The Public Option: A Nonregulatory Alternative to Network Neutrality

Richard T. B. Ma and Vishal Misra, *Member, IEEE*

***Abstract*—Network neutrality and the role of regulation on the Internet have been heavily debated in recent times. Among the various definitions of network neutrality, we focus on the one that *prohibits paid prioritization of content*. We develop a model of the Internet ecosystem in terms of three primary players: consumers, ISPs, and content providers. We analyze this issue from the point of view of the consumer and target the desired system state that maximizes *consumer utility*. By analyzing various structures of an ISP market, we obtain different conclusions on the desirability of regulation. We also introduce the notion of a *Public Option ISP*, an ISP that carries traffic in a network-neutral manner. We find: 1) in a monopolistic scenario, network-neutral regulations might benefit consumers, however the introduction of a Public Option ISP is even better as it aligns the interests of the monopolistic ISP with the consumer utility; and 2) in an oligopolistic scenario, the presence of a Public Option ISP is again preferable to network-neutral regulations, although the presence of competing nonneutral ISPs provides the most desirable situation for the consumers.**

***Index Terms*—Internet economics, Network neutrality, paid prioritization, Public Option, regulatory policy.**

I. Introduction

SINCE 2005, network neutrality has been a hotly debated topic among law and policy makers. The core debate has centered around the argument whether ISPs should be allowed to provide service differentiation and/or user discrimination, with the notion of "user" being either content providers (CPs) or consumers. Proponents of network neutrality, mostly the CPs, have argued that the Internet has been "neutral" since its inception, and that has been a critical factor in the innovation and rapid growth that has happened on it. Opponents of network neutrality, mostly the ISPs, claim that without some sort of service differentiation, ISPs will lose the incentive to invest in the networks, and the end-user experience will suffer. Both camps implicitly or explicitly claim that their approach is beneficial for consumers. A recent Federal Communications Commission (FCC) vote [1] in the US has sided with the proponents, although the ruling leaves some room for service differentiation in wireless networks. The controversy rages on, though, with corporations like Verizon filing lawsuits challenging the ruling and a "toll-tax" dispute between Level3/Netflix and Comcast being cast as a network neutrality issue.

We study the issue explicitly from the consumer's point of view under both monopolistic and oligopolistic scenarios. A lot of arguments for, as well as against, network neutrality live in an idealized world where economies of scale do not exist and monopolies cannot emerge, and therefore perfect competition solves all problems. We believe reality is more nuanced, and hence we examine monopolistic scenarios too.

We use game-theoretic analyses and focus on the *consumer utility* defined in Section II-C. We model the user demand for content and the rate allocation mechanism of the network. The interplay between the two determines the rate equilibrium for traffic flows, based on which the consumer utility can be calculated. Our model of price discrimination, presented in Section III-A, is for the ISPs to offer two classes of service to CPs. The ISP divides its capacity into a premium and an ordinary class, and CPs get charged extra for sending traffic in the premium class. We identify and analyze the strategic games played between ISPs, CPs, and consumers in Section III for a monopolistic scenario, and in Section IV for oligopolistic scenarios. In Section IV-A, we introduce the notion of a *Public Option* ISP, which is neutral to all CPs. The Public Option ISP can be implemented by processes like local loop unbundling [3], which allows multiple telecommunications operators to use connections from the telephone exchange to the customer's premises, in a monopolistic market, and either government or a private organization could run the ISP and still be profitable [13]. Under this framework, our findings are as follows.

- The impact of network neutrality on consumer utility depends on the nature of competition at the ISP level. Concretely, a neutral network might be beneficial for consumers under a monopolistic regime (Section III), whereas a nonneutral network is advantageous for consumers under oligopolistic scenarios (Section IV).
- Introducing a Public Option ISP is advantageous for consumers. In a monopolistic situation, the Public Option ISP offers the best scenario for consumers (Theorem 5), followed by network-neutral regulations, and an unregulated market being the worst.
- In oligopolistic situations, the Public Option ISP is still preferable to network-neutral regulations. However, since the incentive for an ISP to gain market share is aligned with

Manuscript received February 13, 2012; revised October 17, 2012; accepted December 08, 2012; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor C. Dovrolis. This work was supported by Singapore's Agency for Science, Technology and Research (A*STAR) under a research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center, the Ministry of Education of Singapore under AcRF Grant R-252-000-448-133, and the National Science Foundation under Grants CNS-1017934 and CCF-1139915.

R. T. B. Ma is with the Advanced Digital Sciences Center, Illinois at Singapore, Singapore 138632, Singapore, and also with the School of Computing, National University of Singapore, Singapore 117418, Singapore (e-mail: tbma@adsc.com.sg; tbma@comp.nus.edu.sg).

V. Misra is with Department of Computer Science, Columbia University, New York, NY 10027 USA (e-mail: misra@cs.columbia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2012.2237412

maximizing consumer utility (Theorem 6), no regulation is needed to protect the consumers.
- Under an oligopolistic competition, any ISP's optimal pricing and service differentiation strategy, whether network-neutral or not, will be close to the one that maximizes consumer utility (Theorem 6 and Corollary 2). Moreover, under a probable equilibrium where ISPs use homogeneous strategies, their market shares will be proportional to their capacities (Lemma 2), which implies that ISPs do have incentives to invest and expand capacity so as to increase their market shares.

Our paper sheds new light on the network neutrality debate and concretely identifies where and how regulation can help. In particular, our identification of the Public Option ISP is especially important as it provides a solution that combines the best of both worlds, protecting consumer interests without enforcing strict regulations on all ISPs. We start with describing our model in Section II.

II. Three-Party Ecosystem Model

In this section, we study the equilibrium throughput of different CPs when their users compete for the capacity of a bottleneck-access ISP. By understanding this *rate equilibrium*, we will be able to characterize the user utility and further analyze it under different ISP strategies and regulatory policies. We consider a model of the Internet with three parties: CPs, ISPs, and consumers. We focus on a fixed consumer group in a targeted geographic region. We denote M as the number of consumers in the region.[1] Each consumer subscribes to an Internet access service via an ISP. We consider the scenarios where one monopolistic ISP I or a set $\mathcal{I}$ of competing oligopolistic ISPs provides the Internet access for the consumers. Our model does not include the backbone ISPs for two reasons. First, the bottleneck of the Internet is often at the last-mile connection toward the consumers [10], both wired and wireless. We focus on the regional or so-called *eyeball ISPs* that provide the bottleneck last-mile toward the consumers. Second, the recent concern on network neutrality manifests itself in the cases where the last-mile ISPs, e.g., France Telecom and Vodafone, intended to differentiate services and charge CPs, e.g., Apple and Google, for service fees [5]. We denote $\mathcal{N}$ as the set of CPs from which the consumers request content. We denote μ as the last-mile bottleneck capacity toward the consumers in the region. Fig. 1 depicts the contention at the bottleneck among different flows from the CPs. Given a set $\mathcal{N}$ of CPs, a group of M consumers, and a link with capacity μ, we denote the system as a triple $(M, \mu, \mathcal{N})$. We denote λ_i as the aggregate throughput rate from CP i to the consumers. Because consumers initiate downloads and retrieve content from the CPs, we first model the consumer's demand so as to characterize the CPs' throughput rates λ_i's.

A. Consumer Throughput Demand

We denote $\hat{\theta}_i$ as the *unconstrained throughput* for a typical user of CP i. For instance, the unconstrained throughput for the highest-quality Netflix streaming movie is about 5 Mb/s [4], and given an average query page of 20 kB and an average query response time of 0.25 s [2], the unconstrained throughput for a Google search is about 600 kb/s, or just over 1/10 of Netflix. We denote $\alpha_i \in (0, 1]$ as the percentage of consumers that ever access CP i's content, which models the popularity of the content of CP i. We define $\hat{\lambda}_i = \alpha_i M \hat{\theta}_i$ as the unconstrained throughput of CP i. Without contention, CP i's throughput λ_i equals $\hat{\lambda}_i$. However, when the capacity μ cannot support the unconstrained throughput from all CPs, i.e., $\mu < \sum_{i\in\mathcal{N}} \hat{\lambda}_i$, two things will happen: 1) A typical user of CP i obtains throughput $\theta_i < \hat{\theta}_i$ from CP i; and 2) some active users might stop using CP i when θ_i goes below certain threshold, e.g., users of streaming content like Netflix. We denote θ_i as the *achievable throughput* for the consumers downloading content from CP i. We define a demand function $D_i(\theta_i)$ that represents the percentage of consumers that still demand content from CP i under the achievable throughput θ_i.

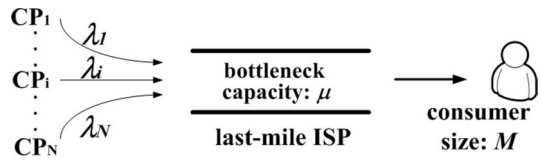

Fig. 1. Capacity contention at the last-mile bottleneck link.

Assumption 1 (User Demand): For any CP i, the demand $D_i(\cdot)$ is a nonnegative, continuous, and nondecreasing function defined on the domain of $[0, \hat{\theta}_i]$ and satisfies $D_i(\hat{\theta}_i) = 1$.

We define the aggregate throughput of a set $\mathcal{N}$ of CPs as $\lambda_{\mathcal{N}}(\theta) = \sum_{i\in\mathcal{N}} \lambda_i(\theta_i)$, where each CP i's rate is defined as

$$\lambda_i(\theta_i) = \alpha_i M D_i(\theta_i)\theta_i. \tag{1}$$

Demand Characterized by Throughput Sensitivity: Users often have different demand patterns for different CPs. For example, the demand for real-time applications decreases dramatically when their throughput drops below a certain threshold where performance cannot be tolerated by users. We can characterize this throughput sensitivity by a positive parameter β_i and consider the demand function

$$D_i(\theta_i) = e^{-\beta_i(\frac{\hat{\theta}_i}{\theta_i}-1)} = e^{-\beta_i(\frac{1}{\omega_i}-1)} \tag{2}$$

where we define $\omega_i = \theta_i/\hat{\theta}_i$ as the percentage of unconstrained throughput achieved for CP i. The user demand decays exponentially with the level of congestion (measured by $(\hat{\theta}_i - \theta_i)/\theta_i$, the ratio of unsatisfied and achieved throughput) scaled by β_i. This demand function distinguishes the CPs via their throughput sensitivity β_i: Larger β_i indicates higher sensitivity to throughput for CP i. Fig. 2 illustrates the demand functions with various values of β_i. To normalize $\hat{\theta}_i$, we plot D_i against ω_i instead of θ_i. We observe that when the achievable throughput θ_i drops linearly, the demand drops sharply for large β_i, e.g., when $\beta_i = 5$, the demand is halved with a 10% drop in throughput from $\hat{\theta}_i$. Large β_i's can be used to model CPs that have stringent throughput requirements, e.g., Netflix,

[1] M can also be interpreted as the *average* or *peak* number of consumers accessing the Internet simultaneously, which will scale with the total number of actual consumers. This does not change the nature of the results we describe subsequently, but gives a more realistic interpretation of the rate equilibrium.

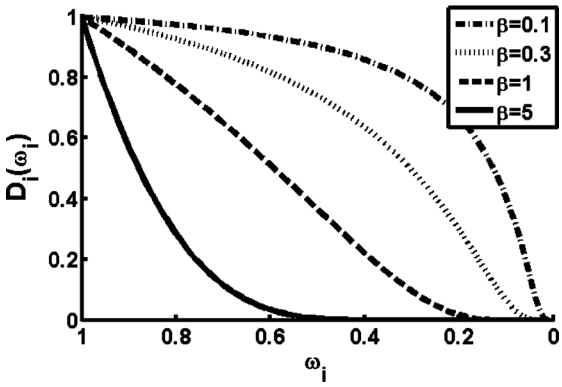

Fig. 2. Demand function $D_i(\omega_i)$.

while small β_i's can be used to model CPs that are less sensitive to throughput, e.g., a Google search query.

B. Rate Allocation Mechanism

When multiple flows share the same bottleneck link, they compete for capacity. We denote d_i as a fixed number of active flows/users of CP i. The rates allocated to the flows depend on the rate allocation mechanism being used in the system.

Definition 1: A *rate allocation mechanism* Θ is a continuous function $\Theta : \mathbb{R}_+^{|\mathcal{N}|+1} \to \mathbb{R}_+^{|\mathcal{N}|}$ that maps any fixed demand $\mathbf{d} = (d_i : i \in \mathcal{N})$ and capacity μ to the rates $\boldsymbol{\theta} = (\theta_i : i \in \mathcal{N})$.

A rate allocation mechanism can be a flow control mechanism, e.g., constant bit rate and variable bit rate mechanisms, under which the rates for each flow are allocated in a centralized manner, or a window-based end-to-end congestion control mechanism, e.g., TCP, under which each flow maintains a sliding window and adapts its size based on implicit feedback, e.g., the round-trip delay. We consider generic rate allocation mechanisms and assume that they obey the physical constraints of the system and satisfy some desirable properties.

Axiom 1 (Feasibility): For all $(\mathbf{d}, \boldsymbol{\mu}) \in \mathbb{R}_+^{|\mathcal{N}|+1}$, $\Theta(\mathbf{d}, \mu) \leq \hat{\theta}$.

Axiom 2 (Pareto Optimality): For all $(\mathbf{d}, \boldsymbol{\mu}) \in \mathbb{R}_+^{|\mathcal{N}|+1}$

$$\lambda_{\mathcal{N}}(\Theta(\mathbf{d}, \mu)) = \min\{\mu, \lambda_{\mathcal{N}}(\hat{\theta})\}.$$

The above axioms constrain that the aggregate rate $\lambda_{\mathcal{N}}$ cannot exceed the capacity μ and the individual rate θ_i cannot exceed its unconstrained throughput $\hat{\theta}_i$. Axiom 2 also characterizes a *work-conserving* property: If congestion can be alleviated without increasing the capacity μ, the rate allocation mechanism Θ would do so by fully utilizing the capacity.

Axiom 3 (Consistency): There exists a family of continuous and nondecreasing functions $\tilde{\Theta}(\eta) = (\tilde{\Theta}_i(\eta) : i \in \mathcal{N})$ such that $\tilde{\Theta}(\eta_1) \neq \tilde{\Theta}(\eta_2)$ for any $\eta_1 \neq \eta_2$, and for all $(\mathbf{d}, \boldsymbol{\mu}) \in \mathbb{R}_+^{|\mathcal{N}|+1}$, $\Theta(\mathbf{d}, \mu) = \tilde{\Theta}(\eta)$ for some η.

Axiom 3 implies that the mechanism allocates the rates among different CPs based on a consistent criteria $\tilde{\Theta}(\eta)$ depending on a system-state variable η, which can be interpreted as a measure of system congestion.

Assumption 2 (Rate Allocation): The network system uses a rate allocation mechanism that satisfies Axioms 1–3.

Instead of focusing on any particular rate allocation mechanism of the system, we work with generic rate allocation mechanisms that satisfy the above assumption throughout this paper. Next, we illustrate some prevalent examples in practice.

End-to-End Congestion Control Mechanisms: Due to the end-to-end design principle of the Internet, congestion control has been implemented by window-based protocols, i.e., TCP and its variations. Mo *et al.* [22] showed that a class of α-proportional fair solutions can be implemented by window-based end-to-end protocols. In fact, any α-proportional fair solution also satisfies Assumption 2. Among the class of α-proportional fair solutions, the max-min fair allocation, a special case of $\alpha = \infty$, is the result of the AIMD mechanism of TCP [7]. A max-min fair mechanism reduces the rate of flows that have the highest achievable rate under congestion and evenly share the capacity among all flows, unless they have reached their unconstrained throughput.

Definition 2: A rate allocation θ is *max-min fair* if it is not possible to increase the rate θ_i while maintaining feasibility, without reducing the throughput of some flow θ_j $(i \neq j)$ with $\theta_j \leq \theta_i$, i.e., for any other feasible allocation θ', $(\exists i \in \mathcal{N})\ \theta'_i > \theta_i \Rightarrow (\exists j \in \mathcal{N})\ \theta'_j < \theta_j < \theta_i$.

Under the max-min fair mechanism, the allocation criteria $\tilde{\Theta}$ can be defined as $\tilde{\Theta}_i(\eta) = \min(\eta, \hat{\theta}_i)$, where $\eta = \max\{\theta_i : i \in \mathcal{N}\}$ serves an indication of the level of system congestion.

Proportional Share Mechanism: A proportional share mechanism reduces the same percentage of the rate of each flow under congestion.

Definition 3: A rate allocation θ is *proportional-fair* if for any CPs i and j, $\theta_i : \theta_j = \hat{\theta}_i : \hat{\theta}_j$.

As shown by Kelly *et al.* [16], this mechanism provides *proportional fairness* that maximizes $\sum_{i \in \mathcal{N}} \hat{\theta}_i \log \theta_i$. The consistency allocation criteria $\tilde{\Theta}$ can be defined as $\tilde{\Theta}_i(\eta) = \eta \hat{\theta}_i$, where $\eta = \theta_i / \hat{\theta}_i$ indicates the system congestion. In practice, a proportional sharing allocation can be a result from weighted fair queueing mechanisms.

C. Rate Equilibrium and the Corresponding Consumer Utility

The demand functions map the achievable throughput to a level of demand; the rate allocation mechanisms map fixed demands to achievable throughput. The interplay between a rate allocation mechanism and the user demand determines the system rate equilibrium, explained by the following theorem.

Definition 4: θ is a *rate equilibrium* of the system $(M, \mu, \mathcal{N})$ if $\boldsymbol{\theta} = \Theta(D(\theta), \mu)$, where $D(\theta) = (D_i(\theta_i) : i \in \mathcal{N})$.

Theorem 1 (Uniqueness of Equilibrium): Under Assumptions 1 and 2, any system $(M, \mu, \mathcal{N})$ has a unique rate equilibrium.

Based on Theorem 1, we denote ϑ as the unique rate equilibrium of a system. Notice that Assumptions 1 and 2 are needed to guarantee the uniqueness of rate equilibrium; otherwise, the system might have multiple or zero equilibrium. We define $\nu = \mu/M$ as the per-capita capacity of the system.

Theorem 2 (Characteristics of Rate Equilibrium): Under Assumptions 1 and 2, for any CP $i \in \mathcal{N}$, its equilibrium throughput ϑ_i can be expressed as $\vartheta_i(M, \mu, \mathcal{N}) = \vartheta_i(\nu, \mathcal{N})$, a nondecreasing and continuous function in the per-capita capacity ν. Also, for any $\mathcal{N}_2 \subset \mathcal{N}_1$, we have

$$\vartheta_i(\nu, \mathcal{N}_2) \begin{cases} = \vartheta_i(\nu, \mathcal{N}_1), & \text{if } \vartheta_i(\nu_1, \mathcal{N}_1) = \hat{\theta}_i \\ \geq \vartheta_i(\nu, \mathcal{N}_1), & \text{otherwise.} \end{cases}$$

Theorem 2 states that the achievable throughput only depends on the per-user capacity ν, not the absolute scale of user population M or system capacity μ. When ν increases, the equilibrium throughput ϑ_i would not be worse off for any CP i. It also compares two systems to different sets of CPs and shows the monotonicity of user throughput: Throughput does not decrease when the set of CP decreases monotonically.

Based on the rate equilibrium ϑ and the resulting throughput $\lambda_i(\vartheta_i)$'s, we define the consumer utility as $CU = \sum_{i\in\mathcal{N}} \phi_i \lambda_i(\vartheta_i)$, where ϕ_i denotes the per-unit traffic utility that the consumers obtain by receiving content from CP i. Each per-unit traffic utility ϕ_i is an exogenous variable and can be derived from communicating with friends, e.g., Skype, watching movies, e.g., Netflix, and obtaining information, e.g., Google. Notice that the single parameter ϕ_i is a linear model for user utility, which might over/underestimate the real utility. However, we can always adjust the demand functions $D_i(\cdot)$'s, which still hold the monotonicity property in Assumption 1, to compensate the difference between the assumed linear utility and the real nonlinear utility. We denote Φ as the per-capita consumer utility defined as

$$\Phi = \frac{CU}{M} = \frac{1}{M}\sum_{i\in\mathcal{N}} \phi_i \lambda_i(\vartheta_i) = \sum_{i\in\mathcal{N}} \phi_i \alpha_i D_i(\vartheta_i)\vartheta_i. \qquad (3)$$

Because the per-capita consumer utility Φ depends on the system rate equilibrium ϑ, it is an endogenous variable.

Corollary 1: Under Assumptions 1 and 2, the per-capita consumer utility Φ can be expressed as $\Phi(M,\mu,\mathcal{N}) = \Phi(\nu,\mathcal{N})$, which is a nondecreasing function in the per-capita capacity ν. In particular, it strictly increases in $\nu \in [0, \sum_{i\in\mathcal{N}} \alpha_i \hat{\theta}_i]$.

Corollary 1 states that the per-capita consumer utility Φ will strictly increase with the system per-capita capacity ν, unless it is already maximized when unconstrained throughput is obtained. Notice that it does not depend on the values of ϕ_i's, but relies on the monotonic traffic demand (Assumption 1) and the consistency (Axiom 3) and work-conserving (Axiom 3) properties of the rate allocation mechanism.

We illustrate the rate allocation using an example of three CPs with demand functions of (2) and parameters $(\alpha_1, \hat{\theta}_1, \beta_1) = (1, 1, 0.1)$, $(\alpha_2, \hat{\theta}_2, \beta_2) = (0.3, 10, 0.3)$ and $(\alpha_3, \hat{\theta}_3, \beta_3) = (0.5, 3, 5)$. CP 1 represents Google-type CPs that are extensively accessed and less sensitive to throughput. CP 2 represents Netflix-type CPs that are more throughput-sensitive and have high unconstrained throughput. CP 3 represents Skype-type CPs that are extremely sensitive to throughput and have medium unconstrained throughput. Fig. 3 illustrates the rates and the corresponding demands of the three CPs under a max-min fair allocation mechanism. We vary the per-capita capacity ν from 0 to 6. We observe that when ν increases from zero, the demand for Google-type content increases first, followed by the demand for Skype-type content, and the demand for Netflix-type content being the last.

Fig. 4 illustrates the rates and the corresponding demands of the three CPs of the previous example under a proportional share mechanism. We observe that when ν increases from zero, the demand and throughput rate for Google-type of content increases sharply, while the demand of Netflix-type content does not start to catch up until the demand of Google-type content reaches around 90%.

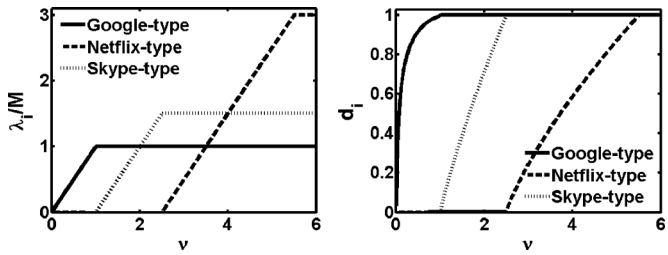

Fig. 3. Equilibrium throughput and demand under max-min fair mechanism.

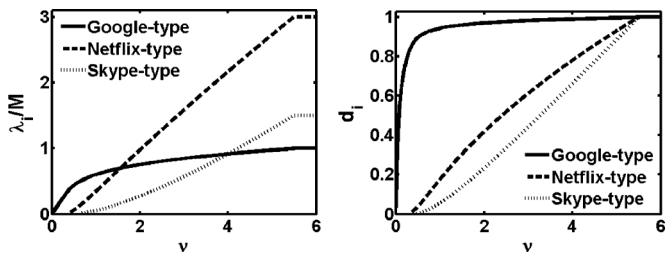

Fig. 4. Equilibrium throughput and demand under proportional mechanism.

In summary, based on minor assumptions on user demand and rate allocation (Assumptions 1 and 2), we derived the rate equilibrium ϑ, based on which the per-capita consumer utility Φ is defined (3). We also derived the monotonic relationship between Φ and the system per-capita capacity ν (Corollary 1). Departed from traditional queueing models for traffic and congestion, our approach faithfully captures the properties of closed-loop Internet protocols like TCP.

III. Monopolistic ISP Analysis

In this section, we start with the scenario where the last-mile capacity is controlled by a single monopolistic ISP I. We analyze the ISP's strategy under which nonneutral service differentiation is allowed and the corresponding best responses of the CPs. We derive the equilibria of the system and analyze the ISP's impact on the system congestion and the utility of the consumers and the ISP itself.

A. Nonneutral Service Differentiation

We assume that the monopolistic last-mile ISP I has a capacity of μ. This ISP can be a retail residential ISP, e.g., Comcast and Time Warner Cable, or a mobile operator, e.g., Verizon and AT&T. Regardless of being a wired or wireless provider, it serves as the last-mile service provider for the consumers. We assume that the ISP is allowed to allocate a fraction $\kappa \in [0,1]$ of its capacity to serve premium CPs and charge them for an extra rate $c \in [0,\infty)$ (dollar per unit traffic) besides ordinary transit charges. For a wired ISP, κ can be interpreted as the percentage of capacity deployed for paid-peering that charge c per unit incoming traffic, and $1-\kappa$ can be interpreted as the percentage of capacity deployed for settlement-free peering where incoming traffic is charge-free. For a wireless ISP, κ can be interpreted as the percentage of capacity devoted for prioritized traffic. The pair of parameters (κ, c) can also be thought of as a type of Paris

Metro Pricing (PMP) [25], [27], where an ordinary and a premium class have capacities of $(1-\kappa)\mu$ and $\kappa\mu$ and charge 0 and c, respectively. In reality, content might be delegated via content distribution networks (CDNs), e.g., Akamai, or backbone ISPs, e.g., Level3 is a major tier-1 ISP that delivers Netflix traffic toward regional ISPs. Therefore, the extra charge c might be imposed on the delivering ISP, e.g., Level3, and then be recouped from the CP, e.g., Netflix, by its delivering ISP. Our model does not assume any form of the implementation.

We denote $\mathcal{O}$ and $\mathcal{P}$ as the disjoint sets of CPs that join the ordinary and premium class, respectively. We denote v_i as CP i's per-unit traffic profit. This profit can be generated by advertising for media clients, e.g., Google, by selling online, e.g., Amazon, or by providing online services, e.g., Netflix and other e-commerce. Our model does not assume how the profit is generated. We define each CP i's utility u_i as

$$u_i(\lambda_i) = \begin{cases} v_i\lambda_i(M,(1-\kappa)\mu,\mathcal{O}), & \text{if } i\in\mathcal{O} \\ (v_i-c)\lambda_i(M,\kappa\mu,\mathcal{P}), & \text{if } i\in\mathcal{P}. \end{cases} \tag{4}$$

We define $IS = c\lambda_{\mathcal{P}}$ as the ISP surplus, i.e., the extra ISP profit generated by prioritizing content, and denote Ψ as the per-capita ISP surplus defined as

$$\Psi = \frac{IS}{M} = \frac{c}{M}\lambda_{\mathcal{P}} = \frac{c}{M}\sum_{i\in\mathcal{P}}\lambda_i(\vartheta_i) = c\sum_{i\in\mathcal{P}}\alpha_i D_i(\vartheta_i)\vartheta_i.$$

Although the rate c is determined by the ISP and can be considered as an exogenous variable, $\lambda_{\mathcal{P}}$ is endogenous, which depends on the CPs' decisions that are affected by the rate c. Therefore, Ψ is an endogenous variable of the system. Notice that our focus is the additional ISP profit earned from the CPs by providing a differentiated service. The ISP surplus does not reflect the ISP's normal operating costs or core revenue from the subscription payments from residential users.

B. Content Provider's Best Response

Given the ISP's decision κ and c, each CP chooses the service class, $\mathcal{O}$ or $\mathcal{P}$, to join. We denote ρ_i as the per-capita throughput over CP i's user base, i.e., $\alpha_i M$ users, defined as

$$\rho_i(\nu,\mathcal{N}) = \frac{\lambda_i(\vartheta_i(\nu,\mathcal{N}))}{\alpha_i M} = D_i(\vartheta_i(\nu,\mathcal{N}))\vartheta_i(\nu,\mathcal{N}). \tag{5}$$

Lemma 1: Given a fixed set $\mathcal{O}$ of CPs in the ordinary class and a fixed set $\mathcal{P}$ of CPs in the premium class, a new CP i's optimal strategy is to join the premium service class, if

$$(v_i-c)\,\rho_i(\kappa\nu,\mathcal{P}\cup\{i\}) \geq v_i\,\rho_i((1-\kappa)\nu,\mathcal{O}\cup\{i\}). \tag{6}$$

With equality, both service classes gives the same utility.

Lemma 1 states that a CP will join the premium service class if that results in higher profit, defined by the per-unit flow profit ($v_i - c$ for the premium class) multiplied by the per-capita throughput ρ_i. The above decision is clear for a CP only if all other CPs have already made their choices. To treat all CPs equally, we model the decisions of all CPs as a simultaneous-move game as part of a two-stage game.

C. Two-Stage Strategic Game

We model the strategic behavior of the ISP and the CPs as a two-stage game, denoted as a quadruple $(M,\mu,\mathcal{N},I)$.

1) *Players:* The monopolistic ISP I and the set of CPs $\mathcal{N}$.
2) *Strategies:* ISP I chooses a strategy $s_I = (\kappa, c)$. Each CP i chooses a binary strategy of whether to join the premium class. The CPs' strategy profile can be written as $s_{\mathcal{N}} = (\mathcal{O},\mathcal{P})$, where $\mathcal{O}\cup\mathcal{P} = \mathcal{N}$ and $\mathcal{O}\cap\mathcal{P} = \emptyset$.
3) *Rules:* In the first stage, ISP I decides $s_I = (\kappa, c)$ and announces it to all the CPs. In the second stage, all the CPs make their binary decisions simultaneously and reach a joint decision $s_{\mathcal{N}} = (\mathcal{O},\mathcal{P})$.
4) *Outcome:* The set $\mathcal{P}$ of the CPs shares a capacity of $\kappa\mu$ and the set $\mathcal{O}$ of the CPs shares a capacity of $(1-\kappa)\mu$. Each CP $i\in\mathcal{O}$ gets a rate λ_i in system $(M,(1-\kappa)\mu,\mathcal{O})$, and each CP $j\in\mathcal{P}$ gets a rate λ_j in system $(M,\kappa\mu,\mathcal{P})$.
5) *Payoffs:* Each CP i's payoff is defined by the utility $u_i(\lambda_i)$ in (4). The ISP's payoff is its surplus $IS = c\lambda_{\mathcal{P}}$ received from the premium class.

Notice that as a consequence of service differentiation, the original system $(M,\mu,\mathcal{N})$ breaks into two independent subsystems $(M,(1-\kappa)\mu,\mathcal{O})$ and $(M,\kappa\mu,\mathcal{P})$. In practice, if the premium service class is underutilized, i.e., $\lambda_{\mathcal{P}} < \kappa\mu$, and if the ISP uses a work-conserving mechanism so that the extra capacity $\kappa\mu - \lambda_{\mathcal{P}}$ in $\mathcal{P}$ would be used by ordinary class, then equivalently, we can think of the ISP's strategy as setting an effective κ that equals $1-\lambda_{\mathcal{P}}/\mu$, or virtually restricting the domain of κ to be upper-bounded by some value less than 1. Effectively, it limits the level of service differentiations and avoids the ordinary class being made a damaged good [12].

If we regard the set of CPs as a single player that chooses a strategy $s_{\mathcal{N}}$, our two-stage game can be thought of as a Stackelberg game [21], where the first-mover ISP can take the best responses of the CPs into consideration and derive its optimal strategy s_I using *backward induction* [21]. Given any fixed strategy $s_I = (\kappa, c)$, the CPs derive their best strategies under a simultaneous-move game, denoted as $(M,\mu,\mathcal{N},s_I)$. We denote $s_{\mathcal{N}}(M,\mu,\mathcal{N},s_I) = (\mathcal{O},\mathcal{P})$ as a strategy profile of the CPs under the game $(M,\mu,\mathcal{N},s_I)$. Technically speaking, when $\kappa = 0$ or 1, there is only one service class. When $\kappa = 0$, we define the trivial strategy profile as $s_{\mathcal{N}} = (\mathcal{N},\emptyset)$; when $\kappa = 1$, although there is not a physical ordinary class, we define the trivial strategy profile as $s_{\mathcal{N}} = (\mathcal{O},\mathcal{N}\backslash\mathcal{O})$, with $\mathcal{O} = \{i : v_i \leq c, i\in\mathcal{N}\}$ which defines the set of CPs that cannot afford to join the premium class. Based on Lemma 1, we can define an equilibrium in the sense of a Nash or competitive equilibrium. To break a tie, we assume that a CP always chooses to join the ordinary service class when both classes provide the same utility.

Definition 5: A strategy profile $s_{\mathcal{N}} = (\mathcal{O},\mathcal{P})$ is a *Nash equilibrium* of a game $(M,\mu,\mathcal{N},s_I)$ if

$$\frac{v_i-c}{v_i}\begin{cases} \leq \rho_i((1-\kappa)\nu,\mathcal{O})/\rho_i(\kappa\nu,\mathcal{P}\cup\{i\}), & \text{if } i\in\mathcal{O} \\ > \rho_i((1-\kappa)\nu,\mathcal{O}\cup\{i\})/\rho_i(\kappa\nu,\mathcal{P}), & \text{if } i\in\mathcal{P}. \end{cases}$$

D. Competitive Equilibrium

Notice that a CP's joining decision to a service class might increase the congestion level and reduce the throughput of flows of that service class. However, if the number of CPs in a service class is big and no single CP's traffic will dominate, an additional CP i's effect will be marginal. Analogous to the

pricing-taking assumption [21] in a competitive market, we make a *throughput-taking assumption* as follows.

Assumption 3 (Throughput Taking): Any CP $i \notin \mathcal{N}$ estimates $\tilde{\rho}_i(\nu, \mathcal{N})$ on its ex-post per-capita throughput $\rho_i(\nu, \mathcal{N} \cup \{i\})$ in the decision-making under a competitive equilibrium.[2]

Based on the above throughput-taking assumption, we can define a competitive equilibrium of the CPs as follows.

Definition 6: A strategy profile $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{P})$ is a *competitive equilibrium* of a game $(M, \mu, \mathcal{N}, s_I)$ if

$$\frac{v_i - c}{v_i} \begin{cases} \leq \rho_i((1-\kappa)\nu, \mathcal{O})/\tilde{\rho}_i(\kappa\nu, \mathcal{P}), & \text{if } i \in \mathcal{O} \\ > \tilde{\rho}_i((1-\kappa)\nu, \mathcal{O})/\rho_i(\kappa\nu, \mathcal{P}), & \text{if } i \in \mathcal{P}. \end{cases} \quad (7)$$

The competitive equilibrium depends on how each CP i calculates $\tilde{\rho}_i = D_i(\tilde{\vartheta}_i)\tilde{\vartheta}_i$, which boils down to an estimation of the ex-post throughput $\tilde{\vartheta}_i$. This estimation depends on the rate allocation mechanism being used. For example, under the max-min fair mechanism, CP i can expect an achievable throughput of $\vartheta_{\mathcal{N}} = \max\{\vartheta_j : j \in \mathcal{N}\}$. Thus, CP i can take this *throughput* as given and estimate that $\tilde{\vartheta}_i = \min\{\hat{\vartheta}_i, \vartheta_{\mathcal{N}}\}$. The competitive equilibrium under the throughput-taking assumption can be regarded as a special type of *congestion equilibrium* [20], where the throughput of the CPs indicates the level of congestion in system.

In practice, because CPs rarely know the characteristics of all other CPs, the *common knowledge* assumption [21] of Nash equilibria might be questionable. Thus, we use competitive equilibria for numerical evaluations.[3] Although Assumption 3 might not be valid if one of the CPs has significant percentage of traffic, our results do not depend on the underlying equilibrium type, and apply for both equilibrium definitions. In the rest of the paper, unless we specifically indicate an equilibrium to be Nash (Definition 5) or competitive (Definition 6), we use the term *equilibrium* to indicate both.

Theorem 3: If $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{P})$ is an equilibrium of a game $(M, \mu, \mathcal{N}, s_I)$, it is also a same type of equilibrium (Nash or competitive) of a game $(\xi M, \xi\mu, \mathcal{N}, s_I)$ for any $\xi > 0$.

Although a game $(M, \mu, \mathcal{N}, s_I)$ might have multiple equilibria, we do not assume that it reaches any particular equilibrium, and our results do not rely on which equilibrium is realized. For any game $(M, \mu, \mathcal{N}, s_I)$ with strategy $s_I = (\kappa, c)$ and the realized equilibrium $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{P})$, the realized per-capita consumer utility Φ is a function of ν, written as

$$\Phi(M, \mu, \mathcal{N}, s_I) = \Phi(\nu, \mathcal{N}, s_I) = \Phi((1-\kappa)\nu, \mathcal{O}) + \Phi(\kappa\nu, \mathcal{P}).$$

E. Monopolistic ISP's Strategy

In order to increase surplus, the ISP's optimal strategy would encourage more CPs to join its premium service class.

Theorem 4: In the game $(M, \mu, \mathcal{N}, I)$, for any fixed $c \geq 0$, strategy $s_I = (\kappa, c)$ is always dominated by $s_I^1 = (1, c)$. If $\lambda_{\mathcal{P}} < \min\{\mu, \sum_{v_i \geq c} \hat{\lambda}_i\}$, s_I is strictly dominated by s_I^1. $s_I = (\kappa, c)$ is also dominated by $s_I' = (\kappa', c)$ with $\kappa' > \kappa$, if equilibrium $(\mathcal{O}', \mathcal{P}')$ under s_I' satisfies $\mathcal{P} \subseteq \mathcal{P}'$.

When the monopoly ISP increases κ, it reduces the congestion level in the premium class and in a new equilibrium, $\mathcal{P}'$ would attract more CPs to join than $\mathcal{P}$. Theorem 4 states that the ISP would have incentives to increase κ so as to maximize its surplus. The effect of increasing κ is twofold: 1) More capacity is allocated to the premium class for sale; and 2) the reduced capacity in the ordinary class makes more CPs switch to the premium class. As a result, one of the optimal strategies of the monopoly is to set $\kappa = 1$. This implies that, if allowed, the selfish ISP will only provide the premium class $\mathcal{P}$ without contributing any capacity for the ordinary class $\mathcal{O}$. Suppose the ISP is allowed to set $\kappa = 1$, we study the optimal rate c and its impact on the consumer utility and its own surplus.

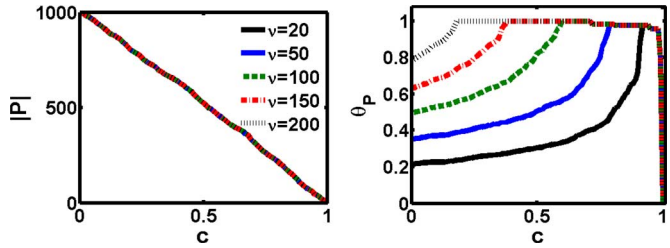

Fig. 5. $|\mathcal{P}|$ and the maximum throughput in premium class $\vartheta_{\mathcal{P}}$ under $\kappa = 1$.

Differing round-trip times, receiver window sizes and loss rates can result in different bandwidths, but to a first approximation, TCP provides a max-min fair allocation of available bandwidth among flows. Although other protocols, e.g., UDP, coexist in the Internet, recent research [17] sees a growing concentration of application traffic, especially video, over TCP. We use the demand function of (2) and the max-min fair mechanism for our numerical simulations. We study a scenario of 1000 CPs, whose α_i, $\hat{\theta}_i$, and v_i are uniformly distributed within $[0, 1]$ and β_i is uniformly distributed within $[0, 10]$. To satisfy all unconstrained throughput for the CPs, the per-capita capacity needs to be around $\nu = 250$. Since throughput-sensitive applications, e.g., Skype, bring more utility to consumers in reality, we consider the consumer utility ϕ_i that is uniformly distributed within $[0, \beta_i]$ (the uniform distribution biases utility toward CPs with high throughput sensitivity while introducing some randomness).[4]

Fig. 5 plots the number of CPs in the premium class $|\mathcal{P}|$ and the maximum achievable throughput $\vartheta_{\mathcal{P}}$ in the premium class. When the premium charge c increases, CPs with v_i smaller than c will be forced to stay away from the premium class, and therefore we observe the monotonic decrease of $|\mathcal{P}|$ in the left figure. When c is small, by increasing c, fewer CPs will be in the premium class. This leads to less congestion and higher $\vartheta_{\mathcal{P}}$. However, when c is large, there is no congestion in the system, and $\vartheta_{\mathcal{P}}$ becomes $\max\{\hat{\theta}_i : i \in \mathcal{P}\}$, which drops sharply when c tends to 1 in the right figure. Fig. 6 plots the per-capita ISP surplus Ψ and consumer utility Φ versus the charge c when the per-capita capacity ν ranges from 20 to 200. We observe three pricing regimes.

1) When c is small, Ψ increases linearly, i.e., $\Psi = c\nu$. This happens when most of the CPs can afford to join the service and the entire capacity is fully utilized, i.e., $\lambda_{\mathcal{P}} = \mu$, resulting in a high level of consumer utility Φ.

[2]This assumption is just for the definition of a competitive equilibrium.

[3]Please refer to [20] for evaluating a competitive equilibrium.

[4]The simulation results illustrate the general qualitative trends. However, our theoretical results do not depend on the particular setting of the experiments.

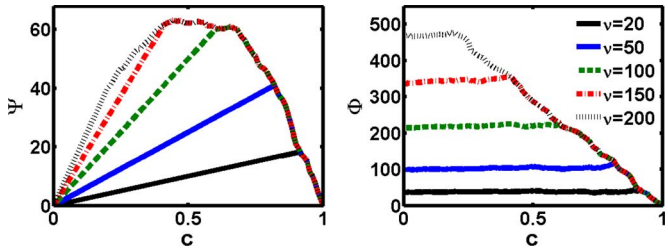

Fig. 6. Per-capita ISP surplus Ψ and consumer utility Φ under $\kappa = 1$.

2) When c is large, Ψ drops sharply. This happens when only a small set of CPs can afford to join the service and the capacity is largely underutilized, i.e., $\lambda_{\mathcal{P}} < \mu$, resulting in a sharp drop in Φ accordingly.
3) When ν is abundant, e.g., $\nu = 200$, there exists a pricing region where Ψ increases sublinearly and Φ decreases. Consequently, the ISP's optimal strategy ($c \approx 0.45$) could intentionally keep more CPs away from the (only) service class and underutilizes the capacity, which hurts the consumer utility Φ.

Fig. 7 illustrates Ψ and Φ under various strategies $s_I = (\kappa, c)$ versus ν ranging up to 500, which doubles the required capacity to serve all unconstrained throughput. For any fixed c in each column, we identify three equilibrium regimes.

1) When ν is small, Ψ increases linearly and Φ increases accordingly. This happens when the premium class capacity is fully utilized, i.e., $\lambda_{\mathcal{P}} = \kappa\mu$.
2) When ν keeps increasing, Ψ starts to decrease and Φ increases at a much slower rate. This happens when the premium class capacity is not fully utilized, i.e., $\lambda_{\mathcal{P}} < \kappa\mu$, and more CPs move from $\mathcal{P}$ to $\mathcal{O}$.
3) When ν is large, Ψ drops to zero for small values of κ, where Φ is maximized. This happens when $\mathcal{O}$'s capacity is abundant enough to serve all CPs' unconstrained throughput, and therefore all CPs use the ordinary class, i.e., $\mathcal{P} = \emptyset$. However, if κ is big, e.g., $\kappa = 0.9$, it guarantees some surplus for the ISP, but reduces the consumer utility from its maximum.

Furthermore, under any fixed ν, we observe that higher κ induces higher surplus for the ISP (Theorem 4), even if that results in an underutilization of the premium class capacity and hurts the consumer utility. When comparing different prices c, we observe that larger values of c make the premium class become underutilized more quickly because fewer CPs can afford to join the premium class when necessary. However, when reaching the turning point where congestion starts to be relieved, κ plays a major role, under which Φ's rate of increase depends on the amount $(1-\kappa)\mu$ of capacity allocated to $\mathcal{O}$.

Regulatory Implications: In the monopolistic scenario, the ISP would maximize κ for the charged service (Theorem 4).

In the case where the system capacity is abundant, i.e., large values for ν, the ISP would allocate more capacity for the premium class than needed, making the premium capacity underutilized. It implies that the ordinary service class would be made a *damaged good* [12], where the ISP would have the incentive to degrade service quality or avoid network upgrades or investments for the noncharged service class. Consequently, the consumer utility is greatly hurt by the ISP's selfish interest. To remedy this problem, the network neutrality principle should be imposed to some extent to protect consumer utility. In other words, the nonneutral service differentiation should be limited. The bottom line is that capacity underutilization should be avoided, which implies that non-work-conserving policies of the ISP should not be allowed. Technically speaking, by imposing a work-conserving policy, we put an upper bound $\kappa(c)$ for the capacity of the premium class, which can be expressed as a function of c. Effectively, the ordinary class would obtain $(1-\kappa(c))\mu$ amount of capacity.

In the case where the system capacity is scarce, i.e., small values for ν, or under a work-conserving policy, although the system capacity would not be underutilized, whether the ISP's pricing strategy is beneficial for consumer utility is still uncertain. In general, an ISP would prefer to set a high price c so as to obtain high surplus $c\lambda_{\mathcal{P}}$ from the premium class. Therefore, the consumer utility depends on whether the CPs in the premium class would provide higher utility for the consumers, i.e., high ϕ_i values for all $i \in \mathcal{P}$. On the one hand, if the price c is too high, it might limit/reject incubative CPs that are potentially beneficial for the consumers, but not yet profitable (low values of v_i). On the other hand, without enough price differentiation, more useful and probably more profitable CPs cannot provide better services so as to increase the consumer utility. In Section IV, we will show that the problem can be solved by introducing a so-called Public Option ISP for ISP competition.

IV. Oligopolistic ISP Analysis

In Section III, we concentrated on a monopolistic ISP I with capacity μ and a strategy $s_I = (\kappa, c)$. In this section, we extend our model to a set $\mathcal{I}$ of oligopolistic ISPs, each $I \in \mathcal{I}$ of which has a capacity μ_I and uses a strategy $s_I = (\kappa_I, c_I)$. We define $\mu = \sum_{I\in\mathcal{I}} \mu_I$ as the total system capacity. Our oligopolistic models have two major differences from the monopolistic model. First, since consumers connect to the Internet via one of the ISPs, they might make strategic decisions on which ISP to subscribe to. We denote M_I as the consumer size of ISP I, where $\sum_{I\in\mathcal{I}} M_I = M$, and $m_I = M_I/M$ as its market share. Second, besides the ISP surplus, a more important objective of any ISP I is to maximize its market share m_I. This is because the core revenue of the last-mile ISPs relies on the subscription payments of the users and the market share is also what the last-mile ISPs can leverage to generate the CP-side surplus in the first place.

Similar to the monopolistic ISP game $(M, \mu, \mathcal{N}, I)$, we denote $(M, \mu, \mathcal{N}, \mathcal{I})$ as the two-stage oligopolistic ISP game, under which the set of ISPs $\mathcal{I}$ choose their strategies $s_{\mathcal{I}} = \{s_I : I \in \mathcal{I}\}$ simultaneously in the first stage, and then the set of CPs $\mathcal{N}$ and the M consumers make their strategic decisions simultaneously in a second-stage game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$. In the second-stage game, we denote $s_M = \{M_I : I \in \mathcal{I}\}$ as the consumers' strategy that determines all ISPs' market shares, and $s_{\mathcal{N}} = \{s_{\mathcal{N}}^I = (\mathcal{O}_I, \mathcal{P}_I) : I \in \mathcal{I}\}$ as the CPs' strategy, where each $s_{\mathcal{N}}^I$ denotes the decision made at ISP I.

We denote Φ_I as the per-capita consumer utility achieved at ISP I, defined as $\Phi_I(M_I, \mu_I, \mathcal{N}, s_I) = \Phi_I(\nu_I, \mathcal{N}, s_I) = \Phi((1-\kappa_I)\nu_I, \mathcal{O}_I) + \Phi(\kappa_I\nu_I, \mathcal{P}_I)$, where $\nu_I = \mu_I/M_I$. We

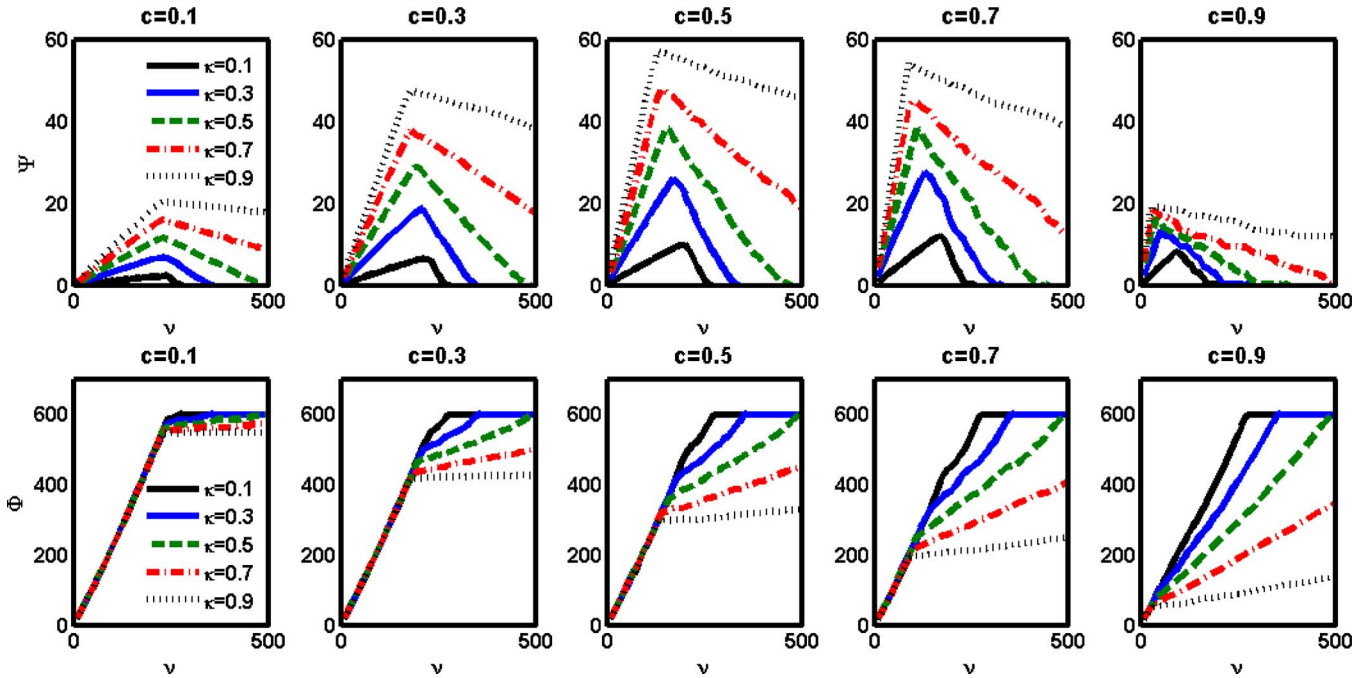

Fig. 7. Per-capita ISP surplus Ψ and consumer utility Φ under various strategies $s_I = (\kappa, c)$ versus the per-capita system capacity ν.

assume that consumers will move toward the ISPs that provide higher per-capita utility Φ_I as follows.

Assumption 4: Under any fixed strategy profile $s_{\mathcal{I}}$ and $s_{\mathcal{N}}$, for any pair of ISPs $I, J \in \mathcal{I}$, consumers will move from I to J if $\Phi_I < \Phi_J$. This process stops when $\Phi_I = \Phi_{\mathcal{I}}\ \forall I \in \mathcal{I}$ for some systemwise per-capita consumer utility $\Phi_{\mathcal{I}}$.

Although consumers might not be totally elastic or/and accessible to all available ISPs in practice, our assumption takes a macro perspective and assumes that if an ISP provides worse user experience on average, there must exist some consumers who can and will move to other better ISPs. If all the users are inelastic, the monopolistic analysis and conclusions in Section III can be applied. If a certain percentage of the users is inelastic, we can decompose the user population into elastic and inelastic components, and the conclusions will be based on a mixed model of a monopolistic and an oligopolistic market. Based on Assumption 4, we can determine the market share of the ISPs and, furthermore, define the equilibrium of the second-stage game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$ as follows.

Definition 7: A strategy profile $(s_M, s_{\mathcal{N}})$ is an equilibrium of the multi-ISP game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$ if: 1) for any $I \in \mathcal{I}$, $s^I_{\mathcal{N}}$ is an equilibrium of the single-ISP game $(M_I, \mu_I, \mathcal{N}, s_I)$; and 2) $\Phi_I = \Phi_J$ for any $I, J \in \mathcal{I}$.

A. Duopolistic ISP Game

We first study a two-ISP game with $\mathcal{I} = \{I, J\}$. Before that, we formally define a Public Option ISP as follows.

Definition 8: A *Public Option* ISP PO is an ISP that uses a fixed strategy $s_{\text{PO}} = (0, 0)$ and does not divide its capacity or charge the CPs.

We assume that ISP J is a Public Option ISP. Fig. 8 illustrates an example of the above duopolistic model, where both ISPs have the same amount of capacity, the CPs choose a service class

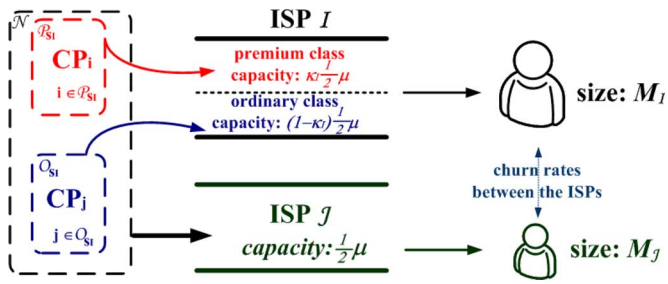

Fig. 8. Duopolistic ISP market model with one Public Option ISP J.

at ISP I and the consumers move between the ISPs. The above setting of the duopolistic game applies for two real scenarios. First, it models the competition between two ISPs, where one of them is actively a Public Option ISP and the other actively manages a nonneutral service differentiation. Second, it also models a situation where a single ISP owns the entire last-mile capacity μ. However, by regulation [3], it is required to lease its capacity to other service providers that do not own the physical line. The leasing ISP might be technologically limited from providing service differentiation on the leased capacity, but actually have customers in the region. For both scenarios, we will answer: 1) whether the nonneutral ISP could obtain substantial advantages over the neutral Public Option ISP (or whether the Public Option could survive under competition), and 2) how the competition is going to impact the consumer utility.

We study the same set of 1000 CPs as in Section III. We further assume that $\mu_I = \mu_J = \mu/2$ in our numerical example. We take the same route to numerically evaluate the competitive equilibria of the CPs under $\kappa_I = 1$.

Fig. 9 plots ISP I's market share m_I, per-capita surplus Ψ_I, defined as $\Psi_I = c\lambda_{\mathcal{P}_I}/M$, and per-capita consumer utility Φ versus ISP I's charge c_I. By the same reasons as before, the

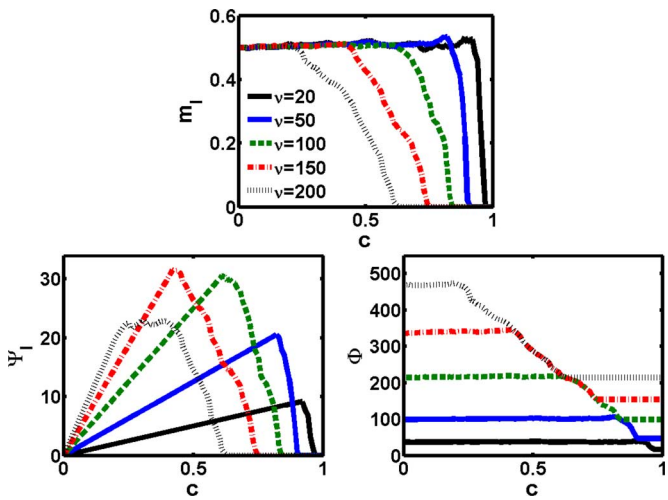

Fig. 9. ISP I's market share m_I, its per-capita surplus Ψ_I, and per-capita consumer utility Φ under its strategy $\kappa_I = 1$.

surplus of I increases linearly when its capacity is fully utilized, i.e., $\lambda_{\mathcal{P}_I} = \kappa_I \mu_I$. However, we observe three differences: 1) After the aggregate throughput in the premium class $\lambda_{\mathcal{P}_I}$ drops below $\kappa_I \mu_I$, Ψ_I drops to zero much steeper than before; 2) Φ does not drop down to zero when c_I increases to 1; and 3) the maximum Ψ_I is lower in the case of $\nu = 200$ than in the case of $\nu = 150$, which means that under $\kappa_I = 1$, capacity expansion could reduce ISP I's surplus from the CPs. All these observations can be explained by checking the market share of ISP I in the upper figure. When c_I increases from zero, the market share m_I remains around 50% until ISP I's capacity becomes underutilized, i.e., $\lambda_{\mathcal{P}_I} < \kappa_I \mu_I$. Afterwards, the market share drops dramatically. This explains that under congestion, i.e., $\lambda_{\mathcal{P}_I} = \kappa_I \mu_I$, by increasing c_I, ISP I restricts the number of CPs in its service class and maintains less congestion, which could result higher consumer utility and, therefore, attract more consumers from ISP J. After $\lambda_{\mathcal{P}_I}$ drops below $\kappa_I \mu_I$, further increase of c_I reduces the number of CPs in the service as well as the total throughput. This reduces consumer utility, and therefore, consumers start to depart from ISP I to J. When c_I reaches 1, no CP survives in I's service class, and all consumers move to ISP J, which guarantees a nonzero consumer utility in equilibrium.

Parallel to Fig. 7, Fig. 10 illustrates the per-capita ISP surplus Ψ_I, consumer utility Φ, and ISP I's market share m_I under various strategies s_I versus ν ranging up to 500. Compared to the monopolistic case, we observe two differences in Ψ_I and Φ: 1) Under any strategy s_I, ISP I's surplus drops sharply to zero after reaching a maximum point where $\lambda_{\mathcal{P}_I}$ drops below $\kappa_I \mu_I$; and 2) the increase of consumer utility does not get affected by ISP I's strategy too much. By observing the market share of ISP I, we identify two capacity regimes. First, when ν is extremely scarce, the differential pricing slightly benefits the consumer; therefore, ISP I can obtain a slightly larger percentage of the market.[5] Second, when the per-capita capacity ν is abundant, ISP I obtains at most an equal share of the market if it uses a small value of κ. Under this case, the capacity under $\mathcal{O}$ can support half of the population's unconstrained throughput and, in fact, the premium class is empty, i.e., $\mathcal{P} = \emptyset$. As a result, ISP I follows the Public Option ISP by using some kind of neutral policy (small κ) and maximizes the consumer utility.

[5]By limiting the number of CPs in $\mathcal{P}$, the proportion of throughput-sensitive traffic is larger, which yields higher consumer utility.

Theorem 5: In the duopolistic game $(M, \mu, \mathcal{N}, \mathcal{I})$, where an ISP J is a Public Option, i.e., $s_J = (0, 0)$, if ISP i's strategy s_I maximizes its market share M_I under an equilibrium $(s_M, s_{\mathcal{N}})$, then the per-capita consumer utility $\Phi_{\mathcal{I}}$ is also maximized under that equilibrium.

Theorem 5 implies that the existence of a Public Option ISP is superior to a network-neutral situation, where $s_I = (0, 0)$. This is because given the freedom of choosing an optimal s_I to maximize market share, ISP I's strategy will induce a maximum consumer utility under $s_J = (0, 0)$.

Based on our results, we answer the previously raised two questions: 1) The nonneutral ISP cannot win substantially over the Public Option ISP, which can still be profitable under the competition, confirming the independent findings from [13]. 2) Regardless of the capacity size, the competition induces higher consumer utility in equilibrium than under network-neutral regulations. The strategic ISP could obtain slightly over 50% of the market. However, if it differentiates services in the way that hurts consumer utility, its market share will drop sharply.

Regulatory Implications: In the duopolistic scenario with one of the ISPs being a Public Option, contrary to the monopolistic case, the nonneutral strategy s_I is always aligned with the consumer utility (Theorem 5). This result shows an interesting alternative to remedy the network neutrality issue under a monopolistic market. Instead of enforcing the ISP to follow network neutrality, the government (or a private organization, if it can be profitable [13]; otherwise, the government would bear a social cost so as to achieve the maximization of consumer utility) can provide the consumers with a Public Option ISP that is neutral to all CPs. Given such a neutral entity in the market, consumers will move to their public option if it provides higher consumer utility than the nonneutral ISP that uses differential pricing to the CPs. Meanwhile, in order to maximize its market share, the nonneutral ISP will adapt its strategy to maximize consumer utility. In conclusion, the introduction of a Public Option ISP is superior to network-neutral regulations under a monopolistic market since its existence aligns the nonneutral ISP's selfish interest with the consumer utility.

B. Oligopolistic ISP Competition Game

After analyzing the duopolistic game between a nonneutral and a Public Option ISP, we further consider a deregulated market under which all ISPs make nonneutral strategies. We consider a multi-ISP game under which each ISP I chooses a strategy s_I to maximize its market share m_I.

We first consider a homogeneous strategy $s = (\kappa, c)$, which can be a preferred strategy of all the ISPs or a regulated strategy imposed by the regulatory authorities.

Lemma 2: If $s_{\mathcal{I}} = \{s_I = s : I \in \mathcal{I}\}$ for some strategy $s = (\kappa, c)$, and $s_{\mathcal{N}}$ is an equilibrium of the single-ISP game

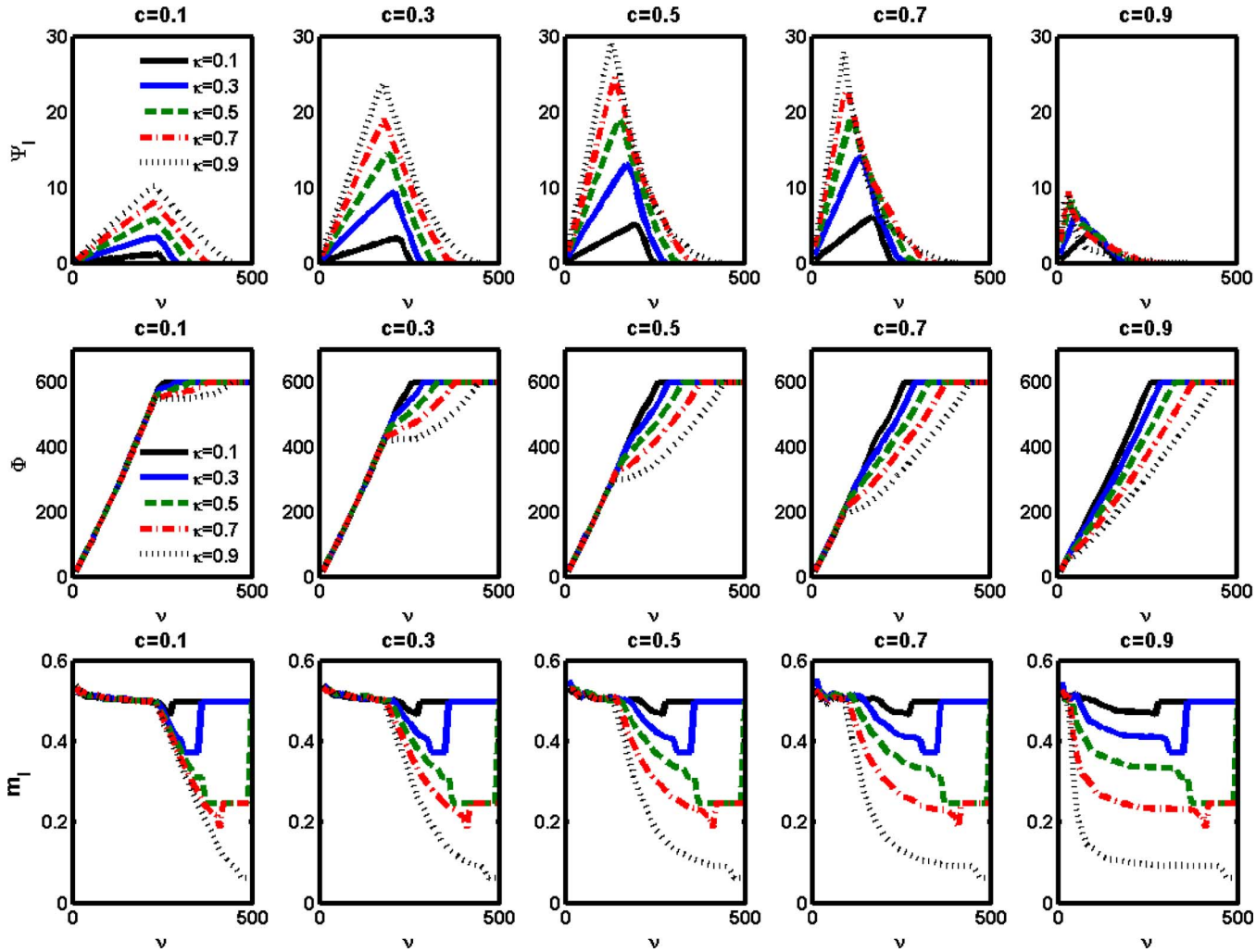

Fig. 10. Per-capita ISP surplus Ψ, consumer utility Φ, and market share m_I under various strategies $s_I = (\kappa, c)$ versus the per-capita capacity ν.

$(M, \mu, \mathcal{N}, s)$, then $\{m_I = \mu_I/\mu, s^I_{\mathcal{N}} = s_{\mathcal{N}} : I \in \mathcal{I}\}$ is an equilibrium of the multi-ISP game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$.

Lemma 2 shows a symmetric equilibrium where market share m_I is proportional to capacity μ_I. It implies that ISPs will have incentives to invest and expend capacity so as to obtain a larger market share. This equilibrium could be reached when ISPs simply mimic one another's strategy.

A further question is whether the competition of market share among the ISPs would induce equilibria where consumer utility is high. To address this issue, we first define

$$\epsilon_{s_I} = \sup\{\Phi(\nu_1, \mathcal{N}, s_I) - \Phi(\nu_2, \mathcal{N}, s_I) : \nu_1 < \nu_2\}. \quad (8)$$

We denote s_{-I} as the strategy profile of the ISPs other than ISP I, and define $\delta_{s_I} = \sup\{m_1 - m_2 : \Phi(\nu_1, \mathcal{N}, s_I) \leq \Phi(\nu_2, \mathcal{N}, s_I)\}$ and $\epsilon_{s_{-I}} = \max\{\epsilon_{s_J} : J \in \mathcal{I}\backslash\{I\}\}$.

Theorem 6: Under any fixed strategy profile s_{-I}, if I's strategy s_I is a best response to s_{-I} that maximizes its market share m_I in the game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$, then s_I is an $\epsilon_{s_{-I}}$ best response for the per-capita consumer utility $\Phi_{\mathcal{I}}$, i.e.,

$$\Phi_{\mathcal{I}} \geq \Phi'_{\mathcal{I}} - \epsilon_{s_{-I}} \qquad \forall s'_I \neq s_I.$$

Moreover, if s_I is a best response that maximizes consumer utility $\Phi_{\mathcal{I}}$ in the game $(M, \mu, \mathcal{N}, s_{\mathcal{I}})$, then s_I is a δ_{s_I} best response for the market share m_I, i.e.,

$$m_I \geq m'_I - \delta_{s_I} \qquad \forall s'_I \neq s_I.$$

Theorem 6 states that, given the fixed strategies of all other ISPs, an ISP's best responses to maximize: 1) its market share, and 2) the consumer utility, are closely aligned. Parallel to Theorem 5, it shows that an ISP's selfish interest is, although not perfectly, aligned with the consumer utility under competition. Technically, the $\epsilon_{s_{-I}}$ imperfection is due to the discontinuity of $\Phi(\nu, \mathcal{N}, s_I)$ in ν. Under a fixed strategy s_I, $\Phi(\nu, \mathcal{N}, s_I)$ is not strictly nondecreasing in ν compared to the result of Corollary 1. The reason is that when ν varies, CPs might move between the service classes. In general, when ν changes a small amount such that $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{P})$ does not change, the monotonicity still holds. However, when ν keeps increasing, CPs will move from $\mathcal{P}$ to $\mathcal{O}$, upon which Φ might drop at the spot. This discontinuity can be characterized by ϵ_{s_I}, which captures the largest vertical distance of a downward gap in the curve $\Phi(\nu, \mathcal{N}, s_I)$. From Fig. 7,

we observe that when $|\mathcal{N}|$ is large, ϵ_{s_I} is quite small, which indicates the general trend of increasing Φ with ν. In fact, when $\epsilon_{s_{-I}}$ approaches zero, Φ becomes nondecreasing, and the objectives of market share maximization and consumer utility maximization converges.

Definition 9: A strategy profile $s_{\mathcal{I}} = \{s_I : I \in \mathcal{I}\}$ is a *market-share Nash equilibrium* of the game $(M, \mu, s_{\mathcal{N}}, \mathcal{I})$ if for any $I \in \mathcal{I}$ and any strategy $s'_I \neq s_I$, the market share m_I satisfies $m_I(s'_I, s_{-I}) \leq m_I(s_I, s_{-I})$. Similarly, $s_{\mathcal{I}}$ is a consumer utility Nash equilibrium of the game $(M, \mu, s_{\mathcal{N}}, \mathcal{I})$ if for any $I \in \mathcal{I}$ and any strategy $s'_I \neq s_I$, the consumer utility $\Phi_{\mathcal{I}}$ satisfies $\Phi_{\mathcal{I}}(s'_I, s_{-I}) \leq \Phi_{\mathcal{I}}(s_I, s_{-I})$.

Corollary 2: If $s_{\mathcal{I}}$ is a market-share Nash equilibrium of the oligopolistic game $(M, \mu, \mathcal{N}, \mathcal{I})$, then it is also a consumer utility $\epsilon_{s_{\mathcal{I}}}$-Nash equilibrium, where $\epsilon_{s_{\mathcal{I}}} = \max\{\epsilon_{s_I} : I \in \mathcal{I}\}$. Conversely, if $s_{\mathcal{I}}$ is a consumer utility Nash equilibrium, then it is also a market-share $\delta_{s_{\mathcal{I}}}$-Nash equilibrium, where $\delta_{s_{\mathcal{I}}} = \max\{\delta_{s_I} : I \in \mathcal{I}\}$.

As a direct consequence of Theorem 6, Corollary 2 addresses that the objectives of maximizing market share and maximizing consumer utility are also closely aligned under Nash equilibria of the oligopolistic game $(M, \mu, \mathcal{N}, \mathcal{I})$.

Regulatory Implications: In the oligopolistic scenario, all ISPs' optimal strategies are closely aligned with the consumer utility. Even if some ISPs use suboptimal decisions, any remaining ISPs' optimal strategy would still nearly maximize the system consumer utility (Theorem 6). This alignment with consumer utility also sustains under Nash equilibria of the multi-ISP competition game (Corollary 2). Under this case, the existence of a Public Option ISP would be suboptimal compared to the efficient Nash equilibria. However, its damage is very limited because the Public Option ISP would be the only one that uses a suboptimal strategy, where all other ISPs can adapt to optimal strategies and more consumers will move from the Public Option to better and nonneutral ISPs. Of course, there is no reason why the Public Option cannot perform the price discrimination that aligns with the consumer utility, which induces an efficient Nash equilibrium in theory. However, implementing a neutral Public Option will avoid mistakes or accidental "collusion" with the existing ISPs in the market. In contrast, if network-neutral regulations are enforced, all ISPs will have to perform a neutral but inefficient strategy, which could reduce the consumer utility substantially. In conclusion, network-neutral regulations are not needed and should not be imposed under a competitive market. However, regulations should enforce the ISPs to be transparent in the sense that ISPs' capacity and strategies should be common knowledge to all ISPs, which would help the market converge to an efficient equilibrium in an easier manner.

V. Related Work

Despite its short history, a lot of work on network neutrality can be found in computer science [11], [24], [6], [13], [27], [18], economics [8], [15], and law [29], [28] literature.

From an economics perspective, Sidak [28] looked at the network neutrality regulation from consumer utility's point of view and argued that differential pricing is essential to the maximization of utility. We also focus on the consumer utility and seek the conditions under which ISPs' strategy would be aligned with consumer utility. Choi *et al.* [8] analyzed the effect of neutral regulations on ISPs' investment incentive and found that expending capacity will decrease the sale price of the premium service. This coincides with our finding under the monopolistic scenario. However, under oligopolistic competitions, we find that ISPs do have incentives to increase capacity so as to maximize market share.

From an engineering perspective, Dhamdhere *et al.* [13] took a profitability perspective and concluded that the ISPs can still survive without violating network neutrality. This supports our proposal of a Public Option ISP that can be implemented and sustained by either a government or a private organization. Crowcroft [11] reviewed various technical aspects and concluded that "perfect" network neutrality has never been and should not be engineered. We share the same view that under competition, network neutrality regulation is not necessary, while under a monopolistic market, a nonregulatory alternative can be a Public Option ISP that incentivizes the existing ISP to maximize consumer utility.

From a modeling point of view, one departure in our approach from previous analyses is the way we model traffic and congestion in the network. Traditionally, the $M/M/1$ formula for delay has been used to abstract out traffic and congestion [8] in economic analyses. Our view is that a more appropriate approach is to more faithfully model closed-loop protocols like TCP that carry most of the traffic on the Internet. Musacchio *et al.* [24] considered advertising CPs and also used a two-stage model under which ISPs move first. Their focus was primarily on a monopolistic ISP. Caron *et al.* [6] modeled differentiated pricing for only two application types. Shetty *et al.* [27] used a similar PMP-like two-class service differentiations and considered capacity planning, regulation as well as differentiated pricing to consumers. Our differentiated pricing focuses on the CP-side, where the CPs choose service classes and consumers choose ISPs. Yuksel *et al.* [30] also used a two-class service model, but focused on transit ISPs and quantified the equivalent overprovisioning cost when best effort is used. Our work focuses on the last-mile eyeball ISPs and consumer utility.

From a regulatory aspect, Wu [29] surveyed the discriminatory practices, e.g., selectively dropping packets, of broadband and cable operators and proposed solutions to manage bandwidth and police ISPs so as to avoid discrimination. Shetty *et al.* [27] proposed a simple regulatory tool to restrict the percentage of capacity the ISPs dedicate to a premium service class. Economides *et al.* [14] compared various regulations for quality of service, price discrimination, and exclusive contracts and drew conclusions on desirable regulation regimes. Ma *et al.* [18], [19] considered the ISP settlement aspect and advocated the use of Shapley value as profit-sharing mechanism to encourage ISPs to maximize social welfare. Our proposal of a Public Option ISP, on the other hand, is an nonregulatory alternative to the network-neutral regulations.

VI. Discussion and Conclusion

In a monopolistic market, the ISP's selfish nonneutral strategy hurts consumer utility. Although network-neutral

regulation might improve consumer utility, we find a better nonregulatory alternative, which is to introduce a Public Option ISP. The existence of a Public Option ISP incentivizes the existing ISP's strategy to be aligned with consumer utility and achieve higher consumer utility than that under network-neutral regulations. In an oligopolistic competition, market forces influence ISPs' nonneutral strategies to be aligned with consumer utility and ISPs will get market shares proportional to their capacities. Although network-neutral regulations are not needed and should not be imposed under oligopolistic scenarios, we envision that the Public Option could be implemented as the safety net, or the last/backup choice, for the consumers if the existing commercial ISPs' strategy hurt consumer utility.

In practice, a Public Option ISP could free commercial providers from neutrality obligations, allowing them to experiment with innovations in service differentiation. As real examples, *municipal broadband* [26] services are provided either fully or partially by local governments in many regions. Although they have received legal resistance from commercial providers in the US, the US FCC has endorsed the public option as a method of bringing broadband to underserved communities [9]. Setting up a Public Option ISP requires initial capital expenditures born by the taxpayers. Further studies are needed to investigate the feasibility and implementation of such a public option under various cost/benefit conditions. A recent study [23] from National Taxpayers Union of the US has shown that mismanagement could lead to unprofitable scenarios.

Theoretically speaking, the existence of a Public Option ISP will be effective if $\mu_{\rm PO} > 0$, regardless of how large its capacity is. This is because, in the idealized game model, we assume that an ISP's sole objective is to maximize its market share. In practice, ISPs will trade off their market share with potential surplus from the CPs, which depends on the characteristics of the CPs, e.g., their profit margin and throughput sensitivity, and the system parameters, e.g., the system capacity and the level of system congestion. Moreover, ISPs might also want to use the CP-side surplus to subsidize their consumers so as to increase their market share. In general, the more ISPs compete freely in a market, the less the market needs a public option and the less capacity we need to deploy for the Public Option ISP to be effective. In the most hostile case where only one monopolistic ISP exists in the market, a Public Option ISP could be effective as long as it has a capacity that is larger than the percentage of consumers that the monopoly cannot afford to lose. For example, if 10% of the market share is critical for the monopoly, implementing 10% of its capacity would be able to at least "steal" 10% of consumers from the monopoly if it follows a network-neutral strategy. If the monopoly applies a worse than neutral strategy for consumer utility, it will lose even more. In that sense, although 10% of the capacity will not be operating optimally, its existence incentivizes the remaining 90% maximizing for consumer utility, which could result in much better consumer utility than requiring the monopoly to follow network-neutral regulations. Last but not least, users might not be able to choose ISPs freely due to various reasons, e.g., lack of transparency of service qualities and contract lock-in. Future studies are needed to capture these effects so as to inform further regulatory policies to support market competition.

In summary, we believe our paper sheds new light on and informs the continuing debate on the role of regulation on the Internet, and our introduction of the Public Option ISP is an important contribution.

APPENDIX

Proof of Theorem 1: Based on Assumption 1, we know that for any $i \in \mathcal{N}$, the throughput $\lambda_i(\theta_i)$ is a strictly increasing and continuous function in θ_i. By Axiom 1, $\lambda_i(\theta_i)$ has a range of $[0, \hat{\lambda}_i]$. We show the uniqueness of the rate equilibrium by the following two cases. We first consider the case where $\mu \geq \sum_{i\in\mathcal{N}} \hat{\lambda}_i$. By Axiom 2, $\lambda_{\mathcal{N}} = \sum_{i\in\mathcal{N}} \lambda_i(\theta_i) = \sum_{i\in\mathcal{N}} \hat{\lambda}_i$. Because $\lambda_i \leq \hat{\lambda}_i$ for all $i \in \mathcal{N}$, we must have $\lambda_i = \hat{\lambda}_i$. Therefore, the unique rate equilibrium must be $\vartheta = \hat{\theta}$.

We then consider the case where $\mu < \sum_{i\in\mathcal{N}} \hat{\lambda}_i$. By Axiom 2, $\lambda_{\mathcal{N}} = \sum_{i\in\mathcal{N}} \lambda_i(\theta_i) = \mu$. Because each $\lambda_i(\theta_i)$ is strictly increasing in θ_i and continuous in the range of $[0, \hat{\lambda}_i]$, there exists a unique $\eta > 0$ such that $\sum_{i\in\mathcal{N}} \lambda_i(\tilde{\Theta}_i(\eta)) = \mu$. We show that $\tilde{\Theta}_i(\eta)$ is a rate equilibrium, i.e., $\Theta(D(\tilde{\Theta}_i(\eta)), \mu) = \tilde{\Theta}_i(\eta)$. Suppose $\Theta(D(\tilde{\Theta}_i(\eta)), \mu) = \tilde{\Theta}_i(\tilde{\eta})$ for some $\tilde{\eta} \neq \eta$. If $\tilde{\eta} < \eta$, by Axiom 3 $\tilde{\Theta}_i(\tilde{\eta}) < \tilde{\Theta}_i(\eta)$. Therefore, $\lambda'_{\mathcal{N}} = \sum_{i\in\mathcal{N}} \lambda_i(\tilde{\Theta}_i(\tilde{\eta})) < \mu$, which violates the Pareto optimality condition of Axiom 2. If $\tilde{\eta} > \eta$, by Axiom 3 $\tilde{\Theta}_i(\tilde{\eta}) > \tilde{\Theta}_i(\eta)$, and therefore $\lambda'_{\mathcal{N}} = \sum_{i\in\mathcal{N}} \lambda_i(\tilde{\Theta}_i(\tilde{\eta})) > \mu$, which is not feasible and violates Axiom 2.

Finally, we show $\theta = \tilde{\Theta}_i(\eta)$ is the *unique* rate equilibrium. By Axiom 3, the allocation must be in the form of $\tilde{\Theta}_i(\tilde{\eta})$ for some $\tilde{\eta}$. By Axiom 2, the allocated solution has to satisfy $\sum_{i\in\mathcal{N}} \lambda_i(\tilde{\Theta}_i(\tilde{\eta})) = \mu$, and therefore $\tilde{\eta} = \eta$. ■

Proof of Theorem 2: To show that we can express $\vartheta_i(M, \mu, \mathcal{N}) = \vartheta_i(\nu, \mathcal{N})$ for all systems $(M, \mu, \mathcal{N})$ with $\nu = M/\mu$, we need to show that $\vartheta_i(M_1, \mu_1, \mathcal{N}) = \vartheta_i(M_2, \mu_2, \mathcal{N})$ if $\nu_1 = M_1/\mu_1 = M_2/\mu_2 = \nu 2$ by two cases. In the first case, $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i \leq \nu_1 = \nu_2$. Under this case, the capacity is more than the aggregate unconstrained throughput, and therefore $\vartheta_i(M_1, \mu_1, \mathcal{N}) = \vartheta_i(M_2, \mu_2, \mathcal{N}) = \hat{\theta}_i$ for all $i \in \mathcal{N}$. In the second case of $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i > \nu_1 = \nu_2$, by Axiom 3, we know that $\vartheta(M_1, \mu_1, \mathcal{N}) = \tilde{\Theta}(\eta_1)$ and $\vartheta(M_2, \mu_2, \mathcal{N}) = \tilde{\Theta}(\eta_2)$ for some η_1 and η_2. We need to show that $\eta_1 = \eta_2$. Since under this case, by Axiom 2, we know that the aggregate throughput $\lambda_{\mathcal{N}}$ saturates the system capacity in both systems. However, if $\eta_1 > \eta_2$ (or $\eta_1 < \eta_2$), by Axiom 3 $\tilde{\Theta}(\eta_1) > \tilde{\Theta}(\eta_2)$ (or $\tilde{\Theta}(\eta_1) < \tilde{\Theta}(\eta_2)$), then either one system does not satisfy the Pareto optimality condition or the other system will violate the feasibility condition that $\lambda_{\mathcal{N}} \leq \mu$. Thus, $\eta_1 = \eta_2$ if $\nu_1 = \nu_2$.

Now, we want to show that $\vartheta_i(\nu, \mathcal{N})$ is nondecreasing in ν. Again, in the case of $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i \leq \nu$, $\vartheta_i(\nu, \mathcal{N}) = \hat{\theta}_i$, which is nondecreasing. In the case of $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i > \nu$, by Axiom 2, $\lambda_{\mathcal{N}} = \mu$, which is equivalent to $\sum_{i\in\mathcal{N}} \alpha_i D_i(\vartheta_i)\vartheta_i = \sum_{i\in\mathcal{N}} \alpha_i D_i(\tilde{\Theta}_i(\eta))\tilde{\Theta}_i(\eta) = \nu$. Thus, when ν is increasing, by Assumption 1 and Axiom 3, the corresponding η is increasing, so as the equilibrium rate $\vartheta_i(\nu, \mathcal{N})$.

Similarly, we can show that for any $\mathcal{N}_2 \subset \mathcal{N}_1$, if $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i \leq \nu$, $\vartheta_i(\nu, \mathcal{N}_1) = \vartheta_i(\nu, \mathcal{N}_2) = \hat{\theta}_i$; if $\sum_{i\in\mathcal{N}} \alpha_i\hat{\theta}_i > \nu$, the corresponding $\eta_2 > \eta_1$, and therefore $\vartheta_i(\nu, \mathcal{N}_2) \geq \vartheta_i(\nu, \mathcal{N}_1)$ for all $i \in \mathcal{N}$. ■

Proof of Corollary 1: If we define $\Phi_i = \phi_i \alpha_i d_i(\vartheta_i)\vartheta_i$ for all $i \in \mathcal{N}$, by definition $\Phi = \sum_{i\in\mathcal{N}} \Phi_i$. Since ϑ_i is a function of ν by Theorem 2, Φ_i and Φ are functions of ν as well. By Assumption 1, Φ_i is an increasing function of ϑ_i. By Theorem 2, ϑ_i is nondecreasing in ν; therefore, Φ_i and then Φ are nondecreasing in ν. By Axiom 2, $\lambda_{\mathcal{N}} = \mu$ when $\nu \in [0, \sum_{i\in\mathcal{N}} \alpha_i \hat{\theta}_i]$, which implies that when ν increases, there must exist some $i \in \mathcal{N}$ with ϑ_i strictly increasing. As a result, Φ_i and Φ have to be strictly increasing as well. ∎

Proof of Lemma 1: When joining the ordinary service class, the throughput of CP i is

$$\begin{aligned}&\lambda_i(M, (1-\kappa)\mu, \mathcal{O}\cup\{i\})\\ &\quad= \alpha_i M \rho_i(M, (1-\kappa)\mu, \mathcal{O}\cup\{i\})\\ &\quad= \alpha_i M \rho_i((1-\kappa)\nu, \mathcal{O}\cup\{i\}).\end{aligned}$$

When joining the premium service class, the throughput is

$$\begin{aligned}\lambda_i(M, \kappa\mu, \mathcal{P}\cup\{i\}) &= \alpha_i M \rho_i(M, \kappa\mu, \mathcal{P}\cup\{i\})\\ &= \alpha_i M \rho_i(\kappa\nu, \mathcal{P}\cup\{i\}).\end{aligned}$$

Therefore, by (4), CP i's utility is

$$u_i = \begin{cases} v_i \alpha_i M \rho_i((1-\kappa)\nu, \mathcal{O}\cup\{i\}), & \text{if } i \in \mathcal{O}\\ (v_i - c)\alpha_i M \rho_i(\kappa\nu, \mathcal{P}\cup\{i\}), & \text{if } i \in \mathcal{P}.\end{cases}$$

By comparing the utilities that can be obtained in the two service classes, we obtain the condition (6). ∎

Proof of Theorem 3: Under the same strategy $s_I = (\kappa, c)$, the new ordinary class $(\xi M, \xi(1-\kappa)\mu, \mathcal{O})$ and the new premium class $(\xi M, \xi\kappa\mu, \mathcal{P})$ have the same per-capita capacity $(1-\kappa)\nu$ and $\kappa\nu$, respectively, as before. By Theorem 2, the new system induces the same throughput ϑ_i as before. As a result, the solution $(\mathcal{O}, \mathcal{P})$ will induce the same values of ρ_i and $\tilde{\rho}_i$, which is an estimate on ρ_i based on ϑ_i. Therefore, both sides of (5) and (7) do not change, and the equilibrium conditions still hold. ∎

Proof of Theorem 4: Under $s_I^1 = (1, c)$, only the premium service class is provided, and CPs will join the premium service class only if $v_i \geq c$. Therefore, the aggregate rate $\lambda_{\mathcal{P}}^1 = \min\{\mu, \sum_{i\in\mathcal{P}_c} \hat{\lambda}_i\}$, where $\mathcal{P}_c = \{i : v_i \geq c\}$. When keeping the same c, $\mathcal{P} \subseteq \mathcal{P}_c$ under any strategy s_I. We have

$$\lambda_{\mathcal{P}} = \min\left\{\kappa\mu, \sum_{i\in\mathcal{P}\subseteq\mathcal{P}_c} \hat{\lambda}_i\right\} \leq \min\left\{\mu, \sum_{i\in\mathcal{P}_c} \hat{\lambda}_i\right\} = \lambda_{\mathcal{P}}^1.$$

The above implies that the surplus $c\lambda_{\mathcal{P}} \leq c\lambda_{\mathcal{P}}^1$. Therefore, s_I is dominated by s_I^1. If $\lambda_{\mathcal{P}} < \min\{\mu, \sum_{i\in\mathcal{P}_c} \hat{\lambda}_i\}$, $\lambda_{\mathcal{P}} < \lambda_{\mathcal{P}}^1$, and therefore s_I is strictly dominated by s_I^1.

Similarly, for any $\kappa' > \kappa$ with $\mathcal{P} \subseteq \mathcal{P}'$, we have

$$\lambda_{\mathcal{P}} = \min\left\{\kappa\mu, \sum_{i\in\mathcal{P}\subseteq\mathcal{P}'} \hat{\lambda}_i\right\} \leq \min\left\{\kappa'\mu, \sum_{i\in\mathcal{P}'} \hat{\lambda}_i\right\} = \lambda_{\mathcal{P}'}.$$

Therefore, s_I is also dominated by s_I'. ∎

Proof of Theorem 5: For any strategy $s_I' \neq s_I$, we have two cases to analyze. First, if $M_I' = M_I$, then $M_J' = M - M_I' = M - M_I = M_J$. Given the same market share for the public option ISP, it induces the same per-capita consumer utility $\Phi_J' = \Phi_J$. In equilibrium, we have $\Phi_{\mathcal{I}}' = \Phi_J' = \Phi_J = \Phi_{\mathcal{I}}$. Second, if $M_I' < M_I$, then $M_J' = M - M_I' > M - M_I = M_J$. Given a larger market share for the public option ISP, the per-capita capacity reduces, i.e., $\nu_J' < \nu_J$. By Corollary 1, the public option will not induce a larger per-capita consumer utility, i.e., $\Phi_J' \leq \Phi_J$. Thus, we have $\Phi_{\mathcal{I}}' = \Phi_J' \leq \Phi_J = \Phi_{\mathcal{I}}$ in equilibrium. ∎

Proof of Lemma 2: When $M_I = \mu_I M/\mu$ and $s_I = s$ for all $I \in \mathcal{I}$, the single-ISP game $(M_I, \mu_I, \mathcal{N}, s)$ is a linearly scaled game of $(M, \mu, \mathcal{N}, s)$. By Theorem 3, we know that $s_{\mathcal{N}}$ is also an equilibrium of the game $(M_I, \mu_I, \mathcal{N}, s_I)$ for any $I \in \mathcal{I}$. By Corollary 1, we know that

$$\begin{aligned}\Phi_I &= \Phi(M_I, \mu_I, \mathcal{N}, s_I)\\ &= \Phi(\nu_I, \mathcal{N}, s) = \Phi(\nu, \mathcal{N}, s) \qquad \forall I \in \mathcal{I}.\end{aligned}$$

The above satisfies the two conditions of an equilibrium in Definition 7 and concludes the proof. ∎

Proof of Theorem 6: If s_I maximizes the market share M_I, for any strategy $s_I' \neq s_I$, we have $M_I' \leq M_I$. Therefore, there exists an ISP $J \neq I$ such that $M_J' \geq M_J$. This implies $\nu_J' \leq \nu_J$. By the definition of ϵ_{s_J} in (8), we have $\Phi_J' - \Phi_J \leq \epsilon_{s_J}$, or equivalently $\Phi_J \geq \Phi_J' - \epsilon_{s_J}$. Thus

$$\Phi_{\mathcal{I}} = \Phi_J \geq \Phi_J' - \epsilon_{s_J} = \Phi_{\mathcal{I}}' - \epsilon_{s_J} \geq \Phi_{\mathcal{I}}' - \epsilon_{s_{-I}}.$$

If s_I maximizes the consumer utility $\Phi_{\mathcal{I}}$, for any strategy $s_I' \neq s_I$, we have $\Phi_{\mathcal{I}}' \leq \Phi_{\mathcal{I}}$. By the definition of δ_{s_I}, we have $m_I' - m_I \leq \delta_{s_I}$, or equivalently $m_I \geq m_I' - \delta_{s_I}$. ∎

Proof of Corollary 2: If $s_{\mathcal{I}}$ is a market share Nash equilibrium, by definition, each s_I is a market share best response of s_{-I}. Therefore, by Theorem 6, we have

$$\Phi_{\mathcal{I}} \geq \Phi_{\mathcal{I}}' - \epsilon_{s_{-I}} \geq \Phi_{\mathcal{I}}' - \epsilon_{s_{\mathcal{I}}} \qquad \forall s_I' \neq s_I$$

which concludes that $s_{\mathcal{I}}$ is a consumer utility $\epsilon_{s_{\mathcal{I}}}$-Nash equilibrium. If $s_{\mathcal{I}}$ is a consumer utility Nash equilibrium, by definition, each s_I is a consumer utility best response of s_{-I}. Therefore, by Theorem 6, we have

$$m_I \geq m_I' - \delta_{s_I} \geq m_I' - \delta_{s_{\mathcal{I}}} \qquad \forall s_I' \neq s_I$$

which concludes that $s_{\mathcal{I}}$ is a market share $\delta_{s_{\mathcal{I}}}$-Nash equilibrium. ∎

REFERENCES

[1] FCC, Washington, DC, USA, "FCC acts to preserve Internet freedom and openness," News Release, Dec. 21, 2010 [Online]. Available: http://www.fcc.gov/document/fcc-acts-preserve-internet-freedom-and-openness

[2] Google, Mountain View, CA, USA, "Google corporate Website," 2012 [Online]. Available: http://www.google.com/corporate/tech.html

[3] "Local loop unbundling," 2012 [Online]. Available: http://en.wikipedia.org/wiki/Local-loop_unbundling

[4] K. Florance, "Netflix technology blog," 2011 [Online]. Available: http://techblog.netflix.com/2011/01/netflix-performance-on-top-isp-networks.html

[5] M. Campbell and J. Browning, "Apple, Google asked to pay up as mobile operators face data flood," *Bloomberg News*, Dec. 8, 2010 [Online]. Available: http://www.bloomberg.com/news/2010-12-07/apple-google-asked-to-pay-up-as-europeean-operators-inundated-by-data.html

[6] S. Caron, G. Kesidis, and E. Altman, "Application neutrality and a paradox of side payments," in *Proc. ACM ReARCH*, Nov. 2010, Article no. 9.
[7] D. M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Comput. Netw. ISDN Syst.*, vol. 17, no. 1, pp. 1–14, Jun. 1989.
[8] J. P. Choi and B.-C. Kim, "Net neutrality and investment incentives," *Rand J. Econ.*, vol. 41, no. 3, pp. 446–471, Autumn, 2010.
[9] K. Claffy, "Workshop on Internet economics (WIE2011) report," *Comput. Commun. Rev.*, vol. 42, no. 2, pp. 110–114, Apr. 2012.
[10] C. Courcoubetis and R. Weber, *Pricing Communication Networks: Economics, Technology and Modelling*. Hoboken, NJ, USA: Wiley, 2003.
[11] J. Crowcroft, "Net neutrality: The technical side of the debate: A white paper," *Comput. Commun. Rev.*, vol. 37, no. 1, pp. 49–56, Jan. 2007.
[12] R. J. Deneckere and R. P. McAfee, "Damaged goods," *J. Econ. Manage. Strategy*, vol. 5, no. 2, pp. 149–174, Jun. 1996.
[13] A. Dhamdhere and C. Dovrolis, "Can ISPs be profitable without violating network neutrality?," in *Proc. ACM NetEcon*, Aug. 2008, pp. 13–18.
[14] N. Economides and J. Tag, "Network neutrality and network management regulation: Quality of service, price discrimination, and exclusive contracts," in *Research Handbook on Governance of the Internet*. London, U.K.: Elgar, 2012.
[15] B. Hermalin and M. L. Katz, "The economics of product-line restrictions with an application to the network neutrality debate," *Inf. Econ. Policy*, vol. 19, no. 2, pp. 215–248, 2007.
[16] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, 1998.
[17] C. Labovitz, D. McPherson, S. Iekel-Johnson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *Proc. ACM SIGCOMM*, New Delhi, India, 2010, pp. 75–86.
[18] R. T. B. Ma, D. Chiu, J. C. Lui, V. Misra, and D. Rubenstein, "Internet Economics: The use of Shapley value for ISP settlement," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 775–787, Jun. 2010.
[19] R. T. B. Ma, D. Chiu, J. C. Lui, V. Misra, and D. Rubenstein, "On cooperative settlement between content, transit and eyeball Internet service providers," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 802–815, Jun. 2011.
[20] R. T. B. Ma and V. Misra, "Congestion equilibrium for differentiated service classes," in *Proc. Allerton Conf. Commun., Control, Comput.*, 2011, pp. 589–594.
[21] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. Oxford, U.K.: Oxford Univ. Press, 1995.
[22] 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.
[23] A. Moylan and B. Mead, "Municipal broadband: Wired to waste," National Taxpayers Union Policy Paper No. 129, Apr. 2012.
[24] J. Musacchio, G. Schwartz, and J. Walrand, "Network neutrality and provider investment incentives," in *Proc. Asilomar Conf.*, Nov. 2007, pp. 1437–1444.
[25] A. Odlyzko, "Paris metro pricing for the Internet," in *Proc. ACM EC*, 1999, pp. 140–147.
[26] C. C. Reinwand, "Municipal broadband the evolution of next generation wireless networks," in *Proc. IEEE Radio Wireless Symp.*, 2007, pp. 273–276.
[27] N. Shetty, G. Schwartz, and J. Walrand, "Internet QoS and regulations," *IEEE/ACM Trans. Netw.*, vol. 18, no. 6, pp. 1725–1737, Dec. 2010.
[28] J. G. Sidak, "A consumer-welfare approach to network neutrality regulation of the Internet," *J. Competition Law Econ.*, vol. 2, no. 3, pp. 349–474, Sep. 2006.
[29] T. Wu, "Network neutrality, broadband discrimination," *J. Telecommun. High Technol. Law*, vol. 2, pp. 141–179, 2003.
[30] M. Yuksel, K. K. Ramakrishnan, S. Kalyanaraman, J. D. Houle, and R. Sadhvani, "Quantifying overprovisioning vs. class-of-service: Informing the net neutrality debate," in *Proce. 19th ICCCN*, 2010, pp. 1–8.

Richard T. B. Ma received the B.Sc. (first-class honors) degree in computer science and M.Phil. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA, in 2010.

During his Ph.D. study, he worked as a Research Intern with the IBM T. J. Watson Research Center, Hawthorne, NY, USA, and Telefonica Research, Barcelona, Spain. He is currently a Research Scientist with the Advanced Digital Science Center, University of Illinois at Urbana–Champaign, USA, and an Assistant Professor with the School of Computing, National University of Singapore, Singapore. His current research interests include distributed systems, network economics, game theory, and stochastic processes.

Vishal Misra (S'98–M'99) received the B.Tech. degree from the Indian Institute of Technology, Bombay, India, in 1992, and the Ph.D. degree from the University of Massachusetts, Amherst, USA, in 2000, both in electrical engineering.

He is an Associate Professor in computer science with Columbia University, New York, NY, USA. His research emphasis is on mathematical modeling of computer systems, bridging the gap between practice and analysis. His recent work includes the areas of Internet economics, peer-to-peer networks, and efficient scheduling policies.

Dr. Misra has served as the Guest Editor for the *Journal of Performance Evaluation* and as TPC Chair and General Chair of the ACM SIGMETRICS conference. He has participated as a member of program committees for conferences such as IEEE INFOCOM, ACM SIGMETRICS, ACM SiGCOMM, IFIP Performance, and IEEE ICNP. He serves on the Board of Directors of ACM SIGMETRICS. He has received an NSF CAREER Award, a DoE CAREER Award, and IBM Faculty Awards.