left[\sum_{n=1}^{\mathcal{Y}_2} (x_i(t) + \mathcal{Y}_2)^s U'_i \left(\frac{c\rho_i}{1} \right) \right] \\ & \leq \frac{1}{h} \mathbb{E} \left[x_i^s(t) \mathcal{Y}_2 (1 + \mathcal{Y}_2)^s U'_i (c\rho_i) \right] \\ & \leq U'_i (c\rho_i) \sum_{n=1}^{\infty} \frac{\omega^n (1+n)^s}{(n-1)!} \mathbb{E} [x_i^s(t)], \end{aligned}$$

where the third step uses the inequality $U'(a) \geq \left(\frac{a}{b}\right)^{-s} U'(b)$ (for $a = c\rho_i$ and $b = \frac{c\rho_i}{x_i(t) + \mathcal{Y}_2}$) proven in Lemma 1. Since $x_i(t)$ is also bounded by a Poisson random variable with mean ωt , it is easy to verify that $\sum_{n=1}^{\infty} \frac{\omega^n (1+n)^s}{(n-1)!} \mathbb{E} [x_i^s(t)] < \infty$, for all t . It then provides an upper bound for the expectation in (23) that is needed for the Dominated Convergence Theorem to hold. To obtain a lower bound for the expectation in (23), we have

$$\begin{aligned} & \frac{1}{h} \mathbb{E} \left[\sum_{n=1}^{x_i(t+h)} U'_i \left(\frac{c\rho_i}{n} \right) - \sum_{n=1}^{x_i(t)} U'_i \left(\frac{c\rho_i}{n} \right) \right] \\ & \geq -\frac{1}{h} \mathbb{E} \left[\sum_{n=0}^{\min(\mathcal{Y}_1, x_i(t))-1} U'_i \left(\frac{c\rho_i}{x_i(t) - n} \right) \right] \\ & \geq -\frac{1}{h} \mathbb{E} \left[\sum_{n=0}^{\min(\mathcal{Y}_1, x_i(t))-1} U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \right] \\ & \geq -\frac{1}{h} \mathbb{E} \left[\mathcal{Y}_1 x_i^s(t) U'_i \left(\frac{c\rho_i}{1} \right) \right] \\ & \geq -U'_i (c\rho_i) \sum_{n=1}^{\infty} \frac{\omega^n}{(n-1)!} \mathbb{E} [x_i^s(t)]. \end{aligned}$$

Since $\sum_{n=1}^{\infty} \frac{\omega^n}{(n-1)!} \mathbb{E} [x_i^s(t)] < \infty$ for all t , this provides the lower bound needed for the Dominated Convergence Theorem to hold. Thus, we can move the limit (as $h \rightarrow 0$) inside the expectation in (23). Let $\mathcal{F}_t = \sigma\{x(u), u \leq t\}$ denote the σ -field generated by the history up to time t . We derive an explicit expression for $\dot{W}(t)$ as shown in equation (25). Next, in order to bound $\dot{W}(t)$ under the utility gap and the information delay conditions (5), we first prove the following lemma, which can be used to provide an upper bound on last term in (25).

Lemma 2: Consider any traffic intensity $\rho \in \tilde{\mathcal{R}}$ and constant $C > 0$. If the suboptimal rate allocation $\phi(t)$ satisfies the utility gap condition in (5), there exists positive constants $\gamma > 0$ and $\epsilon > 0$ such that for all $|r| < C$ and for any network state satisfying

$\max_i \hat{x}_i(t) > \gamma$, the following inequality holds:

$$\sum_{i: \hat{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t) + r} \right) \leq 0. \quad (28)$$

Proof: Under the traffic condition $\rho \in \tilde{\mathcal{R}}$, there exist $\epsilon \geq 0$ and $\delta \geq 0$ such that rate vector $(1 + \epsilon)^4 (1 - \delta)^{-\frac{1}{|1-s|}} \rho$ satisfies the feasible rate constraints, i.e.

$$(1 + \epsilon)^4 (1 - \delta)^{-\frac{1}{|1-s|}} \rho \in \mathcal{R} \quad (29)$$

According to the utility gap condition in (5), we can conclude that for any $\delta \geq 0$, there exists a positive γ_1 such that for all $x(t)$ satisfying $\max_i \hat{x}_i(t) > \gamma_1$,

$$\Delta(\hat{x}(t)) \leq \delta \left| \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)} \right) \right|. \quad (30)$$

To remove the absolute value on the right hand side of (30), we first assume the utility function to be non-negative. Let $u \in \mathcal{R}$ be an arbitrary rate vector and $\delta_0 = (1 - \delta)^{\frac{1}{|1-s|}}$. In view of (2) we obtain the following inequalities, for all $\max_i \hat{x}_i(t) > \gamma_1$,

$$\begin{aligned} 0 & = \Delta(\hat{x}(t)) + \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) \\ & \quad - \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)} \right) \\ & \leq \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - (1 - \delta) \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)} \right) \\ & \leq \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - (1 - \delta) \hat{x}_i(t) U_i \left(\frac{u_i}{\hat{x}_i(t)} \right) \\ & \leq \sum_{i: \hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - \hat{x}_i(t) U_i \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right) \\ & \leq \sum_{i: \hat{x}_i(t) \geq 1} [\phi_i(t) - \delta_0 u_i] U_i \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right) \end{aligned}$$

where the third step follows directly from Lemma 1 by letting $a = u_i$ and $b = \delta_0 u_i$, and the last step holds since the utility function $U_i(\cdot)$ is concave. Choosing $u = \frac{(1+\epsilon)^4}{\delta_0} \rho$, we derive

$$\sum_{i: \hat{x}_i(t) \geq 1} U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) [(1 + \epsilon)^4 \rho_i - \phi_i(t)] \leq 0. \quad (31)$$

When the optimal utility value in (30) is negative, using the same proof technique and choosing $\delta_0 = (1 + \delta)^{-\frac{1}{|1-s|}}$, we can show that the inequality (31) is also satisfied. Note that (31) is almost the same as (28), except for a constant r in the denominator. Toward this end, we make use of the monotonicity of $U'_i(\cdot)$ and the inequality $U'(a) \geq \left(\frac{a}{b}\right)^{-s} U'(b)$ (for $a \geq b > 0$) proven in Lemma 1. For $|r| < C$, we get a chain of inequalities in equation (26), where the fifth step is from (51) and in the third step, K_s is a proper constant depending on C , such that both $(1 + \frac{C}{z})^s \leq 1 + \frac{K_s C}{z}$ and $(1 - \frac{C}{z})^s \geq 1 - \frac{K_s C}{z}$ hold for all $z \geq 1$. Since the feasible rate region \mathcal{R} is bounded, we have $\phi_i(t) \leq \psi$ for some $\psi > 0$. Then, it is easy to show that each term in the last

$$\begin{aligned}
\dot{W}(t) &= \lim_{h \rightarrow 0} \frac{W(t+h) - W(t)}{h} = \lim_{h \rightarrow 0} \frac{\mathbb{E}[\mathbb{E}[V(x(t+h))|\mathcal{F}_t] - V(x(t))]}{h} \\
&= \sum_{i=1}^N \mathbb{E} \left[\sum_{n=1}^{+\infty} \lim_{h \rightarrow 0} \frac{\mathbb{P}(a_i(t, h) - d_i(t, h) = n | \mathcal{F}_t)}{h} \sum_{k=1}^n U'_i \left(\frac{c\rho_i}{x_i(t) + k} \right) \right. \\
&\quad \left. - \sum_{n=1}^{x_i(t)} \lim_{h \rightarrow 0} \frac{\mathbb{P}(a_i(t, h) - d_i(t, h) = -n | \mathcal{F}_t)}{h} \sum_{k=-n+1}^0 U'_i \left(\frac{c\rho_i}{x_i(t) + k} \right) \right] \\
&= \sum_{i=1}^N \mathbb{E} \left[\lambda_i U'_i \left(\frac{c\rho_i}{x_i(t) + 1} \right) - \mu_i \phi_i(t) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x_i(t) \geq 1\}} \right] \tag{25}
\end{aligned}$$

$$\begin{aligned}
&\sum_{i: \hat{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t) + r} \right) \\
&\leq \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t) - C} \right) \right\} \\
&\leq \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) \left(1 - \frac{C}{\hat{x}_i(t)} \right)^s U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \right\} \\
&\leq \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) \left(1 - \frac{K_s C}{\hat{x}_i(t)} \right) U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \right\} \\
&\leq \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \left[\rho_i (1 + \epsilon)^3 + \frac{\phi_i(t) K_s C}{\hat{x}_i(t)} \right] U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \right\} \\
&\leq \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \left[\rho_i (1 + \epsilon)^3 + \frac{\phi_i(t) K_s C}{\hat{x}_i(t)} \right] U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \rho_i (1 + \epsilon)^4 U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \right\} \\
&= \sum_{i: \hat{x}_i(t) \leq C} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) + \sum_{i: \hat{x}_i(t) > C} \left\{ \left[\frac{\phi_i(t) K_s C}{\hat{x}_i(t)} - \epsilon \rho_i (1 + \epsilon)^3 \right] U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \right\} \tag{26}
\end{aligned}$$

$$\begin{aligned}
\dot{W}(t) &= \sum_{i=1}^N \mathbb{E} \left[\lambda_i U'_i \left(\frac{c\rho_i}{x_i(t) + 1} \right) - \mu_i \phi_i(t) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x_i(t) \geq 1\}} \right] \\
&\leq A_1 + \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c\rho_i}{x_i(t) + 1} \right) - \mu_i \phi_i(t) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \right] \\
&\leq A_1 + A_2 + \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i \left(1 + \frac{1}{x_i(t)} \right)^s U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} - \mu_i \phi_i(t) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i \hat{x}_i(t) > \xi - 2C\}} \right] \\
&\leq A_1 + A_2 + \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i \left(1 + \frac{K_s}{x_i(t)} \right) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right] - (1 + \epsilon)^3 \sum_{i: \hat{x}_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c\rho_i}{\hat{x}_i(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i \hat{x}_i(t) > \xi - 2C\}} \right] \\
&\leq A_1 + A_2 + \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i \left(1 + \frac{\epsilon}{2} \right) U'_i \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right] - (1 + \epsilon)^3 \sum_{i: \hat{x}_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c\rho_i}{\hat{x}_i(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i \hat{x}_i(t) > \xi - 2C\}} \right] \tag{27}
\end{aligned}$$

summation in (26) can be upper bounded, i.e.

$$\begin{aligned} & \left[\frac{\phi_i(t)K_s C}{\hat{x}_i(t)} - \epsilon \rho_i (1 + \epsilon)^3 \right] U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \\ & \leq \sup_{1 \leq \hat{x}_i(t) \leq \frac{\psi K_s C}{\epsilon \rho_i (1 + \epsilon)^3}} \frac{\psi K_s C}{\hat{x}_i(t)} U'_i \left(\frac{(1 + \epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \\ & \leq \psi K_s C U'_i \left(\frac{\epsilon \rho_i^2 (1 + \epsilon)^7}{\psi K_s C} \right), \quad \forall i \end{aligned} \quad (32)$$

Combining (32) and Assumption (c) of the utility function in section II, we conclude (26) $\rightarrow -\infty$ as $\max_i \hat{x}_i(t) \rightarrow +\infty$. This means that there exists a constant $\gamma > \gamma_1$, such that for $\max_i \hat{x}_i(t) > \gamma$, we have (26) ≤ 0 . This completes the proof of the lemma. \blacksquare

Now we choose $c = (1 + \epsilon)^4$ in the Lyapunov function (22) and use Lemma 2 to provide an upper bound for the expected drift $\dot{W}(t)$. From the last part of the proof of Lemma 2, we can also conclude that there exists a constant $\gamma_2 > 0$, such that for $\max_i \hat{x}_i(t) > \gamma_2$ we have $\sum_{i: \hat{x}_i(t) \geq 1} \left(\frac{K_s C}{\hat{x}_i(t)} - \frac{\epsilon}{2} \right) U'_i \left(\frac{c \rho_i}{\hat{x}_i(t)} \right) \leq 0$. So if we choose $\xi = \max\{\gamma, \gamma_2\} + 2C$ (which implies $\max_i \hat{x}_i(t) > \max\{\gamma, \gamma_2\}$ when $\max_i x_i(t) > \xi$ and $\|x(t) - \hat{x}(t)\|_1 < 2C$), then we can construct a bounded region, denoted by $\mathcal{G} = \{x(t) : \max_i x_i(t) \leq \xi\}$, within which $\sum_{i=1}^N \lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)+1} \right) \leq A_2 < \infty$ is bounded for all $x(t) \in \mathcal{G}$.

Further, we define another constant $A_1 = \sum_{i=1}^N \lambda_i U'_i(c \rho_i)$.

Define the event $\mathcal{E}_t = \{\|x(t-u) - x(t-\Omega)\|_1 < C, \forall u \in [0, \Omega]\}$, i.e. it is the event that the maximum change of network state within time $t-\Omega$ to t is bounded by C . Since the information delay is bounded by $\tau(t) < \Omega$, event \mathcal{E}_t implies that

$$\begin{aligned} \|x(t) - \hat{x}(t)\|_1 & \leq \|x(t) - x(t-\Omega)\|_1 + \|\hat{x}(t) - x(t-\Omega)\|_1 \\ & < 2C. \end{aligned} \quad (37)$$

If $\max_i \hat{x}_i(t) > \xi - 2C$, the conditions in Lemma 2 are all satisfied under event \mathcal{E}_t . Then, we can bound the expected drift $\dot{W}(t)$ by the equation (27), where the third inequality is direct from Lemma 2. To bound (27), we need to prove the following inequality.

Lemma 3: For sufficiently large $C > 0$ and $A_3 > 0$, the last term in (27) can be bounded by

$$\begin{aligned} & 2A_3 + \sum_{i: \hat{x}_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{\hat{x}_i(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i \hat{x}_i(t) > \xi - 2C\}} \right] \\ & \geq \frac{1}{(1 + \epsilon)^2} \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right]. \end{aligned} \quad (38)$$

Proof: The main idea here is that, when C is large, $\mathcal{P}(\mathcal{E}_t)$ is close to 1. However, the difficult part in proving (38) is that the three product terms inside the expectation on the left hand side are dependent. To handle this, we introduce the network state $x^\Omega(t) = x(t-\Omega)$ at time $t-\Omega$ as an auxiliary variable and bound both sides of (38) with respect to $x^\Omega(t)$, respectively. First, we prove a chain of inequalities in equation (34), where the third inequality uses the fact that $\hat{x} > \xi - 2C \geq \gamma_2$ when $\max_i x_i^\Omega(t) > \xi - C$ and event \mathcal{E}_t occurs. To allow the summation to change from $\{i : x_i^\Omega(t) \geq 1\}$ to $\{i : \hat{x}_i(t) \geq 1\}$ in step 2 above, we define a positive constant A_3 , which upper bounds the terms corresponding to $\hat{x}_i(t) = 0$, i.e. $\sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\{\hat{x}_i(t)=0\}} \right] < A_3$. The constant

A_3 exists because

$$\begin{aligned} & \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\{\hat{x}_i(t)=0\}} \right] \\ & \leq \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i |x_i^\Omega(t)|^s U'_i \left(\frac{c \rho_i}{1} \right) \mathbf{1}_{\{\hat{x}_i(t)=0\}} \right] \\ & = \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i |\hat{x}_i(t) - x_i^\Omega(t)|^s U'_i \left(\frac{c \rho_i}{1} \right) \mathbf{1}_{\{\hat{x}_i(t)=0\}} \right] \\ & < +\infty, \end{aligned} \quad (39)$$

where $|\hat{x}_i(t) - x_i^\Omega(t)|$ can be bounded by a Poisson random variable with mean $\omega \Omega$. Then, inequality (34) establishes a lower bound on the left hand side of (38). To derive an upper bound for the right hand side of (38), we prove a chain of inequalities in (33) using the same argument as above. The first term in (33) can be bounded by equation (33). Recall that $x_i^\Omega(t) = x_i(t-\Omega)$. Given $\mathcal{F}_{t-\Omega}$, $|x_i(t) - x_i^\Omega(t)|$ can be bounded by a Poisson-distributed random variable \mathcal{Y} with mean $\omega \Omega$, because the arrival and departure rates are both bounded by $\omega > 0$. Thus the following limit exists:

$$\begin{aligned} & \lim_{C \rightarrow \infty} \mathbb{E} \left[(1 + |x_i(t) - x_i^\Omega(t)|)^s \mathbf{1}_{\mathcal{E}_t} | \mathcal{F}_{t-\Omega} \right] \\ & \leq \mathbb{E} [(1 + \mathcal{Y})^s] < \infty. \end{aligned} \quad (40)$$

This implies that for fixed $s \geq 0$, we can pick a sufficiently large $C > 0$ such that $\mathbb{E} [(1 + |x_i(t) - x_i^\Omega(t)|)^s \mathbf{1}_{\mathcal{E}_t} | \mathcal{F}_{t-\Omega}] \leq \frac{\epsilon}{2}$. Plugging this result into (35), we obtain

$$\begin{aligned} & \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i x_i(t) > \xi\}} \right] \\ & \leq 2A_3 + \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\frac{\epsilon}{2} \lambda_i U'_i \left(\frac{c \rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\{\max_i x_i^\Omega(t) > \xi - C\}} \right]. \end{aligned} \quad (41)$$

We can prove a chain of inequalities in (36), which further result in

$$\begin{aligned} & \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\frac{\epsilon}{2} \lambda_i U'_i \left(\frac{c \rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\{\max_i x_i^\Omega(t) > \xi - C\}} \right] \\ & \leq \frac{\epsilon}{2 - \epsilon} \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i x_i^\Omega(t) > \xi - C\}} \right]. \end{aligned} \quad (43)$$

For $\epsilon < \frac{1}{2}$, we combining (33), (41), and (43), and derive upper bound for the right hand side of (38), as in (42). The inequality (38) is immediate from (34) and (42). \blacksquare

From (27) and Lemma 3, we derive an upper bound for the expected drift $\dot{W}(t)$:

$$\begin{aligned} & \dot{W}(t) \\ & \leq A_1 + A_2 + 2A_3 - \frac{\epsilon}{2} \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right] \\ & \leq A_1 + 2A_2 + 2A_3 - \frac{\epsilon}{2} \sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)} \right) \right]. \end{aligned} \quad (44)$$

$$\begin{aligned}
\sum_{i: x_i(t) \geq 1} \mathbb{E} \left[\lambda_i U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{\max_i x_i(t) > \xi\}} \right] &\leq 2A_3 + \left(1 + \frac{\epsilon}{2} + \frac{\epsilon}{2-\epsilon} \right) \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i U_i' \left(\frac{c\rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\xi} \mathbf{1}_{\{\max_i x_i^\Omega(t) > \xi - C\}} \right], \\
&\leq 2A_3 + (1 + \epsilon)^2 \sum_{i: x_i^\Omega(t) \geq 1} \mathbb{E} \left[\lambda_i U_i' \left(\frac{c\rho_i}{x_i^\Omega(t)} \right) \mathbf{1}_{\xi} \mathbf{1}_{\{\max_i x_i^\Omega(t) > \xi - C\}} \right]. \quad (42)
\end{aligned}$$

Hence, by rearranging the terms and integrating (44) from $t = 0$ to $t = T$, we obtain

$$\begin{aligned}
\limsup_{T \rightarrow \infty} \frac{1}{T} \int_{t=0}^T \mathbb{E} \left[\sum_{i: x_i(t) \geq 1} \lambda_i U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \right] dt \\
\leq \limsup_{T \rightarrow \infty} \frac{2W(0)}{T\epsilon} + \frac{2A_1 + 4A_2 + 4A_3}{\epsilon} \\
= \frac{2A_1 + 4A_2 + 4A_3}{\epsilon} \quad (45)
\end{aligned}$$

Since function $U_i'(\cdot)$ is a non-negative and non-decreasing function and $\lim_{z \rightarrow \infty} U_i' \left(\frac{1}{z} \right) = \infty$, equation (45) implies the stability of the network, as claimed in the definition of flow-level stability definition (4). The case when $U_i(\cdot)$ is negative can be shown analogously. ■

C. Proof of Proposition 1.

Proof: To prove that the network is unstable when the information delay is zero and the utility gap is on the same order as that of the optimal utility (6), we construct a counter-example, in which the expected total queue-length grows unbounded as time t increases. Consider a network with two classes of flows and a feasible rate region depicted in Figure 5. For an α -fair utility

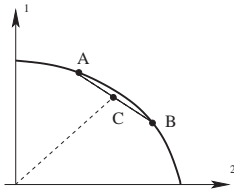


Fig. 5. The feasible rate region under consideration.

with $\alpha = 1/2$, let $\phi_{\text{opt}}(x(t))$ denote the optimal rate allocation for state $x(t)$ at time t . We define a suboptimal rate allocation by

$$\phi(t) = \begin{cases} \phi_{\text{opt}}(x(t)), & \text{if } \phi_{\text{opt}}(x(t)) \text{ does not lie on } \widehat{AB} \\ \phi_A, & \text{otherwise, if } x_1(t) > x_2(t) \\ \phi_B, & \text{otherwise, if } x_1(t) \leq x_2(t) \end{cases} \quad (46)$$

where \widehat{AB} denotes the boundary of the rate region between points A and B , and ϕ_A and ϕ_B are the optimal rate vectors at points A and B respectively. To prove Proposition 1, we notice that $\Delta(x(t)) = 0$ for all $\phi_{\text{opt}}(x(t)) \in \widehat{AB}$. It can be shown that the utility gap is on the same order as the optimal utility, i.e.

$$\begin{aligned}
\limsup_{\max_i x_i(t) \rightarrow \infty} \frac{\Delta(x(t))}{\left| \sum_{i=1}^N x_i(t) U_i \left(\frac{\phi_{\text{opt},i}(t)}{x_i(t)} \right) \right|} \\
= 1 - \liminf_{\max_i x_i(t) \rightarrow \infty} \frac{\sqrt{x_1(t)\phi_1(t)} + \sqrt{x_2(t)\phi_2(t)}}{\sqrt{x_1(t)\phi_{\text{opt},1}(t)} + \sqrt{x_2(t)\phi_{\text{opt},2}(t)}}
\end{aligned}$$

$$\begin{aligned}
&\leq 1 - \liminf_{\max_i x_i(t) \rightarrow \infty} \frac{\sqrt{x_1(t)\phi_{B,1}} + \sqrt{x_2(t)\phi_{A,2}}}{\sqrt{x_1(t)\phi_{A,1}} + \sqrt{x_2(t)\phi_{B,2}}} \\
&\leq 1 - \min \left(\sqrt{\frac{\phi_{B,1}}{\phi_{A,1}}}, \sqrt{\frac{\phi_{A,2}}{\phi_{B,2}}} \right)
\end{aligned}$$

where the second step holds because $\phi_{B,1} \leq \phi_1(t) \leq \phi_{A,1}$ and $\phi_{A,2} \leq \phi_2(t) \leq \phi_{B,2}$ for any rate vector $\phi(t)$ that lies on \widehat{AB} . Hence, the rate allocation policy $\phi(t)$ satisfies condition (6) as claimed.

Next, we choose a point C as the middle point of line \widehat{AB} and show that for small enough $\epsilon > 0$, the network is unstable under traffic intensity $\rho = (1 + \epsilon)\phi_C \in \check{\mathcal{R}}$. Consider a Lyapunov function defined by the weighted sum queue-length

$$V(x) = \mu_1^{-1} w_1 x_1 + \mu_2^{-1} w_2 x_2 \quad (47)$$

where $w_1 = \phi_{B,2} - \phi_{A,2}$ and $w_2 = \phi_{A,1} - \phi_{B,1}$ are two positive constants. Then, formulating the expected Lyapunov function $W(t) = \mathbb{E}[V(x(t))]$ as in Theorem 1, we can show that the expected drift is strictly above zero for traffic intensity $\rho = (1 + \epsilon)\phi_C \in \mathbb{R}$ with a small enough $\epsilon > 0$, i.e.

$$\begin{aligned}
\dot{W}(t) &= \mathbb{E} \left[\sum_{i=1}^2 \frac{\lambda_i}{\mu_i} w_i - w_i \phi_i(t) \right] \\
&= \epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2}) + \mathbb{E}[w_1(\phi_{C,1} - \phi_1(t))] \\
&\quad + \mathbb{E}[w_2(\phi_{C,2} - \phi_2(t))] \\
&\geq \epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2}) > 0
\end{aligned}$$

where the third inequality holds since the suboptimal rate allocation $\phi(t)$ always lies below the straight line AB , whose slope is $-\frac{w_1}{w_2}$. Thus, for the choice of Lyapunov function (47) and the traffic intensity $\rho = (1 + \epsilon)\phi_C \in \mathbb{R}$, the expected drift $\dot{W}(t)$ is strictly above zero by a constant $\epsilon(w_1 \phi_{C,1} + w_2 \phi_{C,2})$. This implies that the network is unstable, since $\lim_{t \rightarrow \infty} W(t) = \infty$ as $t \rightarrow \infty$. ■

D. Proof of Theorem 2.

Proof: We use the same proof technique as in Theorem 1, and show that whenever the traffic condition $\rho \in (1 - \eta)^{\frac{1}{|1-s|}} \check{\mathcal{R}}$ is satisfied, the network is stable under the suboptimal rate allocation policy $\phi(t)$ satisfying (7). Consider the the same Lyapunov function $V(x) = \sum_{i=1}^N \sum_{n=1}^{x_i} U_i \left(\frac{c\rho_i}{n} \right)$ as in the proof of Theorem 1. For the expected Lyapunov function $W(t) = \mathbb{E}[V(x(t))]$, we derive exactly the same drift $\dot{W}(t)$ as in (25). Next, to bound \dot{W} under the utility gap condition (7), we prove the following lemma, which is similar to Lemma 3 in the proof of Theorem 1.

Lemma 4: Consider any traffic intensity $\rho \in (1 - \eta)^{\frac{1}{|1-s|}} \check{\mathcal{R}}$ and constant $C > 0$. If the suboptimal rate allocation $\phi(t)$ satisfies the utility gap condition in (7), there exists positive constants $\gamma > 0$

and $\epsilon > 0$ such that for all $|r| < C$ and for any network state satisfying $\max_i \hat{x}_i(t) > \gamma$, the following inequality holds:

$$\sum_{i:\hat{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t) + r} \right) \leq 0. \quad (48)$$

Proof: Under the traffic intensity $\rho \in (1 - \eta)^{\frac{1}{|1-s|}} \check{\mathcal{R}}$, there exist $\epsilon \geq 0$ and $\delta \geq 0$ such that rate vector $(1 + \epsilon)^4 [1 - (1 + \delta)\eta]^{-\frac{1}{|1-s|}} \rho$ satisfies the feasible rate constraints, i.e.

$$(1 + \epsilon)^4 [1 - (1 + \delta)\eta]^{-\frac{1}{|1-s|}} \rho \in \mathcal{R} \quad (49)$$

According to the utility gap condition in (7), we can conclude that for any $\delta \geq 0$, there exists a positive γ_1 such that for all $x(t)$ satisfying $\max_i \hat{x}_i(t) > \gamma_1$,

$$\Delta(\hat{x}(t)) \leq \eta(1 + \delta) \left| \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)} \right) \right|. \quad (50)$$

To remove the absolute value on the right hand side of (50), we first assume that the utility function is non-negative. Let $u \in \mathcal{R}$ be an arbitrary rate vector and $\delta_0 = [1 - (1 + \delta)\eta]^{-\frac{1}{|1-s|}}$. In view of (2) we obtain the following inequalities, for all $\max_i \hat{x}_i(t) > \gamma_1$,

$$\begin{aligned} 0 &= \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) + \Delta(\hat{x}(t)) \\ &\quad - \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)} \right) \\ &\leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - [1 - (1 + \delta)\eta] \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)} \right) \\ &\leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - [1 - (1 + \delta)\eta] \hat{x}_i(t) U_i \left(\frac{u_i}{\hat{x}_i(t)} \right) \\ &\leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(t)}{\hat{x}_i(t)} \right) - \hat{x}_i(t) U_i \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right) \\ &\leq \sum_{i:\hat{x}_i(t) \geq 1} [\phi_i(t) - \delta_0 u_i] U'_i \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right) \end{aligned}$$

where the third step follows directly from Lemma 1 by letting $a = u_i$ and $b = \delta_0 u_i$, and the last step holds since the utility function $U_i(\cdot)$ is concave. Choosing $u = \frac{(1+\epsilon)^4}{\delta_0} \rho$, we derive

$$\sum_{i:\hat{x}_i(t) \geq 1} U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) [(1+\epsilon)^4 \rho_i - \phi_i(t)] \leq 0. \quad (51)$$

When the optimal utility value in (50) is negative, using the same proof technique and choosing $\delta_0 = (1 + (1 + \delta)\eta)^{-\frac{1}{|1-s|}}$, we can show that the inequality (51) is also satisfied. The rest of proof follows directly from the proof of Lemma 2. We make use of the monotonicity of $U'_i(\cdot)$ and the inequality $U'(a) \geq (\frac{a}{b})^{-s} U'(b)$ (for $a \geq b > 0$). For any $|r| < C$, we get

$$\sum_{i:\hat{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) - \phi_i(t) U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t) + r} \right)$$

$$\begin{aligned} &\leq \sum_{i:\hat{x}_i(t) \geq 1} \rho_i (1 + \epsilon)^3 U'_i \left(\frac{\rho_i}{C} \right) \\ &\quad + \sum_{i:\hat{x}_i(t) \geq 1} \left[\frac{\phi_i(t) K_s C}{\hat{x}_i(t)} - \epsilon \rho_i (1 + \epsilon)^3 \right] U'_i \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \end{aligned} \quad (52)$$

Due to Assumption (c) of the utility function in section II, we conclude (52) $\rightarrow -\infty$ as $\max_i \hat{x}_i(t) \rightarrow +\infty$. This means that there exists a constant $\gamma > \gamma_1$, such that for $\max_i \hat{x}_i(t) > \gamma$, we have (52) ≤ 0 . This completes the proof of the lemma. \blacksquare

Now we choose $c = (1 + \epsilon)^4$ in the Lyapunov function and use Lemma 3 to provide an upper bound for the expected drift $\dot{W}(t)$. The rest of the derivation is exactly the same as the proof of Theorem 1 and thus will not be elaborated here. To conclude, we obtain

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \int_{t=0}^T \mathbb{E} \left[\sum_{i:\hat{x}_i(t) \geq 1} \lambda_i U'_i \left(\frac{c \rho_i}{x_i(t)} \right) \right] dt \leq \frac{2A_1 + 4A_2 + 4A_3}{\epsilon}$$

Because function $U'_i(\cdot)$ is a non-negative and non-decreasing function and $\lim_{z \rightarrow \infty} U'_i \left(\frac{1}{z} \right) = \infty$, the last equation above implies the stability of the network under the traffic intensity $\rho \in (1 - \eta)^{\frac{1}{|1-s|}} \mathcal{R}$. The case when $U_i(\cdot)$ is negative can be shown analogously. \blacksquare

E. Proof of Theorem 3.

Proof: As the first step, we form the Lagrangian for the optimization problem (12) as

$$\mathcal{L}(\phi, p, q) = \sum_{i=1}^N \phi_i + p^T (c - R\phi) + q(V(\phi) + \Delta - U_{\text{opt}})$$

where $V(\phi) = \sum_{i=1}^N x_i^\alpha \phi_i^{1-\alpha} / (1-\alpha)$ is the achievable utility of a rate allocation ϕ . Vector p and scalar q are Lagrangian multipliers for the two constraints in (12) respectively. At the optimal point of (12), the KKT conditions for optimality are given by

$$R\phi = c, \quad V(\phi) = U_{\text{opt}} - \Delta \quad (53)$$

$$R^T p - q \frac{dV(\phi)}{d\phi} - \frac{d(\sum_{i=1}^N \phi_i)}{d\phi} = 0 \quad (54)$$

From the implicit function theorem, variables ϕ , p and q can be viewed as implicit functions of Δ , which is uniquely defined by the KKT conditions (53) and (54). We define a vector $y = [\phi; p; q]$ and a residual

$$G(y, \Delta) = \begin{pmatrix} R^T p - q \frac{dV(\phi)}{d\phi} - \mathbf{1} \\ R\phi - c \\ U_{\text{opt}} - \Delta - V(\phi) \end{pmatrix} \quad (55)$$

where $\mathbf{1}$ is a $N \times 1$ vector consisting of all one's. Then the KKT conditions can be rewritten as $G(y, \Delta) = 0$. The first order derivative of the residual $G(y, \Delta)$ can be obtained as follows

$$\frac{\partial G}{\partial y} = \begin{pmatrix} qD & R^T & -\frac{dV(\phi)^T}{d\phi} \\ R & 0 & \mathbf{0} \\ -\frac{dV(\phi)}{d\phi} & \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} qD & \hat{R}^T \\ \hat{R} & \mathbf{0} \end{pmatrix} \quad (56)$$

and $\frac{\partial G}{\partial \Delta} = (\mathbf{0} \quad \mathbf{0} \quad -1)^T$, where $\hat{R} = [R; -dV(\phi)/d\phi]$ is an extended routing matrix and D is a diagonal matrix of the form

$$D = \alpha \cdot \text{diag} \{[\phi_1^{-\alpha-1}; \dots; \phi_N^{-\alpha-1}]\} \quad (57)$$

Let $e = [0, 0, \dots, 1]^T$. From the implicit function theorem, we obtain

$$\begin{aligned} \frac{dy}{d\Delta} &= - \left(\frac{\partial G}{\partial y} \right)^{-1} \frac{\partial G}{\partial \Delta} \\ &= \begin{pmatrix} qD & \hat{R}^T \\ \hat{R} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ e \end{pmatrix} \\ &= \begin{pmatrix} D^{-1} \hat{R}^T (\hat{R} D^{-1} \hat{R}^T)^{-1} e \\ * \end{pmatrix} \end{aligned} \quad (58)$$

This implies

$$\frac{\partial T}{\partial \Delta} = \frac{d(\sum_{i=1}^N \phi_i)}{d\phi} \cdot \frac{d\phi}{d\Delta} = \mathbf{1}^T D^{-1} \hat{R}^T (\hat{R} D^{-1} \hat{R}^T)^{-1} e$$

Moreover, considering (56), we obtain

$$(\hat{R} D^{-1} \hat{R}^T)^{-1} = H^T \cdot K \cdot H \quad (59)$$

where K and H are axillary matrices given by

$$H = \begin{pmatrix} I & 0 \\ -(\frac{x}{\phi^\alpha})^T & \mathbf{D}^{-1} R^T (R D^{-1} R^T)^{-1} & 1 \end{pmatrix} \quad (60)$$

and

$$K = \begin{pmatrix} (R D^{-1} R^T)^{-1} & 0 \\ 0 & \frac{1}{(\frac{x}{\phi^\alpha})^T \cdot A^{-1} \cdot (\frac{x}{\phi^\alpha})} \end{pmatrix} \quad (61)$$

where the vector division and power x^α/ϕ^α are component-wise. The matrix A is defined as shown in Theorem 3. Plugging the expression for \hat{R} and performing some matrix manipulation, we derive the desired result. ■

F. Proof of Theorem 4.

Proof: we first notice that optimization problem (16) can be rewritten as

$$\begin{aligned} Z(\Delta) &= \min_{\phi} Z \\ \text{s.t.} \quad & R\phi \preceq Zc, \quad \phi \succeq 0 \\ & \sum_{i=1}^N x_i^\alpha \frac{\phi_i^{1-\alpha}}{1-\alpha} \geq U_{\text{opt}} - \Delta \end{aligned} \quad (62)$$

Its Lagrangian is then given by

$$\mathcal{L}(\phi, Z, p, q) = Z + p^T (R\phi - Zc) + q(U_{\text{opt}} - V(\phi) - \Delta)$$

At the optimal point of (62), the KKT conditions for optimality are given by

$$R\phi = Zc, \quad V(\phi) = U_{\text{opt}} - \Delta \quad (63)$$

$$R^T p - q \frac{dV(\phi)}{d\phi} - \frac{d(\sum_{i=1}^N \phi_i)}{d\phi} = 0, \quad p^T c = 1 \quad (64)$$

From the implicit function theorem, variables ϕ , Z , p and q can be viewed as implicit functions of Δ , which is uniquely defined

by the KKT conditions (63) and (64). We define a vector $y = [\phi; p; q; Z]$ and a residual

$$G(y, \Delta) = \begin{pmatrix} R^T p - q \frac{dV(\phi)}{d\phi} \\ R\phi - Zc \\ U_{\text{opt}} - \Delta - V(\phi) \\ 1 - p^T c \end{pmatrix} \quad (65)$$

Then the KKT conditions (63) and (64) are equivalent to $G(y, \Delta) = 0$. From the implicit function theorem, we have

$$\frac{dZ}{d\Delta} = - \left(\frac{\partial G}{\partial y} \right)^{-1} \frac{\partial G}{\partial \Delta}, \quad (66)$$

Plugging $\frac{\partial G}{\partial y}$ and $\frac{\partial G}{\partial \Delta}$ into above and performing some matrix manipulations, we can derive the result in Theorem 4. ■

REFERENCES

- [1] F. P. Kelly, A. Maulloo and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237-252, 1998.
- [2] T. Bonald and L. Massoulié, "Impact of Fairness on Internet Performance," in *Proceedings of ACM Sigmetrics*, pp. 82-91, June 2001.
- [3] T. Bonald, M. Massoulié, A. Proutiere and J. Virtamo, "A Queuing Analysis of Max-Min Fairness, Proportional Fairness and Balanced Fairness," *Special Issue of Queueing Systems: Queueing Models for Fair Resource Sharing*, vol. 53, pp. 65-84, June 2006.
- [4] J. Liu, A. Proutiere, Y. Yi, M. Chiang and H. V. Poor, "Flow-Level Stability of Data Networks with Non-convex and Time-varying Rate Regions", in *Proceedings of ACM Sigmetrics*, pp. 239-250, June 2007.
- [5] H. Q. Ye, "Stability of Data Networks Under an Optimization-Based Bandwidth Allocation," *IEEE Transactions on Automatic Control*, vol. 48, no. 7, pp. 1238-1242, July 2003.
- [6] X. Lin and B. Shroff, "The Impact of Imperfect Scheduling on Cross-Layer Congestion Control in Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 14, no. 2, pp. 302-315, April 2006.
- [7] X. Lin, N. B. Shroff and R. Srikant, "On the Connection-Level Stability of Congestion-Controlled Communication Networks," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2317-2338, May, 2008.
- [8] P. Giaccone, B. Prabhakar and D. Shah, "Towards simple, high-performance schedulers for high-aggregate bandwidth switches", in *Proceedings of IEEE Infocom*, pp. 1160-1169, June 2002.
- [9] R. Bucho and H.J. Kushner, "Control of Mobile Communication Systems With Time-Varying Channels via Stability Methods", in *IEEE Transactions on Automatic Control*, vol.49, no.11, pp.1954-1962, November 2004.
- [10] A. Eryilmaz, R. Srikant and J. Perkins, "Stable Scheduling Policies for Fading Wireless Channels", *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411-424, April 2005.
- [11] A. Tang, J. Wang and S. Low, "Counter-intuitive Behaviors in Networks under End-to-end Control," *IEEE /ACM Transactions on Networking*, vol. 14, no. 2, pp. 355-368, April 2006.
- [12] J. Mo and J. Walrand, "Fair End-to-End Window-based Congestion Control," *IEEE ACM Transactions on Networking*, vol. 8, no. 5, pp. 556-567, October 2000.
- [13] T. Lan, X. Lin, M. Chiang, and R. B. Lee, "How bad is suboptimal rate allocation?," *Proc. IEEE INFOCOM*, Phoenix, AZ, April 2008.
- [14] T. Bonald, S. Borst, N. Hegde and A. Proutiere, "Wireless data networks in multicell scenarios", in *Proceedings of ACM Sigmetrics*, pp. 378-380, June 2004.
- [15] L. Massoulié and J. Roberts, "Bandwidth sharing: objectives and algorithms", *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 320-328, June 2002.
- [16] S. Kunniyur and R. Srikant, "End-to-end congestion control: utility functions, random losses and ECN marks", *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 689-702, October 2003.