Regulating Monopolistic ISPS without Neutrality

Tang, Jing; Ma, Richard T. B.

2014

Tang, J., & Ma, R. T. B. (2014). Regulating Monopolistic ISPs without Neutrality. In Proceedings of The 22nd IEEE International Conference on Network Protocols (ICNP), 374-384.

<https://hdl.handle.net/10356/81935>

<https://doi.org/10.1109/ICNP.2014.61>

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/ICNP.2014.61>].

Downloaded on 26 Apr 2023 19:39:50 SGT

Regulating Monopolistic ISPs without Neutrality

Jing TANG

MAGIC, Interdisciplinary Graduate School
Nanyang Technological University, Singapore
Email: tang0311@ntu.edu.sg

Richard T. B. MA

School of Computing, National University of Singapore
and Advanced Digital Sciences Center, University of Illinois
Email: tbma@comp.nus.edu.sg

Abstract—Net neutrality has been heavily debated as a potential Internet regulation. Advocates have expressed concerns about the pricing power of ISPs, which might be used to discriminate Content Providers (CPs), and consequently destroy innovations at the edge of the Internet and hurt the user welfare. However, without service differentiation, ISPs do not have incentives to expand infrastructure capacities and provide quality of services, which will eventually impair the future Internet.

Although competition among ISPs would alleviate the problem and reduce the need for regulations, the problem is more severe in monopolistic markets. We study the service differentiation offered by a monopolistic ISP and find that its profit-optimal strategy makes an ordinary service “damaged good”, which hurts the welfare of CPs. Instead of imposing net neutrality regulations, we propose a flexible and lenient policy framework that generalizes net neutrality regulations. We find that a stringent regulation is needed when 1) the ISP’s capacity is abundant, 2) the profit distribution of CPs is concentrated, or 3) the utility of CPs and their users are not positively correlated. We believe that by allowing the ISPs to differentiate services under a well-designed policy constraint, the utility of the Internet ecosystem could be greatly improved.

I. INTRODUCTION

A decade after the privatization of the Internet in the mid-nineties, Internet Service Providers (ISPs) started to perform service differentiation (or discrimination) on data traffic in practice. This becomes a matter of great concern to the regulatory authorities and their goals to protect the *universal core connecting service* of the Internet and its user welfare. In 2005, Madison River Communications blocked VoIP traffic [17] and was fined \$15,000 by the U.S. Federal Communications Committee (FCC). Meanwhile, *net neutrality* [25] was proposed as a potential regulation for the Internet. Later in 2008, the FCC issued an order censuring Comcast from interfering with subscribers’ use of peer-to-peer applications. As the major opponents of net neutrality, ISPs, e.g., Verizon and Comcast, argued the necessity of maintaining a non-neutral network and filed lawsuits to challenge the rulings. As a turning point in 2010, the U.S. federal court, in a unanimous decision, found that the FCC lacked the power to enforce these *network neutral* rules and vacated the 2008 order issued by the FCC [11].

Although net neutrality is still being heavily debated among policy and law makers, the debate has mostly centered around the argument whether service differentiation should be allowed for the Internet transport services, e.g., IP transit and content delivery services. ISPs argue that due to network congestion and security, network management is needed to differentiate

traffic and maintain a more efficient network. From an economic perspective, the revenue model of the two-sided Internet market [1] is not balanced: revenues from online services of the CPs are growing more than twice as fast as those from the Internet access ISPs [20], whose market capitalizations have stagnated as investors weigh high capital requirements against continued margin pressure. ISPs feel that CPs are free-riding on their invested infrastructure. Without service differentiation, ISPs will not have incentives to invest and expand their infrastructure capacity so as to provide better quality services. This raises serious challenges regarding the viability of the current Internet model in the future. Nevertheless, CPs worry that, without net neutrality, ISPs might have too much pricing power to charge them and negatively discriminate their traffic.

In this work, we focus on monopolistic access markets, e.g., broadband or mobile markets in rural areas where ISPs have fewer incentives to deploy capacities due to low customer densities and natural monopolies emerge as a result of high deployment costs. Since monopolies have high pricing power over users, regulations are most in need in these markets. We consider a type of paid prioritization where the ISP partitions bandwidth capacity into a chargeable *premium service class* and a free *ordinary service class*. CPs are allowed to choose the service class and will be charged by a uniform price determined by the ISP for the traffic sent using the premium service. This paid prioritization reflects the current practices of access ISPs which adopt paid peering, e.g., Netflix pays to peer with Comcast directly [10], and settlement-free peering. Nevertheless, this service differentiation is non-forcible, i.e., CPs do not have to pay the ISP by using the ordinary service, and non-discriminative, i.e., traffic sent in the same service class will be charged equally and treated in a network-neutral manner. We explicitly model the congestion of each service class to capture the negative network effect (or network externality) among the traffic of CPs in a single service class, and study the CPs’ optimal choices of service which depend on both the premium service’s price and the congestion of both service classes. Via backward induction, we further characterize the ISP’s optimal capacity allocation and pricing decisions, and analyze their impacts on the welfare of CPs and users. We identify the policy regime under which the monopoly should be regulated and design a desirable policy to regulate the monopolistic ISP which is less stringent than imposing net neutrality regulations. In particular, our contributions and findings include:

- We characterize the CPs’ optimal choices of service by a novel *partition equilibrium* and prove its uniqueness property (Theorem 2).
- Without regulation, an ISP is always better off when

allocating more capacity to the premium service (Theorem 3), which makes the free ordinary service “*damaged good*”.

- We design a policy framework to regulate ISPs’ service differentiation. It generalizes net neutrality regulations, and therefore, is more lenient and efficient. Under this framework, we characterize an ISP’s optimal capacity allocation and the monotonicity properties of the resulting partition equilibrium (Theorem 4 and 5).
- We study the policy’s impact on the profit of ISP and the utility of CPs and their users. We find that although a lenient regulation increases an ISP’s profit, more stringent policies are needed when the ISP’s capacity is abundant or the CPs’ profit distribution is concentrated.
- The welfare of users depends on the correlation between the users’ value on the CPs and their profit. We find that when this correlation is positive, more lenient policy could improve the user welfare and vice versa.

Intuitively, Internet’s welfare can be increased if valuable and congestion-sensitive content are prioritized appropriately. Our framework and results provide new insights into the regulations of monopolistic access markets.

II. RELATED WORK

Tim Wu first proposed the concept of net neutrality in [25], which 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. Economides et al. [9] compared various regulations for quality of service, price discrimination, and exclusive contracts, and drew conclusions on desirable regulation regimes. Our work focuses on a practical paid prioritization structure of a monopolistic ISP. Shetty et al. [22] proposed a simple regulatory tool to restrict the percentage of capacity the ISPs dedicate to a premium service class. We propose to impose a less restrictive framework under which ISPs are allowed to freely differentiate services and set prices, as long as the difference in the induced congestion of the services is upper-bounded by some policy constraint. To maximize social welfare, Ma et al. [13] considered the ISP settlement aspect and advocated the use of Shapley value as profit-sharing mechanism to encourage ISPs. This work focuses on the interactions between the ISPs and the CPs, or in other words, whether price and service differentiation for the CPs would induce a better profit sharing and welfare for the Internet ecosystem. Ma and Misra [15] advocated the use of a *Public Option ISP* to compete with monopolistic ISPs rather than imposing the network-neutral regulations. We focus on the scenarios where the introduction of such a competing ISP is practically impossible. Our conclusion is that net neutrality regulations are still not desirable, because our less stringent policy framework generalizes net neutrality regulations and could be more efficient.

From a modeling perspective, our model of paid prioritization can be regarded as a generalization of the Paris Metro Pricing (PMP) scheme first proposed in Odlyzko [19], where

both service classes have the same capacity. The formula for M/M/1 queueing delay has been used to abstract out traffic and congestion, e.g., Choi and Kim [5], in economic analyses. Our novel congestion model, i.e., the *partition equilibrium*, is a refinement of the Nash equilibrium. In particular, the partition equilibrium possesses a structural property that guarantees its uniqueness and makes our analysis of welfare tractable. Musacchio et al. [18] also focused on a monopolistic ISP and considered advertising CPs in a two-stage model. Caron et al. [3] modeled differentiated pricing for only two application types. Shetty et al. [22] 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. [26] also used a two-class service model, but focused on transit ISPs and quantified the equivalent over provisioning cost under best-effort. We focus on a monopolistic access ISP and the welfare of different parties. Podlesny et al. [21] designed differentiated network services that enhance throughput rates and queueing delays through controlling the buffer size. We differentiate services by allocating each class different capacity rather than controlling the traffic flows.

From an economics perspective, Sidak [23] looked at net neutrality regulations from consumer welfare’s point of view and argued that differential pricing is essential to the maximization of utility. We not only focus on the consumer utility, but also the welfare of the ISP and the CPs. From an engineering perspective, Crowcroft [6] reviewed various technical aspects and concluded that “perfect” net neutrality has never been and should not be engineered. We share the same view that even under a monopolistic market, net neutrality regulation is not necessary, and a more flexible and efficient regulation framework can be used instead.

III. SYSTEM MODEL

We consider an access market consisting of a monopolistic ISP I and a nonempty set $\mathcal{N}$ of CPs. We denote μ as the ISP’s capacity and λ_i as the throughput of each CP $i \in \mathcal{N}$. Due to the statistical multiplexing nature of data networks such as the Internet, higher volume of data traffic induces higher network congestion, which negatively affect the performance of users and causes a negative network effect (or network externality) among the throughput of the CPs. We denote ϕ as a metric of network congestion and characterize the throughput of any CP i as a function $\lambda_i(\phi)$ of the congestion ϕ .

Assumption 1: $\lambda_i(\phi) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is continuous, strictly decreasing in ϕ and satisfies $\lim_{\phi \rightarrow \infty} \lambda_i(\phi) = 0$.

Assumption 1 states that the throughput of any CP diminishes as the network congestion becomes more severe. We define $\lambda_i^{max} \triangleq \lambda_i(0)$ ¹ as the maximum throughput of CP i when network congestion does not exist, i.e., $\phi = 0$. Without loss of generality, we express the throughput λ_i as $\lambda_i(\phi) = \lambda_i^{max} \omega_i(\phi)$, where each $\omega_i(\phi) : \mathbb{R}_+ \mapsto [0, 1]$ defines a discount of CP i ’s maximum throughput λ_i^{max} .

¹In detail, we can write $\lambda_i(\phi) = m_i(\phi)x_i(\phi)$, where $m_i(\phi)$ and $x_i(\phi)$ are the number of active users and the per-user throughput under congestion ϕ . Thus, λ_i^{max} can be interpreted as the product of $m_i(0)$, the maximum number of users, and $x_i(0)$, the maximum per-user throughput under no congestion.

A. Single-Class Network-Neutral Model

We start with a neutral network where a set of CPs share the ISP's capacity. We denote the system as a pair of parameters $(\mu, \mathcal{N})$ indicating its capacity and the set of sharing CPs. We denote λ as the aggregate system throughput, defined as $\lambda(\phi) \triangleq \sum_{i \in \mathcal{N}} \lambda_i(\phi)$ for any congestion ϕ . We characterize the congestion ϕ of the system by a function $\phi \triangleq \Phi(\lambda, \mu)$, which depends on the throughput λ and capacity μ as follows.

Assumption 2: $\Phi(\lambda, \mu) : \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is continuous, strictly increasing in λ and strictly decreasing in μ . It satisfies $\lim_{\lambda \rightarrow 0} \Phi(\lambda, \mu) = 0 \ \forall \mu > 0$, $\lim_{\mu \rightarrow \infty} \Phi(\lambda, \mu) = 0$ and $\lim_{\mu \rightarrow 0} \Phi(\lambda, \mu) = +\infty \ \forall \lambda > 0$.

Assumption 2 states that the system congestion ϕ decreases when the system capacity μ increases or the accommodated traffic λ reduces, and vice-versa.

We define an *equilibrium* of the system as follows.

Definition 1: A congestion level ϕ is an equilibrium of the system $(\mu, \mathcal{N})$ if $\phi = \Phi(\lambda(\phi), \mu)$.

Theorem 1: Under Assumption 1 and 2, for any non-empty set $\mathcal{N}$ of CPs and capacity μ , the system $(\mu, \mathcal{N})$ has a unique equilibrium $\phi(\mu, \mathcal{N})$, which satisfies

$$\phi(\mu', \mathcal{N}') \geq \phi(\mu, \mathcal{N}), \quad \forall \mu' \leq \mu \text{ and } \mathcal{N}' \supseteq \mathcal{N}.$$

Theorem 1 shows that the system congestion ϕ is uniquely determined and it becomes more severe if the system capacities shrink or the traffic demand from CPs becomes higher. By the uniqueness of system congestion, we denote $\phi(\mu, \mathcal{N})$ as the unique steady-state congestion of system $(\mu, \mathcal{N})$, and denote $\lambda_i(\phi)$ as the corresponding aggregate throughput of the CPs.

Remark: The convergence to an equilibrium happens naturally in practice. If the instantaneous congestion is higher than ϕ , some users will back off, which leads to lower traffic demand and throughput; if the instantaneous congestion is lower than ϕ , better performance will attract more users to demand for throughput and increase the system congestion.

B. PMP-type of Differentiated Service Model

We assume that the ISP is allowed to allocate its total capacity μ into two separate service classes: a premium class with capacity $q\mu$ and an ordinary class with capacity $(1-q)\mu$. Traffic sent via the premium class will be charged by a rate p (dollars per unit traffic) and traffic sent via the ordinary class is free of charge. This type of differentiation is similar to the Paris Metro Pricing (PMP) [19], where the capacity of both service classes are the same, i.e., $q = 0.5$, while the ordinary class is also charged lower than that of the premium class. Similar to the PMP scheme, CPs choose the service class they want to use and the utilities of CPs are affected by other CPs' choices of service classes via the negative network effect, i.e., network congestion. We focus on this type of service differentiation for three reasons. First, since CPs get to choose the service class to use, the ISP does not actively discriminate CPs and all CPs under the same service class are treated neutrally. Second, this type of scheme can be easily implemented by priority-based protocols like DiffServ [2] and separation of capacities for different service classes is seen in similar contexts such as

paid peering versus public peering among ISPs. Third, the de facto neutral network can be considered as special cases, i.e., $q = 0$ or $(p, q) = (0, 1)$, under which a free single-class service is provided to the CPs. Therefore, it provides a convenient and unified framework to compare various performance and welfare metrics under a network-neutral system with those under non-neutral systems.

We define $\mu_H(q) \triangleq q\mu$ and $\mu_L(q) \triangleq (1-q)\mu$ as the capacity of the two service classes. We denote $\mathcal{H}$ and $\mathcal{L}$ as the disjoint sets of CPs that choose to use the premium and ordinary service class, respectively. Given any ISP's capacity allocation decision q and CPs' choices $(\mathcal{H}, \mathcal{L})$, the two-class differentiated system can be regarded as two independent single-class systems. In particular, we denote $\phi_{\mathcal{H}}$ and $\phi_{\mathcal{L}}$ as the congestion in both service classes, which are the single-class congestion of $\phi_{\mathcal{H}} = \Phi_H(\mu_H, \mathcal{H})$ and $\phi_{\mathcal{L}} = \Phi_L(\mu_L, \mathcal{L})$.

We denote V_i and U_i as the utility of CP i and their users, respectively, defined as

$$\begin{cases} V_i \triangleq (v_i - p)\lambda_i(\phi_{\mathcal{H}}) & \text{and } U_i \triangleq u_i\lambda_i(\phi_{\mathcal{H}}) & \text{if } i \in \mathcal{H}; \\ V_i \triangleq v_i\lambda_i(\phi_{\mathcal{L}}) & \text{and } U_i \triangleq u_i\lambda_i(\phi_{\mathcal{L}}) & \text{if } i \in \mathcal{L}, \end{cases}$$

where v_i denotes the per-unit traffic profit for CP i and u_i denotes the per-unit traffic utility for its users. We denote V and U as the total utility of the CPs and users, and denote T as the ISP's profit, i.e., the monetary transfer from the CPs to the ISP under the premium service, defined as

$$V \triangleq \sum_{i \in \mathcal{N}} V_i, \quad U \triangleq \sum_{i \in \mathcal{N}} U_i, \quad \text{and} \quad T \triangleq p \sum_{i \in \mathcal{H}} \lambda_i(\phi_{\mathcal{H}}).$$

Notice that although the ISP also receives revenue from end-users, because we focus on monopolistic access markets, we assume that the user-side revenue is stable. Here, the profit T models the extra revenue that can be earned through the service differentiation from the CPs. We define $W \triangleq V + T$ as the system welfare. In general, the ISP wants to maximize its profit T , each CP i wants to maximize its utility V_i , while a regulatory authority might want to optimize the system welfare W and user utility U . To study the desirable regulatory regimes, we first need to understand the behaviors of the ISP and CPs. Next, we proceed to a game-theoretic analysis of the strategic decisions of the ISP and CPs, and their impact on various performance and welfare metrics.

IV. GAME-THEORETIC ANALYSIS

We define $s_I \triangleq (p, q)$ as the ISP's service differentiation strategy and $\mathcal{S} \triangleq \mathbb{R}_+ \times [0, 1]$ as its strategy space. Because the sets $\mathcal{H}$ and $\mathcal{L}$ reflect the CPs' strategic decisions, we define $s_{\mathcal{N}} \triangleq (\mathcal{H}, \mathcal{L})$ as the CPs' strategy profile. Because the CPs' preferences over different service classes depend on the ISP's strategy s_I , we model the interaction among the ISP and CPs as a dynamic two-stage game where the ISP chooses strategy s_I in the first stage and the CPs make their collective decision $s_{\mathcal{N}}$ simultaneously in the second stage. We use backward induction [16] to analyze this dynamic game and first tackle the second-stage simultaneous-move game of the CPs given any fixed strategy s_I of the ISP as follows.

A. Second-Stage Simultaneous-Move Game of CPs

We consider a fixed ISP strategy $s_I = (p, q)$. When $q = 0$ or 1, the system only has one service class effectively and the CPs' decisions can be intuitively characterized as

$$\begin{cases} \mathcal{H} = \emptyset \\ \mathcal{L} = \mathcal{N} \end{cases} \text{ if } q = 0; \begin{cases} \mathcal{H} = \{i \in \mathcal{N} : v_i > p\} \\ \mathcal{L} = \{i \in \mathcal{N} : v_i \leq p\} \end{cases} \text{ if } q = 1.$$

To break a tie when both service classes derive the same amount of utility to a CP, we assume that the CP will choose to use the ordinary service class to avoid payments to the ISP. For any non-degenerated case of $q \in (0, 1)$, a Nash equilibrium of the CPs can be defined as follows.

Definition 2 (Nash Equilibrium): $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$ is a Nash equilibrium if

$$\begin{cases} (v_i - p)\omega_i(\phi_{\mathcal{H}}) > v_i\omega_i(\phi_{\mathcal{L} \cup \{i\}}), & \forall i \in \mathcal{H}, \\ (v_i - p)\omega_i(\phi_{\mathcal{H} \cup \{i\}}) \leq v_i\omega_i(\phi_{\mathcal{L}}), & \forall i \in \mathcal{L}. \end{cases} \quad (1)$$

The Nash equilibrium prescribes an outcome of the game where no player has an incentive to deviate unilaterally. However, this delicate solution concept assumes ‘‘common knowledge’’ [16] among the players, i.e., all players know the structure of the game, know that their rivals know it, know that their rivals know that they know it, and so on. In practice, ISPs normally do not disclose private information, e.g., available resources and routing strategies, and therefore, this ‘‘common knowledge’’ assumption might not be valid. Moreover, the strategy space of the game increases exponentially with the system scale, and solving Nash equilibrium has been shown to be computationally expensive [4], [7] even under two players. This makes the Nash equilibrium computationally intractable for large-scale networks like the Internet.

To solve this technical challenge, we focus on a family of CPs whose throughput characterization ω_i satisfies the following regularity condition [14].

Assumption 3: For any CP $i \in \mathcal{N}$ and any ϕ_1 and ϕ_2 with $\omega_i(\phi_2) > 0$, $\omega_i(\phi_1)/\omega_i(\phi_2) = F_i(G(\phi_1, \phi_2))$, where $F_i(\cdot)$ is an increasing function and $G(\phi_1, \phi_2)$ is nonincreasing in ϕ_1 and nondecreasing in ϕ_2 .

Assumption 3 can be interpreted as there exist a common metric $G(\phi_{\mathcal{H}}, \phi_{\mathcal{L}})$, i.e., a congestion gap, between the two service classes $\mathcal{H}$ and $\mathcal{L}$ such that each CP i 's achievable throughput ratio $\lambda_i(\phi_{\mathcal{H}}) : \lambda_i(\phi_{\mathcal{L}}) = \omega_i(\phi_{\mathcal{H}}) : \omega_i(\phi_{\mathcal{L}})$ can be evaluated as a function F_i of that common metric $G(\phi_{\mathcal{H}}, \phi_{\mathcal{L}})$.

Two examples of $\omega_i(\cdot)$ that satisfy the above assumption are $\omega_i(\phi) = e^{-\beta_i \phi}$ and $\omega_i(\phi) = \beta_i^\phi$ (for $\beta_i \leq 1$), where β_i serves as a differentiating parameter that reflects CP i 's sensitivity to congestion. In particular, $G(\phi_1, \phi_2) = \phi_2 - \phi_1$ for both examples and $F_i(x) = e^{\beta_i x}$ and $F_i(x) = \beta_i^{-x}$ (for $\beta_i \leq 1$) for the two examples, respectively.

We denote ζ_i as a priority of CP i defined as

$$\zeta_i = \begin{cases} F_i^{-1}\left(\frac{v_i - p}{v_i}\right) & \text{if } v_i > p; \\ -\infty & \text{otherwise.} \end{cases} \quad (2)$$

Lemma 1: Under Assumption 3, $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$ is a Nash equilibrium if $\zeta_i \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}}) < \zeta_j$ for any $i \in \mathcal{L}$ and $j \in \mathcal{H}$.

Lemma 1 provides a sufficient condition for a Nash equilibrium. It states that if we can partition the CPs based on their priorities ζ_i such that the metric $G(\phi_{\mathcal{H}}, \phi_{\mathcal{L}})$ separates the values ζ_i of the two service classes, then the resulting partition $(\mathcal{H}, \mathcal{L})$ is a Nash equilibrium. Because the condition $\zeta_i \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}}) < \zeta_j$ for any $i \in \mathcal{L}$ and $j \in \mathcal{H}$ is a sufficient condition of Nash equilibrium, it might not be necessary.

As an illustration, we consider a system consisting of two CPs $\mathcal{N} = \{1, 2\}$. Both CPs have the same throughput function $\lambda_1(\phi) = \lambda_2(\phi) = e^{-\phi}$, but different per-unit traffic profits $v_1 = 1.0$ and $v_2 = 0.9$. Suppose the congestion function satisfies $\Phi(\lambda, \mu) = \lambda/\mu$ and the ISP has a capacity of $\mu = 2$ and sets its strategy to be $s_I = (p, q) = (0.1, 0.5)$, which evenly partitions capacity between the two service classes and charge 0.1 for the premium class. In this example, $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L}) = (\{1\}, \{2\})$ and $s'_{\mathcal{N}} = (\mathcal{H}', \mathcal{L}') = (\{2\}, \{1\})$ are the only Nash equilibria. However, the sufficient condition of Lemma 1 does not hold for either of the Nash equilibria. This example also shows that the uniqueness of Nash equilibrium cannot be guaranteed in general. However, since CP 1's priority ζ_1 is higher than CP 2's priority ζ_2 , $s_{\mathcal{N}} = (\{1\}, \{2\})$ could be a more plausible Nash equilibrium. Inspired by this structural property of Lemma 1, we propose a refined equilibrium concept, i.e., *partition equilibrium*, whose uniqueness can be guaranteed as follows.

Definition 3 (Partition Equilibrium): Under Assumption 3, we define $h = \arg \min_i \{\zeta_i : i \in \mathcal{H}\}$ to be the CP that has the lowest priority in the set $\mathcal{H}$ of CPs and $l = \arg \max_i \{\zeta_i : i \in \mathcal{L}\}$ to be the CP that has the highest priority in the set $\mathcal{L}$ of CPs. $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$ is a *partition equilibrium* if it satisfies the following three conditions:

- 1) $\zeta_h > \zeta_l$ if both $\mathcal{H} \neq \emptyset$ and $\mathcal{L} \neq \emptyset$,
- 2) $(v_h - p)\omega_h(\phi_{\mathcal{H}}) > v_h\omega_h(\phi_{\mathcal{L} \cup \{h\}})$ if $\mathcal{H} \neq \emptyset$,
- 3) $(v_l - p)\omega_l(\phi_{\mathcal{H} \cup \{l\}}) \leq v_l\omega_l(\phi_{\mathcal{L}})$ if $\mathcal{L} \neq \emptyset$.

The definition of the partition equilibrium describes the structure of the equilibrium as a partition of the set of CPs, where the CP h (the lowest priority CP in $\mathcal{H}$, if $\mathcal{H} \neq \emptyset$) and the CP l (the highest priority CP in $\mathcal{L}$, if $\mathcal{L} \neq \emptyset$) will not have incentives to move to the other service class as specified by the Nash equilibrium condition (1). Although the congestion level in the premium class is lower than that of the ordinary class in general, there might exist pathological cases where $\zeta_h < \zeta_l$ constitutes a ‘‘prisoner's dilemma’’ type of alternative equilibrium. By imposing the condition $\zeta_h > \zeta_l$ in the definition of a partition equilibrium, we can guarantee the uniqueness of the equilibrium as the following result.

Theorem 2 (Uniqueness of Partition Equilibrium): Under an ISP's strategy $s_I = (p, q)$, if ζ_i s are all distinct, then there exists a unique partition equilibrium $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$.

Notice that the requirement for distinct ζ_i s in Theorem 2 is just a technical assumption. In practice, if the characteristics of CPs are different, the values of ζ_i are different in general. For CPs with identical values of ζ_i , we can conceptually aggregate them as a single CP in determining the partition equilibrium. By Theorem 2, we denote the quantities under the unique partition equilibrium as a function of the ISP's strategy (p, q) , for example, $\mathcal{H}(p, q)$ and $T(p, q)$ denote the set of CPs in the premium class and the ISP's profit under

the partition equilibrium, respectively. In particular, we denote φ_L and φ_H as the congestion of the service classes under a partition equilibrium, i.e., $\varphi_L \triangleq \phi_{\mathcal{L}}(p, q)$ and $\varphi_H \triangleq \phi_{\mathcal{H}}(p, q)$.

To visualize the CPs under partition equilibria and get a better understanding of their behavior, we evaluate a scenario of 1000 CPs, whose maximum throughput λ_i^{max} are uniformly distributed within $[0, 1]$. We consider gamma distributions of CPs' profit v_i with the same mean value 0.5, e.g., $\Gamma(0.5, 1)$ and $\Gamma(5, 0.1)$. We consider the form $\omega_i(\phi) = e^{-\beta_i \phi}$ for the CPs and normalize the maximum of β_i to be 1 and make each β_i uniformly distributed within $[0, 1]$. Notice that when $\omega_i(\phi) = e^{-\beta_i \phi}$, the *congestion elasticity of throughput*, defined as $\epsilon_\phi \triangleq \frac{d\omega_i}{d\phi} \frac{\phi}{\omega_i} = -\beta_i \phi$, is linear in the congestion ϕ . It captures how sensitive the change in throughput caused by changes in congestion via the linear coefficient parameter β_i .

Fig. 1. Partition equilibrium ($\mathcal{H}, \mathcal{L}$) under different ISP's strategies $s_I = (p, q)$, capacities μ and CPs' profit distributions F_v .

Figure 1 visualizes the CPs under partition equilibria. In each subfigure, we fixed the profit distribution F_v of the CPs and the capacity μ and price p of the ISP, and we vary q to be 0.25, 0.5 and 0.75. On the x and y-axis, we vary the sensitivity β_i and profit v_i , respectively. Thus, each point in the subfigures represents a CP. The three curves from bottom to top (red, green and blue) show the partition of 1000 CPs into $\mathcal{H}$ (the upper-right partition) and $\mathcal{L}$ (the lower-left partition). By comparing the upper-left and upper-right subfigures, we observe that when the price p increases from 0.5 to 1, the partition curves move upwards, because more CPs cannot afford the premium class and move to the ordinary service class. By comparing the upper-left and lower-left subfigures, we observe that when the capacity μ increases from 10 to 100, the partition curves move rightwards, because the ordinary class is less congested and more CPs stay with it. By comparing the upper-left and lower-right subfigures, we observe that when the profit distribution F_v changes from $\Gamma(0.5, 1)$ to $\Gamma(5, 0.1)$, the partition curves move closer to each other, because the profit distribution becomes more concentrated and it is more difficult to differentiate CPs.

Fig. 2. Congestion φ_H (upper) and φ_L (lower) when ISP varies p .

Figure 2 plots the congestion levels φ_H (upper subfigures) and φ_L (lower subfigures). For each column, we fix the capacity μ and profit distribution F_v . In each subfigure, we vary the ISP's price p on the x-axis and plot three curve with $q = 0.25, 0.5$ and 0.75 , respectively. Intuitively, under fixed allocation q , when p increases, more CPs will move from the premium class to the ordinary class and therefore, the congestion level φ_H reduces while φ_L increases. However, we also observe that φ_H is not strictly decreasing with p . This could happen when a CP moves from the ordinary class to the premium class, which triggers some CP in the premium class move back to the ordinary class and the resulting congestion levels become non-monotonic. In each subfigure, we observe that a larger value of q will provide lower (higher) congestion for the $\mathcal{H}$ ($\mathcal{L}$) service class. By comparing the left and middle columns, we observe that when the ISP's capacity μ increases from 10 to 100, the congestion level of both service classes decrease. By comparing the left and right columns, we observe that when the CPs' profit distribution F_v changes from $\Gamma(0.5, 1)$ to $\Gamma(5, 0.1)$, the congestion level of the premium class drop much faster to zero, while that of the ordinary class becomes higher. This is because when the profit v_i is more concentrated, when p increases, fewer CPs can afford using the premium service, and therefore, making it less congested.

B. The First-Stage Strategy of the ISP

In the previous subsection, we studied the strategies of CPs and the resulting partition equilibrium under any fixed ISP's strategy $s_I = (p, q)$. In this subsection, we study the optimal strategy (p, q) that maximizes the profit $T(p, q)$ of the ISP. As we observed in Figure 2 that a higher value of q provides lower congestion to the premium service class, resulting higher throughput for CPs in $\mathcal{H}$. The following theorem shows that it also increases the profit T of the ISP.

Theorem 3: For any fixed capacity μ and price p , we have $T(p, q') \geq T(p, q)$ for all $q' > q$.

Theorem 3 states that under any fixed price p , allocating more capacity to $\mathcal{H}$ does not reduce the ISP's profit. It implies that the ISP always has incentive to allocate all its capacity μ to the premium class, i.e., to set $q = 1$, so as to maximize its

profit T . However, when $q = 1$, $\mathcal{H} = \{i \in \mathcal{N} : v_i > p\}$, and therefore, CPs with profit lower than p cannot obtain service.

Fig. 3. ISP's profit T under different ISP's strategies $s_I = (p, q)$, capacities μ and CPs' profit distributions F_v .

Figure 3 plots the ISP's profit T on the y-axis when q varies along the x-axis. In each subfigure, we fix the capacity μ and profit distribution F_v , and plot three curves with $p = 0.5, 1$ and 1.5 . We observe that the ISP's profit increases with q monotonically as indicated by Theorem 3. By comparing the left and middle subfigures, we observe that when the capacity μ increases, the ISP's profit T increases. By comparing the left and right subfigures, we observe that when the CPs' profit distribution F_v becomes more concentrated, a lower price, i.e., $p = 0.5$, achieves higher profit for the ISP. This is because when the profit v_i is more concentrated, the higher price will exclude too many CPs from the premium service class, reducing the total throughput and the ISP's profit.

Fig. 4. ISP's profit T , CPs' utility V and system welfare W under $q = 1$.

Figure 4 plots the ISP's profit T , the CPs' utility V and the system welfare W , as p varies along the x-axis and q is set to 1. We vary μ and F_v across the three subfigures. We observe that both T and W show a single-peak pattern, while V decreases with ISP's price p . By comparing the left and middle subfigures, we observe that when μ increases, all T , V and W increase; however, when the ISP maximizes its profit, the corresponding system welfare W deviates from its maximum more, even lower than that of a neutral setting of $p = 0$, under a large capacity μ . By comparing the left and right subfigures, we observe that when the CPs' profit distribution F_v becomes more concentrated, a higher ISP price reduces the welfare metrics and the optimal ISP price also decreases, which confirms our observation in Figure 3.

V. REGULATIONS ON SERVICE DIFFERENTIATION

We have seen that the ISP's optimal strategy allocates all its capacity to the premium service class (Theorem 3) and charges a price which might reduce the system welfare below what can be achieved under a neutral network (Figure 4). Consequently,

the ordinary class is made to be a "damaged good" [8] such that CPs are forced to choose the premium class, because the ordinary service class effectively does not exist.

An easy solution is to impose a stringent regulation such that the ISP is not allowed to perform service differentiation, i.e., restricting $q = 0$ or equivalently $(p, q) = (0, 1)$. A more flexible solution could allow ISPs to charge CPs such that the resulting partition equilibrium is effectively neutral as follows.

Definition 4 (Effectively Neutral Strategy): An ISP's strategy $s_I = (p, q)$ is *effectively neutral* if $s_I \in \mathcal{S}^0$ where

$$\mathcal{S}^0 = \{(p, q) \in \mathcal{S} : q = 0 \text{ or } (q = 1 \text{ and } \mathcal{H}(p, q) = \mathcal{N})\}.$$

Definition 4 still does not allow the ISP to partition its capacity and actually restricts the price p such that all CPs would use the single-class service. However, as observed in Figure 4, even under $q = 1$, a non-neutral price $p > 0$ might induce a higher system welfare than that under a neutral system, i.e., $(p, q) = (0, 1)$. The reduction in system welfare comes from the selfish ISP behavior that "destroys" the ordinary service class. Instead of imposing effective neutrality as above, we consider less restrictive regulations as follows.

Definition 5 (Regulated ISP Strategies): An ISP's strategy $s_I = (p, q)$ satisfies a policy constraint g if $s_I \in \mathcal{S}_g$ where the constraint set $\mathcal{S}_g$ is defined as

$$\mathcal{S}_g = \{(p, q) \in \mathcal{S} : (p, q) \in \mathcal{S}^0 \text{ or } G(\varphi_H, \varphi_L) \leq g\}.$$

Definition 5 imposes a policy constraint on the ISP's service differentiation strategy (p, q) such that the congestion gap $G(\varphi_H, \varphi_L)$ between the two service classes are upper-bounded by g ; otherwise, the ISP must maintain a neutral network with strategy $s_I \in \mathcal{S}^0$. By setting a small value of g , the policy is equivalent to a net neutrality regulation; however, when setting a large value for g , the policy gives ISPs more freedom to differentiate the congestion levels of the two service classes.

A. ISP Behavior under Policy Constraints

Influenced by a policy constraint g , given any fixed price p , the allocation decision q of the ISP has the following property.

Theorem 4 (Optimal Allocation): For any fixed price p and policy constraint g , if $T(p, q) > 0$ for some $(p, q) \in \mathcal{S}_g$, then there exists a unique $q^*(p) \in \mathcal{S}_g$ that maximizes the ISP's profit under the policy constraint g , satisfying $(1 - q^*(p))(G(\varphi_H, \varphi_L) - g) = 0$.

Theorem 4 intuitively states that if $q = 1$ does not violate the policy constraint g , the optimal allocation satisfies $q^*(p) = 1$; otherwise, $q^*(p)$ is chosen such that the policy constraint is tight, i.e., $G(\varphi_H, \varphi_L) = g$. When p is too big, no CP can afford the premium service, resulting $\mathcal{H} = \emptyset$. Under such a situation, there might exist multiple q that satisfy $(p, q) \in \mathcal{S}_g$ and maximize T at 0. Consequently, we need $T(p, q) > 0$ for some $(p, q) \in \mathcal{S}_g$ to guarantee uniqueness of $q^*(p)$.

Figure 5 plots the optimal allocation $q^*(p)$ and the corresponding ISP's profit T and congestion levels φ_H and φ_L in

Fig. 5. Optimal allocation $q^*(p)$ and the resulting T , φ_H and φ_L under $F_v = \Gamma(5, 0.1)$ and $\mu = 10$.

the four subfigures as the price p varies along the x-axis. In each subfigure, $\mu = 10$ and $F_v = \Gamma(5, 0.1)$, and we plot four curves with policy constraints $g = 2, 4, 8$ and 16 , respectively. When p is small, we observe that $q^*(p)$ equals 1, because although the policy constraint g is not met, $\mathcal{H} = \mathcal{N}$ and therefore, $(p, q^*(p))$ is effectively neutral. When p starts to increase, we observe that $q^*(p)$ decreases accordingly so as to meet the policy constraint g , which can be seen from the difference in φ_H and φ_L in the lower subfigures. When p becomes very large, we also observe that although the policy constraint is again not satisfied, $q^*(p)$ drops down to zero such that the strategy $(p, q^*(p))$ becomes neutral again. In the upper-right subfigure, we observe that the ISP's profit follows a single-peak pattern and more lenient policies provide higher profits for the ISP intuitively. The upper-left subfigure shows that when the optimal allocation $q^*(p)$ is less than 1, it is nonincreasing in p so as to keep the congestion gap $G(\varphi_H, \varphi_L)$ at a constant g indicated by Theorem 4. This monotonicity can be shown formally as follows.

Theorem 5 (Partition Monotonicity under Regulations):

Under a fixed policy constraint g , for any price p , we denote $(p, q^*(p)) \in \mathcal{S}_g$ as the ISP's optimal strategy and $(\mathcal{H}^*(p), \mathcal{L}^*(p))$ as the corresponding optimal partition equilibrium. For any $p' > p$, if $\zeta_i(p) \geq \zeta_j(p)$ implies $\zeta_i(p') \geq \zeta_j(p')$ for all $i, j \in \mathcal{N}$, then we must have

$$q^*(p') \leq q^*(p), \quad \mathcal{H}^*(p') \subseteq \mathcal{H}^*(p) \quad \text{and} \quad \mathcal{L}^*(p') \supseteq \mathcal{L}^*(p).$$

Theorem 5 states that when the ISP charges a higher price p , because it is constrained by the policy g , its optimal allocation $q^*(p)$ needs to adjust downwards. Intuitively, when price p increases and capacity $\mu_H = q\mu$ decreases, the premium service class would become less attractive and therefore, the set $\mathcal{H}$ of CPs will decrease in general. Technically, when price p changes, the priority ζ_i of each CP i also changes, which might change the preferences of service classes for some CPs locally. However, as the number of CPs is often large in practice, the monotonicity properties hold in general.

Fig. 6. Partition equilibrium $(\mathcal{H}^*, \mathcal{L}^*)$ under various policy constraints g and profit distribution F_v , when the ISP uses optimal strategy $s_I^*(p) = (p, q^*(p))$.

Similar to Figure 1, Figure 6 visualizes the 1000 CPs under various partition equilibria when the policy constraint g and profit distribution F_v vary in the three subfigures. In each subfigure, we vary the ISP's price p to be 0.5, 1 and 1.5 and choose the allocation strategy to be $q^*(p)$. The three curves from bottom to top (red, green and blue) show the partition of 1000 CPs into $\mathcal{H}^*$ (the upper-right partition) and $\mathcal{L}^*$ (the lower-left partition). This observation confirms with the result of Theorem 5 where the set $\mathcal{L}$ grows monotonically with p . By comparing the left and middle subfigures, we observe that when the policy constraint becomes more lenient, the partition curves move left and the premium class $\mathcal{H}$ accepts more CPs under the partition equilibrium. By comparing the left and right subfigures, we observe that when the CPs' profit distribution F_v changes, the partition curves do not move. Because under the same policy constraint g , the congestion gap $G(\varphi_H, \varphi_L)$ between the service classes remains the same; meanwhile, the priority ζ_i of CP i remains the same when the price p is fixed, therefore, as a consequence of Lemma 1, the partition that divides the CPs should remain the same as well. This observation shows that by controlling the policy constraint g , we could effectively influence how the CPs are partitioned, regardless of the profit distribution of the CPs. Notice that the above partition of CPs holds for any capacity μ that result in a non-neutral allocation, i.e., $q^* \in (0, 1)$, such that the congestion gap $G(\varphi_H, \varphi_L) = g$ is maintained.

B. Welfare Analysis and Desirable Regulations

By specifying a policy constraint, we can influence the ISP's service differentiation decisions. From a regulatory perspective, we would like to understand how these regulations affect the utilities of different parties, so as to derive the desirable regulations for the monopolistic ISP markets.

Figure 7 plots the ISP's profit T , the CPs' utility V and the system welfare W , as p varies along the x-axis and q is set to be the optimal allocation $q^*(p)$ under a given policy constraint. F_v is set to be $\Gamma(0.5, 1)$ and $\Gamma(5, 0.1)$ in the upper and lower subfigures, respectively. In each row, we vary the policy g and capacity μ across the three subfigures. We observe that the T curves show a single-peak pattern: 1) when p is small, higher price could increase ISP's profit as CPs still want to use the premium service, and 2) when p is large, high prices restrain CPs to use the premium class and under the policy constraint, the ISP has to allocate less capacity for $\mathcal{H}$ and eventually, maintain a single-class ordinary service class and receive zero profit. By comparing the upper and

Fig. 7. ISP's profit T , CPs' utility V and system welfare W under the ISP's optimal strategy $s_I^*(p) = (p, q^*(p))$.

lower subfigures, we observe that when the profit distribution F_v is more concentrated, the profit T drops to zero much earlier. In particular, in the lower subfigures, we observe that when T reaches zero, V and W jump back to the level of a neutral system. The left column shows that under a stringent policy, i.e., $g = 2$, T , V and W are close to those under neutrality systems. By comparing the left and middle columns, we observe that when the policy becomes more lenient from $g = 2$ to $g = 8$, it could induce higher ISP's profit and system welfare than those under a neutral system, although at a price of reducing the CPs' utility. By comparing the middle and right columns, we observe that when the system capacity μ expands from 10 to 100, although the welfare metrics increase in general, the lenient policy, i.e., $g = 8$, induces an aggressive strategy of the ISP under which the system welfare is lower than that of a neutral system.

Theorem 6 (Profit Monotonicity under Regulations):

The ISP's optimal profit $T^*(g)$ under policy g , defined as $T^*(g) \triangleq \sup\{T(p, q) : (p, q) \in \mathcal{S}_g\}$, is nondecreasing in g .

Theorem 6 intuitively states that when the policy is more lenient, the ISP could obtain higher profits. However, as we have seen in Figure 7 that a lenient policy might induce an aggressive ISP decision which hurt the system welfare. Next, we study how the policy affects the ISP's profit, CPs' utility and system welfare, when the ISP maximizes its profit T . Given a fixed policy constraint g , we denote $p^*(g)$ as an optimal price such that $(p^*, q^*(p^*))$ maximizes T .

Figure 8 plots the ISP's profit T , the CPs' utility V and the system welfare W , as g varies along the x-axis and the ISP performs its optimal strategy $s_I^* = (p^*(g), q^*(p^*(g)))$. We observe that T increases with g when policy becomes more lenient as shown in Theorem 6; however, when g is too large, the system welfare W shows a decreasing trend. By comparing the left and middle subfigures, we observe that although the

Fig. 8. ISP's profit T , CPs' utility V and system welfare W under the ISP's optimal strategy $(p^*, q^*(p^*))$.

increase in capacity μ improves the welfare metrics in general, higher ISP profit might correspond to lower system welfare. By comparing the middle and right subfigures, we observe that when the profit distribution of CPs is more concentrated, a more lenient policy g might induce a system welfare lower than that under a neutral system, i.e., under $g = 0$.

From a regulatory perspective, a policy maker might want to impose an optimal policy under which the induced system welfare W is maximized. We denote g^* as such an optimal policy. Because net neutrality regulation can be enforced by imposing $g = 0$, an optimal policy g^* will induce no less system welfare than that under a neutral system. Figure 8 implies that an optimal policy g^* might depend on both the system capacity μ and the profit distribution F_v .

Fig. 9. Optimal policy g^* under different ISP's capacity μ and CPs' profit distribution F_v .

Figure 9 plot the optimal policy g when the system capacity μ varies along the x-axis and the profit distribution F_v varies in the four curves. In each curve, we observe that the optimal policy g^* decreases as μ increases, which implies that regulations should be more stringent when the system capacity becomes more abundant. When fixing the capacity μ , we observe that when the profit distribution F_v becomes more concentrated, the optimal policy g^* decreases, which implies that regulations should be again more stringent. This is because when profit distribution is concentrated, it is difficult and inefficient to differentiate the CPs, and therefore, a more stringent policy should be imposed to restrict ISP's service differentiation.

C. User Utility and Its Correlation with CP Utility

By now, we have focused on the utility V of the CPs rather than the utility U of the users. In practice, we could use V as a proxy to estimate U , because the user utility u_i can often be estimated by v_i , the profit of CP i for providing content and services to their users. Because U is defined by $U = \sum_{i \in \mathcal{H}} u_i \lambda_i$ and u_i does not affect the throughput λ_i or any state of the system, the user utility U is influenced by the throughput λ_i of the CPs via their decisions of service classes. As CPs with higher profit v_i have higher priorities ζ_i , the user utility U depends on the correlation between the CP's profit v_i and user's utility u_i . To formally define the correlation between v_i and u_i , we use Spearman's ρ correlation coefficient [24], which is a nonparametric rank metric of correlation [12]. Suppose we sort the values of v_i s and u_i s and we denote r_v^i and r_u^i as the rank of v_i and u_i , respectively. We denote $\bar{r}_u$ and $\bar{r}_v$ to be the average rank of the u_i s and v_i s. Spearman's ρ correlation coefficient is defined by

$$\rho = \frac{\sum_{i \in \mathcal{N}} (r_u^i - \bar{r}_u)(r_v^i - \bar{r}_v)}{\sqrt{\sum_{i \in \mathcal{N}} (r_u^i - \bar{r}_u)^2 \sum_{i \in \mathcal{N}} (r_v^i - \bar{r}_v)^2}} \in [-1, 1].$$

To understand the desirable policy for maximizing the user utility and its relationship with the correlation between u_i and v_i , we evaluate the induced user utility U under different policy g , when the correlation coefficient ρ varies. We still conduct experiments on the same set of 1000 CPs as previous evaluations. We assume that the ISP performs an optimal service differentiation strategy s_I^* that maximizes its profit under any policy g , and the values of v_i and u_i are correlated although having the same marginal distributions.

Fig. 10. Comparison of the user utility U under different policy g and capacity μ , when the ISP uses optimal strategy s_I^* .

Figure 10 plots the user utility U under different policy g and correlation coefficient ρ , when the ISP uses an optimal strategy s_I^* . In each subfigure, we fix the capacity μ and marginal distribution $F_v = F_u$, and vary the policy g to be 0, 4 and 16. The result under $g = 0$ can be seen as the user utility under a neutral network. For each policy g , we generate 500 experiments with different values of ρ as shown in the x-axis. We observe that under a neutral network where $g = 0$, the user utility U does not depend on the correlation coefficient ρ between the distributions F_v and F_u . However, under the ISP's optimal strategy s_I^* , user utility will be increased if the distribution F_u is positively correlated with the profit distribution F_v of the CPs, and vice-versa. Intuitively, this is because CPs with higher profit v_i have higher priority ζ_i , they have higher chance to use the premium service class and receive higher throughput λ_i . As

v_i s and u_i s are more positively correlated, the user utility $U = \sum_{i \in \mathcal{N}} u_i \lambda_i$ will increase as expected. In practice, the CPs are more profitable when their content/services are more valuable to their users, and therefore, allowing the ISP to perform certain level of service differentiation could provide higher user welfare U than that under a neutral network as shown in the left subfigure. However, by comparing the left and right subfigures, we observe that when the capacity μ increases, although the overall user utility increases, U is mostly lower under a non-neutral policy $g > 0$ than that under $g = 0$ in the right subfigure. This result tells that a more stringent policy is needed to protect the user utility when the system capacity is more abundant.

VI. CONCLUSIONS

In this paper, we consider the problem of regulating monopolistic access ISPs without imposing net neutrality regulations. In particular, we consider PMP-type of service differentiation provided by the ISPs and derive the CPs' decisions under a unique partition equilibrium. Based on this solution concept, we find that the ISP's profit-optimal strategy always allocate all capacity to the premium service class and damage the free ordinary service class. To resolve this issue, we propose a framework to restrict the ISP's service differentiation strategy such that the congestion gap between the two service classes should be upper-bounded by policy; otherwise, a neutral network should implemented. Because our framework generalizes the net neutrality regulation, it is more flexible in practice and is more efficient than net neutrality when an appropriate policy constraint is chosen. To optimize the utility of the Internet ecosystem and its users, we find that a more stringent regulation is needed when 1) the ISP's capacity is abundant, 2) the profit distribution of CPs is concentrated, or 3) the utility of CPs and their users are not positively correlated. In conclusion, we believe that our flexible policy framework could be used to more effectively regulate monopolistic ISPs than enforcing net neutrality regulations.

ACKNOWLEDGEMENTS

The authors would like to thank anonymous reviewers for helpful comments. This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR) and Ministry of Education of Singapore AcRF grant R-252-000-526-112.

REFERENCES

- [1] M. Armstrong. Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691, 2006.
- [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] S. Caron, G. Kesidis, and E. Altman. Application neutrality and a paradox of side payments. In *Proceedings of the Re-Architecting the Internet Workshop*, ReARCH '10, pages 1–6, New York, NY, USA, 2010. ACM.
- [4] X. Chen and X. Deng. Settling the complexity of two-player nash equilibrium. In *Foundations of Computer Science. FOCS '06. 47th Annual IEEE Symposium on*, pages 261–272, 2006.
- [5] J. P. Choi and B. C. Kim. Net neutrality and investment incentives. *The RAND Journal of Economics*, 41(3):446–471, 2010.

- [6] J. Crowcroft. Net neutrality: the technical side of the debate: a white paper. *SIGCOMM Comput. Commun. Rev.*, 37(1):49–56, Jan. 2007.
- [7] C. Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, Jun 2013.
- [8] R. J. Deneckere and R. P. McAfee. Damaged goods. *Journal of Economics and Management Strategy*, 5(2):149–174, 1996.
- [9] N. Economides and J. Tag. Network neutrality and network management regulation: Quality of service, price discrimination, and exclusive contracts. *NET Institute Working Paper*, 2011.
- [10] S. Higginbotham. Netflix and comcast have reached a peering agreement. <http://gigaom.com/2014/02/21/comcast-netflix-peering/>, 2014.
- [11] C. Kang. Court rules for comcast over fcc in net neutrality case. *Washington Post*, April 7, 2010.
- [12] M. G. Kendall and J. D. Gibbons. *Rank correlation methods*. Edward Arnold. Oxford University Press, 1990.
- [13] R. T. B. Ma, D. M. Chiu, J. C. S. Lui, V. Misra, and D. Rubenstein. Internet economics: the use of shapley value for isp settlement. *Networking, IEEE/ACM Transactions on*, 18(3):775–787, June 2010.
- [14] R. T. B. Ma and V. Misra. Congestion and its role in network equilibrium. *Selected Areas in Communications, IEEE Journal on (JSAC)*, 30(11):2180–2189, 2012.
- [15] R. T. B. Ma and V. Misra. The public option: a non-regulatory alternative to network neutrality. *Networking, IEEE/ACM Transactions on*, 21(6):1866–1879, Dec 2013.
- [16] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic theory*. Oxford University Press, 1995.
- [17] D. McCullagh. Telco agrees to stop blocking voip calls. *CNET News*, March 3, 2005.
- [18] J. Musacchio, J. Walrand, and G. Schwartz. Network neutrality and provider investment incentives. In *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*, pages 1437–1444, 2007.
- [19] A. Odlyzko. Paris metro pricing for the internet. In *Proceedings of the 1st ACM conference on Electronic commerce*, EC '99, pages 140–147, New York, NY, USA, 1999. ACM.
- [20] M. Page, L. Rossi, and C. Rand. A viable future model for the Internet. *A. T. Kearney Paper*, December 2010.
- [21] M. Podlesny and S. Gorinsky. Leveraging the rate-delay trade-off for service differentiation in multi-provider networks. *Selected Areas in Communications, IEEE Journal on (JSAC)*, 29(5):997–1008, May 2011.
- [22] N. Shetty, G. Schwartz, and J. Walrand. Internet QoS and regulations. *Networking, IEEE/ACM Transactions on*, 18(6):1725–1737, Dec. 2010.
- [23] J. G. Sidak. A consumer-welfare approach to network neutrality regulation of the internet. *Journal of Competition Law and Economics*, 2(3):349–474, 2006.
- [24] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.
- [25] T. Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 2:141, 2003.
- [26] M. Yuksel, K. Ramakrishnan, S. Kalyanaraman, J. Houle, and R. Sadhvani. Quantifying overprovisioning vs. class-of-service: Informing the net neutrality debate. In *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on*, pages 1–8, 2010.

APPENDIX

Proof of Theorem 1: By Assumption 1, $\lambda_i(\phi)$ is continuous and strictly decreasing in ϕ . Thus, $\lambda(\phi) = \sum_{i \in \mathcal{N}} \lambda_i(\phi)$ is continuous and strictly decreasing in ϕ under system $(\mu, \mathcal{N})$. Further, by Assumption 2, when $\mu \rightarrow \infty$ or $\mu \rightarrow 0$, there exists a unique congestion $\Phi(\lambda(\phi), \mu)$. When $0 < \mu < \infty$, $\Phi(\lambda(\phi), \mu)$ is continuous and strictly decreasing in ϕ . Let $f(\phi) = \Phi(\lambda(\phi), \mu) - \phi$, which is a continuous and strictly decreasing function in ϕ . $f(\cdot)$ has a maximum value of $f(0) = \Phi(\lambda(\phi), \mu) > 0$ and $\lim_{\phi \rightarrow \infty} f(\phi) = -\infty$. Therefore, there exists a unique equilibrium $\phi(\mu, \mathcal{N})$ that satisfies $f(\phi) = 0$.

To simplify the notation, we define $\phi' = \phi(\mu', \mathcal{N}')$ and $\phi = \phi(\mu, \mathcal{N})$. Either $\lambda(\phi') \geq \lambda(\phi)$ or $\lambda(\phi') < \lambda(\phi)$. When $\lambda(\phi') \geq \lambda(\phi)$, by Assumption 2, we have $\phi' = \Phi(\lambda(\phi'), \mu') \geq \Phi(\lambda(\phi), \mu) = \phi$ since $\mu' \leq \mu$. When $\lambda(\phi') < \lambda(\phi)$, we know $\lambda(\phi') = \sum_{i \in \mathcal{N}'} \lambda_i(\phi') < \lambda(\phi) = \sum_{i \in \mathcal{N}} \lambda_i(\phi) < \sum_{i \in \mathcal{N}'} \lambda_i(\phi)$ for any $\mathcal{N}' \supseteq \mathcal{N}$, which also implies $\phi' > \phi$. ■

Proof of Lemma 1: For any CP with $v_i \leq p$, we know $i \in \mathcal{L}$ under Nash equilibrium. So, we check $-\infty = \zeta_i \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}})$. For any CP with $v_i > p$, if $i \in \mathcal{L}$ that $\zeta_i \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}})$, we have $(v_i - p)/v_i = F_i(\zeta_i) \leq F_i(G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}})) = \omega_i(\phi_{\mathcal{L}})/\omega_i(\phi_{\mathcal{H}}) < \omega_i(\phi_{\mathcal{L}})/\omega_i(\phi_{\mathcal{H} \cup \{i\}})$, which satisfies the definition of Nash equilibrium. Similarly, for any CP with $v_j > p$, if $j \in \mathcal{H}$ that $\zeta_j > G(\phi_{\mathcal{L}}, \phi_{\mathcal{H}})$, we have $(v_j - p)/v_j > \omega_j(\phi_{\mathcal{L} \cup \{j\}})/\omega_j(\phi_{\mathcal{H}})$, which also satisfies the definition of Nash equilibrium. ■

Proof of Theorem 2: We first prove that the system has at most one partition equilibrium under strategy $s_I = (p, q)$. Suppose there are two partition equilibrium $(\mathcal{H}, \mathcal{L}) \neq (\mathcal{H}', \mathcal{L}')$ under $s_I = (p, q)$. Without loss of generality, assume there is a CP $k \in \mathcal{L}$ such that $k \in \mathcal{H}'$. Thus, we have $\zeta_k \leq \zeta_l$ and $\zeta_k \geq \zeta_{h'}$, rewriting as $\zeta_{h'} \leq \zeta_k \leq \zeta_l$. Further, for all $i' \in \mathcal{L}'$, we have $\zeta_{i'} \leq \zeta_{l'} < \zeta_{h'} \leq \zeta_l$, which indicates all CPs in $\mathcal{L}'$ are in $\mathcal{L}$. Thereby, $\mathcal{L}' \cup \{h'\} \subseteq \mathcal{L}$, and similarly $\mathcal{H}' \supseteq \mathcal{H} \cup \{l\}$. By Theorem 1, we have $\phi_{\mathcal{L}' \cup \{h'\}} \leq \phi_{\mathcal{L}}$ and $\phi_{\mathcal{H}'} \geq \phi_{\mathcal{H} \cup \{l\}}$. Then $G(\phi_{\mathcal{L}}, \phi_{\mathcal{H} \cup \{l\}}) \leq G(\phi_{\mathcal{L}' \cup \{h'\}}, \phi_{\mathcal{H}'})$ by the definition of function $G(\cdot)$. When $v_l > p$, by Definition 3, we have $F_l(\zeta_l) = (v_l - p)/v_l \leq \omega_l(\phi_{\mathcal{L}})/\omega_l(\phi_{\mathcal{H} \cup \{l\}}) = F(G(\phi_{\mathcal{L}}, \phi_{\mathcal{H} \cup \{l\}}))$. Hence $\zeta_l \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H} \cup \{l\}})$, which is also applied to $v_l \leq p$ since $\zeta_l = -\infty$. And similarly, $\zeta_{h'} > G(\phi_{\mathcal{L}' \cup \{h'\}}, \phi_{\mathcal{H}'})$. Thus, $\zeta_l \leq G(\phi_{\mathcal{L}}, \phi_{\mathcal{H} \cup \{l\}}) \leq G(\phi_{\mathcal{L}' \cup \{h'\}}, \phi_{\mathcal{H}'}) < \zeta_{h'}$, which contradicts $\zeta_{h'} \leq \zeta_l$.

Then we show there exactly exists one partition equilibrium $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$. We denote $N = |\mathcal{N}|$ as the total number of CPs and sort CPs based on ζ_i in an ascending order, which indicates $\zeta_i < \zeta_j$ for all $1 \leq i < j \leq N$. We denote $\mathcal{H}(m)$ as the set of CPs in higher class that $i \in \mathcal{H}(m)$ for all $i \geq m$ and $\mathcal{L}(m)$ as the set of CPs in lower class that $i \in \mathcal{L}(m)$ for all $i \leq m$, where $m = 1, 2, \dots, N$. Thus $\phi_{\mathcal{H}(m)}$ is decreasing in m and $\phi_{\mathcal{L}(m)}$ is increasing in m , which indicates $G(\phi_{\mathcal{L}(m)}, \phi_{\mathcal{H}(m)})$ is decreasing in m . Let $f(m) = \zeta_m - G(\phi_{\mathcal{L}(m)}, \phi_{\mathcal{H}(m)})$, which is increasing in m . If $f(N) \leq 0$, let $m^* = N + 1$. Otherwise, find the smallest value of m^* that $f(m^*) > 0$, let $i \in \mathcal{H}$ and $j \in \mathcal{L}$ for all $i \geq m^*$ and $j < m^*$. Further, let $h = m^*$ and $l = m^* - 1$, we have $(v_h - p)/v_h = F_h(\zeta_h) > F_h(G(\phi_{\mathcal{L}(h)}, \phi_{\mathcal{H}(h)})) = F_h(G(\phi_{\mathcal{L} \cup \{h\}}, \phi_{\mathcal{H}})) = \omega_h(\phi_{\mathcal{L} \cup \{h\}})/\omega_h(\phi_{\mathcal{H}})$ if $h \leq N$. Similarly, we have $(v_l - p)/v_l \leq \omega_l(\phi_{\mathcal{L}})/\omega_l(\phi_{\mathcal{H} \cup \{l\}})$ if $l \geq 1$. Therefore, $s_{\mathcal{N}} = (\mathcal{H}, \mathcal{L})$ is a partition equilibrium. ■

Proof of Theorem 3: First, we prove $\mathcal{H}(p, q') \supseteq \mathcal{H}(p, q)$ and $\mathcal{L}(p, q') \subseteq \mathcal{L}(p, q)$ for all $q' > q$. For any fixed price $p > 0$, we just need to prove $\zeta_{h'} \leq \zeta_h, \forall q' > q$, where $\zeta_h = \min_i \{\zeta_i : i \in \mathcal{H}(p, q)\}$ and $\zeta_{h'} = \min_i \{\zeta_i : i \in \mathcal{H}(p, q')\}$. Suppose $\zeta_{h'} > \zeta_{l'} \geq \zeta_h$, to simplify the notation, we define $\mathcal{H} = \mathcal{H}(p, q)$, $\mathcal{L} = \mathcal{L}(p, q)$, $\mathcal{H}' = \mathcal{H}(p, q')$ and $\mathcal{L}' = \mathcal{L}(p, q')$, then $\mathcal{H}' \cup \{l'\} \subseteq \mathcal{H}$ and $\mathcal{L}' \supseteq \mathcal{L} \cup \{h\}$. By the definition of partition equilibrium, $\zeta_{l'} < G(\phi_{\mathcal{L}'}, \phi_{\mathcal{H}' \cup \{l'\}})$

and $\zeta_h \geq G(\phi_{\mathcal{L} \cup \{h\}}, \phi_{\mathcal{H}})$. By Theorem 1, $\phi(\mathcal{L}', (1 - q')\mu) \geq \phi(\mathcal{L} \cup \{h\}, (1 - q)\mu)$ and $\phi(\mathcal{H}' \cup \{l'\}, q'\mu) \leq \phi(\mathcal{H}, q\mu)$ for all $q' > q$. By the definition of function $G(\cdot)$, we have $\zeta_{l'} \leq G(\phi_{\mathcal{L}'}, \phi_{\mathcal{H}' \cup \{l'\}}) \leq G(\phi_{\mathcal{L} \cup \{h\}}, \phi_{\mathcal{L}}) < \zeta_h$, which show a contradiction.

Let $\lambda_{\mathcal{H}}$ refer to the aggregate throughput of premium class $\mathcal{H}$. Either $\phi' = \Phi(\lambda_{\mathcal{H}'}, q'\mu) \geq \Phi(\lambda_{\mathcal{H}}, q\mu) = \phi$ or $\phi' < \phi$. When $\phi' = \Phi(\lambda_{\mathcal{H}'}, q'\mu) \geq \Phi(\lambda_{\mathcal{H}}, q\mu) = \phi$, also $q'\mu > q\mu$, then we have $\lambda_{\mathcal{H}'} \geq \lambda_{\mathcal{H}}$ by the property of function Φ . When $\phi' < \phi$, $\lambda_{\mathcal{H}'} = \sum_{i \in \mathcal{H}'} \lambda_i(\phi') > \sum_{i \in \mathcal{H}'} \lambda_i(\phi) \geq \sum_{i \in \mathcal{H}} \lambda_i(\phi)$. Thus, $T(p, q') = p\lambda_{\mathcal{H}'} \geq p\lambda_{\mathcal{H}} = T(p, q)$ for all $q' > q$, where the equality holds only if $\mathcal{H} = \mathcal{H}' = \emptyset$. ■

Proof of Theorem 4: Suppose $s_I = (p, q^*(p))$ maximizes ISP's profit under policy constraint g that $(1 - q^*(p))(G(\varphi_H, \varphi_L) - g) < 0$, where $\mathcal{H}^*$ and $\mathcal{L}^*$ are the corresponding partitions. We know $s_I = (p, 1) \in \mathcal{S}_g$ if $\mathcal{L}^* = \emptyset$, where $T(p, 1) > T(p, q^*(p))$. Thus, $\mathcal{L}^* \neq \emptyset$. Denote $G(\phi(q\mu, \mathcal{H}^*), \phi((1 - q)\mu, \mathcal{L}^*))$ as a function of q with fixed $\mathcal{H}^*$, $\mathcal{L}^*$ and μ , which is increasing in q . We can increase $q^*(p)$ to $\hat{q}$ that $G(\phi(\hat{q}\mu, \mathcal{H}^*), \phi((1 - \hat{q})\mu, \mathcal{L}^*)) = g$. Further, using the result in the proof of Theorem 3, we have $\mathcal{H}(p, \hat{q}) \supseteq \mathcal{H}(p, q^*(p))$ and $\mathcal{L}(p, \hat{q}) \subseteq \mathcal{L}(p, q^*(p))$. Thus, $G(\varphi_H(p, \hat{q}), \varphi_L(p, \hat{q})) \leq G(\phi(q\mu, \mathcal{H}^*), \phi((1 - q)\mu, \mathcal{L}^*)) = g$, which indicates $s_I = (p, \hat{q}) \in \mathcal{S}_g$. Since $\mathcal{H}^* \neq \emptyset$, by Theorem 3, $T(p, \hat{q}) > T(p, q^*(p))$, which shows a contradiction. ■

Proof of Theorem 5: When $\zeta_i(p) \geq \zeta_j(p)$ implies $\zeta_i(p') \geq \zeta_j(p')$ for all $i, j \in \mathcal{N}$, we must have either $\mathcal{H}(p', q) \subseteq \mathcal{H}(p, q)$ or $\mathcal{H}(p', q) \supset \mathcal{H}(p, q)$ by the definition of partition equilibrium. Suppose $\mathcal{H}(p', q) \supset \mathcal{H}(p, q)$, then $\varphi_H(p', q) > \varphi_H(p, q)$, which is equal to $\Phi(\lambda_{\mathcal{H}}(p', q), q\mu) > \Phi(\lambda_{\mathcal{H}}(p, q), q\mu)$. Then by the property of function $\Phi(\cdot)$, we have $\lambda_{\mathcal{H}}(p', q) > \lambda_{\mathcal{H}}(p, q)$, which indicates $\sum_{i \in \mathcal{H}(p', q)} \lambda_i(\varphi_H(p', q)) > \sum_{i \in \mathcal{H}(p, q)} \lambda_i(\varphi_H(p, q))$. Thus, we have $\varphi_H(p', q) < \varphi_H(p, q)$, which shows a contradiction. Then we obtain that $\mathcal{H}(p', q) \subseteq \mathcal{H}(p, q)$ for any $p' > p$. Thus, $\varphi_H(p', q) \leq \varphi_H(p, q)$ and $\varphi_L(p', q) \geq \varphi_L(p, q)$, which implies $G(\varphi_H(p', q), \varphi_L(p', q)) \geq G(\varphi_H(p, q), \varphi_L(p, q))$. Therefore, when $s_I^*(p') = (p', q^*(p')) \in \mathcal{S}_g$, we have $s_I = (p, q^*(p)) \in \mathcal{S}_g$ for any $p' > p$, which means $q^*(p) \geq q^*(p')$.

Further, using the result in the proof of Theorem 3, we get $\mathcal{H}(p', q^*(p')) \subseteq \mathcal{H}(p', q^*(p)) \subseteq \mathcal{H}(p, q^*(p))$ and $\mathcal{L}(p', q^*(p')) \supseteq \mathcal{L}(p', q^*(p)) \supseteq \mathcal{L}(p, q^*(p))$ for any $p' > p$. ■

Proof of Theorem 6: By Definition 5, when $s_I = (p, q) \in \mathcal{S}_g$, we have $(p, q) \in \mathcal{S}^0$ or $G(\varphi_H(p, q), \varphi_L(p, q)) \leq g$ under a policy constraint g . Then under a more lenient policy constraint g' that $g' > g$, we have $(p, q) \in \mathcal{S}^0$ or $G(\varphi_H(p, q), \varphi_L(p, q)) \leq g < g'$, which implies $s_I = (p, q) \in \mathcal{S}_{g'}$, i.e., $\mathcal{S}_g \subseteq \mathcal{S}_{g'}$ for any $g' > g$. Thus, $T^*(g) = \sup\{T(p, q) : (p, q) \in \mathcal{S}_g\} \leq \sup\{T(p, q) : (p, q) \in \mathcal{S}_{g'}\} = T^*(g')$ for any $g' > g$. ■