Paid Prioritization and Its Impact on Net Neutrality

Jingjing Wang
Department of Information Engineering
The Chinese University of Hong Kong
Email: wjj010@ie.cuhk.edu.hk

Richard T. B. Ma
National University of Singapore
Email: tbma@comp.nus.edu.sg

Dah Ming Chiu
Department of Information Engineering
The Chinese University of Hong Kong
Email: dmchiu@ie.cuhk.edu.hk

***Abstract*—The net neutrality debate has been centered at the question: whether price and service differentiation should be allowed for the Internet? We focus on a monopoly market, where regulation is often required, and study the type of service differentiation where an option of paid prioritization is provided for the Content Providers (CPs) by an Internet Service Provider (ISP). We study the ISP's pricing strategy and the corresponding CPs' responses. Based on the higher level CPs' choices of service classes and the lower level traffic equilibrium, we analyze the utility of the ISP and the CPs as well as the social welfare. By comparing the induced social welfare under different settings, we find that ISP's optimal pricing leads to an efficient differentiation among the CPs such that the social welfare is highly optimized. We also identify the conditions under which the ISP would have a strong incentive to expand its capacity when the market grows. In conclusion, our results support the use of priority-based pricing and service differentiation rather than imposing net neutrality regulations.**

I. Introduction

Net neutrality (also known as network neutrality) has been heatedly debated in recent years among policy and law makers. At the center of the debate, Internet Service Providers (ISPs) and Content Providers (CPs) argue whether service and price differentiation should be allowed for the Internet transport services, e.g., IP transit and content delivery services. ISPs argue that a non-neutral network is more beneficial for the Internet ecosystem. First, they argue that due to network congestion and security, network management is needed to differentiate traffic and maintain a more efficient network. Second, without service and price differentiation, ISPs will not have incentives to expand their infrastructure capacity and provide better quality services, which will impair the future development of the Internet. Third, they argued that the revenue model of the two-sided Internet market is not balanced: ISPs only earn fixed monthly payments from the end-user side; however, CPs earn much more from the online services as well as advertising for their media customers. In this two-sided market, ISPs do not obtain a share from the content side revenue, and therefore, get the feeling that the CPs are free-riding on their invested infrastructure. On the other hand, CPs think the boom in services at the edge of the network over the past decade should be credited to net neutrality. If net neutrality is abandoned, ISPs may have too much pricing power to charge CPs and also obtain chances to favor specific CPs, leading an end to the Internet boom.

The provision of price and service differentiation will change both the economic structure and traffic demand for the Internet. Charging the CPs for premium services will create new money flows from CPs to ISPs, and therefore, it is worth studying how these two parties would respond to this new economic structure. End users will also perceive varied delays for contents from CPs that use different service classes. We are curious about how the traffic demand would be reshaped due to the service differentiation.

To understand the changes caused by price and service differentiation, our model includes three parties: a monopoly ISP, a set of CPs and their end users, which interact in the following sequence: 1) the ISP sets the price for a priority-based premium service; 2) CPs choose which service to use; 3) end users adjust the traffic demand based on received quality. We model the service differentiation based on an M/M/1 priority queueing model in Section II and analyze the CPs' choices of service classes in Section III. This type of priority-based price and service differentiation is also referred to as *paid prioritization* in this work and analyzed via backward induction. In Section IV, we evaluate the system performance, i.e., delay and throughput, and analyze the ISP's pricing strategy and its impact on the ISP's profit, the CPs' utility and the social welfare, through which we obtain some insights and implications on the net neutrality debate. Our contributions and findings include:

- Under any fixed CPs' choices of service class, we prove the uniqueness of a lower-level traffic equilibrium (Theorem 1) and its monotonicity properties (Theorem 2).
- For a network neutral case, we prove a unique dominant-strategy equilibrium (Theorem 4) for the CPs' strategies.
- Under paid prioritization, we characterize the *congestion equilibrium* of the CPs' strategies (Theorem 5 and 6).
- By evaluating the utility of the ISP and the CPs and the resulting social welfare, we find that
    1) Paid prioritization induces a higher social welfare than that under a neutral network.
    2) Under the ISP's optimal price, the social welfare is highly optimized but the CPs' total utility is reduced due to the monetary transfer from the CPs to the ISP.
- By evaluating the ISP's capacity expansion decisions under different system scales, we find that
    1) The ISP's optimal price is small when the system is less congested or has a small scale.
    2) The ISP will be incentivized to expand capacity with

ISBN 978-3-901882-58-6 © 2014 IFIP

the growth of demand, if its capacity cost is low and the system scale is neither too small nor too large.

Intuitively, the social welfare increases in a non-neutral network because high-valued, delay-sensitive contents could be better served. When a system is not congested, CPs may find an ordinary service sufficient; when a system is small, the loss in statistical multiplexing might outperform the benefit of prioritization. Both cases would naturally lead to a lower optimal price for the ISP. We believe that our modeling framework and results provide new insights to the net neutrality debate.

II. Priority-Based Service and Traffic Equilibrium

The recent debate on network neutrality manifests itself in the cases where the last-mile ISPs intended to differentiate services and charge CPs, e.g., Apple and Google, for service fees [3]. We want to understand whether or not a non-neutral treatment of contents for different CPs is beneficial for the Internet ecosystem as a whole. In practice, the bottleneck of the Internet is often at the last-mile connection towards the end users [7], both wired and wireless. We focus on a monopolistic last-mile ISP with bottleneck capacity μ. This is the case where market competition does not exist and regulations might be most in need. Because latency arises from the queueing delays in routers, we use the M/M/1 queueing framework to characterize network traffic and quantify the varied average delays of each service class, similar to prior work [4] [11]. In practice, priority-based differentiation at a packet level could also be implemented via feasible schemes like DiffServ [2] for the Internet.

We consider a set $\mathcal{N}$ of CPs, from which end-users request for content via the ISP. We assume that the ISP provides differentiated services for the CPs: an ordinary class ($\mathcal{L}$-class) and a premium class ($\mathcal{H}$-class). Traffic sent in $\mathcal{H}$-class will have a higher priority and be sent before those queued in $\mathcal{L}$-class. For traffic in the same service class, they are processed in an FIFO manner. Under this two-class M/M/1 priority queueing system, all the CPs get to choose whether or not to use the premium service. We denote $s_{\mathcal{N}} = \{\mathcal{H}, \mathcal{L}\}$ as a strategy profile of the CPs, where $\mathcal{H}$ and $\mathcal{L}$ denote the set of CPs in $\mathcal{H}$-class and $\mathcal{L}$-class, satisfying $\mathcal{H} \cap \mathcal{L} = \emptyset$ and $\mathcal{H} \cup \mathcal{L} = \mathcal{N}$. For each CP $i \in \mathcal{N}$, we denote λ_i as its throughput achieved at the ISP, defined as $\lambda_i = \Lambda_i(\phi_i)$. ϕ_i is a congestion/quality metric of the service experienced by CP i's end users. In our context, ϕ_i denotes the average queueing delay for CP i at the ISP. We denote λ_i^{max} as the maximum achievable throughput of CP i.

Assumption *1:* For any CP $i \in \mathcal{N}$, its throughput $\Lambda_i(\phi_i)$ is exogenous, continuous, strictly decreasing and satisfies

$$\lim_{\phi_i \to 0} \Lambda_i(\phi_i) = \lambda_i^{max}, \quad \text{and} \quad \lim_{\phi_i \to +\infty} \Lambda_i(\phi_i) = 0.$$

Assumption 1 states that the CP's throughput decreases as the service quality is degraded. This effect comes by two reasons: 1) network quality affects the throughput naturally, and 2) worse performance also discourages the demand from end-users, which in return affects the aggregate throughput of the CP. In particular, λ_i^{max} can be considered as the throughput when the network congestion does not exist and the maximum number of interested users are using CP i at a full speed. This maximum throughput depends on how popular the content is, i.e., how many users of the CP, and the content's maximum throughput, e.g., 5Mb/s is enough for high-quality Netflix streaming movies.

We denote $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{|\mathcal{N}|})$ as the throughput of all the CPs. For any CPs' joint decision $s_{\mathcal{N}}$, we define $\lambda_{\mathcal{H}} = \sum_{i \in \mathcal{H}} \lambda_i$ and $\lambda_{\mathcal{L}} = \sum_{i \in \mathcal{L}} \lambda_i$ as the aggregate throughput for each service class. Based on the M/M/1 priority queueing theory [9], we derive queueing the delays $\phi_{\mathcal{H}}$ and $\phi_{\mathcal{L}}$ of the $\mathcal{H}$- and $\mathcal{L}$-class, respectively, as follows:

$$\begin{cases} \phi_{\mathcal{H}} = \Phi_{\mathcal{H}}(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}}) = \dfrac{1}{\mu - \lambda_{\mathcal{H}}}, \\ \phi_{\mathcal{L}} = \Phi_{\mathcal{L}}(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}}) = \dfrac{\mu}{(\mu - \lambda_{\mathcal{H}} - \lambda_{\mathcal{L}})(\mu - \lambda_{\mathcal{H}})}. \end{cases} \tag{1}$$

By defining $\boldsymbol{\phi} = (\phi_{\mathcal{H}}, \phi_{\mathcal{L}})$ as the system queueing delay, the above equations can be written in a matrix form as:

$$\boldsymbol{\phi} = \Phi(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}}) \tag{2}$$

Under any strategy profile $s_{\mathcal{N}}$ of the CPs and the resulting queueing delay $\boldsymbol{\phi}$, we can characterize each CP i's throughput by Assumption 1 as follows

$$\lambda_i = \begin{cases} \Lambda_i(\phi_{\mathcal{H}}) & \text{if } i \in \mathcal{H}, \\ \Lambda_i(\phi_{\mathcal{L}}) & \text{if } i \in \mathcal{L}; \end{cases} \tag{3}$$

or in a matrix form as

$$\boldsymbol{\lambda} = \Lambda(\boldsymbol{\phi}, s_{\mathcal{N}}). \tag{4}$$

Assumption 1 characterizes how throughput responses to delays and Equation (1) characterizes how the delays are affected by throughput in return. The following Definition 1 describes the steady-state of the system delays.

Definition *1:* For any fixed strategy profile $s_{\mathcal{N}}$ and system capacity μ, a delay vector $\boldsymbol{\phi}$ is an equilibrium if it satisfies

$$\begin{cases} \phi_{\mathcal{H}} = \Phi_{\mathcal{H}}(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}}), \\ \phi_{\mathcal{L}} = \Phi_{\mathcal{L}}(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}}), \end{cases} \quad \text{where } \boldsymbol{\lambda} = \Lambda(\boldsymbol{\phi}, s_{\mathcal{N}})$$

or in a matrix form as

$$\boldsymbol{\phi} = \Phi(\Lambda(\boldsymbol{\phi}, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) \tag{5}$$

Definition 1 states that the system delay $\boldsymbol{\phi}$ in equilibrium would induce the amount of CP traffic, i.e., $\boldsymbol{\lambda} = \Lambda(\boldsymbol{\phi}, s_{\mathcal{N}})$, that causes exactly that amount of delays, i.e., $\boldsymbol{\phi} = \Phi(\boldsymbol{\lambda}, \mu, s_{\mathcal{N}})$, in both service classes.

Theorem *1:* For any fixed strategy profile $s_{\mathcal{N}}$ and system capacity μ, there always exists a unique equilibrium $\boldsymbol{\phi}$.

Theorem 1 states the existence and uniqueness of a lower-level traffic equilibrium. From Equation (1), one may observe that the delay of $\mathcal{H}$-class $\varphi_{\mathcal{H}}$ does not depend on the CPs in $\mathcal{L}$-class *while the delay of $\mathcal{L}$-class does depend on the CPs in $\mathcal{H}$-class*. In fact, we can extend the result for any number of priority classes based on the same technique.

By Theorem 1, we define $\boldsymbol{\varphi}(\mu, s_{\mathcal{N}}) = (\varphi_{\mathcal{H}}, \varphi_{\mathcal{L}})$ as the unique equilibrium of a system $(\mu, s_{\mathcal{N}})$. We simplify the notation as $\varphi(s_\mathcal{N})$ when a fixed capacity μ is considered. We also use $\varphi(\mathcal{H},\mathcal{L})$ to denote $\varphi(\{\mathcal{H},\mathcal{L}\})$ for the convenience of illustration when we look further into the service classes. Next, we study the properties for the system equilibrium.

Theorem *2:* For any $s_\mathcal{N}$ and $s'_\mathcal{N}$ with $\mathcal{H},\mathcal{H}' \notin \emptyset$ and $\mathcal{H}' \subset \mathcal{H}$, the unique equilibrium $\boldsymbol{\varphi}(\mu, s_\mathcal{N}) = \big(\varphi_\mathcal{H}, \varphi_\mathcal{L}\big)$ satisfies

$$\varphi_\mathcal{H}(\mu, s'_\mathcal{N}) < \varphi_\mathcal{H}(\mu, s_\mathcal{N}) < \varphi_\mathcal{L}(\mu, s_\mathcal{N}), \quad \forall\ \mu > 0;$$

$$\varphi_\mathcal{H}(\mu_1, s_\mathcal{N}) > \varphi_\mathcal{H}(\mu_2, s_\mathcal{N}), \quad \forall\ 0 < \mu_1 < \mu_2.$$

Theorem *3:* For all $\mathcal{H},\mathcal{L} \notin \emptyset$, such that $\mathcal{H}\cap\mathcal{L} = \emptyset$ and $\mathcal{H}\cup\mathcal{L} = \mathcal{N}$, the $\mathcal{L}$-class delay satisfies $\varphi_\mathcal{L}(\emptyset,\mathcal{N}) < \varphi_\mathcal{L}(\mathcal{H},\mathcal{L})$.

Theorem 2 intuitively states that in equilibrium, the delay of $\mathcal{H}$-class is always smaller than that of $\mathcal{L}$-class, and when the system's capacity μ increases, the delay of $\mathcal{H}$-class always decreases. However, this is not always true for $\mathcal{L}$-class. Theorem 3 implies that any partition of the set $\mathcal{N}$ of CPs will cause a longer delay in $\mathcal{L}$-class although the number of CPs in $\mathcal{L}$-class is smaller. Theorem 2 and 3 also infer that $\varphi_\mathcal{H}(\mathcal{N},\emptyset)$ is a delay upper-bound for $\mathcal{H}$-class and $\varphi_\mathcal{L}(\emptyset,\mathcal{N})$ is a delay lower-bound for $\mathcal{L}$-class, where both bounds effectively capture the same single-class delay, i.e., $\varphi_\mathcal{H}(\mathcal{N},\emptyset) = \varphi_\mathcal{L}(\emptyset,\mathcal{N})$.

III. CPs' Strategies and Equilibrium

In this section, we study CPs' choices of service classes which are fixed as $s_\mathcal{N}$ in Section II. We assume that the ISP charges c for per unit traffic sent in $\mathcal{H}$-class, which will induce a new economic structure of the system. CPs now have to trade off between an extra payment to the ISP and a higher traffic demand in $\mathcal{H}$-class due to the smaller induced delay. For each CP i, we denote v_i as its per-unit traffic valuation, e.g., the profit generated by advertising for clients (Google), e-commerce (Amazon) or online services (Netflix). We consider the case of a fixed μ in this section and $\varphi(s_\mathcal{N})$ is in short for $\varphi(\mu, s_\mathcal{N})$. We also denote u_i as CP i's utility, which depends on the strategies of the CPs $s_\mathcal{N}$ as follows:

$$u_i = \left\{ \begin{array}{cl} v_i\Lambda_i\big(\varphi_\mathcal{L}(s_\mathcal{N})\big) & \text{if } i \in \mathcal{L}, \\ (v_i - c)\Lambda_i\big(\varphi_\mathcal{H}(s_\mathcal{N})\big) & \text{if } i \in \mathcal{H}. \end{array} \right.$$

Given a fixed capacity μ and an ISP charge c, the set $\mathcal{N}$ of CPs choose their service classes strategically so as to maximize their own utilities. Under this game-theoretic model, the special case of $c = 0$ will induce a neutral network, where choosing the premium service is a *dominant strategy* for all CPs and the system only has a single service class in effect.

Theorem *4:* When $c = 0$, $s_\mathcal{N} = (\mathcal{H},\mathcal{L}) = (\mathcal{N},\emptyset)$ is the unique dominant strategy equilibrium.

When $c > 0$, each CP needs to tradeoff between the per-unit traffic profit (v_i for $\mathcal{L}$-class and $v_i - c$ for $\mathcal{H}$-class) and the achieved throughput ($\Lambda_i(\varphi_\mathcal{H})$ and $\Lambda_i(\varphi_\mathcal{L})$). In general, dominant strategy equilibrium often does not even exist. Instead, a Nash equilibrium of this simultaneous-move game can be defined as follows.

Definition *2:* Any strategy profile $s_\mathcal{N} = (\mathcal{H},\mathcal{L})$ is a Nash equilibrium if it satisfies

$$\frac{v_i - c}{v_i} \left\{ \begin{array}{ll} > \dfrac{\Lambda_i\Big(\varphi_\mathcal{L}\big(\mathcal{H}\backslash\{i\}, \mathcal{L}\cup\{i\}\big)\Big)}{\Lambda_i\big(\varphi_\mathcal{H}(s_\mathcal{N})\big)} & \text{if } i \in \mathcal{H}, \\ \\ \le \dfrac{\Lambda_i\big(\varphi_\mathcal{L}(s_\mathcal{N})\big)}{\Lambda_i\Big(\varphi_\mathcal{H}\big(\mathcal{H}\cup\{i\}, \mathcal{L}\backslash\{i\}\big)\Big)} & \text{if } i \in \mathcal{L}. \end{array} \right.$$

When applying the Nash equilibrium concept, a "common knowledge" assumption is often needed, i.e., every CP knows all the other CPs, and they are aware of the fact that they know each other, and etc. In practice, the number of CPs is large and it is unrealistic to assume that all CPs know the characteristics of all their competitors. Moreover, although each CP only makes a binary decision in the strategic game, the total strategy space has a size of $2^{|\mathcal{N}|}$, which makes the evaluation of the Nash equilibrium computationally expensive. To resolve this problem, analogous to the "price-taking" [14] assumption of competitive equilibria in classic economics, we can make a similar "congestion-taking" assumption for the CPs.

Assumption *2:* For any service class $\mathcal{X}$, any CP $i \notin \mathcal{X}$ uses $\varphi_\mathcal{X}$ as an estimate of the ex-post congestion $\varphi_{\mathcal{X}\cup\{i\}}$ in its decision making, and we define $\varphi_\mathcal{X} = 0$ if $\mathcal{X} = \emptyset$.

Based on the above "congestion-taking" assumption, we adopt the concept of "congestion equilibrium" [12] as follows.

Definition *3:* Any strategy profile $s_\mathcal{N} = (\mathcal{H},\mathcal{L})$ is a congestion equilibrium if it satisfies

$$\frac{v_i - c}{v_i} \left\{ \begin{array}{ll} > \Lambda_i\big(\varphi_\mathcal{L}(s_\mathcal{N})\big)/\Lambda_i\big(\varphi_\mathcal{H}(s_\mathcal{N})\big) & \text{if } i \in \mathcal{H}, \\ \\ \le \Lambda_i\big(\varphi_\mathcal{L}(s_\mathcal{N})\big)/\Lambda_i\big(\varphi_\mathcal{H}(s_\mathcal{N})\big) & \text{if } i \in \mathcal{L}. \end{array} \right.$$

Notice that the "congestion-taking" might not be hold for influential CPs such as Google and Netflix in practice; however, Vernon Smith, behavioral economist and Nobel laureate, has shown in his pioneering empirical work [18] that for a modest scale of supply and demand, e.g., even three players in both the supply and demand sides, the system (a handful of powerful CPs in our context) also adapt itself to its competitive equilibrium which makes the condition required by the classical competitive equilibrium unnecessary.

Although any CP's throughput function $\Lambda_i(\phi_i)$ can be quite general under Assumption 1, the above setting does not yet capture the traffic characteristics of the CPs. In particular, we consider the following class of the throughput functions

$$\Lambda_i(\phi_i) = \lambda_i^{max} e^{-\alpha_i\phi_i}, \qquad (6)$$

where each CP i's throughput is characterized by a delay-sensitivity parameter α_i on the exponent term. The bigger the α_i is, the more sensitive CP i's throughput is to the delay. Therefore, big values of α_i can be used to model inelastic traffic, e.g., realtime streaming content, and small values of α_i can be used to model elastic traffic, e.g., file download.

Another way to understand α_i is by evaluating the economic metric *demand elasticity of delay* ϵ_i defined by

$$\epsilon_i = \frac{d\Lambda_i(\phi_i)}{d\phi_i}\frac{\phi_i}{\Lambda_i(\phi_i)} = -\alpha_i\phi_i,$$

which captures the ratio of the percentage change in demand caused by the percentage change in the delay. The above equation clearly shows that any CP's demand elasticity of delay is proportional to its value of α_i.

By substituting (6) into Definition 3, we can characterize the congestion equilibrium by the following Theorem 5.

Theorem *5:* For any fixed ISP charge c, a strategy profile $s_{\mathcal{N}}$ is a congestion equilibrium only if

$$\varphi_{\mathcal{L}}(s_{\mathcal{N}}) - \varphi_{\mathcal{H}}(s_{\mathcal{N}}) \begin{cases} > \beta_i(c) & \text{if } i \in \mathcal{H}, \\ \\ \leq \beta_i(c) & \text{if } i \in \mathcal{L}, \end{cases}$$

where $\beta_i(c)$ is defined as

$$\beta_i(c) = \begin{cases} \dfrac{1}{\alpha_i}\ln\left(\dfrac{v_i}{v_i - c}\right) & \text{if } v_i > c, \\ \\ +\infty & \text{otherwise .} \end{cases}$$

Theorem 5 reveals the structural property of a congestion equilibrium for the upper-level CP choices. For each CP i, we can calculate a priority $\beta_i(c)$ such that the CPs with higher priorities (smaller values of $\beta_i(c)$) will be more likely to end up using the premium service.

IV. Performance Evaluation and Implications

In this section, we evaluate the system performance, the utilities of each party and the social welfare via extensive and carefully designed experiments based on the lower-level traffic equilibrium (Section II) and the upper-level congestion equilibrium of the CPs' choices (Section III) together.

We define the ISP profit and the CPs' aggregate utility as

$$U^{ISP} = c\lambda_{\mathcal{H}} = c\sum_{i\in\mathcal{H}}\Lambda_i(\varphi_{\mathcal{H}}) \quad\text{and}\quad U^{CP} = U^{CP}_{\mathcal{H}} + U^{CP}_{\mathcal{L}},$$

where $U^{CP}_{\mathcal{H}} = \sum_{i\in\mathcal{H}} u_i = \sum_{i\in\mathcal{H}}(v_i - c)\Lambda_i(\varphi_{\mathcal{H}})$ and $U^{CP}_{\mathcal{L}} = \sum_{i\in\mathcal{L}} u_i = \sum_{i\in\mathcal{L}} v_i\Lambda_i(\varphi_{\mathcal{L}})$ define the aggregate CP utility in $\mathcal{H}$- and $\mathcal{L}$-class, respectively. We define the social welfare as $U = U^{ISP} + U^{CP}$, the sum of the ISP's profit and the CPs' aggregate utility. We define $\lambda^{max} = \sum_{i\in\mathcal{N}}\lambda^{max}_i$ as the sum of CPs' maximum throughput, which is the maximum traffic demand from the end-users. Under a random setting, we denote $\mathbb{E}\big(\lambda^{max}\big)$ as the mean of this maximum demand.

Although the ISP and CP utilities depend on the values of c and v_i, the following Theorem 6 states that the equilibrium $s_{\mathcal{N}}$ does not change when these values scale linearly.

Theorem *6:* If a strategy profile $s_{\mathcal{N}}$ is a Nash equilibrium or congestion equilibrium for a system with ISP charge c and CP valuations $\{v_i : i \in \mathcal{N}\}$, it is also the same type of equilibrium for any linearly scaled system with $\tilde{c} = kc$ and $\tilde{v_i} = kv_i, i \in \mathcal{N}$ for all $k > 0$. In particular, we have $\tilde{\boldsymbol{\varphi}} = \boldsymbol{\varphi}$, $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}$, $\tilde{U}^{CP} = kU^{CP}$, $\tilde{U}^{ISP} = kU^{ISP}$ and $\tilde{U} = kU$.

By Theorem 6, we could normalize the maximum valuations of CPs to be 1 and choose each v_i as a uniform random variable in $[0,1]$ without loss of generality. We consider a system of $|\mathcal{N}| = 1000$ independent CPs. The CP delay sensitivity parameter α_i and the maximum throughput λ^{max}_i are assumed to be uniformly distributed in $[0,10]$ and $[0, 2\bar{\lambda}_{max}]$, where $\bar{\lambda}_{max}$ denotes the mean of λ^{max}_i. Inspired by the structural property of the congestion equilibrium in Theorem 5, we calculate the congestion equilibrium of the CPs' strategies as follows. Initially, all CPs stay in $\mathcal{L}$-class. We update the CP strategy profile $s_{\mathcal{N}}$ by allowing CPs to move sequentially from $\mathcal{L}$-class to $\mathcal{H}$-class in the ascending order of $\beta_i(c)$ until $s_{\mathcal{N}}$ reaches the congestion equilibrium of Definition 3.

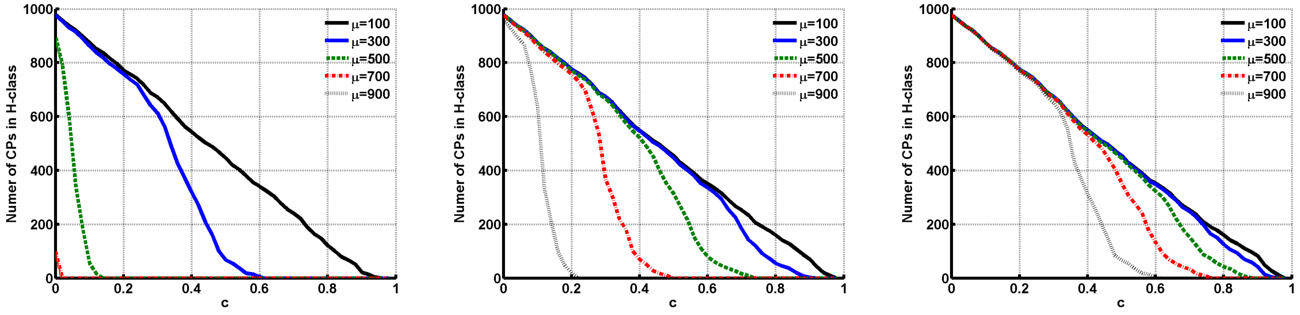
A. System Performance Evaluation

We evaluate various performance metrics of the system, i.e., $|\mathcal{H}|$, $\varphi_{\mathcal{H}}$, $\varphi_{\mathcal{L}}$, $\lambda_{\mathcal{H}}$ and $\lambda_{\mathcal{L}}$ in Fig. 1-5. In each of these figures, we vary $\mathbb{E}\big(\lambda^{max}\big)$ to be 500, 1000 and 1500 in the left, middle and right sub-figures. In each sub-figure, we vary the ISP charge c from 0 to 1 on the x-axis and plot five curves with $\mu = 100, 300, 500, 700$ and 900, respectively.

We start with the boundary cases where $c = 0$ or $c = 1$. We observe that $\mathcal{H} = \mathcal{N}$ under $c = 0$ and $\mathcal{L} = \mathcal{N}$ under $c = 1$ as shown in Fig. 1. When $c = 0$, by Theorem 4, the delay in the premium class is always lower and therefore, all the CPs choose to use it. When $c = 1$, all CPs' valuations are lower than or equal to c, and therefore, no CP can afford to use the premium service. In both cases, the system effectively has only one service class and maintains a neutral network. In particular, when c approaches 1, $\varphi_{\mathcal{H}}$ approaches the M/M/1 theoretic lower bound $1/\mu$ in Fig. 2 and $\varphi_{\mathcal{L}}$ also approaches its lower bound that equals the maximized value for $\varphi_{\mathcal{H}}$ (by Theorem 3) in Fig. 3.

For $c \in (0,1)$, we could make the following observations.

- For any fixed μ and $\mathbb{E}\big(\lambda^{max}\big)$, $|\mathcal{H}|$ and $\varphi_{\mathcal{H}}$ decrease monotonically as c increases in Fig. 1 and Fig. 2, respectively. Although $\varphi_{\mathcal{L}}$ does not decrease monotonically, it also have a decreasing trend with c in general. Besides, because $\varphi_{\mathcal{L}}$ is around two orders of magnitude larger than $\varphi_{\mathcal{H}}$, the trend of the difference in delay $\varphi_{\mathcal{L}} - \varphi_{\mathcal{H}}$ is similar to that of $\varphi_{\mathcal{L}}$.
- For any fixed μ and $\mathbb{E}\big(\lambda^{max}\big)$, $\lambda_{\mathcal{H}}$ decreases and $\lambda_{\mathcal{L}}$ increases as c increases. Furthermore, $\lambda_{\mathcal{H}}$ decreases much slower until $\lambda_{\mathcal{H}}$ has a sharp decline while $\lambda_{\mathcal{L}}$ increases much faster until $\lambda_{\mathcal{L}}$ gets saturated. We can understand that before $\lambda_{\mathcal{H}}$ decreases sharply, the aggregate throughput can be compensated by the decreasing delay although $|\mathcal{H}|$ gets smaller so that $\lambda_{\mathcal{H}}$ does not drop too quickly. At the same time, $\varphi_{\mathcal{L}}$ is still very high so that $\lambda_{\mathcal{L}}$ can not increase too much although $|\mathcal{L}|$ gets bigger.
- Under any fixed capacity μ and charge c, $\varphi_{\mathcal{H}}$ and $\varphi_{\mathcal{L}}$ increases generally as $\mathbb{E}\big(\lambda^{max}\big)$ increases across the three sub-figures in Fig. 2 and 3.
- Under any fixed capacity μ and charge c, $\lambda_{\mathcal{H}}$ increases as $\mathbb{E}\big(\lambda^{max}\big)$ increases across the three sub-figures in Fig. 4. However, in Fig. 5 for any fixed μ and c, 1) when c is

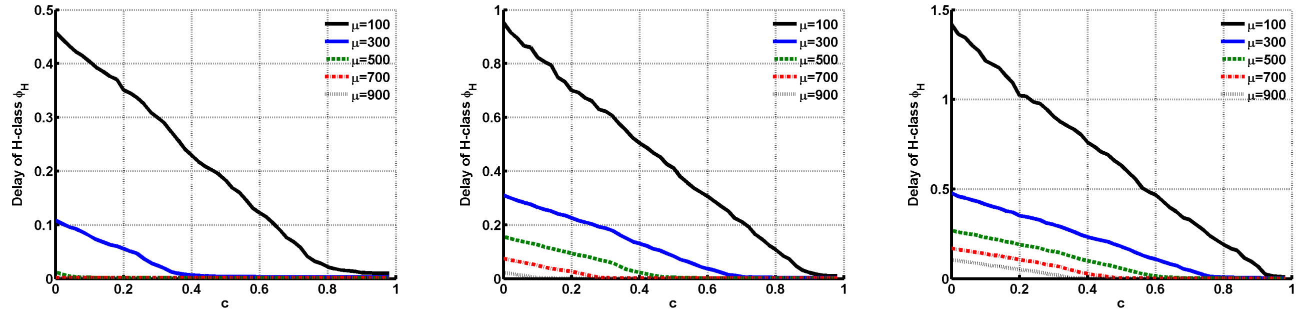


(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 1. Number of CPs in $\mathcal{H}$-class in Equilibrium $|\mathcal{H}|$ under different $\mathbb{E}(\lambda^{max})$.

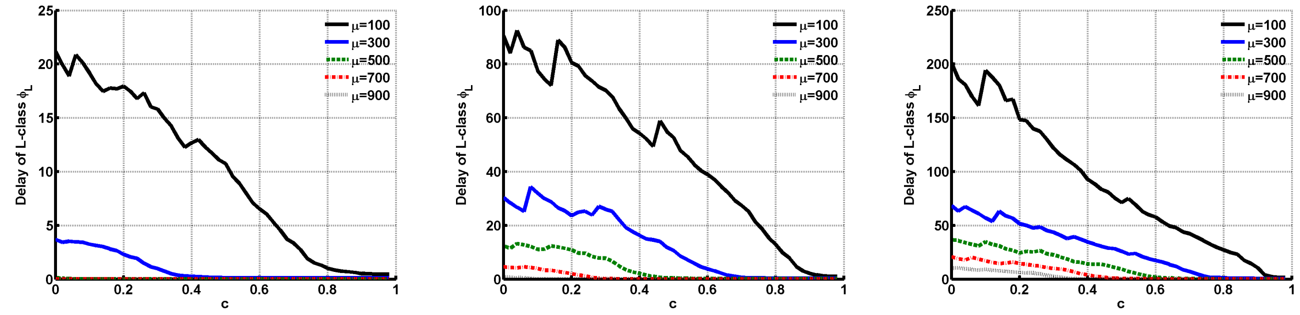


(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 2. Delay of $\mathcal{H}$-class $\varphi_{\mathcal{H}}$ under different demand $\mathbb{E}(\lambda^{max})$.



(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 3. Delay of $\mathcal{L}$-class $\varphi_{\mathcal{L}}$ under different demand $\mathbb{E}(\lambda^{max})$.

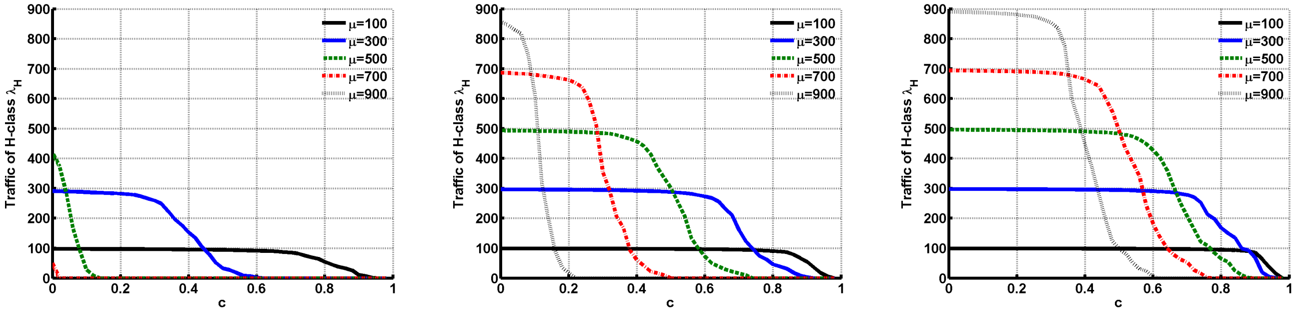
high and μ is large, $\lambda_{\mathcal{L}}$ increases with $\mathbb{E}(\lambda^{max})$ because most CPs cannot afford the premium service and μ is large enough to accommodate more traffic in $\mathcal{L}$-class; 2) when c is high and μ is fixed, $\lambda_{\mathcal{L}}$ will get saturated due to the limitation of μ; 3) when c is low, $\lambda_{\mathcal{L}}$ decreases with the demand because CPs will choose $\mathcal{H}$-class when faced with a cheap charge and high congestion.

- When either the capacity μ gets bigger or the demand $\mathbb{E}(\lambda^{max})$ gets smaller, $|\mathcal{H}|$ decreases sharper with c. This can be understood as the premium service deserve cheaper when the system becomes less congested.

B. ISP Profit, CP Utility, Social Welfare

After analyzing the system performance, we further look into the derived utilities of different parties, i.e., the ISP profit U^{ISP}, the aggregate CP utility U^{CP} and the social welfare $U = U^{ISP} + U^{CP}$ in Fig. 6-8, respectively.

In Fig. 6, we plot the ISP's profit U^{ISP} on the y-axis. In each subfigure, we fix the demand $\mathbb{E}(\lambda^{max})$ and plot how U^{ISP} changes with different settings of μ and c. For each fixed capacity μ, we observe a single peak curve for the U^{ISP}: when c is too small, the ISP cannot earn much despite of the heavy traffic $\lambda_{\mathcal{H}}$ in $\mathcal{H}$-class; when c is too large, the ISP still cannot earn much due to the small number of CPs

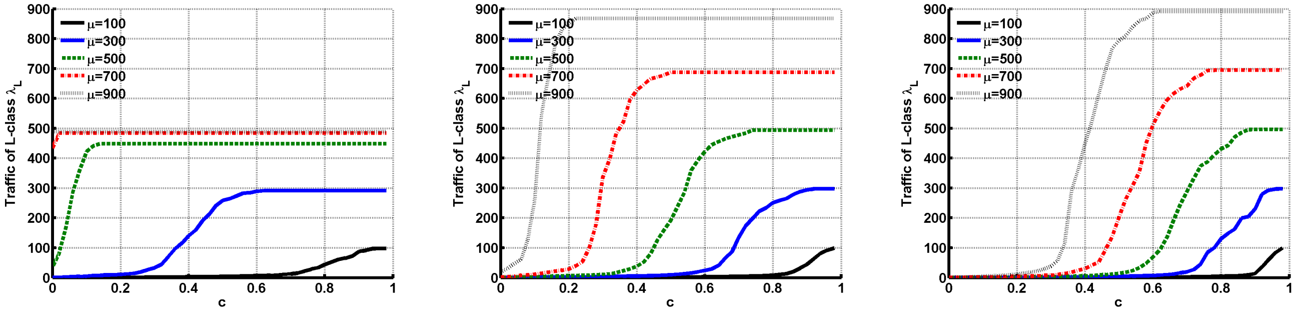


(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 4. Traffic of $\mathcal{H}$-class $\lambda_{\mathcal{H}}$ under different demand $\mathbb{E}(\lambda^{max})$.

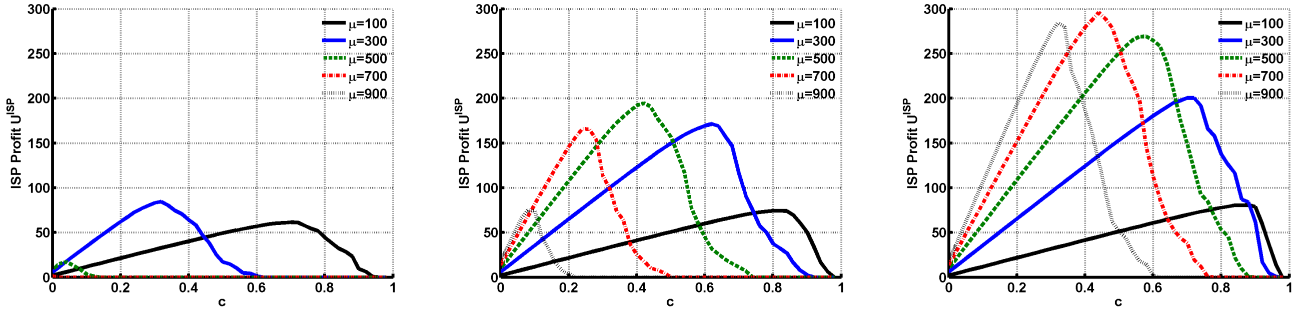


(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 5. Traffic of $\mathcal{L}$-class $\lambda_{\mathcal{L}}$ under different demand $\mathbb{E}(\lambda^{max})$.



(a) $\mathbb{E}(\lambda^{max}) = 500$

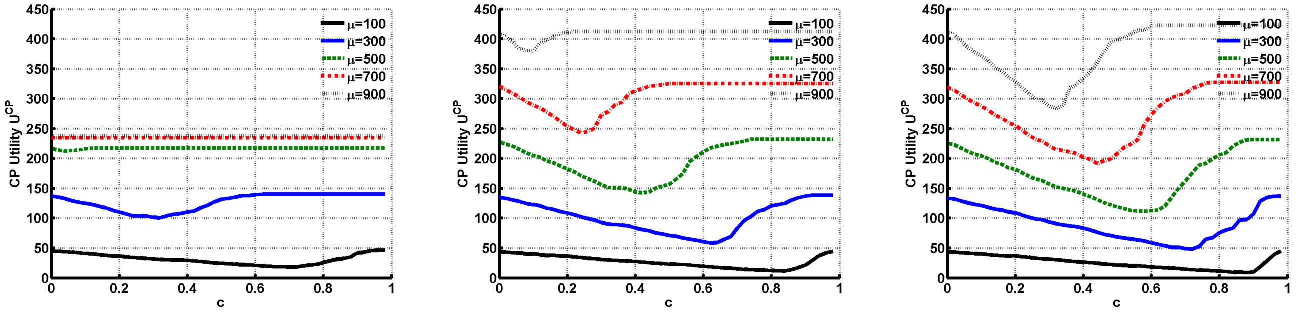
(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 6. ISP Profit U^{ISP} under different demand $\mathbb{E}(\lambda^{max})$.

in $\mathcal{H}$-class, which only induces small amount of traffic $\lambda_{\mathcal{H}}$. We also observe that the ISP's optimal price, denoted as c^*, increases when the capacity μ decreases. However, the optimal μ to maximize U^{ISP} cannot be too large or small, because a small μ limits the amount of traffic $\mathcal{H}$-class can serve; however, a too large μ discourages the CPs to use the premium service due to the lower delay in $\mathcal{L}$-class. Across the three sub-figures, we observe that for a fixed μ, when the demand $\mathbb{E}(\lambda^{max})$ increases, the optimal ISP price becomes higher due to congestion. Also, when μ is larger, U^{ISP} gets larger because the ISP can serve larger demand when μ increases.

In Fig. 7, we plot the aggregate CP utility U^{CP} on the y-axis. We observe that the CP utility increases with the capacity μ for the fixed c and demand $\mathbb{E}(\lambda^{max})$. Each CP utility curve has a valley where the charge c is optimal for the ISP profit (corresponding to the peak in Fig. 6). These valleys show the utility transfer from the CP-side to the ISP under differentiated services in comparison with the neutral cases under $c = 0$ or $c = 1$. We also observe that the optimal ISP charge c^* decreases when its capacity μ expands. Across the three sub-figures, when c is high and μ is small, U^{CP} might decrease with the demand $\mathbb{E}(\lambda^{max})$. This is because when the system gets more congested, the gap $\varphi_{\mathcal{L}} - \varphi_{\mathcal{H}}$ becomes larger resulting in that more CPs choose $\mathcal{H}$-class and the ISP gets

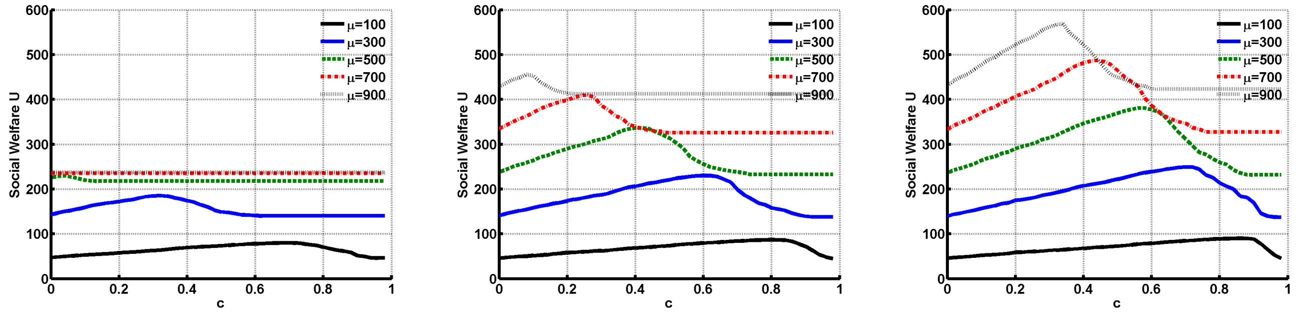


(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 7. CP Profit U^{CP} under different demand $\mathbb{E}(\lambda^{max})$.



(a) $\mathbb{E}(\lambda^{max}) = 500$

(b) $\mathbb{E}(\lambda^{max}) = 1000$

(c) $\mathbb{E}(\lambda^{max}) = 1500$

Fig. 8. Social Welfare U under different demand $\mathbb{E}(\lambda^{max})$.

more payment from the CPs despite of the increase of $\lambda_{\mathcal{H}}$.

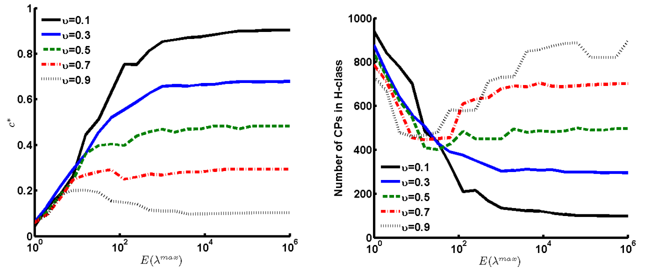
In Fig. 8, we plot the social welfare $U = U^{ISP} + U^{CP}$ on the y-axis. Similar to U^{ISP}, U also shows a single-peak shape. This implies that the priority-based differentiation or paid prioritization provides better social welfare than that of a neutral single-class system (under $c = 0$ or $c = 1$). Similar to U^{CP}, U increases when the ISP capacity μ expands. Interestingly, the valley of U^{CP} corresponds to the peak of U. This implies that when the ISP maximizes its profit by choosing an optimal charge c^*, it effectively differentiates the CPs with different valuations (v_i) and delay sensitivities (α_i) such that CPs with higher valuation and sensitivity, i.e., the CPs with small values of $\beta_i(c)$, would use the premium service and get high throughput (also by Theorem 5). Across the three sub-figures, we also observe that U increases with the demand $\mathbb{E}(\lambda^{max})$ and its increase is steeper when the capacity μ is larger, as it could accommodate more traffic.

Implications on net neutrality: By comparing the derived utilities of different parties and the social welfare, we find that the ISP's optimal pricing strategy is aligned with social welfare: effective differentiation for profit maximization is also good for prioritizing CPs with higher valuations. In this sense, priority based service differentiation is better for social welfare than a neutral network. However, it is true that the optimal social welfare is achieved by sacrificing CPs with lower valuations and biased towards the ISP for profit distribution. Thus, from a fairness perspective, policy makers might want to regulate the price not to be too high so as to balance the social welfare and fairness among different parties.

C. ISP Optimal Pricing and Investment Incentives

In this subsection, we continue to study the ISP investment incentives if it is allowed to provide paid prioritization. We define $\nu = \mu/\mathbb{E}(\lambda^{max})$ to be the capacity for per-maximum traffic demand. We consider the case that the ISP would expand its capacity to meet a growing demand $\mathbb{E}(\lambda^{max})$ by keeping a fixed ratio ν. Under this setting, we could 1) study how the ISP would react towards the market growth and varied system scale by adjusting its optimal charge c^* and capacity μ; 2) obtain trends and insights for the real Internet ecosystem, the scale of which could be larger than the settings used in previous subsections. We vary $\mathbb{E}(\lambda^{max})$ from 10^0 to 10^6 and plot five curves with $\nu = 0.1$, 0.3, 0.5, 0.7 and 0.9 to simulate the situations where the ISP maintains from a fairly congested system ($\nu = 0.1$) to a much less congested system ($\nu = 0.9$).

We search for the ISP's optimal price c^* corresponding to the single peak of the ISP profit for each given μ. In Fig. 9, we plot c^* and $|\mathcal{H}|$ on the y-axis in the left and right sub-figures, respectively. From the left sub-figure, we observe that when the system scale is small, i.e., when $\mathbb{E}(\lambda^{max})$ and μ are both small, the optimal charge of the ISP is very small and most of the CPs will be in $\mathcal{H}$-class. This can be explained by the fact that with a small-scale M/M/1 system, it is not so



(a) ISP Optimal Price c^* (b) Number of CPs in $\mathcal{H}$-class $|\mathcal{H}|$

Fig. 9. ISP's optimal price and the number of CPs in $\mathcal{H}$-class

efficient to do price and service differentiation. However, when the system scale becomes large, the optimal price of the ISP increases and is stabilized at a certain level. This is because the statistical multiplexing effect makes the system more efficient and the ISP is able to differentiate the CPs better. In particular, when the ISP provisions a larger capacity, it results in a lower optimal price because the system becomes less congested and CPs have less incentives to use the premium service. However for the ISP, there is also a tradeoff between charging a high c^* and inducing a larger $\lambda_{\mathcal{H}}$ so as to maximize its profit U^{ISP}. Eventually, the optimal capacity planning ν for the ISP also depends on the cost for deploying more capacity. When the cost of capacity is high, the ISP will not have incentives to maintain a system with a high value of ν, i.e., a system with low congestion. This is because a high value of ν will induce a low optimal price c^*, and if c^* is lower than the capacity cost, it is not sustainable for the ISP to expand capacity.

Implications on ISP investment: Under paid prioritization, ISPs do have incentives to expand capacity when demand increases. However, how fast the ISP will expand its capacity compared to the growth of demand depends on the system scale and its capacity cost. In particular, the ISP does not have a strong incentive to expand capacity, i.e., to keep a low value of ν, if its capacity expanding cost is high or the system scale is either too small or too large.

V. Related Works

To address net neutrality rigorously, economists have analyzed this issue from various perspectives. Hermalin and Katz [10] regarded the realization of net neutrality equivalent to the imposition of a single product quality requirement. Economides and Tag [8] addressed the various regulations combined with quality service, differentiated pricing and exclusive contracts through a two-sided market model. Their discussion on cross-side externality acts as the rationale for the government intervention. Njorogel *et al.* [16] study this problem through a game theoretical two-sided model and reach conclusions on the investment incentives. Musacchio *et al.* [15] compared one-sided and two-sided pricing in terms of social welfare and tried to address the question whether the non-directly connected ISPs should charge CPs or not.

Unlike the analysis from economical view which only consider the basic money and utility exchange as well as the positive externalities (network effects), the following works make another step further by incorporating network characteristics into the analysis to model the traffic demand. Choi and Kim [6] first adopted the priority queueing framework to capture the network's response towards neutral and non-neutral cases. Hotelling model was used for the competition between CPs under differentiated service priorities. We extend the two-CP model to a more general one and allow CPs to choose their preferred service class. Jan *et al.* [11] and Altman *et al.* [1] also adopt queueing delays to model the congestion externality. Different from their work, we analyze the problem in a microeconomic way and model the congestion externality that affects individual CPs' throughput by which the analysis is more detailed.

Our work is also related to Ma et al. [13], which proposed to use a Public Option ISP as a better alternative for net neutrality regulations. In terms of modeling, we similarly characterize the CP traffic and derive a unique traffic equilibrium. The difference is that their model is based on a Paris Metro Pricing [5], [17] framework, where the differentiated service classes are based on capacity sharing as characterized in [5]. Our model however does not separate the capacity physically and therefore, the two service classes in our model are interdependent. Also, contrast to the result of Ma et al. [13] where a monopoly ISP has an incentive to make the lower-class a "damaged good" and the higher-class non-work-conserving, under the priority based service differentiation, our system is always work-conserving and the ISP's optimal pricing strategy is highly aligned with social welfare.

Moreover from the law perspective, Wu [19] focused on the discriminatory issues brought by the violation of net neutrality. Yoo [20] also focus on the economics of congestion to propose a new analytical framework for assessing such restrictions.

VI. Conclusions

In this work, we study paid prioritization, i.e., the priority-based price and service differentiation, which we find is beneficial for the Internet ecosystem. Our service differentiation is based on the M/M/1 priority queueing delay and we established a two-level equilibrium model: the lower-level traffic equilibrium and the higher-level CP choice equilibrium. The lower-level equilibrium models the end-user traffic demand responded to network congestion and the higher-level equilibrium captures the CPs' business incentives and decisions. Based on our model, we evaluate the system performance and the utilities of different parties. We find that the ISP's optimal pricing strategy is highly aligned with the social welfare, although the utility of CPs is reduced due to the payment to the ISP. The results imply that price and service differentiation is beneficial for the Internet compared to a neutral network; however, policy makers might still need to carefully regulate the ISP to provide better fairness among the ISP and the CPs. We also investigated the optimal pricing of ISP under different system scales and identified the conditions under which the ISP will have incentives to expand capacity with the growth of traffic demand.

Acknowledgement

This work was supported by Singapores 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-252-000-526-112.

References

[1] E. Altman, A. Legout, and Y. Xu. Network non-neutrality debate: An economic analysis. *NETWORKING 2011, Lecture Notes in Computer Science*, pages 68–81.

[2] D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, and S. Blake. An architecture for differentiated services. *IETF Request for Comment 2475, December 1998.*

[3] M. Campbell and J. Browning. Apple, Google asked to pay up as mobile operators face data flood. *Bloomberg News*, December 7 2010.

[4] S. Caron, G. Kesidis, and E. Altman. Application neutrality and a paradox of side payments. *Proceedings of the ACM ReARCH '10*, November 2010.

[5] C.-K. Chau, Q. Wang, and D.-M. Chiu. On the viability of paris metro pricing for communication and service networks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9, 2010.

[6] J. P. Choi and B.-C. Kim. Net neutrality and investment incentives. *The Rand Journal of Economics*, 41(3):446–471, Autumn 2010.

[7] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. John Wiley & Sons Ltd., 2003.

[8] N. Economides and J. Tag. Network neutrality and network management regulation: Quality of service, price discrimination, and exclusive contracts. *Research Handbook on Governance of the Internet. London: Edward Elgar*, 2012.

[9] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory, 3rd Ed, Wiley*. 1998.

[10] B. Hermalin and M. L. Katz. The economics of product-line restrictions with an application to the network neutrality debate. *Information Economics & Policy*, 19(2):215–248, 2007.

[11] J. Krmer and L. Wiewiorra. Network neutrality and congestion sensitive content providers: Implications for content variety, broadband investment, and regulation. *Information Systems Research*, 23, Dec. 2012.

[12] R. T. B. Ma and V. Misra. Congestion equilibrium and its role in network equilibrium. *IEEE Journal on Selected Areas of Communications*, 30(11), December 2012.

[13] R. T. B. Ma and V. Misra. The public option: a nonregulatory alternative to network neutrality. *IEEE/ACM Transactions on Networking*, 21(6), December 2013.

[14] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic theory*. Oxford University Press, 1995.

[15] J. Musacchio, G. Schwartz, and J. Walrand. Network neutrality and provider investment incentives. *Asilomar Conference*, pages 1437–1444, November 2007.

[16] P. Njoroge, A. E. Ozdaglar, N. E. Stier-Moses, and G. Y. Weintraub. Investment in two sided markets and the net neutrality debate. *Columbia Business School DRO (Decision, Risk and Operations) Working Paper No. 2010-05.*, 2010.

[17] A. Odlyzko. Paris metro pricing for the Internet. *Proceedings of ACM EC'99*, pages 140–147, 1999.

[18] V. L. Smith. An experimental study of competitive market behavior. *Journal of Political Economy*, 70, Apr. 1962.

[19] T. Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 141, 2005.

[20] C. S. Yoo. Network neutrality and the economics of congestion. *Georgetown Law Journal*, 94, Jun. 2006.

Appendix
Proofs of Selected Theorems

Proof of Theorem 1: The proof is equivalent to show that there is a unique solution to $\Phi\Big(\Lambda(\boldsymbol{\phi}, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}\Big) - \boldsymbol{\phi} = \mathbf{0}$. $\phi_{\mathcal{H}}$ is independent from $\phi_{\mathcal{L}}$ so that the $\mathbb{R}^2 \to \mathbb{R}^2$ mapping can be reduced to one $\mathbb{R} \to \mathbb{R}$ and another $\mathbb{R}^2 \to \mathbb{R}^2$ mapping:

$$\begin{cases} \Phi_{\mathcal{H}}\big(\Lambda_{\mathcal{H}}(\phi_{\mathcal{H}}, \mathcal{H}), \mu\big) - \phi_{\mathcal{H}} = 0 \\ \Phi_{\mathcal{L}}\Big(\Lambda(\boldsymbol{\phi}, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}\Big) - \phi_{\mathcal{L}} = 0 \end{cases} \tag{7}$$

We first show the uniqueness of $\phi_{\mathcal{H}}$. We have

$$\begin{gathered} \lim_{\phi_{\mathcal{H}} \to +\infty} \Phi_{\mathcal{H}}(\Lambda_{\mathcal{H}}(\phi_{\mathcal{H}}, \mathcal{H}), \mu) - \phi_{\mathcal{H}} = -\infty, \\ \lim_{\phi_{\mathcal{H}} \to \phi'_{\mathcal{H}} s.t. \lambda_{\mathcal{H}}(\phi'_{\mathcal{H}}) = \mu} \Phi_{\mathcal{H}}(\Lambda_{\mathcal{H}}(\phi_{\mathcal{H}}, \mathcal{H}), \mu) - \phi_{\mathcal{H}} = +\infty, \end{gathered} \tag{8}$$

Function $f(\phi_{\mathcal{H}}) = \Phi_{\mathcal{H}}(\Lambda_{\mathcal{H}}(\phi_{\mathcal{H}}), \mu, \mathcal{H}) - \phi_{\mathcal{H}}$ is continuous so that there exists a $\phi^*_{\mathcal{H}}$ s.t. $f(\phi^*_{\mathcal{H}}) = 0$. Since $f(\phi_{\mathcal{H}})$ is strictly decreasing with $\phi_{\mathcal{H}}$, $\phi^*_{\mathcal{H}}$ is the unique solution to $f(\phi_{\mathcal{H}}) = 0$. We only consider the congested case, i.e. $\sum_{i \in \mathcal{N}} \lambda_i^{max} > \mu$. Therefore, this achieved unique solution $\phi^*_{\mathcal{H}}$ has to be greater than zero.

The second part is to find a unique solution $\phi = (\phi_{\mathcal{H}}, \phi_{\mathcal{L}})$ to $\Phi_{\mathcal{L}}(\Lambda(\phi, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) - \phi_{\mathcal{L}} = 0$ when $\phi_{\mathcal{H}} = \phi^*_{\mathcal{H}}$. Suppose $\phi_{\mathcal{L}} = k\phi^*_{\mathcal{H}}$ and the problem turns to find a $k > 1$ s.t. $\Phi_{\mathcal{L}}(\Lambda(\phi^*, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) - k\phi^*_{\mathcal{H}} = 0$, $\phi^* = (\phi^*_H, k\phi^*_H)$. Similarly, there is $\lim_{k \to +\infty} \Phi_{\mathcal{L}}(\Lambda(\phi^*, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) - k\phi^*_{\mathcal{H}} = -\infty$. Define $g(k) = \Phi_{\mathcal{L}}(\Lambda(\phi^*, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) - k\phi^*_{\mathcal{H}}$, $\lambda_{\mathcal{H}}(\phi_{\mathcal{H}}) = \sum_{i \in \mathcal{H}} \Lambda_i(\phi_{\mathcal{H}})$ and $\lambda_{\mathcal{L}}(\phi_{\mathcal{L}}) = \sum_{i \in \mathcal{L}} \Lambda_i(\phi_{\mathcal{L}})$. We have

$$g(k) = (\frac{\mu}{\mu - \lambda_{\mathcal{H}}(\phi^*_{\mathcal{H}}) - \lambda_{\mathcal{L}}(k\phi^*_{\mathcal{H}})} - k)\phi^*_{\mathcal{H}} \tag{9}$$

Define $\delta = \mu - \lambda_{\mathcal{H}}(\phi^*_{\mathcal{H}})$. $\phi^*_{\mathcal{H}} > 0$ s.t. $\delta > 0$. δ does not change with k and it is determined by $\mathcal{H}$ alone. Besides $\lambda_{\mathcal{L}}(k\phi^*_{\mathcal{H}})$ increases as k decreases so that $\lim_{k \to +\infty} \lambda_{\mathcal{L}}(k\phi^*_{\mathcal{H}}) = 0$. Therefore, there exist a $K > 0$ which is large enough such that $\lambda_{\mathcal{L}}(k\phi^*_{\mathcal{H}}) \leq \delta$ if $k \geq K$. Here we can get $\lim_{k \to K^+} g(k) = +\infty$. Then there must exist a $K^+ \leq k^* < +\infty$ which makes $g(k^*) = 0$. Since $g(k)$ is also a strictly decreasing function with k, k^* is the unique solution to $g(k) = 0$. In particular,

$$k^* = \frac{\mu}{\mu - \lambda_{\mathcal{H}}(\phi^*_{\mathcal{H}}) - \lambda_{\mathcal{L}}(k^*\phi^*_{\mathcal{H}})} > 1. \tag{10}$$

Now we have proved there exists a unique $\phi_{\mathcal{L}} = k^*\phi^*_{\mathcal{H}}$ which satisfies $\Phi_{\mathcal{L}}(\Lambda(\phi, s_{\mathcal{N}}), \mu, s_{\mathcal{N}}) - \phi_{\mathcal{L}} = 0$. ■

Proof of Theorem 3: Suppose the sum of the traffic for $(\emptyset, \mathcal{H} \cup \mathcal{L})$ are $\lambda^t_{\mathcal{H}}$ and $\lambda^t_{\mathcal{L}}$. The sum of traffic for $(\mathcal{H}, \mathcal{L})$ are $\lambda^s_{\mathcal{H}}$ and $\lambda^s_{\mathcal{L}}$. First because $\mathcal{H}$ is independent from $\mathcal{L}$, there is $\varphi_{\mathcal{H}}(\mathcal{H} \cup \mathcal{L}, \forall) = \varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L})$. Due to the monotonicity of $\varphi_{\mathcal{H}}$, we have the following inequalities $\varphi_{\mathcal{H}}(\emptyset, \forall) < \varphi_{\mathcal{H}}(\mathcal{H}, \forall) < \varphi_{\mathcal{H}}(\mathcal{H} \cup \mathcal{L}, \forall)$. Thus we have $\varphi_{\mathcal{H}}(\mathcal{H}, \forall) < \varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L})$. In particular, it is $\varphi_{\mathcal{H}}(\mathcal{H}, \mathcal{L}) < \varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L})$ here. Since λ_H is strictly decreasing, there is $\lambda^s_{\mathcal{H}} > \lambda^t_{\mathcal{H}}$.

We have $\varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L}) = \frac{1}{\mu - \lambda^t_{\mathcal{H}} - \lambda^t_{\mathcal{L}}}$ and $\varphi_{\mathcal{L}}(\mathcal{H}, \mathcal{L}) = \frac{1}{\mu - \lambda^s_{\mathcal{H}} - \lambda^s_{\mathcal{L}}} \; \frac{\mu}{\mu - \lambda^s_{\mathcal{H}}}$. Suppose $\varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L}) > \varphi_{\mathcal{L}}(\mathcal{H}, \mathcal{L})$. $\frac{1}{\mu - \lambda^t_{\mathcal{H}} - \lambda^t_{\mathcal{L}}} > \frac{1}{\mu - \lambda^s_{\mathcal{H}} - \lambda^s_{\mathcal{L}}}$ because $\frac{\mu}{\mu - \lambda^s_{\mathcal{H}}} > 1$. Further $\lambda^t_{\mathcal{H}} + \lambda^t_{\mathcal{L}} > \lambda^s_{\mathcal{H}} + \lambda^s_{\mathcal{L}}$. We have proved that $\lambda^s_{\mathcal{H}} > \lambda^t_{\mathcal{L}}$ so that the only case is $\lambda^t_{\mathcal{J}} > \lambda^s_{\mathcal{J}}$. However because $\lambda_{\mathcal{J}}$ is also strictly decreasing and we have assumed $\varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L}) > \varphi_{\mathcal{L}}(\mathcal{H}, \mathcal{L})$, there should be $\lambda^t_{\mathcal{L}} < \lambda^s_{\mathcal{L}}$ which is a contradiction. Therefore we have proved that $\varphi_{\mathcal{L}}(\emptyset, \mathcal{H} \cup \mathcal{L}) < \varphi_{\mathcal{L}}(\mathcal{H}, \mathcal{L})$. ■