Stability and Benefits of Suboptimal Utility Maximization

Tian Lan, *Member, IEEE*, Xiaojun Lin, *Member, IEEE*, Mung Chiang, *Senior Member, IEEE*, and
Ruby B. Lee, *Fellow, IEEE, ACM*

Abstract—Network utility maximization has been widely used to
model resource allocation and network architectures. However, in
practice, often it cannot be solved optimally due to complexity rea-
sons. Thus motivated, we address the following two questions in
this paper: 1) Can suboptimal utility maximization maintain queue
stability? 2) Can underoptimization of utility objective function in
fact benefit other network design objectives? We quantify the fol-
lowing intuition: A resource allocation that is suboptimal with re-
spect to a utility maximization formulation maintains maximum
flow-level stability when the utility gap is sufficiently small and in-
formation delay is bounded, and it 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
reduce link saturation. These results provide a theoretical support
for turning attention from optimal but complex solutions of net-
work optimization to those that are simple even though suboptimal.

Index Terms—Network utility optimization, queueing analysis,
stability analysis.

I. INTRODUCTION

THE FRAMEWORK of network utility maximiza-
tion (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 pub-
lications 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 queueing system where service rate depends on
the resource allocation (e.g., rate control) policy employed by

the network. For convex NUM and under the assumptions of
Poisson arrivals, exponentially distributed files sizes, 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 traffic intensity lies within the rate re-
gion; see, e.g., [4]–[7]. 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. The impact of suboptimal solution is not well
studied in existing literature. It is of practical importance to
sharpen our understanding for two reasons.

- All rate-control algorithms require a nonnegligible amount
  of execution time before they can reach an optimal rate
  allocation that maximizes system utility. Most distributed
  rate-control algorithms cannot reach the optimal rate
  allocation in any finite number of iterations. Alternative
  decompositions also lead to different convergence behav-
  iors [2], [3]. The time for the iterations to converge is
  often so long that network states, e.g., the composition of
  user population, change many times before convergence
  occurs. As a result, practical rate-allocation algorithms
  are subject to a positive, and possibly random, time delay.
  Furthermore, since we cannot afford to run the algorithms
  until it converges, each instance of the utility maximization
  problem can only be solved suboptimally.
- In wireless networks, a scheduling problem has an expo-
  nential computational complexity despite the fact that the
  rate region of the system is convex. For example, when
  the feasible rate region of a network is obtained by time-
  sharing among different subsets of users, a nonconvex mul-
  tiuser/link scheduling problem still needs to be solved in
  order to find the exact rate region achieved by time-sharing
  [8]. Such high computational complexity further increases
  the amount of time that is needed to reach an optimal rate
  allocation. In addition, many cross-layer optimization al-
  gorithms that implement rate schedulers as an inner loop
  require the scheduling iterations to stop at some suboptimal
  point, for example, due to a timescale separation assump-
  tion. There are many theoretical studies that investigate the
  use of low-complexity and even distributed scheduling al-
  gorithms. However, with these scheduling algorithms, the
  rate-allocation algorithm will either take longer conver-
  gence time [9] or will not converge to an optimal rate allo-
  cation [8]. In either case, if we are limited to a finite number
  of iterations, suboptimality in the rate allocation becomes
  the only realistic outcome.

T. Lan was with the Department of Electrical Engineering, Princeton Uni-
versity, Princeton, NJ 08544 USA. He is now with the Department of Electrical
and Computer Engineering, The George Washington University, Washington,
DC 20052 USA (e-mail: tlan@gwu.edu).

X. Lin is with the School of Electrical and Computer Engineering, Purdue
University, West Lafayette, IN 47907 USA.

M. Chiang and R. B. Lee are with Department of Electrical Engineering,
Princeton University, Princeton, NJ 08544 USA.

The gap between elegant theory and useful practice thus leads 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 focus on suboptimal utility maximization, and then quantify effects of information delay and utility gap on flow-level stability and on other important network performance metrics such as link saturation.

In [10], the authors show that for a class of rate-allocation algorithms based on so-called dual solutions, the optimal stability region can be achieved even if an algorithm does not converge to the optimal rate allocation at any time. Similar observations have also been made in switching [11] and scheduling [12] 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 the optimal utility, and 2) the time delay of the queue-length information. We study stability as a function of both utility gap and information delay. Our results apply to a class of general NUM formulations, in which flow-level queueing models are not first-order Markov, thus making our proof technique of independent interest to general flow-level queueing models. 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. However, when the utility gap is proportional 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.

When information delay is bounded, since suboptimal rate allocations with a small enough utility gap are capable of achieving the maximum stability, we investigate the potential *benefits* of allowing such a utility gap, i.e., the upside of underoptimizing utility objectives. It is clear that by deliberately underoptimizing a utility, we can improve network performance in other metrics. What remains unclear is precisely how much improvement we can possibly achieve by underoptimizing utility objectives with a given allowable gap. We formulate the potential performance improvement as a function of utility gap and derive a first-order approximation of the tradeoff curve based on a local sensitivity (shadow price) analysis. This formulation generalizes that in [13], which focuses on how network performance can be affected by the choice of α-fair utilities and assumes that optimality always holds. Our result not only illustrates potential benefits of underoptimizing a utility, but also quantitatively characterizes a *tradeoff* between sacrificing utility value and improving other network performance metrics, e.g., link saturation. Our analysis can be easily extended beyond the class of α-fair utilities.

The results in this paper explore a new perspective to look at suboptimal solutions of utility maximization problems. We show that suboptimal rate-allocation policies may not always be inferior in performance. More precisely, by underoptimizing a utility and allowing a certain optimization gap, we can still retain maximum flow-level stability and obtain network performance improvements in other metrics. The rest of the paper is organized as follows. In Section II, we introduce a class of utility functions considered in this paper and define 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 maximum flow-level stability is provided. In Section III-B, when utility gap is proportional to the maximum utility, we show that the resulting achievable stability region can be strictly smaller, and we further obtain a lower bound for all achievable stability regions. In Section IV, we analyze a tradeoff between utility gap and link saturation. Results based on a sensitivity analysis are derived to measure the benefits of underoptimizing α-fair utility. Simulation results are provided at the end of Sections III and IV, respectively. For smoother flow of the main results, we collect all proofs in the appendixes.

Throughout this paper, we use the following notations. Vectors are denoted in small letters, e.g., x, with their ith component denoted by x_i. Matrices are denoted by capitalized letters, e.g., A, with A_{ij} denoting its $\{i, j\}$th component. Vector inequalities denoted by $x \succeq y$ are considered componentwise. The superscript $(\cdot)^T$ 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 $\check{\mathcal{R}}$ for its interior constructed by removing all Pareto-boundary points.

II. UTILITY MAXIMIZATION AND GAP

Consider a communication network shared by a set of data flows, which belong to N distinct flow classes. We refer to the vector $x = [x_1, \ldots, x_N]$ as the network state, where x_i denotes the number of flows of class i that remain in the system. The problem of rate allocation is to determine the total rate allocated to class-i flows at state x, denoted by $\phi_i(x)$. Rate $\phi_i(x)$ is equally shared by all class-i flows, each assigned a rate $\phi_i(x)/x_i$. We refer to the vector $\phi(x) = [\phi_1(x), \ldots, \phi_N(x)]$ as the rate allocation at state x. Let $\mathcal{R} \subset \mathbb{R}_+^N$ be a set of all possible rate-allocation vectors. Rate allocation $\phi(x)$ is restricted by $\phi(x) \in \mathcal{R}$, which means that the network can support the rate vector $\phi(x)$. In this paper, we only require the set $\mathcal{R}$ to be convex, compact, and coordinate-convex,[1] which holds in many settings, e.g., [5] and [8].

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

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

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

a) $U_i(z) \geq 0 \ \forall z$ and $U_i(0) = 0$, or $U_i(z) \leq 0 \ \forall z, i$.
b) $U_i(z)$ is concave and monotonically increasing.
c) $\lim_{z \to 0} U_i'(z) = \infty, \forall i$.
d) There exists s, s.t. $\frac{z U_i''(z)}{U_i'(z)} \geq -s, \forall z, i$.

---

[1] A set $\mathcal{R}$ is coordinate-convex when the following is true: If $x \in \mathcal{R}$, then $y \in \mathcal{R}$ for all y componentwisely less than x.

Assumptions a) and b) are commonly used in the literature [6]. The assumption that all utility functions take the same sign is a technical condition that is required to remove the absolute value in Condition (7) and prove Lemma 4 (see Appendix-C). Assumption c) can be interpreted as one that prevents starvation since it implies that the slope of the utility function increases to infinity as the rate 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 well-used α-fair utility functions [14], defined by

$$U_i(z) = \begin{cases} \frac{z^{1-\alpha}}{1-\alpha}, & \alpha > 0 \text{ and } \alpha \neq 1 \\ \log z, & \alpha = 1 \end{cases} \quad (2)$$

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. For example, maximizing the α-fair utility corresponds to maximizing weighted throughput as $\alpha \to 0$, weighted proportional fairness as $\alpha = 1$, minimum potential delay as $\alpha = 2$, and max-min fairness as $\alpha \to \infty$ [5].

It may be impractical to solve the rate-allocation problem (1) optimally for all network states. In this paper, we consider a more general scenario where rate allocations are not optimal and thus could possibly reduce network performance. This work is motivated by 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, a practical rate-allocation vector $\phi(\hat{x})$ can at best correspond to the optimal rate allocation $\phi_{\text{opt}}(\hat{x})$ for some vector $\hat{x}$, where $\hat{x} \neq x$ is the network state *observed* by a practical rate-allocation policy. Second, due to computational overhead or requirement for distributed computation, even given the perfect network state $\hat{x}$, a practical rate-allocation policy may still not be able to solve the NUM problem (1) optimally. There may exist a utility gap due to suboptimality of the rate-allocation policy. We quantify suboptimality of a practical rate allocation $\phi(\hat{x})$ with respect to state $\hat{x}$ by a utility gap as follows:

$$\Delta(\hat{x}) = \sum_{i:\hat{x}_i \geq 1} \hat{x}_i U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}_i)}{\hat{x}_i}\right) - \hat{x}_i U_i\left(\frac{\phi_i(\hat{x})}{\hat{x}_i}\right). \quad (3)$$

Utility gap $\Delta(\hat{x})$ measures the difference between suboptimal rate allocations and the optimal allocation caused only by the imperfect computation of a rate-allocation algorithm. Given certain conditions on utility gap $\Delta(\hat{x})$ and a model for observed network states $\hat{x}$, in Section III we will characterize the stability region of an arbitrary suboptimal rate-allocation policy. In Section IV, we will formulate and analyze a tradeoff between utility gap and two network performance metrics.

III. INFORMATION DELAY AND STABILITY

Consider a network where class-i flows arrive according to a Poisson process of intensity $\lambda_i \geq 0$ and have i.i.d. exponential file sizes of mean $1/\mu_i$. A flow is considered to have left the network when its file transfer is completed. Let $\rho_i = \lambda_i/\mu_i$ be the traffic intensity of class-i flows. We formulate a stochastic process of the network state, denoted by $x(t)$. To model observed network state $\hat{x}(t)$ at time t, we introduce an information

delay process $\tau(t)$, such that observed state of class-i flows is given by $\hat{x}_i(t) = x_i(t - \tau_i(t))$, for an information delay $\tau_i(t)$. Since network information may arrive out of order, information delay $\tau_i(t)$ is not necessarily increasing over time t. Therefore, the rate allocation at time t is given by

$$\phi(\hat{x}(t)) = \phi(x_i(t - \tau_i(t))). \quad (4)$$

We say that flow-level stability holds under rate allocations $\phi(\hat{x}(t))$ if there exists a positive nondecaying function $f(\cdot)$ with $\lim_{z \to \infty} f(z) = \infty$, such that the resulting queue-length process $x(t)$ satisfies

$$\limsup_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{E}\left[\sum_{i=1}^N f(x_i(t))\right] dt < \infty. \quad (5)$$

The stability condition in (5) is usually referred to as *stability in the mean* [12], [17]. If we further assume that the rate-allocation policy and the information delay make the queueing system an aperiodic Markov chain with a single communicating class (which is the case where there is no information delay and utility gap), then the *stability in the mean* property in (5) further implies that the Markov chain is positive recurrent [17].

Remark 1: For ease of exposition, in the above model, we have assumed that the actual rate allocation is a given function $\phi(\cdot)$ of observed network states $\hat{x}$. The result of this paper could also be extended to the case where $\phi(\cdot)$ is replaced by a function $\phi_t(\cdot)$ that also varies with time. In that case, $\Delta(\hat{x})$ in (3) can be defined as a supremum of the utility gaps caused by a time-varying $\phi_t(\cdot)$ over t.

When there is no utility gap or information delay, flow-level stability has been studied in [4]–[7] using first-order Markov models. If a feasible rate region $\mathcal{R}$ is compact and convex, and an optimal rate allocation $\phi_{\text{opt}}(x(t))$ that maximizes problem (1) with an α-fair utility is implemented at each time t, it is proven that such a policy achieves the maximum stability region $\check{\mathcal{R}}$ (i.e., the interior constructed by removing all Pareto-boundary points from feasible rate region $\mathcal{R}$). In other words, a sufficient and necessary condition for stability is that the traffic intensity vector must belong to $\check{\mathcal{R}}$.

However, due to information delay, rate allocation $\phi(\hat{x}(t))$ depends on previous network states at time $t - \tau_i(t)$, for $i = 1, \ldots, N$. Therefore, the usual method of flow-level stability analysis in [4]–[6], which requires a first-order Markov model of the queue-length process, is inadequate. In this paper, we consider nonoptimal rate-allocation policies and prove stability by evaluating an expected Lyapunov drift. Let $h > 0$ be a small time interval. The evolution of the ith queue is described by the following equation:

$$x_i(t + h) = x_i(t) + a_i(t, h) - d_i(t, h) \quad (6)$$

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 departing flows. Due to utility gap, the departure rate now depends on the suboptimal rate allocation $\phi(\hat{x}(t))$, given an observed state $\hat{x}(t) = x_i(t - \tau_i(t))$. In this section, we will derive a sufficient condition for achieving maximum stability. When the condition is not satisfied, we prove that the resulting achievable stability region may be strictly smaller than the feasible rate region. The main results are stated in Theorems 1 and 2.

A. Sufficient Condition for Maximum Stability

Theorem 1: For an arbitrary suboptimal rate allocation $\phi(\hat{x})$, if information delay $\tau(t)$ is uniformly bounded by a constant $\Omega > 0$ and the order of utility gap [which is nonnegative due to the definition in (3)] caused by imperfectness of the rate-allocation algorithm is less than that of the optimal utility when the number of active flows grows large, i.e.,

$$\limsup_{\max_i \hat{x}_i \to \infty} \frac{\Delta(\hat{x})}{\left| \sum_{i:\hat{x}_i \geq 1} \hat{x}_i U_i \left(\frac{\phi_{\text{opt},i}(\hat{x})}{\hat{x}_i} \right) \right|} = 0 \qquad (7)$$

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

There are two key difficulties in the proof. First, to account for utility gap, we need to derive a relationship between traffic intensity and the suboptimal rate allocation $\phi(\hat{x}(t))$, which is of a stronger form than those used in [5] and [6] (see Lemma 4 in Appendix-B). Second, information delay leads to a further gap between $\phi(\hat{x}(t))$ and $\phi(x(t))$. We need to carefully bound the effect of this gap, especially when the difference between $\hat{x}(t)$ and $x(t)$ is large (see detailed comment after Lemma 5 in Appendix-B). We refer readers to Appendix-B for the detailed Proof of Theorem 1.

Remark 2: Theorem 1 establishes a sufficient condition for achieving maximum stability: It shows that it is not necessary to solve the optimal solution to the utility optimization problem (1) and to obtain perfect information on network states $x(t)$. Condition (7) could be used to construct stopping rules for designing suboptimal rate-control policies that can achieve maximum stability. We note that such a stopping rule could be designed without knowledge of the optimal rate allocation ϕ_{opt}. For example, consider a rate controller based on a dual algorithm for solving the NUM problem (1). The optimal utility is upper-bounded by the objective value of the dual problem, which can be easily calculated from current dual prices. Similarly, current primal variables can be used to generate a feasible rate allocation and to compute achievable utility values. Therefore, if the controller stops whenever the difference between the dual objective value and the achievable utility values is smaller than a threshold Γ, then we can show that the resulting utility gap (between the optimal utility and the achievable utility value) is bounded by $\Delta(\hat{x}) \leq \Gamma$. Condition (7) is then satisfied since

$$\limsup_{\max_i \hat{x}_i \to \infty} \frac{\Delta(\hat{x})}{\left| \sum_{i:\hat{x}_i \geq 1} \hat{x}_i U_i \left(\frac{\phi_{\text{opt},i}(\hat{x})}{\hat{x}_i} \right) \right|}$$
$$\leq \limsup_{\max_i \hat{x}_i \to \infty} \frac{\Gamma}{\left| \sum_{i:\hat{x}_i \geq 1} \hat{x}_i U_i \left(\frac{\phi_{\text{opt},i}(\hat{x})}{\hat{x}_i} \right) \right|} = 0.$$

Hence, maximum stability region can be achieved by this stopping rule according to Theorem 1. Although the above example assumes a centralized controller to check the stopping rule, we envision that distributed versions of the stopping rule are also possible, which we leave for future work. The condition in (7) is useful because it provides a guideline for designing such stopping rules for general networks.

B. Lower Bound on the Achievable Stability Region

When Condition (7) in Theorem 1 is not satisfied and utility gap is on the same order as that of the optimal utility, the resulting 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 $\phi(\hat{x})$ such that its utility gap is on the same order as the order of the optimal utility and its information delay is zero, i.e., for some constant $\eta \in (0, 1)$

$$\limsup_{\max_i \hat{x}_i \to \infty} \frac{\Delta(\hat{x})}{\left| \sum_{i:\hat{x}_i \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x})}{\hat{x}_i} \right) \right|} \leq \eta \qquad (8)$$

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

Proposition 1 implies that if utility gap is large, there exists a suboptimal rate-allocation policy whose achievable stability region is strictly smaller than $\check{\mathcal{R}}$, regardless of 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 (8) is satisfied?

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

Theorem 2: For an arbitrary suboptimal rate allocation $\phi(\hat{x})$, if information delay is uniformly bounded by a constant and the order of utility gap is the same as that of the optimal utility, i.e.,

$$\limsup_{\max_i \hat{x}_i \to \infty} \frac{\Delta(\hat{x})}{\left| \sum_{i=1}^{N} \hat{x}_i U_i \left(\frac{\phi_{\text{opt},i}(\hat{x})}{\hat{x}_i} \right) \right|} \leq \eta \qquad (9)$$

then the resulting achievable stability region is lower-bounded by $(1 - \eta)^{\frac{1}{1-s}} \check{\mathcal{R}}$, where s is the parameter defined in Assumption d) in Section II. There also exists a suboptimal rate-allocation policy satisfying (9) whose stability region is exactly $(1-\eta)^{\frac{1}{1-s}} \check{\mathcal{R}}$. Therefore, the lower bound is tight.

According to Lemma 1, if a utility function satisfies Assumptions a)–d) and is positive, then we have $U(a) \geq (\frac{a}{b})^{|1-s|} U(b)$, i.e., the utility function $U(\cdot)$ shows a polynomial growth rate with exponent $|1 - s|$. This implies that when the order of utility gap is the same as that of optimal utility, stability depends on s since utility value has a polynomial growth rate with the exponent $|1 - s|$.

Remark 3: Theorem 2 provides a lower bound for achievable stability regions. Of course, under Condition (9), 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 worst case, i.e., there exists a suboptimal rate-allocation policy with zero information delay and its stability region is exactly $(1 - \eta)^{\frac{1}{1-s}} \mathcal{R}$. Proposition 1 and Theorem 2 together characterize the stability of a broad class of suboptimal rate-allocation policies.

C. Numerical Examples

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
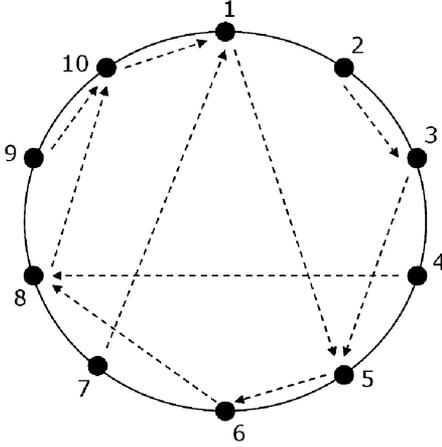
Fig. 1.   Ring network with 10 users and 10 flow classes.

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 $\phi_{\text{opt}}(x(t))$ for each time t and then perturb it randomly to construct a set of suboptimal rate allocations, resulting in constant information delay and utility gap

$$\tau_0 = \tau_i(t) \quad \Delta_0 = \Delta(\hat{x}(t)) \qquad \forall t. \qquad (10)$$

According to Remark 1, since both utility gap and information delay are constants, suboptimal rate-allocation policy $\phi(\,\cdot\,)$ constructed above will achieve the maximum stability region, which is constructed by link-capacity constraints $\mathcal{R} = \{\phi \in \mathbb{R}_+^N : R\phi < \mathbf{1}, \phi \succeq 0\}$, where R is $L \times N$ routing matrix: $R_{li} = 1$ if class-i flows use link l, and $R_{li} = 0$ otherwise. The rate region is a polytope determined by a set of linear link-capacity constraints.

Fig. 2 illustrates flow-level stability of the ring network under different suboptimal rate-allocation policies for $(\Delta_0, \tau_0) = \{(0,0), (5,0), (0,2), (5,2)\}$, respectively, by plotting average total queue length versus traffic load. In this simulation, we assume that flow arrival rates for all flow classes are equal, i.e., $\rho_i = \rho_0$ for $i = 1, \ldots, 10$. 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. Fig. 2 also shows that average queue length of a suboptimal policy $\phi(\hat{x}(t))$ approaches that of the optimal rate-allocation policy when utility gap and information delay go down to zero.

IV. UTILITY GAP AND NETWORK PERFORMANCE

Section III showed that a suboptimal rate-allocation policy may still achieve the maximum stability region. Since each utility function is designated to capture one particular network objective, allowing a nonzero utility gap (or, equivalently, underoptimizing a utility) gives us freedom to potentially improve other network performance objectives, such as the maximum link saturation discussed in this section. More precisely, there exists a tradeoff between the 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 an approximation of the tradeoff curve based on local sensitivity analysis. To obtain closed-form solutions, we focus on α-fair utility functions in this section, although our result can be extended to general concave utility functions.
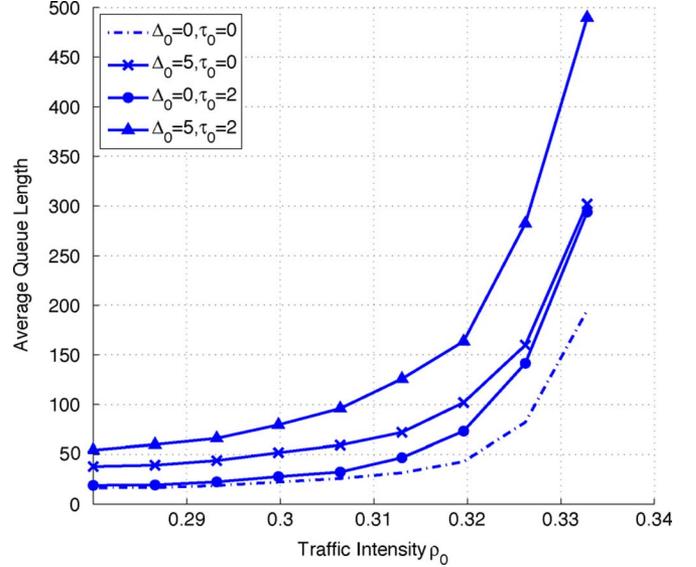

Fig. 2.   Average total queue length of the ring network for four different rate-allocation policies. It is shown that the three suboptimal rate-allocation policies with constant utility gap and information delay still stabilize the network for traffic intensity $\rho_0 < \frac{1}{3}$, although their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

Our approach is different from [13], which is restricted to a throughput-fairness tradeoff for *optimal* solutions only. In contrast, in this section, we address the following question pertaining to *suboptimality*: By deliberately underoptimizing a utility with gap Δ, what is the maximum performance improvement we can possibly achieve?

A. Model and Analysis

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

$$\max_{\phi} \quad \sum_{i=1}^{N} x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha}$$
$$\text{s.t.} \quad R\phi \preceq c, \ \phi \succeq 0. \qquad (11)$$

Let ϕ_{opt} be the optimal rate allocation that solves the maximization problem (11). We say that a rate allocation ϕ underoptimizes an α-fair utility by a gap Δ if

$$\Delta = U_{\text{opt}} - \sum_{i=1}^{N} x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha} \qquad (12)$$

where $U_{\text{opt}} = \sum_{i=1}^{N} x_i^{\alpha} \phi_{\text{opt},i}^{1-\alpha} / (1-\alpha)$ is the optimal utility value achieved by rate allocation ϕ_{opt}. Since α-fair utility is designated for achieving fairness, underoptimizing it with a gap Δ relaxes maximization problem (11). Thus, it gives freedom to potentially improve other network performance objectives, such as maximum link saturation. However, it is unclear how much

performance improvement we can achieve by underoptimizing an α-fair utility with a given allowable gap. For example, if we prepare to sacrifice utility by 5%, how much reduction of link saturation can be expect in return? We formulate this tradeoff and provide a local sensitivity analysis based on examining the Karush–Kuhn–Tucker (KKT) conditions.

We consider maximum link saturation defined by

$$Z = \max_{l \in L} \frac{\sum_i R_{il} \phi_i}{c_l} \qquad (13)$$

as a target network performance metric. By underoptimizing a utility, it is possible to reduce maximum link saturation and balance network traffic over all links, an important goal in the operation of large networks by the Internet service providers. Moreover, reducing the maximum link saturation could potentially minimize the occurrence of "bottleneck" links in a network, reduce packet delay, and make the network more robust to link-capacity fluctuation and traffic bursts.

To characterize the optimal tradeoff between utility gap and maximum link saturation, we compute the minimum Z that can be achieved by underoptimizing an α-fair utility with a designated utility gap. This tradeoff is formulated as function $Z(\Delta)$ as follows:

$$Z(\Delta) = \min_{\phi} \quad \max_{l \in L} \frac{\sum_i R_{il} \phi_i}{c_l}$$
$$\text{subject to} \quad R\phi \preceq c, \ \phi \succeq 0$$
$$\sum_{i=1}^{N} x_i^{\alpha} \frac{\phi_i^{1-\alpha}}{1-\alpha} \geq U_{\text{opt}} - \Delta. \qquad (14)$$

Remark 4: For the tradeoff defined by (14), it is easy to see that increasing utility gap relaxes its constraint set and leads to a smaller optimal objective value. Thus, maximum link saturation $Z(\Delta)$ is monotonically decreasing over utility gap Δ. Furthermore, it is easy to verify that the optimization problem (14) is convex.

Now, we conduct a local sensitivity analysis for the saturation-utility tradeoff defined by (14). Again, we make the assumption that the active constraint set in (14) is unchanged when utility gap Δ is perturbed locally. The main result is summarized in the next theorem. We denote Z_0 as the link saturation at $\Delta = 0$.

Theorem 3: When the utility gap is small, the saturation-utility tradeoff can be approximated using its first-order expansion

$$Z - Z_0 = \left[\frac{dZ}{d\Delta} \bigg|_{\Delta=0} \right] \Delta + o(\Delta). \qquad (15)$$

The first-order derivative (shadow price) is given by

$$\frac{dZ}{d\Delta} \bigg|_{\Delta=0} = -\frac{1}{c^T (RD^{-1}R^T)^{-1} c} \qquad (16)$$

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

B. Numerical Examples

Here, we plot the saturation-utility tradeoff curve and its first-order approximation obtained in Theorem 3 for the ring network described in Section III-C. Since all links have unit capacities, the feasible rate region is given by $\mathcal{R} = \{\phi : R\phi \leq \mathbf{1}, \phi \succeq \mathbf{0}\}$,
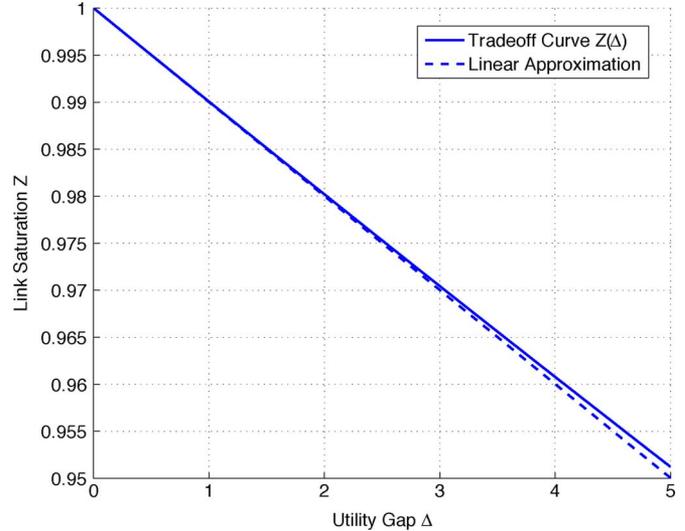

Fig. 3. Saturation-utility tradeoff curve and its first-order approximation.

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 convex optimization problem (14) for different utility gap Δ to obtain the exact tradeoff curve $Z(\Delta)$, which is plotted in Fig. 3 using solid lines. The number of active flows is chosen to be $x_i = 10, \forall i$. A proportional fairness utility function corresponding to $\alpha = 1$ is employed.

When utility gap Δ is small, the saturation-utility tradeoff can be approximated by its first-order expansions in (15). Using the closed-form solution in Theorem 3, we compute the first-order gradient $\frac{dZ}{d\Delta}|_{\Delta=0} = -0.010$. Thus, the tradeoff curve can be approximated by

$$T(\Delta) \approx Z(0) - 0.01\Delta. \qquad (17)$$

Fig. 3 shows that the saturation-utility tradeoff defined in (14) can be well approximated by its first-order expansion, given by the closed-form expression in Theorem 3. This tradeoff curve allows us to predict how much performance improvement we can possibly achieve by underoptimizing the utility with a designated small utility gap. For example, if we underoptimize the utility by 1%, i.e., $\Delta = 1\%|U_{\text{opt}}| = 0.312$, it is clear from (17) that a link saturation reduction of 0.31% ($Z - Z_0 = -0.0031$) could be expected in return. This result not only illustrates the potential benefits of underoptimizing an α-fair utility, but also quantitatively characterizes a tradeoff between sacrificing utility value and achieving network performance improvement. Whether this particular tradeoff is worth making or not depends on operator's preference, but it is important to provide the choices of tradeoffs through 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-saturation tradeoff are quantified with closed-form expressions in this paper. There are open questions to be addressed in studying the impact of suboptimal solutions to network optimization, e.g., in characterizing degradation of fairness due to

utility gap and in conducting a global sensitivity analysis, before we fully understand "how bad suboptimal rate allocation is."

<center>APPENDIX
PROOFS</center>

From Assumptions a)–d) in Section II, in Appendix-A we first prove two lemmas, which are useful throughout the rest of the proofs. Appendix-B gives the proof of stability in Theorem 1. It makes use of Lemma 3 (whose proof is omitted due to space limitation and can be found in [15]) and Lemmas 4–6 (which are proven in Appendix-C–E, respectively). The Proof of Theorem 2 in Appendix-G is very similar to that of Theorem 1, except that Lemma 4 is replaced by Lemma 7 under a different utility gap condition. Finally, the saturation-utility tradeoff is proven in Appendix-H.

A. Useful Lemmas

We first collect three useful properties, which will be used to bound the difference between rate allocations by their utility values or first-order derivatives.

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

Proof: To prove part i), from Assumption d), we have $\frac{U''(z)}{U'(z)} \geq -\frac{s}{z}$. Choose $a \geq b > 0$ and integrate both sides of the inequality from b to a. We obtain $U'(a) \geq (\frac{a}{b})^{-s} U'(b)$. By switching a and b, we can prove part ii). To show part iii), if the utility function is negative [i.e., case 2 in Assumption a)], we fix b in the inequality in part i) and integrate it from $a = b$ to $a = +\infty$, i.e.,

$$b^s U'(b) \int_b^{+\infty} \frac{1}{y^s} dy \leq U(\infty) - U(b) \leq -U(b) \qquad (18)$$

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 from a to $b \geq a$, 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. ∎

Lemma 2: For $C \geq 0$, there exist a constant K_c such that $(1+\frac{C}{z})^s \leq 1+\frac{K_c C}{z}$ holds for all $z \geq 1$, and $(1-\frac{C}{z})^s \geq 1-\frac{K_c C}{z}$ holds for all $z \geq C$.

Proof: The proof is straightforward by comparing the first-order derivatives with respect to z. ∎

B. Proof of Theorem 1

Proof: We first sketch the main steps of the proof. To prove stability, we define a Lyapnov function $V(x(t))$ and analyze its

expectation $W(t) = \mathbb{E}[V(x(t))]$ as a function of time.[2] We first derive an expression for $\dot{W}(t)$, the drift of the expected value of the Lyapunov function. Here, we need to use the Dominated Convergence Theorem in order to exchange the order of a limit and an expectation (Lemma 3). Then, using Lemmas 4–6, we upper-bound the drift $\dot{W}(t)$ by a negative function of the network state $x(t)$, plus some positive constants. Finally, integrating the drift $\dot{W}(t)$ and its upper bound from time $t = 0$ to $t = T$ establishes the stability condition in (5) and completes the proof. For ease of presentation, in this section we present the main flow of the proof with statements of Lemmas 3–6. The detailed proofs are summarized in Appendix-C–F. We remind the readers that we will use the lemmas in Appendix-A.

Consider the following Lyapunov function:

$$V(x(t)) = \sum_{i=1}^{N} V_i(x_i(t)) = \sum_{i=1}^{N} \sum_{n=1}^{x_i(t)} \frac{1}{\mu_i} U_i' \left(\frac{c\rho_i}{n}\right) \qquad (19)$$

where $c > 0$ is a constant defined later in the proof. Let $\mathcal{F}_t = \sigma\{x(u), u \leq t\}$ denote the σ-field generated by the history up to time t. For $h > 0$, we derive

$$\lim_{h \to 0} \frac{W(t+h) - W(t)}{h}$$
$$= \lim_{h \to 0} \frac{\mathbb{E}[\mathbb{E}[V(x(t+h)) \mid \mathcal{F}_t] - V(x(t))]}{h}$$
$$= \lim_{h \to 0} \mathbb{E}\left[\sum_{i=1}^{N} \sum_{n=-X_i(t)}^{+\infty} \frac{\mathbb{P}(x_i(t+h) - x_i(t) = n \mid \mathcal{F}_t)}{h\mu_i} \right.$$
$$\left. \cdot (V_i(x_i(t) + n) - V_i(x_i(t))) \right]. \qquad (20)$$

In order to move the limit (as $h \to 0$) inside the expectation and the summation on the right-hand side of (20), we will make use of the Dominated Convergence Theorem, which requires the following lemma.

Lemma 3: There exists an integrable function $g(x_i(t))$ such that, for all $0 < h < 1$ and $m \geq 0$

$$\left| \sum_{n=-X_i(t)}^{+\infty} \frac{\mathbb{P}(x_i(t+h) - x_i(t) = n \mid \mathcal{F}_t)}{h\mu_i} \right.$$
$$\left. \cdot (V_i(x_i(t) + n) - V_i(x_i(t))) \right| \leq g(x_i(t)). \qquad (21)$$

This lemma provides the bound needed for the Dominated Convergence Theorem to hold. Its proof can be found in [15]. We can move the limit inside the expectation on the right-hand side of (20). Due to the orderliness property of Poisson process (which implies that arrivals do not occur simultaneously), we can take the limit (as $h \to \infty$) and narrow down the conditional probability terms in (20) to get

$$\dot{W}(t) = \sum_{i=1}^{N} \mathbb{E}[\lambda_i(V_i(x_i(t) + 1) - V_i(x_i(t)))]$$
$$- \sum_{i=1}^{N} \mathbb{E}[\phi_i(t)\mu_i(V_i(x_i(t)) - V_i(x_i(t) - 1))].$$

---

[2]A similar problem with feedback delay is treated in [12], although the model there does not involve any flow-level dynamics.

Using $\rho_i = \lambda_i/\mu_i$ and plugging in the Lyapunov in (19) into the above function, we obtain

$$\dot{W}(t) = \sum_{i=1}^{N} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\right] \tag{22}$$

$$-\sum_{i=1}^{N} \mathbb{E}\left[\phi_i(\hat{x}(t))U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x_i(t)\geq1\}}\right]. \tag{23}$$

In the following proof, we will bound (22) and (23) separately.

Derive a bound for (23). In order to bound $\dot{W}(t)$ by a negative function of $x(t)$, we first derive a bound for the second summation (23) and replace the rate allocation $\phi(\hat{x}(t))$ by a function of traffic intensity ρ_i. The resulting bound will then have a form similar to (22) so that we can compare the difference. This bound for (23) is stated in the following lemma:

Lemma 4: For any traffic intensity $\rho \in \check{\mathcal{R}}$ and constant $C > 0$, if the suboptimal rate allocation $\phi(\hat{x}(t))$ satisfies the utility gap condition in (7), there exist 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:

$$\sum_{i:\hat{x}_i(t)\geq1} \rho_i(1+\epsilon)^3 U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) - \phi_i(\hat{x}(t))U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)+r}\right) \leq 0 \tag{24}$$

where $c = (1 + \epsilon)^4$ is a constant.

We choose $r = x(t) - \hat{x}(t)$ in (24). Note that in order to use Lemma 4 to bound (23), we need the network state to satisfy $\max_i \hat{x}_i(t) > \gamma$ and the quantity $|x(t) - \hat{x}(t)|$ to be bounded by C. Toward this end, we define the event $\mathcal{E}_t = \{\|x(t - u) - x(t - \Omega)\|_1 < C/2, \forall u \in [0, \Omega]\}$, i.e., it is the event that the maximum change of network state within time $t - \Omega$ to t is bounded by $C/2$. Since the information delay is bounded by $\tau(t) < \Omega$, event $\mathcal{E}_t$ implies the following:

$$\|x(t) - \hat{x}(t)\|_1 \leq \|x(t) - x(t-\Omega)\|_1 + \|\hat{x}(t) - x(t-\Omega)\|_1$$
$$\leq C.$$

Therefore, we can bound (23) by

$$-\sum_{i=1}^{N} \mathbb{E}\left[\phi_i(\hat{x}(t))U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x_i(t)\geq1\}}\right]$$

$$\leq A_1 - \sum_{i:\hat{x}_i(t)\geq1} \mathbb{E}\left[\phi_i(\hat{x}(t))U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{E}_t}\mathbf{1}_{\{\max_i \hat{x}_i(t)>\gamma\}}\right]$$

$$\leq A_1 - (1+\epsilon)^3 \sum_{i:\hat{x}_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)\mathbf{1}_{\mathcal{E}_t}\mathbf{1}_{\{\max_i \hat{x}_i(t)>\gamma\}}\right]. \tag{25}$$

Note that $\hat{x}_i(t) \leq C$ if $x_i(t) = 0$ under event $\mathcal{E}_t$. To change the summation from $\{i : x_i(t) \geq 1\}$ to $\{i : \hat{x}_i(t) \geq 1\}$ in the first step, we have added

$$A_1 = \sum_{i=1}^{N} \psi U_i'\left(\frac{c\rho_i}{C}\right) \tag{26}$$

where we choose $\phi_i(\hat{x}(t)) \leq \psi$ since the feasible rate region $\mathcal{R}$ is compact. The last step of (25) follows from Lemma 4. In order to compare the difference to (22), we still need to replace the $\hat{x}(t)$ on the right-hand side by $x(t)$. Let $\mathcal{G} = \{x(t) :$

$\max_i x_i(t) \leq \xi\}$ be a bounded region with $\xi = \gamma + 2C$. We can prove the following result, which further bounds the last step (25) by a term similar to (22).

Lemma 5: There exists $A_2, A_3 > 0$ such that the right-hand side of (25) can be further bounded by

$$-(1+\epsilon)^2 \sum_{i:\hat{x}_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)\mathbf{1}_{\mathcal{E}_t}\mathbf{1}_{\{\max_i \hat{x}_i(t)>\gamma\}}\right]$$

$$\leq A_2 + 4A_3 - \sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right]. \tag{27}$$

The intuition behind Lemma 5 is that when $\mathcal{E}_t$ occurs, the absolute difference between $x(t)$ and $\hat{x}(t)$ is bounded by C. If in addition the network state is large, then the relative difference between $x(t)$ and $\hat{x}(t)$ will be small, and hence the corresponding values of $U_i'(\cdot)$ will be close to each other. Proving Lemma 4 turns out to be nontrivial. The challenge is that $\hat{x}(t)$ and $\mathcal{E}_t$ in (27) are dependent. To handle this difficulty, we consider the σ-field $\mathcal{F}_{t-\Omega} = \sigma\{x(u), u \leq t - \Omega\}$. If we introduce the network state $x^\Omega(t) = x(t-\Omega)$ as an auxiliary variable, then $\hat{x}(t)$ and $\mathcal{E}_t$ can be bounded with respect to $x^\Omega(t)$, separately. For details, please refer to Appendix-D.

Derive a bound for (22). We next provide a corresponding bound for (22) to compare to (27). Note that region $\mathcal{G}$ is compact. As part of the Proof of Lemma 5, A_2 is chosen in (49) such that

$$\sum_{i=1}^{N} \rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}\}} \leq A_2 < \infty. \tag{28}$$

Applying A_2 to (22), we have

$$\sum_{i=1}^{N} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\right]$$

$$\leq A_1 + \sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\right]$$

$$\leq A_1 + A_2 + \sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right] \tag{29}$$

where A_1 is added to change the summation to $\{i : x_i(t) \geq 1\}$. To further bound (29), we make use of Lemma 1-(ii), with $a = c\rho_i/x(t)$ and $b = c\rho_i/(x(t)+1)$

$$\sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)+1}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right]$$

$$\leq \sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i\left(1+\frac{1}{x_i(t)}\right)^s U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right]$$

$$\leq \sum_{i:x_i(t)\geq1} \mathbb{E}\left[\rho_i\left(1+\frac{K_1}{x_i(t)}\right) U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right] \tag{30}$$

where K_1 in the last step is the constant defined in Lemma 2. Finally, we use the following result to further bound the last step of (30).

Lemma 6: For any $K_1, C, \epsilon > 0$ there exists ξ such that, for any $\max_i \hat{x}_i(t) > \xi$, we have

$$\sum_{i:\hat{x}_i(t) \geq 1} \rho_i \left(\frac{K_1 C}{\hat{x}_i(t)} - \frac{\epsilon}{2} \right) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t)} \right) \leq 0. \qquad (31)$$

Combining Lemma 6 for $C = 1$ with (29) and (30), we derive the following upper bound for (22):

$$\sum_{i=1}^{N} \mathbb{E} \left[\rho_i U_i' \left(\frac{c\rho_i}{x_i(t) + 1} \right) \right]$$
$$\leq A_1 + A_2 + \sum_{i:x_i(t) \geq 1} \mathbb{E} \left[\rho_i \left(1 + \frac{\epsilon}{2} \right) U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right] \qquad (32)$$

Prove stability. We define another constant A_0 by

$$A_0 = 2A_1 + (2 + \epsilon) A_2 + (4 + 4\epsilon) A_3.$$

Replacing (22) and (23) by their upper bounds, we can bound the expected drift $\dot{W}(t)$ by

$$\dot{W}(t) \leq A_0 - \frac{\epsilon}{2} \sum_{i:x_i(t) \geq 1} \mathbb{E} \left[\rho_i U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \mathbf{1}_{\{x(t) \in \mathcal{G}^c\}} \right]$$
$$\leq A_0 + A_2 - \frac{\epsilon}{2} \sum_{i:x_i(t) \geq 1} \mathbb{E} \left[\rho_i U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \right]. \qquad (33)$$

where inequality (28) is used in the last step and a constant A_2 is added. Rearranging the terms and integrating (33) from $t = 0$ to $t = T$, we obtain

$$\limsup_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} \mathbb{E} \left[\sum_{i:x_i(t) \geq 1} \rho_i U_i' \left(\frac{c\rho_i}{x_i(t)} \right) \right] dt$$
$$= \limsup_{T \to \infty} \frac{2W(0) - 2W(T)}{T\epsilon} + \frac{2A_0 + 2A_2}{\epsilon}$$
$$\leq \frac{2A_0 + 2A_2}{\epsilon} \qquad (34)$$

where the last step uses the fact that $W(T) \geq 0$ is positive. Since function $U_i'(\cdot)$ is a nonnegative and nondecreasing, and $\lim_{z \to \infty} U_i'(1/z) = \infty$, (34) implies the stability of the network, as claimed in the stability definition (5). ∎

C. Proof of Lemma 4

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

$$\Delta(\hat{x}(t)) \leq \delta \left| \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)} \right) \right|. \qquad (35)$$

To remove the absolute value on the right-hand side of (35), we first assume that the utility function is nonnegative. Plugging the expression of utility gap $\Delta(\hat{x}(t))$ into (35), we have

$$0 \leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)} \right)$$
$$- (1 - \delta) \hat{x}_i(t) U_i \left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)} \right). \qquad (36)$$

Since $\phi_{\text{opt},i}(\hat{x}_i(t))$ is the optimal rate allocation for the NUM problem (1) at state $\hat{x}_i(t)$, no rate vector $u \in \mathcal{R}$ can achieve a higher utility value. Choose $\epsilon > 0$ and $\delta > 0$ such that

$$u \triangleq (1+\epsilon)^4 (1-\delta)^{-\frac{1}{|1-s|}} \rho \in \mathcal{R}. \qquad (37)$$

Such $\epsilon > 0$ and $\delta > 0$ exist due to the traffic condition $\rho \in \check{\mathcal{R}}$. Let $\delta_0 = (1-\delta)^{\frac{1}{|1-s|}}$. We then have

$$0 \leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)} \right) - (1-\delta) \hat{x}_i(t) U_i \left(\frac{u_i}{\hat{x}_i(t)} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq 1} \hat{x}_i(t) U_i \left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)} \right) - \hat{x}_i(t) U_i \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq 1} [\phi_i(\hat{x}(t)) - \delta_0 u_i] U_i' \left(\frac{\delta_0 u_i}{\hat{x}_i(t)} \right)$$
$$= \sum_{i:\hat{x}_i(t) \geq 1} [\phi_i(\hat{x}(t)) - (1+\epsilon)^4 \rho_i] U_i' \left(\frac{(1+\epsilon)^4 \rho_i}{\hat{x}_i(t)} \right) \qquad (38)$$

where the second step uses Lemma 1-(iii) with $a = u_i/\hat{x}_i(t)$ and $b = \delta_0 u_i/\hat{x}_i(t)$, the third step uses the concavity of the utility function $U_i(\cdot)$, and the last step is from (37). Note that a similar expression for (38) can also be shown if the utility function is negative. Specifically, when the utility value in (35) is negative, using the same proof technique and choosing $\delta_0 = (1+\delta)^{-\frac{1}{|1-s|}}$, we can show that the inequality (38) is also satisfied. Refer to [15] for the proof.

Note that (38) is almost the same as (24), except for a constant r in the denominator and an additional $(1+\epsilon)$ factor. Let $c = (1+\epsilon)^4$. For $|r| < C$, we have

$$\sum_{i:\hat{x}_i(t) \geq 1} -\phi_i(\hat{x}(t)) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t) + r} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq C} -\phi_i(\hat{x}(t)) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t) + r} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq C} -\phi_i(\hat{x}(t)) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t) - C} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq C} -\phi_i(\hat{x}(t)) \left(1 - \frac{C}{\hat{x}_i(t)} \right)^s U_i' \left(\frac{c\rho_i}{\hat{x}_i(t)} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq C} -\phi_i(\hat{x}(t)) \left(1 - \frac{K_c C}{\hat{x}_i(t)} \right) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t)} \right)$$
$$\leq \sum_{i:\hat{x}_i(t) \geq C} - \left((1+\epsilon)^4 \rho_i - \frac{K_c C \phi_i(\hat{x}(t))}{\hat{x}_i(t)} \right) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t)} \right) \qquad (39)$$

In the first step of (39) some negative terms are dropped. The second step makes use of the monotonicity of $U_i'(\cdot)$. The third step uses Lemma 1-(i). The fourth step uses Lemma 2. The last step is from (38). Since $\phi_i(\hat{x}(t)) \leq \psi$, using the proof in Lemma 6, we can show that for large enough γ and $\max_i \hat{x}_i(t) > \gamma$

$$\sum_{i:\hat{x}_i \geq C} - \left((1+\epsilon)^3 \epsilon \rho_i - \frac{K_c C \psi}{\hat{x}_i(t)} \right) U_i' \left(\frac{c\rho_i}{\hat{x}_i(t)} \right) \leq 0. \qquad (40)$$

Applying (40) to (39), we derive

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

which completes the Proof of Lemma 4. ∎

D. Proof of Lemma 5

Proof: To prove (27), we use network state $x^\Omega(t) = x(t-\Omega)$ at time $t-\Omega$ as an auxiliary variable and bound both sides of (27) with respect to $x^\Omega(t)$, respectively.

Bound the left-hand side of (27). Define the event $\mathcal{X}_t^\Omega = \{\max_i x_i^\Omega(t) > \xi - C\}$, i.e., it is the event that the maximum queue length at time $t-\Omega$ is greater than $\xi - C$. We start with

$$\sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq A_3 + \sum_{i:\hat{x}_i(t)\geq 1, x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right] \quad (41)$$

where a constant A_3 is added to change the summation from $\{i : x_i^\Omega(t) \geq 1\}$ to $\{i : \hat{x} \geq 1\}$. Such A_3 satisfies

$$\sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\{\hat{x}_i(t)=0\}}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i \left|x_i^\Omega(t)\right|^s U_i'\left(\frac{c\rho_i}{1}\right) \mathbf{1}_{\{\hat{x}_i(t)=0\}}\right]$$

$$= \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i \left|\hat{x}_i(t) - x_i^\Omega(t)\right|^s U_i'\left(\frac{c\rho_i}{1}\right) \mathbf{1}_{\{\hat{x}_i(t)=0\}}\right]$$

$$\leq A_3 < +\infty \quad (42)$$

where A_3 in the last step exists because $|\hat{x}_i(t) - x_i^\Omega(t)|$ can be bounded by a Poisson random variable with mean $\omega\Omega$. The second step uses Lemma 1-(ii). To proceed from (41), we have

$$\sum_{i:\hat{x}_i(t)\geq 1, x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)+C}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i \left(1+\frac{C}{\hat{x}_i(t)}\right)^s U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i \left(1+\frac{K_c C}{\hat{x}_i(t)}\right) U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i \left(1+\frac{\epsilon}{2}\right) U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]. \quad (43)$$

The first step uses the monotonicity of $U'(\cdot)$ and the condition $|x_i^\Omega(t) - \hat{x}_i(t)| < C$ under event $\mathcal{E}_t$. The second step above uses Lemma 1-(ii). The third step is due to $(1+\frac{C}{z})^s \leq 1 + \frac{K_c C}{z}$ in Lemma 2. The last step follows from the Proof of Lemma 6.

Applying the result to (41), we derive an upper bound for the left-hand side of (27) as follows:

$$\sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right] - A_3$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i(1+\frac{\epsilon}{2}) U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i(1+\frac{\epsilon}{2}) U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \mathbf{1}_{\mathcal{E}_t} \mathbf{1}_{\{\max_i \hat{x}_i(t)\geq\gamma\}}\right]$$

$$(44)$$

where the last step uses $\xi = \gamma + 2C$ and

$$\mathcal{E}_t \cap \mathcal{X}_t^\Omega \subset \left\{\left\|x_i^\Omega(t) - \hat{x}_i(t)\right\|_1 < \frac{C}{2}, \max_i x_i^\Omega(t) > \xi - C\right\}$$

$$\subset \{\max_i \hat{x}_i(t) > \gamma\}. \quad (45)$$

Derive a bound for the right-hand side of (27). Define the event $\mathcal{X}_t = \{x(t) \in \mathcal{G}^c\} = \{\max_i x_i(t) > \xi\}$, i.e., it is the event that the maximum queue length at time t is greater than ξ. Using the argument for (41) and (43), we first change the summation on the right-hand side of (27) from $\{i : x_i(t) \geq 1\}$ to $\{i : x_i^\Omega(t) \geq 1\}$ by adding A_3 [which is defined in (42)]

$$\sum_{i:x_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right) \mathbf{1}_{\mathcal{X}_t}\right]$$

$$\leq A_3 + \sum_{i:x_i^\Omega(t)\geq 1, x_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right) \mathbf{1}_{\mathcal{X}_t}\right]. \quad (46)$$

To bound (46), we define the event $\mathcal{L}_i = \{x_i(t) > x_i^\Omega(t)\}$ and evaluate (46) over three events: $\mathcal{E}_t^c \cap \mathcal{L}_i, \mathcal{E}_t^c \cap \mathcal{L}_i^c$, and $\mathcal{E}_t$, separately. For $\mathcal{E}_t^c \cap \mathcal{L}_i$, note that $\mathcal{E}_t^c \cap \mathcal{L}_i$ implies that the increase of $x(t)$ in the interval $t-\Omega$ to t is larger than C. The number of arrivals in this interval must also be larger than C. Since the arrival process is Poisson, when C is large, we can bound the probability of this event by a small number. Hence, the contribution to the expectation in (46) will be small. Specifically, let $a_i(t)$ be the arrival within time $t-\Omega$ to t. We have

$$\sum_{i:x_i^\Omega(t)\geq 1, x_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right) \mathbf{1}_{\mathcal{X}_t} \mathbf{1}_{\mathcal{E}_t^c} \mathbf{1}_{\mathcal{L}_i}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1, x_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right) \mathbf{1}_{\mathcal{E}_t^c} \mathbf{1}_{\mathcal{L}_i}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i \left(\frac{x_i(t)}{x_i^\Omega(t)}\right)^s U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t^c} \mathbf{1}_{\mathcal{L}_i}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i \left(1+\frac{a_i(t)}{x_i^\Omega(t)}\right)^s U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t^c} \mathbf{1}_{\mathcal{L}_i}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i(1+a_i(t))^s U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathbf{1}_{\mathcal{E}_t^c} \mathbf{1}_{\mathcal{L}_i}\right]$$

$$\leq \sum_{i:x_i^\Omega(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right) \mathcal{K}\right]$$

$$(47)$$

where the second step uses Lemma 1-(ii), the third step uses the monotonicity of $U'(\cdot)$ and $x_i(t) - x_i^\Omega(t) \leq a_i(t)$ (since $a_i(t)$

is the arrival within time $t - \Omega$ to t), and the quantity $\mathcal{K}$ in the last step is chosen as

$$\mathcal{K} = \mathbb{E}\left[(1 + a_i(t))^s \, \mathbf{1}_{\mathcal{E}_t^c}\big|\mathcal{F}_{t-\Omega}\right]. \tag{48}$$

The conditional expectation is taken over the σ-field $\mathcal{F}_{t-\Omega} = \sigma\{x(u), u \le t - \Omega\}$, generated by the history up to time $t - \Omega$. Given $\mathcal{F}_{t-\Omega}$, $a_i(t)$ is Poisson-distributed with mean $\lambda_i \Omega$. Therefore, using $\mathbb{E}[(1 + a_i(t))^s] < \infty$, we can use the Dominant Convergence Theorem to show

$$\lim_{C \to \infty} \mathcal{K} = \lim_{C \to \infty} \mathbb{E}\left[(1 + a_i(t))^s \left(1 - \mathbf{1}_{\mathcal{E}_t}\right)\big|\mathcal{F}_{t-\Omega}\right]$$
$$= \mathbb{E}\left[(1 + a_i(t))^s (1 - \lim_{C \to \infty} \mathbf{1}_{\mathcal{E}_t})\big|\mathcal{F}_{t-\Omega}\right] = 0.$$

This implies that we can make $C > 0$ to be sufficiently large so that $\mathcal{K} \le \frac{\epsilon}{4}$. Applying this result to (47), we conclude

$$\sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathcal{K}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\frac{\epsilon}{4}\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\right]$$
$$\le \frac{\epsilon}{4}A_2 + \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\frac{\epsilon}{4}\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\right] \tag{49}$$

where A_2 defined in (49) bounds $\sum_{i:x_i^\Omega(t)\ge1} \rho_i U_i'(\frac{c\rho_i}{x_i^\Omega(t)}) \le A_2$, over the compact region $(\mathcal{X}_t^\Omega)^c = \{\max_i x_i^\Omega(t) \le \xi - C\}$.

Similarly, we evaluate (46) over $\mathcal{E}_t^c \cap \mathcal{L}_i^c$

$$\sum_{i:x_i^\Omega(t)\ge1, x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t}\mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_i^c}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t}\mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_i^c}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{E}_t^c}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbb{E}[\mathbf{1}_{\mathcal{E}_t^c}|\mathcal{F}_{t-\Omega}]\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathcal{K}\right]$$
$$\le \frac{\epsilon}{4}A_2 + \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\frac{\epsilon}{4}\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\right]. \tag{50}$$

The first step uses the monotonicity of $U'(\cdot)$ and $x_i(t) \le x_i^\Omega(t)$ under $\mathcal{L}_i^c$. The fourth step is due to $\mathbb{E}[\mathbf{1}_{\mathcal{E}_t^c}|\mathcal{F}_{t-\Omega}] \le \mathcal{K}$ according to (48). The last step uses the inequality (49).

Finally, we evaluate (46) over $\mathcal{E}_t$. Note that when $\mathcal{E}_t$ occurs, the difference between $x(t)$ and $x^\Omega(t)$ is small. We will use this idea to replace the $x(t)$ in the right-hand side of (46) by $x^\Omega(t)$. Recognizing

$$\mathcal{E}_t \cap \mathcal{X}_t \subset \left\{\|x^\Omega(t) - x_i(t)\|_1 < \frac{C}{2}, \max_i x_i(t) > \xi\right\}$$
$$\subset \left\{\max_i x_i^\Omega(t) > \xi - C\right\} = \mathcal{X}_t^\Omega \tag{51}$$

we obtain

$$\sum_{i:x_i^\Omega(t)\ge1, x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t}\mathbf{1}_{\mathcal{E}_t}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1, x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t) + C}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i \left(1 + \frac{C}{x_i^\Omega(t)}\right)^s U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\right]$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i \left(1 + \frac{\epsilon}{2}\right) U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\right] \tag{52}$$

where the second step uses the monotonicity of $U'(\cdot)$, the third step uses Lemma 1-(ii), and the fourth step is from Lemmas 2 and 6 [similar to the argument in (43)].

Note that $\mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_i} + \mathbf{1}_{\mathcal{E}_t^c}\mathbf{1}_{\mathcal{L}_i^c} + \mathbf{1}_{\mathcal{E}_t} = 1$. Therefore, putting (49), (50), and (52) together and plugging the result into (46), we derive

$$\sum_{i:x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t}\right] - A_3 - \frac{\epsilon}{2}A_2$$
$$\le \sum_{i:x_i^\Omega(t)\ge1, x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t}\right] - \frac{\epsilon}{2}A_2$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\left(\mathbf{1}_{\mathcal{E}_t} + \frac{\epsilon}{2}\mathbf{1}_{\mathcal{E}_t} + \frac{\epsilon}{2}\right)\right]. \tag{53}$$

We still need to relate the last constant $\epsilon/2$ above to $\mathbf{1}_{\mathcal{E}_t}$. Toward this end, we can show the following the inequality, which is obtained in a similar way as the third and the last lines of (50):

$$\sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t^c}\right]$$
$$\le \frac{\epsilon}{4}A_2 + \frac{\epsilon}{4}\sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\right] \tag{54}$$

We add $\sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}[\rho_i U_i'(\frac{c\rho_i}{x_i^\Omega(t)})\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}]$ to both sides of (54) and rearrange the terms

$$\left(1 - \frac{\epsilon}{4}\right)\sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\right]$$
$$\le \frac{\epsilon}{4}A_2 + \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\right]. \tag{55}$$

Since $\frac{1}{2} < 1 - \frac{\epsilon}{4}$ for $\epsilon < 1$, we replace $1 - \frac{\epsilon}{4}$ by $\frac{1}{2}$ on the left-hand side of (55), and plug the result into the last line of (53) to obtain

$$\sum_{i:x_i(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\mathcal{X}_t}\right] - A_3 - \epsilon A_2$$
$$\le \sum_{i:x_i^\Omega(t)\ge1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i^\Omega(t)}\right)\mathbf{1}_{\mathcal{X}_t^\Omega}\mathbf{1}_{\mathcal{E}_t}\left(1 + \frac{3}{2}\epsilon\right)\right] \tag{56}$$
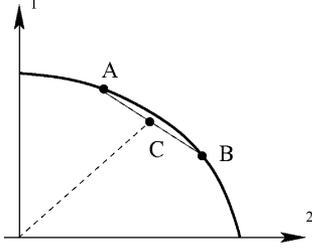
Fig. 4. Feasible rate region under consideration.

where it uses $\frac{\epsilon^2}{4} < \frac{\epsilon}{2}$ for $\epsilon < 1$. Equation (56) establishes a lower bound for the right-hand side of (27).

Prove (27). To show inequality (27), we combine the two bounds (44) and (56) to get

$$\sum_{i:x_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\mathbf{1}_{\{x(t)\in\mathcal{G}^c\}}\right]$$
$$\leq \epsilon A_2 + \left(2+\frac{3\epsilon}{2}\right)A_3 + \left(1+\frac{\epsilon}{2}\right)\left(1+\frac{3\epsilon}{2}\right)$$
$$\cdot \sum_{i:\hat{x}_i(t)\geq 1} \mathbb{E}\left[\rho_i U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)\mathbf{1}_{\mathcal{E}_t}\mathbf{1}_{\{\max_i \hat{x}_i(t)>\gamma\}}\right].$$
(57)

Finally, note that $\epsilon < 1$, $\frac{3}{2}\epsilon < 2$, and $(1+\frac{\epsilon}{2})(1+\frac{3\epsilon}{2}) < (1+\epsilon)^2$. Inequality (27) is immediate from (57). ∎

E. Proof of Lemma 6

Proof: We first show that there exists a constant A_4, such that

$$\sum_{i:\hat{x}_i(t)\geq 1} \rho_i\left(\frac{K_1 C}{\hat{x}_i(t)} - \frac{\epsilon}{4}\right)U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) \leq A_4. \quad (58)$$

Since the left-hand side of (58) is nonnegative only if $\hat{x}_i(t) \leq 4K_1 C/\epsilon$, we have

$$\sum_{i:\hat{x}_i(t)\geq 1} \rho_i\left(\frac{K_1 C}{\hat{x}_i(t)} - \frac{\epsilon}{4}\right)U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)$$
$$\leq \sum_i \max_{1\leq\hat{x}_i(t)\leq\frac{4K_1 C}{\epsilon}} \frac{\rho_i K_1 C}{\hat{x}_i(t)}U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)$$
$$\leq \sum_i \frac{\rho_i K_1 C}{1}U_i'\left(\frac{c\rho_i\epsilon}{4K_1 C}\right) \triangleq A_4 \quad (59)$$

where negative terms are dropped in the first step, and the second step uses the monotonicity of $U'(\cdot)$ and $\hat{x}_i(t) \geq 1$. Plugging this into the left-hand side of (31), we have

$$\sum_{i:\hat{x}_i(t)\geq 1} \rho_i\left(\frac{K_1 C}{\hat{x}_i(t)} - \frac{\epsilon}{2}\right)U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right)$$
$$\leq A_4 - \sum_{i:\hat{x}_i(t)\geq 1} \frac{\rho_i\epsilon}{4}U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right). \quad (60)$$

According to Assumption c) in Section II, for $\xi \to \infty$ and $\max_i \hat{x}_i(t) > \xi$, we have (60) $\to -\infty$. Therefore, there exists large enough ξ, such that (60) ≤ 0. ∎

F. 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 (8), we construct a counterexample 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 (which is convex and compact, e.g., an ellipse), depicted in Fig. 4. For an α-fair utility 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

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

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 $\phi_{\text{opt}}(x(t)) \notin \widehat{AB}$. It can be shown that the utility gap is on the same order as the optimal utility, i.e.,

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

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 $\phi(\hat{x}(t))$ that lies on $\widehat{AB}$. Hence, the rate-allocation policy $\phi(\hat{x}(t))$ satisfies condition (8) as claimed.

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

$$V(x) = \mu_1^{-1}w_1 x_1 + \mu_2^{-1}w_2 x_2 \quad (61)$$

where $w_1 = \phi_{B,2} - \phi_{A,2}$ and $w_2 = \phi_{A,1} - \phi_{B,1}$ are two positive constants. Then, we can formulate the expected Lyapunov function $W(t) = \mathbb{E}[V(x(t))]$ and derive an expression for the drift $\dot{W}(t)$ using the same manner as in (22) and (23). In the following, we 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.:

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

where the third inequality holds since the suboptimal rate allocation $\phi(\hat{x}(t))$ always lies below the straight line AB, whose slope

is $-\frac{w_1}{w_2}$. Thus, for the choice of Lyapunov function (61) 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 $\lim_{t\to\infty} W(t) = \infty$ as $t \to \infty$. ∎

G. Proof of Theorem 2

Proof: The proof is almost the same as that of Theorem 1, except that we need to prove a result similar to Lemma 3, but with traffic intensity $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\check{\mathcal{R}}$. For the Lyapunov function $V(x) = \sum_{i=1}^{N}\sum_{n=1}^{x_i} U_i'(\frac{c\rho_i}{n})/\mu_i$, we derive the same drift $\dot{W}(t)$ as in (22) and (23). To bound $\dot{W}$ under the new utility gap condition (9), we prove the following lemma, which generalizes Lemma 4 to traffic intensity $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\check{\mathcal{R}}$.

Lemma 7: Let $c = (1+\epsilon)^4$ for any traffic intensity $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\check{\mathcal{R}}$ and constant $C > 0$. If the suboptimal rate allocation $\phi(\hat{x}(t))$ satisfies the utility gap condition in (9), there exist 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:

$$\sum_{i:\hat{x}_i(t)\geq 1} \rho_i(1+\epsilon)^3 U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right) - \phi_i(\hat{x}(t))U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)+r}\right) \leq 0. \quad (62)$$

Proof: From the utility gap condition in (9), we conclude that for any $\delta \geq 0$, there exists a positive γ such that, for $\hat{x}(t)$ satisfying $\max_i \hat{x}_i(t) > \gamma$, we have

$$\Delta(\hat{x}(t)) \leq \eta(1+\delta)\left|\sum_{i:\hat{x}_i(t)\geq 1} \hat{x}_i(t)U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}(t))}{\hat{x}_i(t)}\right)\right|. \quad (63)$$

We first assume that the utility function is nonnegative. Plugging the expression of utility gap $\Delta(\hat{x}(t))$ into (63), we obtain

$$0 \leq \sum_{i:\hat{x}_i(t)\geq 1} \hat{x}_i(t)U_i\left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)}\right) - [1-(1+\delta)\eta]\hat{x}_i(t)U_i\left(\frac{\phi_{\text{opt},i}(\hat{x}_i(t))}{\hat{x}_i(t)}\right). \quad (64)$$

Since $\phi_{\text{opt},i}(\hat{x}_i(t))$ maximizes the NUM problem (1) at state $\hat{x}_i(t)$, we can replace $\phi_{\text{opt},i}(\hat{x}_i(t))$ above by the following choice of rate vector:

$$u = (1+\epsilon)^4[1-(1+\delta)\eta]^{-\frac{1}{|1-s|}}\rho \in \mathcal{R}. \quad (65)$$

Such $\epsilon > 0$ and $\delta > 0$ exist due to the traffic condition $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\check{\mathcal{R}}$. Let $\delta_0 = [1-(1+\delta)\eta]^{\frac{1}{|1-s|}}$. Similar to (38) in the Proof of Lemma 3, we have

$$0 \leq \sum_{i:\hat{x}_i(t)\geq 1} \hat{x}_i(t)U_i\left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)}\right) - [1-(1+\delta)\eta]\hat{x}_i(t)U_i\left(\frac{u_i}{\hat{x}_i(t)}\right)$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} \hat{x}_i(t)U_i\left(\frac{\phi_i(\hat{x}(t))}{\hat{x}_i(t)}\right) - \hat{x}_i(t)U_i\left(\frac{\delta_0 u_i}{\hat{x}_i(t)}\right)$$

$$\leq \sum_{i:\hat{x}_i(t)\geq 1} [\phi_i(\hat{x}(t)) - \delta_0 u_i]U_i'\left(\frac{c\rho_i}{\hat{x}_i(t)}\right). \quad (66)$$

The first step uses Lemma 1-(iii), and the last step uses the concavity of the utility function $U_i(\cdot)$. Note that (66) is exactly the same as (38) in the Proof of Lemma 3. Therefore, the rest of the proof below (38) can be copied directly to prove Lemma 5 for traffic intensity $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\check{\mathcal{R}}$. ∎

Except for Lemma 5, other derivations in the Proof of Theorem 1 can be applied directly. To summarize, we conclude

$$\limsup_{T\to\infty} \frac{1}{T}\int_{t=0}^{T} \mathbb{E}\left[\sum_{i:x_i(t)\geq 1} \lambda_i U_i'\left(\frac{c\rho_i}{x_i(t)}\right)\right] dt \leq \frac{2A_0 + 2A_2}{\epsilon}.$$

Because function $U_i'(\cdot)$ is a nonnegative and nondecreasing function and $\lim_{z\to\infty} U_i'(\frac{1}{z}) = \infty$, the last equation above implies the stability of the network under the traffic intensity $\rho \in (1-\eta)^{\frac{1}{|1-s|}}\mathcal{R}$. ∎

H. Proof of Theorem 3

Proof: We first notice that optimization problem (14) can be rewritten as

$$Z(\Delta) = \min_{\phi} \quad Z$$
$$\text{s.t.} \quad R\phi \preceq Zc, \ \phi \succeq 0$$
$$\sum_{i=1}^{N} x_i^{\alpha}\frac{\phi_i^{1-\alpha}}{1-\alpha} \geq U_{\text{opt}} - \Delta. \quad (67)$$

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 (67), the KKT conditions for optimality are given by

$$R\phi = Zc, \quad V(\phi) = U_{\text{opt}} - \Delta \quad (68)$$
$$R^T p - q\nabla_\phi V(\phi) = 0, \quad p^T c = 1. \quad (69)$$

From the implicit function theorem, variables ϕ, Z, p, and q can be viewed as implicit functions of Δ, which are uniquely defined by the KKT conditions (68) and (69). We define a vector $y = [\phi; p; q; Z]$ and a residual

$$G(y, \Delta) = \begin{pmatrix} R^T p - q\nabla_\phi V(\phi) \\ R\phi - Zc \\ U_{\text{opt}} - \Delta - V(\phi) \\ 1 - p^T c \end{pmatrix}. \quad (70)$$

Then, the KKT conditions (68) and (69) are equivalent to $G(y, \Delta) = 0$.

From the implicit function theorem, we have

$$\frac{dZ}{d\Delta} = -(\nabla_y G)^{-1}\nabla_\Delta G. \quad (71)$$

Plugging $\nabla_y G$ and $\nabla_\Delta G$ into the above and performing some matrix manipulations, we can derive the result in Theorem 4. ∎

REFERENCES

[1] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, Mar. 1998.

[2] D. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Trans. Autom. Control*, vol. 52, no. 12, pp. 2254–2269, Dec. 2007.

[3] J. W. Lee, M. Chiang, and R. A. Calderbank, "Jointly optimal congestion and contention control in wireless ad hoc networks," *IEEE Commun. Lett.*, vol. 10, no. 3, pp. 216–218, Mar. 2006.

[4] T. Bonald and L. Massoulie, "Impact of fairness on internet performance," in *Proc. ACM SIGMETRICS*, Jun. 2001, pp. 82–91.

[5] T. Bonald, M. Massoulie, A. Proutiere, and J. Virtamo, "A queuing analysis of max-min fairness, proportional fairness and balanced fairness," *Queue. Syst.*, vol. 53, pp. 65–84, Jun. 2006.

[6] H. Q. Ye, "Stability of data networks under an optimization-based bandwidth allocation," *IEEE Trans. Autom. Control*, vol. 48, no. 7, pp. 1238–1242, Jul. 2003.

[7] 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 *Proc. ACM SIGMETRICS*, Jun. 2007, pp. 239–250.

[8] X. Lin and B. Shroff, "The impact of imperfect scheduling on cross-layer congestion control in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[9] A. Eryilmaz, A. Ozdaglar, D. Shah, and E. Modiano, "Distributed cross-layer algorithms for the optimal control of multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 638–651, Apr. 2010.

[10] X. Lin, N. B. Shroff, and R. Srikant, "On the connection-level stability of congestion-controlled communication networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2317–2338, May 2008.

[11] P. Giaccone, B. Prabhakar, and D. Shah, "Towards simple, high-performance schedulers for high-aggregate bandwidth switches," in *Proc. IEEE INFOCOM*, Jun. 2002, pp. 1160–1169.

[12] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 411–424, Apr. 2005.

[13] A. Tang, J. Wang, and S. Low, "Counter-intuitive behaviors in networks under end-to-end control," *IEEE/ACM Trans. Netw.*, vol. 14, no. 2, pp. 355–368, Apr. 2006.

[14] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[15] T. Lan, X. Lin, M. Chiang, and R. B. Lee, "Stability and benefits of suboptimal utility maximization," Tech. rep., 2009 [Online]. Available: http://www.seas.gwu.edu/~tlan/paper/Suboptimality-v11-14pages.pdf

[16] T. Bonald, S. Borst, N. Hegde, and A. Proutiere, "Wireless data networks in multicell scenarios," in *Proc. ACM SIGMETRICS*, Jun. 2004, pp. 378–380.

[17] P. R. Kumar and S. P. Meyn, "Stability of queueing networks and scheduling policies," *IEEE Trans. Autom. Control.*, vol. 40, no. 2, pp. 251–260, Feb. 1995.

Tian Lan (S'03–M'05) received the B.A.Sc. degree from Tsinghua University, Beijing, China, in 2003, the M.A.Sc. degree from the University of Toronto, Toronto, ON, Canada, in 2005, and the Ph.D. degree from Princeton University, Princeton, NJ, in 2010, all in electrical engineering.

He is currently an Assistant Professor of electrical and computer engineering with the George Washington University, Washington, DC. His research interests are in wireless communications, optimization, distributed algorithms, and network security.

Dr. Lan received the 2008 IEEE Signal Processing Society Best Paper Award and the 2009 IEEE GLOBECOM Best Paper Award.

Xiaojun Lin (S'02–M'05) received the B.S. degree in electronics and information systems from Zhongshan University, Guangzhou, China, in 1994, and the M.S. and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2000 and 2005, respectively.

He is currently an Assistant Professor of electrical and computer engineering with Purdue University. His research interests are in the analysis, control, and optimization of communication networks.

Dr. Lin received the 2005 Best Paper of the Year Award from the *Journal of Communications and Networks* and the IEEE INFOCOM 2008 Best Paper Award. His paper was also one of two runner-up papers for the Best Paper Award at IEEE INFOCOM 2005. He received the National Science Foundation (NSF) CAREER Award in 2007.

Mung Chiang (S'00–M'03–SM'08) received the B.S., M.S., and Ph.D. degrees from Stanford University, Stanford, CA, in 1999, 2000, and 2003, respectively.

He is an Associate Professor of electrical engineering and an affiliated faculty of applied and computational mathematics and of computer science with Princeton University, Princeton, NJ. He was an Assistant Professor with Princeton University from 2003 to 2008. His research areas include optimization, distributed control, and stochastic analysis of communication networks, with applications to the Internet, wireless networks, broadband access, and content distribution.

Dr. Chiang has received the following awards: PECASE, TR35, ONR YIP, NSF CAREER, MPS YRA runner-up, Hertz Fellow, and Princeton Wentz Junior Faculty.

Ruby B. Lee (S'73–M'80–F'03) received the A.B. degree with distinction in computer science and comparative literature from Cornell University, Ithaca, NY, in 1973, and the M.S. degree in computer science and Ph.D. degree in electrical engineering with a minor in computer science from Stanford University, Stanford, CA, in 1975 and 1980, respectively.

She is the Forrest G. Hamrick Professor of Engineering and Professor of Electrical Engineering with Princeton University, Princeton, NJ, with an affiliated appointment in the Computer Science Department. She is the Director of the Princeton Architecture Laboratory for Multimedia and Security (PALMS). Prior to joining the Princeton faculty in 1998, she served as Chief Architect at Hewlett-Packard, Palo Alto, CA, responsible at different times for processor architecture, multimedia architecture, and security architecture. She was a key architect of the PA-RISC architecture used in HP workstations and servers. Her current research includes security-aware computer architecture, multicore security, cache-based software side-channel attacks, advanced bit permutations, no-overhead crypto, and secure cloud computing.

Prof. Lee is a Fellow of the Association for Computing Machinery (ACM). She is Associate Editor-in-Chief of *IEEE Micro*.