Net Neutrality, Network Capacity, and Innovation at the Edges*

Jay Pil Choi† Doh-Shin Jeon‡ Byung-Cheol Kim§

May 22, 2015

Abstract

We study how net neutrality regulations affect a high-bandwidth content provider(CP)'s investment incentives to enhance its quality of services (QoS) in content delivery to end users. We find that the effects crucially depend on whether the CP's entry decision is constrained by the Internet service provider (ISP)'s network capacity. With limited capacity, prioritized delivery services are complementary to the CP's investments and can facilitate entry of congestion-sensitive content; however, this creates more congestion for other existing content. By contrast, if capacity is relatively large, prioritized services reduce QoS investment as they become substitutes, but improve traffic management. These results are qualitatively robust to the extension of the ISP's endogenous choice of network capacity.

JEL codes: D4, K2, L1, L5, O3
Key words: Net neutrality, network capacity, entry, quality of services (QoS), content provider, queuing, congestion

---

*We thank Marc Bourreau, Jane Choi, Ben Hermalin, Jeroen Hinloopen, Joshua Gans, Bruno Jullien, Martin Peitz, Wilfried Sand-Zantman, Glenn Woroch, and seminar participants at 2015 AEA meeting at Boston, 2014 EEA-ESEM at Toulouse, 2014 IIOC at Northwestern Univ., 2014 NET Conference at UC Berkeley, 2014 ICT Conference Paris at Telecom ParisTech, 2013 Midwest Economic Theory Conference at Univ. of Michigan, and Georgia Institute of Technology for helpful comments. We gratefully acknowledge financial support from the NET Institute (www.NETinst.org) through the 2013 summer grant program. An earlier version of this paper was circulated as NET Institute Working Paper #13-24 under the title "Asymmetric Neutrality Regulation and Innovation at the Edges: Fixed vs. Mobile Networks." The usual disclaimer applies.

†Michigan State University, 220A Marshall-Adams Hall, East Lansing, MI 48824-1038. E-mail: choijay@msu.edu.

‡Toulouse School of Economics and CEPR, Manufacture de Tabacs, 21 allees de Brienne - 31000 Toulouse, France. E-mail: dohshin.jeon@gmail.com.

§Corresponding author. School of Economics, Georgia Institute of Technology, 221 Bobby Dodd Way, Atlanta, GA 30332-0225. E-mail: byung-cheol.kim@econ.gatech.edu.

1 Introduction

Net neutrality is the principle that all packets on the Internet must be treated equally in their delivery without discrimination and charges regardless of its content source, destination, and type. The "open Internet" order in 2010 represents the U.S. Federal Communication Commission's (FCC) initial attempt at securing net neutrality, and has served as a focal guideline for neutrality regulations.[1] However, the FCC's order has faced legal challenges by major ISPs such as Comcast and Verizon Communications. The United States Court of Appeals for the District of Columbia Circuit concurred with the ISPs and ruled that the FCC overstepped its authority.[2] In response, on February 26, 2015, the FCC adopted new rules for broadband Internet service by reclassifying high-speed Internet service as a "telecommunications service," rather than an information service.[3] This seemingly technical maneuver allows the FCC to circumvent the legal issue of its authority over Internet service and treat the service as a public utility under Title II of the Telecommunication Act. However, the battle is not yet over, as two lawsuits were filed immediately after the FCC's new rules and more legal challenges to the new regulations are expected.[4] In this process, the issue of net neutrality has emerged as the most important and controversial regulatory agenda since the inception of the Internet.[5]

The extant literature on network neutrality has mainly focused on the expansion of Internet service providers (ISPs)' network capacity as innovation at the "core." In this paper, we focus on the implications of neutrality regulation on innovation incentives at the "edges" to reflect the growing importance of CPs' investments in the modern Internet ecosystem.[6] More specifically, the ISPs' capacity expansion making bigger "pipelines" is not the only solution to resolving the congestion problem. In fact, major content providers such as Google, Netflix, and Amazon have

[1] FCC 10-201, *In the Matter of Preserving the Open Internet, Broadband Industry Practices* (the "FCC Order"), published in Fed. Reg. Vol. 76, No. 185, Sept. 23, 2011, went into effect on November 20, 2011.

[2] See *Comcast Corp. v. FCC* (600 F.3d 642) and *Verizon v. FCC* (740 F.3d 623).

[3] FCC 15-24, *In the Matter of Protecting and Promoting the Open Internet*, released on March 12, 2015.

[4] See "First Lawsuits Filed Against the F.C.C.'s 'Net Neutrality' Rules" by Rebecca R. Ruiz in *The New York Times* on March 23, 2015.

[5] When the FCC decided to open up for public debate regarding new rules for the open Internet, it received a total of approximately 3.7 million public comments, making net neutrality by far the most-commented issue in agency history. See "FCC received a total of 3.7 million comments on net neutrality" by Jacob Kastrenakes in *The Verge* on September 16, 2014.

[6] Networks constitute the "core"of the Internet while content, applications, and devices are at the "edge." See Reggiani and Valletti (2012) for more discussion on this.

developed various measures to improve the quality of service (QoS) for their content and applications, independent of the ISP's network infrastructure. For example, they have pursued alternative technological solutions such as content distribution (or delivery) networks (CDN)[7] and advanced compression technology to ensure a sufficient quality of service, without asking for preferential treatment of their own content (Xiao, 2008).[8] From an end user's perspective, the fundamental goal is to enjoy highest quality of service at a minimum fee; the channel through which this is achieved, either through ISP's capacity investment or CP's CDN investments, is of little interest to end users. Researchers have seldom studied how these new technological changes relate to regulatory decisions, yet regulators and policy-makers need to understand how the network regulations would affect the content providers' investments in alternative technology solutions to ensure their quality of services, independent of the ISPs (Maxwell and Brenner, 2012).

Reflecting technology advances at the edges of the Internet, we develop a theoretical model to analyze the effects of net neutrality regulation on innovation incentives of major content providers. To be consistent with the FCC's interpretation, we characterize neutrality regulation as not allowing for paid prioritization under which the ISPs can allocate some traffic into a prioritized lane for a premium charge. In this setting, we find that the effects of net neutrality regulation substantially depends on the relative size of the ISPs' network capacity vis-à-vis major content providers' bandwidth usage.

The intuition is as follows. With a limited network capacity, the paid prioritization can facilitate the entry of a congestion-sensitive content provider while the entry may not be made under neutral networks because the content provider may find it too costly to invest up to its desired QoS. For this case, the prioritization complements innovation at the edges. The newly available content would generate additional value to the network. However, the entry of new content does not necessarily result in a higher social welfare. This is because the new content can consume a substantial portion of the existing network capacity, which increases the congestion for other content. Such a negative externality becomes more pronounced with a limited capacity network. Indeed, the surplus from new content can be outweighed by the efficiency loss from the

[7] "CDN is to cache frequently accessed content in various geographical locations, and redirect access request of such content to the closer place. (...) [B]y moving content closer to end users, CDN can dramatically reduce delay, delay variation, and packet loss ratio for users' applications and thus their perception of network QoS (Xiao, 2008 p.117)."

[8] It is well known that the innovative video compression technologies have contributed to better content delivery for live-streaming video applications. In addition, third-party commercial CDN providers such as Akamai and Internap have rapidly expanded their businesses to provide a high QoS for content providers.

elevated congestion for other content.

In contrast, if the network capacity is large enough, prioritized delivery and QoS investment turn into substitutes. Consider a high network capacity case in which the entry of new content is no longer a focal issue. That is, suppose that the high-bandwidth content providers enter even without the prioritized service. The prioritization then presents a different trade-off. On the positive side, the prioritization results in more efficient traffic management by assigning the faster delivery service to the more delay-sensitive content, which is referred to as the "traffic management effect." The prioritization thus enhances *static* efficiency. However, the availability of the prioritized service may dampen content providers' incentives to invest in QoS because the paid prioritization can provide an alternative technological solution to achieve their desired level of QoS. We refer to this under-investment problem as the "QoS investment effect." In other words, the prioritization may yield a negative effect on social welfare by weakening *dynamic* incentives for QoS investment.[9] The social welfare depends on the relative magnitude of these two forces, and we consider it more applicable to the fixed network where the entry of content providers has not been treated as a serious concern, relative to the mobile network where the network capacity can be a constraint on the entry decision.

We extend the model to allow for the ISP's investment in network capacity prior to the entry of the major CP. This extension confirms and even strengthens the main insight obtained for a given capacity. When a major CP's entry critically depends on the ISP's network capacity, the ISP's incentive to induce entry by investing in capacity is suboptimal regardless of the neutrality regulation regime. Intuitively, this problem is much more severe under neutral networks in which the ISP's investment incentive is independent of the surplus created by the major CP than under non-neutral networks in which the ISP can internalize the surplus to some degree. Provided that the entry occurs anyway, however, the ISP invests less under non-neutrality to enhance its bargaining position to such an extent that the major CP finds its entry unprofitable without purchasing a prioritized delivery service from the ISP. By contrast, under neutral networks the ISP's investment is simply to reduce waiting time for non-major CPs. Overall, these findings suggest that neutrality regulations can be adverse to the entry of major CPs when the network capacity is limited and can be a constraint on entry, whereas non-neutrality may reduces the ISP's incentive to invest in

[9]Consistent with this insight, Xiao (2008) claims that major content providers have increased their pursuit of quality of service through technological solutions rather than prioritization after the F.C.C.'s intensive efforts to apply network neutrality regulations.

capacity when the extensive margin is not an issue. This result is consistent with Choi and Kim (2010) who show that the ISP's investment incentives may be weaker under non-neutral networks in order to sell the priority at a higher price by making the prioritized service more valuable.

As several comprehensive reviews on the literature of net neutrality are available (e.g., Lee and Wu (2009), Schuett (2010), Lee and Hwang (2011), and Krämer, Wiewiorra, and Weinhardt (2012)), we briefly provide a selective review of notable works in relation to this paper.

The main focus of the extant studies has been investment incentives for the ISPs on its "last mile" network capacity. In particular, proponents and opponents of the regulation collide head-to-head on whether the content providers' alleged free-riding would have a chilling effect on the ISPs' incentives to upgrade their "pipelines." Academic research on this issue includes Musacchio, Schwartz, and Walrand (2009), Choi and Kim (2010), Cheng, Bandyopadhyay and Guo (2011), Economides and Hermalin (2012), Krämer and Wiewiorra (2012), and Njoroge et al. (2013). A related issue is the content providers' hold-up concern that may result in no entry or less investment in content. This concern arises because investments by high-value content providers may be expropriated *ex post* by Internet service providers who can play as gatekeepers with paid prioritization services. For studies along this avenue, we refer to Bandyopadhyay, Guo, and Cheng (2009), Choi and Kim (2010), Grafenhofer (2010), Reggiani and Valletti (2012), and Bourreau, Kourandi, and Valletti (2012). We depart from the existing literature on investment incentives by exploring new, but highly important, innovation channels adopted by major content providers such as Google, Amazon, Microsoft, and Netflix.[10]

Beyond investment incentives, economists have studied how network neutrality would affect consumer and social welfare from various perspectives, mostly in the framework of two-sided markets. The pricing dynamics of the Internet business have been studied by Pehnelt (2008), Lee and Wu (2009), Economides and Tåg (2012), and Economides and Hermalin (2012). Several studies acknowledge the importance of market competition in evaluating the effects of net neutrality. A partial list includes Becker, Carlton, and Sider (2010) and Bourreau *et al.* (2012) who focus on the competition between ISPs on the consumer side, whereas Mialon and Banerjee (2013) are more concerned with the market structure of the content side. Hermalin and Katz (2007) analyze the

[10]From an end user's perspective the goal is to enjoy the highest quality of service at a minimum fee. In this regard, it does little matter how this goal is achieved, either through the ISP's capacity investments or CPs' CDN investments. Indeed, there seems to be a consensus on the basic premise that end-users' quality of service must be the primary goal of a desirable network ecosystem (Xiao (2008), Altman et al. (2012), and Guo, Cheng, and Bandyopadhyay (2013)).

effects of network neutrality on consumer and social surplus by viewing neutrality regulations as a product line restriction in vertically differentiated product space. Taking a similar approach, Choi, Jeon, and Kim (2015) develop a model of second-degree price discrimination in a two-sided market to study how the business models of content providers affect social welfare with and without the regulation.

As common in the literature, we also adopt the two-sided market approach with a monopoly platform; however, we focus on positive externality from CPs' investments in alleviating network congestion and negative externality from new CPs' entry to an already congestive network, and investigate how the interplay of such externalities affects consumer and social welfare. In this regard, Peitz and Schuett (2015) is closely related to our paper. They consider so-called *congestion control techniques* that decrease packet losses during delivery to users with an "inflation of traffic" by sending multiple redundant packets. This practice may be privately optimal but aggravates the congestion problem on the network. They introduce the tragedy of common property resources into the net neutrality discussion and show that net neutrality regulation may lead to socially inefficient inflation of traffic whereas the socially optimal allocation can be achieved with tiered pricing. In contrast, our paper investigates the effects of net neutrality regulation on CPs' investment incentives in CDN or compression technologies, which *decreases* the packet size of individual content and generates a *positive* spillover to the network. Considering the importance of innovations at the edges of the Internet, it seems very important to offer formal analyses to understand CPs' innovation incentives and the resulting externalities associated with such innovations. We also find Economides and Hermalin (2015) related to this paper. Offering a new explanation for why ISPs offer plans with download caps, they show that congestion externality can induce ISPs to use download limits as a mechanism to restrict the aggregate bandwidth usage.[11] Both models deal with quality decisions by CPs and address how congestion externalities affect social welfare, but through different channels.

The remainder of our paper is organized as follows. We present our model in Section 2 including a generalized queuing model which describes how prioritization and CPs' investment affects congestion. In Section 3, we first show that the first-best outcome is characterized by discrimination across content types with different sensitivities to delay. This implies that net

[11]Recently, direct payments from consumers to content providers have received more attention by researchers. Gans (2015) and Economides and Hermalin (2015) consider a setting in which consumers need to pay content-specific prices to content providers, whereas Choi *et al.*(2015) consider micropayments in a reduced form represented as CPs' business models.

neutrality regulation can be justified only as a second-best policy when a social planner cannot directly control content providers' entry and investment decisions. Then, we analyze the QoS investment decisions by the major content providers under neutral and non-neutral network regimes. In Section 6, we analyze the ISP's endogenous choice of network capacity and its impact on our findings from the preceding analysis. Section 7 presents some extensions of our model. We consider heterogeneous consumers, a discontinuous QoS in congestion, and multiple major CPs. We wrap up with concluding remarks in Section 8.

2 The Model

2.1 ISP, CPs, and Consumers

We consider a monopolistic broadband Internet service provider (ISP) who is in charge of last mile delivery of online content to end-users.[12] Since we are primarily interested in major content providers' independent investment incentives to improve quality of service, we consider two types of content providers: one major content provider (henceforth, simply referred to as 'MCP') such as Google, Netflix, Disney, and Amazon Instant Video, and a continuum of other non-major content providers (simply, 'NCPs') whose mass is normalized to one. This distinction allows us to focus on MCP's investment decision to improve QoS for a successful content business; the MCP's relatively large scale of operation justifies the costly investment.

There is a continuum of homogeneous consumers whose mass is normalized to one. Let v and V denote the consumer's intrinsic utility from consuming MCP's content and NCPs' respectively. Each consumer experiences a disutility from delays of content delivery due to network congestion. We adopt an additive utility specification in which the net surplus decreases in the average waiting time for both types of content, respectively denoted by w and W. Then, a consumer earns the gross utility:

$$\begin{cases} u(w) = v - kw & \text{for MCP} \\ \\ U(W) = V - W & \text{for NCPs} \end{cases} \tag{1}$$

where $k \geq 1$ measures the relative sensitivity of MCP's content to delays compared to NCPs' of which sensitivity is normalized to one. Normalizing the mass of consumers to one, $u(w)$ and $U(W)$

[12] In reality, the Internet is a network of networks with multiple network service providers. It is not uncommon that an originating ISP may not be the same as a terminating ISP for complete delivery of content, with several interconnected network providers being involved along a transit route. Choi et al. (2015) addresses an equivalence in the choice of network qualities between interconnected ISPs and a monopoly ISP.

also represent the entire surplus from each type of content.

We assume MCP can extract the entire surplus $u(w)$ by charging consumers for delivered content in the absence of a priority service under net neutrality, but it negotiates with the ISP over the price of the priority service in a non-neutral network.[13] For NCPs' content, we introduce a parameter $\beta \in [0,1]$ to denote the ISP's share of the total surplus generated by NCPs' content delivery. In other words, the ISP receives $\beta U(W)$ from providing delivery services for NCPs' content; the rest of the surplus, $(1-\beta)U(W)$, is shared among NCPs and end users. The parameter β can be seen as the ISP's ability to extract rent from NCPs and end users via connection fees. Alternatively, one may regard β as a measure of the extent to which the ISP internalizes any externality inflicted on NCPs and end users by its decisions. If $\beta = 0$, the ISP will not take into account any potential effects on NCPs' content traffic when the ISP deals with MCP. By contrast, if $\beta = 1$, the ISP will fully internalize the externality. As will be clearer later, the parameter β plays an important role in assessing the welfare effects of net neutrality regulations. The private and the social planner's incentives coincide when $\beta = 1$ because the ISP fully internalizes any externality created in its dealing with MCP. However, for any $\beta < 1$, there may be a discrepancy between the ISP's optimal decision and the social planner's, with the potential for discrepancy more pronounced with a lower β.

2.2 Network Congestion, CP's Investment and QoS Improvement

Users initiate the Internet traffic through their "clicks" on desired content and become final consumers of the delivered content. As a micro-foundation to model network congestion, we adopt the standard M/M/1 queuing system which is considered a good approximation to congestion in real computer networks.[14]

Let μ denote the ISP's network capacity. Each consumer demands a wide range of content from both MCP and NCPs. The content request rate follows a Poisson process, which represents the intensity of content demand. For NCPs' content, we normalize the arrival rate of the Poisson distribution and the size of packets for each content to one. Since the mass of the NCP is one, the overall demand parameter (i.e., the total volume of traffic) for NCPs' content is also normalized to one. By contrast, we envision MCP as one discrete player operating a content network platform

[13] In Section 7.1, we relax the assumption of full rent extraction by the MCP and show that this simplification does not change our results qualitatively.

[14] Choi and Kim (2010), Cheong et al. (2011), Bourreau et al. (2012), Krämer and Wiewiorra (2012) adopt the M/M/1 queuing model to analyze network congestion.

that provides a continuum of content whose aggregate packet size is given by λ. Then, we can interpret λ as the sheer volume of MCP's content or a measure of the relative traffic volume of MCP's content vis-à-vis NCPs' aggregate traffic volume.[15] The total traffic volume for the ISP thus amounts to $1+\lambda$ and we need the condition of $\mu > 1+\lambda$ for a meaningful analysis of network congestion; otherwise, the waiting time becomes infinity.

The MCP can make an investment of $h \geq 0$ to enhance the quality of service in its content delivery. As discussed earlier, the investment can take various forms, such as compression technology to reduce packet-size or content delivery networks (CDN) that shorten the delivery distance by installing content servers at local data centers so that end-users' demands are served by the closest data center.[16] The common objective of all such investments is to speed up content delivery to enhance the user experience. We thus model them simply as an investment in a compression technology that would reduce the traffic volume of the major CP's content from λ to $a\lambda$, where $a = \frac{1}{1+h} \in (0,1]$; more investment leads to a smaller packet size for MCP's content. Therefore, its delivery speed increases even without the ISP's capacity expansion. No investment ($h = 0$) corresponds to $a = 1$. We assume the investment cost is increasing and convex in the investment level, i.e., $c'(h) > 0$ and $c''(h) > 0$, and satisfies the Inada condition of $c(0) = 0$ and $c'(0) = 0$ with a fixed cost of operation $F > 0$ upon entry.

We consider two network regimes: neutral and non-neutral networks. Consistent with the literature and regulatory obligations, we take the availability of a paid prioritized service as the defining characteristic that distinguishes the two network regimes. In the neutral regime, there is no paid prioritization: all traffic is treated equally with every packet being served according to the *best-effort* principle on a first come, first served basis. In the non-neutral regime, ISPs are allowed to provide a two-tiered service with the paid priority class packets delivered first.

In the neutral network, both MCP's and NCPs' content are delivered with the same speed. More specifically, each user in the M/M/1 queuing system faces the following total waiting time

[15] For instance, if MCP's content mass is ξ and the packet size for each content is m, then we have $\lambda = \xi \cdot m$.

[16] According to Xiao (2008), there are at large three different types of delays that account for the total delay from one end of the network to the other: (1) end-point delay, (2) propagation delay, and (3) link (or access) delay. Increasing speed of bottleneck links can be the most effective approach to address (3), whereas caching or content delivery networks (CDN) helps to reduce (2). The ISP's capacity expansion at the last mile helps to reduce (1). While the total delay is collectively affected by all these different types of delays, end-users typically cannot distinguish what type of delay affected their perceived quality of service.

for the major CP's content:

$$w_n(a,\mu) = \underbrace{\frac{1}{\mu-(1+a\lambda)}}_{\text{waiting time per packet}} \times \underbrace{a\lambda}_{\text{total packet size}}. \tag{2}$$

The total volume of traffic (packet size) amounts to $1+a\lambda$ (one for NCPs' content and $a\lambda$ for MCP's content with compression[17]), and thus the average waiting time per packet is given by $\frac{1}{\mu-(1+a\lambda)}$ for both types of content. With the packet size of $a\lambda$ for the major CP's content, the total waiting time is computed as (2). With no investment in the compression technology ($h=0$, or $a=1$), the average waiting time reduces to $\frac{1}{\mu-(1+\lambda)}$ as in the standard M/M/1 queuing system. Similarly, for the non-major CP's content, we can derive the total waiting time as

$$W_n(a,\mu) = \frac{1}{\mu-(1+a\lambda)} \times 1. \tag{3}$$

because the total packet size for NCPs' content is one.

Without neutrality obligations, the ISP may adopt a paid prioritization in which MCP can purchase the premium service at some price to send its content ahead of NCPs' packets in queue so that the waiting time for the prioritized packets is given by

$$w_d(a,\mu) = \frac{1}{\mu-a\lambda} \times a\lambda. \tag{4}$$

The faster delivery of the prioritized packets is achieved at the expense of NCPs' content. Once the priority service is introduced, the non-prioritized content is delivered at a slower speed; the waiting time for the "basic" service in the non-neutral network is given by

$$W_d(a,\mu) = \frac{\mu}{\mu-(1+a\lambda)}\frac{1}{\mu-a\lambda} \times 1. \tag{5}$$

In what follows, when there is no confusion, we often suppress the dependence of a on h with $w_r(h,\mu) = w_r(a(h),\mu)$ and $W_r(h,\mu) = W_r(a(h),\mu)$, where $r = n, d$.

[17]Strictly speaking, queuing happens at the package level but compression at data-level so that the packet size is different from the packet quantity. However, here we use them interchangeably because simple normalization can make this conversion possible. Specifically, let $\lambda = a \cdot \frac{D_{MCP}}{MTU}$ where D_{MCP} denotes the average data intensity of MCP and MTU stands for the maximum transmission unit. If we normalize the MTU and the average data intensity of NCPs both to one, our notations make the two concepts interchangeable.

2.3 Generalized Queuing System and Its Properties

Using (3)-(6), we can derive the following set of properties that are not only intuitive but also serve collectively as an important micro-foundation for our analysis.

Property 1 The major content provider's investment to enhance its own quality of service generates positive spillover into other content in both neutral and non-neutral networks: i.e.,

$$\frac{\partial W_n}{\partial h} < 0 \quad \text{and} \quad \frac{\partial W_d}{\partial h} < 0.$$

Intuitively, less use of bandwidth from one content provider means more network capacity for other content in a given network capacity.

Property 2 For a given pair of (a, μ), the prioritization makes the waiting time for prioritized major CP's content shorter, and the waiting time for non-major content longer than the respective ones in the neutral network: i.e.,

$$w_d(a, \mu) < w_n(a, \mu) \quad \text{and} \quad W_d(a, \mu) > W_n(a, \mu).$$

Property 3 For a given pair of (a, μ), the total waiting time is equal regardless of the network regimes: i.e.,

$$w_n(a, \mu) + W_n(a, \mu) = w_d(a, \mu) + W_d(a, \mu).$$

This result is an extended version of the waiting cost equivalence characterized in Choi and Kim (2010), Bourreau et al. (2012), Krämer and Wiewiorra (2012) in a more generalized queuing system that allows for a content provider's investment for QoS enhancement and its spillover effects. Intuitively, the total waiting time must depend on the network capacity and the total packet size to be delivered whether or not a subset of the packets is prioritized.

Property 4 For a given pair of (a, μ), prioritizing the major CP's traffic reduces the total delay cost: i.e., $kw_n(a, \mu) + W_n(a, \mu) > kw_d(a, \mu) + W_d(a, \mu)$ for any $k > 1$.

This is because the major CP's content is assumed to be more sensitive to congestion ($k > 1$) and the prioritization allocates more congestion-sensitive content to the faster lane. Formally, this property is proved by applying Properties 2 and 3:

$$[kw_n(a, \mu) + W_n(a, \mu)] - [kw_d(a, \mu) + W_d(a, \mu)] = (k - 1)[w_n(a, \mu) - w_d(a, \mu)] > 0.$$

2.4 Decision and Bargaining Timings

In the neutral network, MCP's decisions are straightforward since it does not involve a bargaining situation with the ISP.

N-1. For a given ISP's network capacity μ, the major CP makes a decision on whether to enter the market. If MCP enters, it chooses its investment level h.

N-2. For a given (μ, h), content is delivered to consumers and the payoffs are accordingly realized.

In the non-neutral network, we need an additional stage in which the major CP and the ISP bargain over the price of the prioritized service.

D-1. For a given μ, the CP and the ISP bargain over the price of the prioritized service.

D-2. With an agreement on the price of the prioritized service, MCP makes its entry and investment decisions taking the prioritized service into account. Without a mutual agreement, the prioritized service is not introduced and, as in the neutral regime, all traffic is delivered without any preferential treatment under the best effort principle. The MCP's entry and investment decisions remain the same as in the neutral regime.

D-3. Given (μ, h) and a priority class, content is delivered to consumers and the payoffs are realized.

We assume MCP's investment is *not contractible* in that the MCP and the ISP can agree only on the priority price, but the investment decision is solely left to MCP.

3 Optimal QoS Investment and Network Regimes

3.1 Benchmark: The First-best

We first characterize the first-best outcome (given a network capacity μ) in which the social planner can control MCP's entry and QoS investment decisions as well as the network regime. In our setup, the comparison of alternative network regimes is meaningful only when MCP's entry is relevant. If there is no entry, the determination of the network regime in the first-best outcome is vacuous because there is only one type of content provider. We thus focus on the case in which the social planner induces the entry of MCP. Denote the socially optimal QoS investment level in each network regime by h_r^{FB} for $r = n, d$ that is characterized as follows:

$$h_r^{FB} = \arg\min_{h_r} \Psi_r(h) = kw_r(h) + W_r(h) + c(h). \tag{6}$$

Then, we can establish the following intuitive result.

Proposition 1 (First-Best Comparison) *Suppose that the social planner induces the entry of the major CP. Then, for $k > 1$, the first-best non-neutral network is always superior in welfare to the first-best neutral network.*

Proof. *See the Appendix.* ■

Proposition 1 tells us that the first-best outcome always entails a non-neutral network when MCP's entry is socially desirable because it allows a more efficient traffic management compared to a neutral network (Property 4). This result suggests that net neutrality regulation can be justified only as a second-best policy when the entry and the investment decisions are left to the private parties. In fact, our subsequent analysis reveals that a second-best neutral network can provide a higher social welfare than a second-best non-neutral network.

3.2 Neutral Networks

Consider a neutral network in which all packets are equally treated based on the first-come-first-served principle. As usual, we proceed with backward induction and distinguish two subgames depending on whether or not MCP has entered. Assuming MCP's entry, the content provider's optimal choice of h is to maximize its profit:

$$\max_{h \geq 0} \pi_n = v - k w_n(h, \mu) - c(h) - F,$$

where $w_n(h, \mu) = \frac{\lambda}{(\mu - 1)(1+h) - \lambda}$ from (2). The first-order condition with respect to h becomes

$$\left. \frac{\partial \pi_n}{\partial h} \right|_{h_n^*} = \frac{k\lambda(\mu - 1)}{[(\mu - 1)(1 + h) - \lambda]^2} - c'(h) = 0, \tag{7}$$

for an interior solution h_n^*. The marginal benefit of the investment decreases in the ISP's network capacity, which is easily confirmed by the cross-partial derivative $\frac{\partial}{\partial \mu}\left(\frac{\partial \pi_n}{\partial h}\right) < 0$. Let $\pi_n^*(\mu) \equiv \pi_n(h_n^*(\mu), \mu)$ denote the maximized profit of MCP at the optimal investment level $h_n^*(\mu)$ for a given network capacity μ. By the Envelope Theorem, we find the MCP obtains a higher profit as the network capacity increases:

$$\frac{d\pi_n^*}{d\mu} = \frac{\partial \pi_n}{\partial \mu} = -k \frac{\partial w_n(h_n^*, \mu)}{\partial \mu} = k \frac{\lambda(1 + h_n^*)}{[(\mu - 1)(1 + h_n^*) - \lambda]^2} > 0. \tag{8}$$

This relationship implies that a threshold network capacity $\underline{\mu}_n$ exists such that $\pi_n^*(\mu) \geq 0$ if and only if $\mu \geq \underline{\mu}_n$. In other words, MCP makes an investment only when the ISP's capacity is above

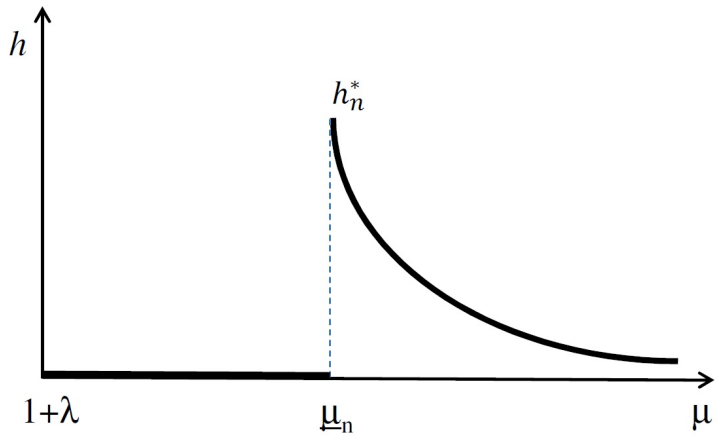

Figure 1: Optimal QoS Investment in the Neutral Network

this threshold level. For a sufficiently low capacity $\mu < \underline{\mu}_n$, the investment cost is too high to justify entry into the content service market. Hence, there is a discontinuity in MCP's investment at the threshold value $\underline{\mu}_n$: no investment for $\mu < \underline{\mu}_n$ but $h_n^* > 0$ for $\mu \geq \underline{\mu}_n$.

Furthermore, we analyze how the (interior) optimal investment h_n^* changes with the capacity level for $\mu > \underline{\mu}_n$ and establish the following lemma:

Lemma 1 *The MCP's QoS investment decreases in the ISP's network capacity μ, i.e., $\frac{\partial h_n^*}{\partial \mu} < 0$ for $\mu \geq \underline{\mu}_n$.*

Proof. *See the Appendix.* ■

We can illustrate the optimal QoS investment in the neutral network as in Figure 1: $h_n^* = 0$ for $\mu < \underline{\mu}_n$ and then $h_n^* > 0$ and $\frac{\partial h_n^*}{\partial \mu} < 0$ for $\mu \geq \underline{\mu}_n$.

3.3 Non-neutral Networks

Now consider a non-neutral network in which MCP has an option to buy a prioritized delivery service at a negotiated price. One benefit of such an arrangement is that MCP can achieve the same quality of service with a lower investment in the compression technology due to a preferential treatment of its content delivery. The analysis for the non-neutral network proceeds similarly as in the neutral network. Suppose that MCP and ISP agree on a price of the prioritized service. We define MCP's profit gross of any payout for the priority as

$$\pi_d \equiv u - kw_d(h, \mu) - c(h) - F, \tag{9}$$

where $w_d(h,\mu) = \frac{\lambda}{\mu(1+h)-\lambda}$. The first-order condition for MCP's optimal investment decision with the prioritized service (h_d^*) yields the following equation:

$$\left.\frac{\partial \pi_d}{\partial h}\right|_{h_d^*} = \frac{k\lambda\mu}{[\mu(1+h)-\lambda]^2} - c'(h) = 0. \tag{10}$$

Defining $\pi_d^*(\mu) \equiv \pi_d(h_d^*(\mu),\mu)$, we can show that the maximized profit increases in the network capacity, i.e.,

$$\frac{d\pi_d^*}{d\mu} = \frac{\partial \pi_d}{\partial \mu} = -k\frac{\partial w_d(h_d^*,\mu)}{\partial \mu} = k\frac{\lambda(1+h)}{[\mu(1+h)-\lambda]^2} > 0,$$

and the optimal investment decreases in the capacity, $\frac{\partial h_d^*}{\partial \mu} < 0$.[18]

While the investment decision h_d^* is independent of β, the price of prioritization must be affected by the level of β because the paid prioritization will make the ISP earn less from NCPs' content due to increased delay for non-prioritized content. The ISP would ask for compensation from MCP for the loss via the priority price. The ISP's incentive to provide the prioritized service would be higher as β becomes smaller. In this section, we analyze the case of $\beta = 0$, in which MCP's entry is facilitated to the maximum extent, and relegate the analysis of $\beta > 0$ to the next section.[19] In particular, if $\beta = 0$, the ISP and the major content provider will agree on some price of prioritization whenever $\pi_d^*(\mu) > 0$. As MCP's profit $\pi_d^*(\mu)$ strictly increases with μ as in the neutral network, there will be another threshold capacity $\underline{\mu}_d$ such that $\pi_d^*(\mu) \geq 0$ if and only if $\mu \geq \underline{\mu}_d$. Again, MCP's investment discretely jumps up at the threshold $\underline{\mu}_d$, then decreases with μ for $\mu > \underline{\mu}_d$. Because $\pi_d^*(\mu) > \pi_n^*(\mu)$ and $\pi_d^*(\mu)$ increases in μ, we must have $\underline{\mu}_n > \underline{\mu}_d$.

The last step needed to compare h_n^* and h_d^* is to verify that the marginal benefit of the QoS investment is greater in the neutral network compared to that in the non-neutral network. The reason is that the marginal benefit from reducing the content delivery size increases with the severity of congestion in the network, as is shown below.

$$\frac{\partial \pi_n}{\partial h} > \frac{\partial \pi_d}{\partial h} \text{ because we have } \left|w_n'(h)\right| = \frac{\lambda(\mu-1)}{[(\mu-1)(1+h)-\lambda]^2} > \frac{\lambda\mu}{[\mu(1+h)-\lambda]^2} = \left|w_d'(h)\right|.$$

Consequently, we establish the following lemma:

Lemma 2 *The major CP reduces its QoS investment with the purchase of the prioritization service, i.e., $h_n^*(\mu) > h_d^*(\mu)$, for all $\mu > \underline{\mu}_n$.*

[18] The proof is omitted as it is similar to the process leading to Lemma 1 in Section 3.2.

[19] We formally derive this result in the next section (see Lemma 5).

3.4 Network Capacity and QoS Investments

Based on Lemmas 1-2, we can summarize the major CP's optimal investment decisions for $\beta = 0$ in the following Proposition.

Proposition 2 *Suppose* $\beta = 0$.

(i) For a limited network capacity of $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$*, a paid prioritization and MCP's investment are "complements" in that prioritization induces MCP to enter and make a positive investment, whereas the major CP does not enter in the neutral network.*

(ii) For a larger capacity $\mu > \underline{\mu}_n$*, prioritization and MCP's investment are "substitutes" in that purchasing prioritization reduces the major CP's QoS investment, compared to the investment that would be made in the neutral network.*

We illustrate the optimal QoS investments in both network regimes in Figure 2. The upward arrow for the range of $\underline{\mu}_d < \mu < \underline{\mu}_n$ depicts the greater QoS investment under the non-neutral network compared to the neutral network. The downward arrow when $\mu > \underline{\mu}_n$ shows the smaller investment with the paid prioritization. Intuitively, the prioritization reduces the QoS investment incentives because it provides an alternative technological solution to achieve the desired level of QoS.

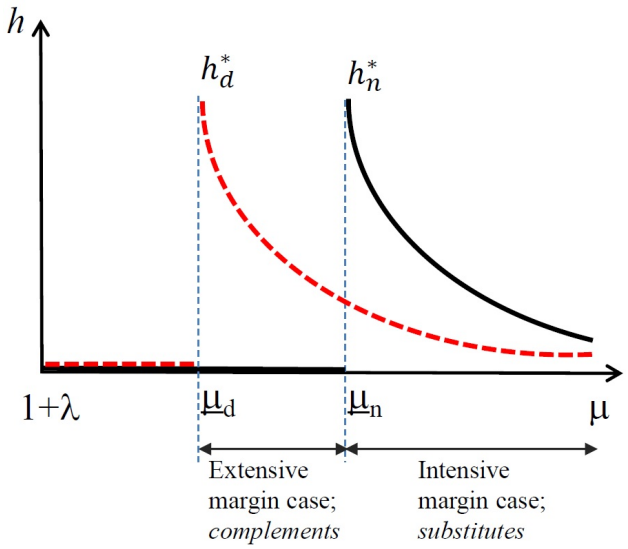

Figure 2: Optimal QoS investments, Net Neutrality, and Network Capacity

Our analysis shows that MCP's entry crucially depends on the ISP's network capacity. In the remainder of the paper, we refer the limited capacity case of $\mu \in (\underline{\mu}_d, \underline{\mu}_n)$ to as "extensive margin

case" and the high capacity case of $\mu > \underline{\mu}_n$ to as "intensive margin case" in the sense that MCP's entry is a focal issue in the former but not in the latter.

4 The Extensive Margin Case

In this section let us analyze the effects of net neutrality regulation on various participants when MCP makes no entry under the neutral regime because $\pi_n^*(\mu) < 0$ for $\mu < \underline{\mu}_n$ but the entry becomes possible in the non-neutral network (at least for $\beta = 0$).

4.1 Effects of MCP's Entry on Congestion

Under a non-neutral network, MCP's entry has two countervailing effects. On one hand, the new content generates a positive surplus $\pi_d^*(\mu) > 0$, which can be shared by the content provider and the ISP according to their respective bargaining powers. On the other hand, the entry exacerbates the network congestion through following two channels: The additional bandwidth taken by the new MCP's content means more congestion for a given network capacity. In addition, the prioritized delivery of MCP's content means a slower delivery for NCPs' content. Formally, we examine the difference in waiting time for the non-major CPs' content with the introduction of a two-tiered service, ΔW, that can be decomposed into two parts.

$$\Delta W \equiv W_d(h_d^*(\mu), \mu) - W_n(\phi, \mu) = \underbrace{[W_n(h_d^*(\mu), \mu) - W_n(\phi, \mu)]}_{(+) \text{ due to new content entry}} + \underbrace{[W_d(h_d^*(\mu), \mu) - W_n(h_d^*(\mu), \mu)]}_{(+) \text{ due to different priority classes}}, \tag{11}$$

where ϕ stands for no entry of the MCP. The first bracketed term in (11) measures the increase in delivery time even in the absence of prioritization due to increased traffic volume with the entry of the major CP — the Internet "pipe" now needs to be shared with the major CP. The second one captures the non-major content's waiting time increase due to the prioritization for a given QoS investment h_d. On both accounts, NCPs suffer from longer delivery time, i.e., $\Delta W > 0$. We confirm this intuition formally by showing that

$$\Delta W = \frac{a_d^* \lambda \left(2\mu - a_d^* \lambda - 1\right)}{\left[\mu - (1 + a_d^* \lambda)\right] \left(\mu - a_d^* \lambda\right) (\mu - 1)} > 0 \text{ for any } a_d^* \in (0, 1].$$

4.2 Effects of Prioritization on the ISP and Social Welfare

We now examine the ISP's incentives to provide the prioritized service in the non-neutral regime and the overall welfare effects of this prioritization. The prioritized service will be provided to MCP

and its price will be agreed upon between the ISP and MCP if their joint profits increase with the service. The joint profits under the neutral regime will be given by $\Pi_n(\phi, \mu, \beta) = \beta[V - W_n(\phi, \mu)]$ because there is no entry in the neutral network. With a priority service in the non-neutral network, their joint profits are given by $\Pi_d(h, \mu, \beta) = \pi_d(h, \mu) + \beta[V - W_d(h, \mu)]$. The change in joint profits due to introduction of the prioritization can be written as follows:[20]

$$\Delta\Pi^E(\mu, \beta) \equiv \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(\phi, \mu, \beta) = \pi_d^*(\mu) - \beta\Delta W(\mu), \tag{12}$$

where the superscript E in $\Delta\Pi^E$ denotes that we consider the case that the MCP's entry (extensive margin) generates the key trade-offs. As (12) clearly shows, the MCP's entry generates the value of $\pi_d^*(\mu)$ but the ISP must bear the loss of $\beta\Delta W(\mu)$ due to its negative effects on NCPs. Recalling $\Delta\Pi^m(\mu, \beta = 1)$ equals to the change in social welfare from the MCP's entry, one can immediately see that any private incentives to introduce a prioritized service under $\beta < 1$ exceeds the socially optimal incentives. Specifically, the discrepancy between the private and social incentives is measured by $(1-\beta)\Delta W(\mu)$, which is inversely related to β. If $\beta = 1$, the ISP completely internalizes the effects on consumers and NCPs, with the private and social incentives coinciding.

Recognizing the crucial role of the network capacity in measuring the effects of a prioritization on social welfare and private incentives to introduce it, we examine how $\Delta\Pi^E(\mu, \beta)$ changes with μ for a given β. First, we analyze the effects of a marginal change in network capacity on the waiting time for MCP's content in the non-neutral network (w_d) as follows:

$$\frac{dw_d(h_d^*(\mu), \mu)}{d\mu} = \frac{\partial w_d(h_d^*(\mu), \mu)}{\partial \mu} + \frac{\partial w_d(h_d^*(\mu), \mu)}{\partial h}\frac{\partial h_d^*}{\partial \mu}. \tag{13}$$

A small increase in μ has a direct positive effect on the quality of service, represented by the first term on the RHS in (13). However, an indirect negative effect counteracts: the MCP responds to a higher network capacity by reducing its QoS investment. We find that the positive direct effect dominates the negative indirect effect.

Lemma 3 *The waiting time in the non-neutral network unambiguously decreases as the network capacity increases regardless of priority classes: (i)* $\frac{dw_d(h_d^*(\mu),\, \mu)}{d\mu} < 0$ *and (ii)* $\frac{dW_d(h_d^*(\mu),\, \mu)}{d\mu} < 0$.

Proof. *See the Appendix.* ■

From (12) and Lemma 3-(i), it is clear that $\Delta\Pi^E(\mu, \beta)$ strictly increases in μ if $\beta = 0$ for

[20]Note that ΔW does not depend on β because h_d^* is independent of β.

any $k \geq 1$; by continuity, this result holds for a small β. Even if β is not that small, the private incentives to introduce a prioritization still increases with μ if the delay sensitivity parameter for the MCP's content, measured by k, is sufficiently large.

Lemma 4 *There exists $\overline{k}(\mu, \beta)$ such that for $k \geq \overline{k}(\mu, \beta)$, $\Delta\Pi^m(\mu, \beta)$ strictly increases in μ.*

Proof. *See the Appendix.* ■

The intuition for Lemma 4 is simple. As k increases, the social benefit of being able to assign a fast lane to MCP's content also increases while the negative effect of the entry on NCPs' content remains constant. Because the ISP takes only a part of the negative externality into account, if β is small enough, the ISP will find it better to offer the prioritization and the MCP will make the entry even when k is very close to one. We define $\underline{\mu}_d(\beta)$ as the cutoff capacity above which MCP enters.[21] The MCP's investment $h_d^*(\mu)$ is independent of the parameter β given its entry and ΔW is also a constant regardless of the degree of β, the joint profits of the ISP and MCP conditional on MCP's entry strictly decreases with β for a given μ. Lemma 5 summarizes this result:

Lemma 5 *Suppose $k \geq \overline{k}(\mu, \beta)$. Then, $\underline{\mu}_d(\beta)$ strictly increases with β.*

Any entry under $\mu < \underline{\mu}_d(1)$ is socially harmful; Lemma 5 tells us that such an excessive entry can occur for any $\beta < 1$. This is because the coalition of the ISP and MCP does not fully internalize the negative externality of increased congestion onto the non-major content. Using Lemmas 4-5, we obtain the following Proposition.

Proposition 3 *Consider the extensive margin case with $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$. Suppose $k \geq \overline{k}(\mu, \beta)$.*

(i) Given a β, a paid prioritization induces MCP's entry as long as $\mu \geq \underline{\mu}_d(\beta)$, where $\underline{\mu}_d(\beta)$ strictly increases with β starting from $\underline{\mu}_d(0) = \underline{\mu}_d$.

(ii) If $\underline{\mu}_d(1) < \underline{\mu}_n$, MCP enters in a non-neutral network though the entry is not socially desirable for $\mu \in (\underline{\mu}_d(\beta), \underline{\mu}_d(1))$. By contrast, the entry is socially efficient for $\mu \in (\underline{\mu}_d(1), \underline{\mu}_n)$ and the prioritization makes this socially desirable entry possible.[22]

Figure 3 illustrates Proposition 2-(ii): when $\underline{\mu}_d(1) < \underline{\mu}_n$, MCP makes a socially inefficient entry due to the paid prioritization for the shaded range of μ. Remarkably, the socially excessive

[21] With this notation, we confirm that the cutoff capacity defined for $\beta = 0$ in Section 3 is expressed as $\underline{\mu}_d(0) = \underline{\mu}_d$.

[22] If $\underline{\mu}_n < \underline{\mu}_d(1)$, there is no socially efficient entry due to the prioritization.

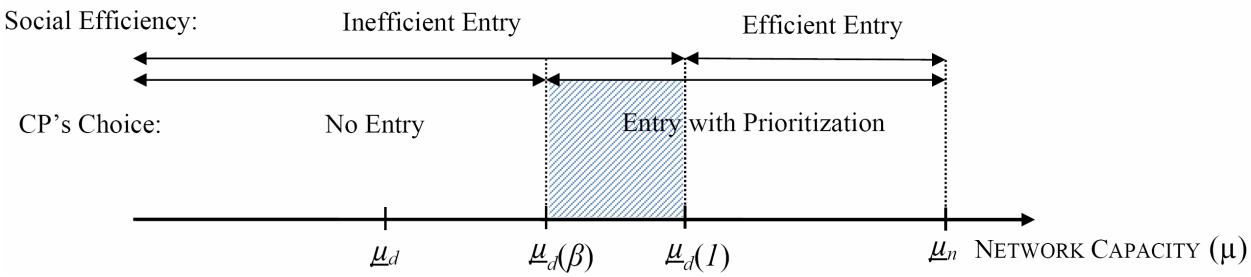

Figure 3: Social Efficiency and Private Entry Decision

entry is not an eventual outcome all the time; in Section 7.1 we show an insufficient entry can arise if one considers consumer heterogeneity and/or competition between ISPs.

5 The Intensive Margin Case

In this section we consider a network in which the network capacity is large enough to induce the major content provider's entry regardless of the network regimes, i.e., $\mu \geq \underline{\mu}_n$. In other words, now the MCP's content is available without a prioritized service. Recall that the joint payoffs of the ISP and MCP are given in the network regime $r = n, d$ as follows:

$$\Pi_r(h, \mu, \beta) = \pi_r(h, \mu) + \beta[V - W_r(h, \mu)] \tag{14}$$

where n stands for neutral networks and d for non-neutral (discriminatory) networks. Hence, the prioritization will be adopted if the two parties find the non-neutral regime better than the neutral treatment: i.e.,

$$\Delta\Pi^I(\mu, \beta) = \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta) > 0, \tag{15}$$

where the superscript I indicates that we are considering the *intensive margin* case and $\Delta\Pi^I(\mu, \beta) > 0$ means a higher joint payoff under the non-neutral network. For the intensive margin case, we find a different trade-off that we describe below from the trade-off reported for the extensive margin case.

5.1 A New Trade-off: Traffic Management vs. Under-investment

The effects of the prioritization on the joint payoff can be decomposed into two parts: (1) static traffic management effect and (2) dynamic QoS investment effect. Formally, it yields that

$$\Delta\Pi^I(\mu, \beta) = \underbrace{[\Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta)]}_{\text{Traffic Management Effect } (+)} + \underbrace{[\Pi_n(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta)]}_{\text{QoS Investment Effect } (-)}. \tag{16}$$

The first term in (16) is always positive and we refer it to as the "traffic management effect": for a given QoS investment level h, prioritizing the MCP's traffic reduces the total delay cost because MCP's content is more sensitive to congestion ($k > 1$). Precisely, we have

$$\begin{aligned} \text{Traffic Management Effect} &= \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta) \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] + \beta\left[W_n(h_d^*(\mu)) - W_d(h_d^*(\mu))\right] \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] - \beta\left[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))\right] \\ &= (k-\beta)[w_n(h_d^*) - w_d(h_d^*)] > 0 \end{aligned} \tag{17}$$

where the third equality in (17) is obtained from Property 3, $w_n(h_d^*)+W_n(h_d^*) = w_d(h_d^*)+W_d(h_d^*)$.[23] The inequality is clear because $k > 1 \geq \beta$ and Property 2, $w_n(h_d^*) > w_d(h_d^*)$.

We refer the second square bracket in (16) to as the "QoS investment effect": the availability of a prioritized service will decrease the MCP's investment level from $h_n^*(\mu)$ to $h_d^*(\mu)$ (Lemma 2), which in turn affects the resulting joint payoff. To determine the sign of this term, let $h_n^J(\mu, \beta)$ denote the MCP's investment choice that maximizes the joint profit of the two parties in the neutral regime, i.e.,

$$h_n^J(\beta) = \arg\max_h \Pi_n(h, \mu, \beta) \ = \pi_n(h, \mu) + \beta[U - W_n(h, \mu)] \tag{18}$$

which is alternatively defined as

$$h_n^J(\beta) = \arg\min_h \ k w_n(h) + c(h) + \beta W_n(h). \tag{19}$$

From the profit maximization problem (18), we can see that the MCP's private optimal choice, $h_n^*(\mu)$ that maximizes $\pi_n(h, \mu)$, fail to incorporate its positive externality on the NCPs' content. The joint decision on the QoS investment internalize such externality only part of it, $\beta W_n(h)$. Thus, the joint decision yields so-called under-investment problem. This logic can be earned from the cost minimization problem (19) in which MCP chooses $h_n^*(\mu)$ to minimize $k w_n(h)+c(h)$, but ignores the additional cost $\beta W_n(h)$. In sum, in either interpretation, we find that a under-investment problem persists in that $h_n^*(\mu) < h_n^J(\mu, \beta)$ for any $\beta > 0$. This result combined with Lemma 2, $h_d^*(\mu) <$

[23] The traffic management effect can be derived directly from Property 4 when consumer utility and ISP's profit are specified as linear functions in the waiting costs and the QoS investment is evaluated at h_d^*.

$h_n^*(\mu)$, [24] proves that the under-investment problem is more serious in the non-neutral network compared to the neutral network: formally, it means the QoS investment effect must be negative:

$$\text{QoS Investment Effect} = \Pi_n(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta) < 0. \tag{20}$$

5.2 Effects of Prioritization on Social Welfare

We now analyze a social planner's incentive to introduce the paid prioritization and compare it with the private incentive. We consider the constrained (second-best) social optimum in which the social planner can choose the network regime only, but the QoS investment is left to the MCP's decision. The social welfare in each regime coincides with the joint payoff of the ISP and MCP when $\beta = 1$ for (14), which is given by

$$S_r(\mu) = \Pi_r(h_r^*(\mu), \mu, \beta = 1) = \pi_r(h_r^*(\mu), \mu) + [U - W_r(h, \mu)], \tag{21}$$

where $r = n, d$. Let $\Delta S(\mu)$ be the effect of the prioritization service on social welfare:

$$\begin{aligned} \Delta S(\mu) &= S_d(\mu) - S_n(\mu) \\ &= \Delta\Pi^I(\mu, \beta) + (1-\beta)\left[W_n(h_n^*) - W_d(h_d^*)\right] \end{aligned} \tag{22}$$

As is clearly seen, if $\beta = 1$, the private incentive to adopt the prioritization service is perfectly aligned with the social incentive (i.e., $\Delta S(\mu) = \Delta\Pi^I(\mu, 1)$). For any uninternalized externality with $\beta < 1$, however, we have socially excessive adoption of the paid prioritization as the ISP and MCP would not fully internalize the effect of increased delay on NCPs' content which is represented by

$$\Delta S(\mu) - \Delta\Pi^f(\mu, \beta) = (1-\beta) \underbrace{\left[W_n(h_n^*) - W_d(h_d^*)\right]}_{\text{externality on NCP's content}} \tag{23}$$

We can further decompose the externality term in (23) and show $\Delta S(\mu) < \Delta\Pi^f(\mu, \beta)$ as follows:

$$W_n(h_n^*) - W_d(h_d^*) = [W_n(h_n^*) - W_n(h_d^*)] + [W_n(h_d^*) - W_d(h_d^*)] < 0 \tag{24}$$

The first term of (24) has a negative sign because of Lemma 2 ($h_n^* > h_d^*$), and the second term also takes a negative value following Property 2. Lastly, we notice that the discrepancy between

[24] Note that the objective function $[kw_n(h) + \beta W_n(h)] + c(h)$ in the minimization problem is a convex function of h because each component of $w_n(h)$, $W_n(h)$, and $c(h)$ is also convex in h. The convexity of the objective function warrants the clear comparison.

the social incentives and the private incentives is inversely related to β. Interestingly, we find that if the discrepancy reaches its maximum ($\beta = 0$), the ISP and MCP will always find it profitable to adopt the prioritization in the non-neutral network regardless of whether the neutrality regulation would give higher social welfare. To see this, we verify the following:

$$\begin{aligned} \Delta\Pi^I(\mu, \beta = 0) &= \pi_d(h_d^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &\geq \pi_d(h_n^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &= w_n(h_n^*(\mu), \mu) - w_d(h_n^*(\mu), \mu) > 0, \end{aligned} \tag{25}$$

where the first (weak) inequality comes from the revealed preference argument and the last inequality from Property 2. Proposition 4 summarizes thus far findings for the intensive margin case.

Proposition 4 *Consider a network with $\mu > \underline{\mu}_n$ in which MCP always enters. Then, we find*

(i) A prioritization service involves a trade-off between the positive efficient traffic management effect and the negative QoS investment effect.

(ii) A prioritization will be adopted only when the gain from the better traffic management exceeds the loss from the diminished QoS investment.

(iii) If β is close to zero, the ISP and MCP always find the prioritization profitable in the non-neutral networks.

(iv) In general, there exist socially excessive incentives to adopt a prioritization service, i.e., $\Delta S(\mu) \leq \Delta\Pi^I(\mu, \beta)$.

5.3 Net Neutrality as a Second-Best Policy

We close this section by offering a numerical example that illustrates net neutrality regulation as a second-best policy. According to Proposition 1, in the first-best world, non-neutrality yields the higher welfare than neutrality because of the traffic management effect (there is no QoS investment effect in the first-best). However, as the below example shows, this is no longer the case in the second-best world because the under-investment problem is more severe in a non-neutral network than in a neutral network. For explicit derivations, let us consider a cost function of $c(h) = h^2$ and set the values of parameters $\mu = 3, \lambda = 2$, $u = 5$, $U = 3$, and $F = 0$ for $k \in \{1, 2, 3\}$. Table 1 shows the contrast between the first-best and the second-best outcomes.

Table 1: First-Best vs. Second-Best

k	h_n^*	h_d^*	h_n^{FB}	h_d^{FB}	$S_d^* - S_n^*$	$S_d^{FB} - S_n^{FB}$
1	0.693	0.406	1.145	1.145	-1.213	0.000
2	0.874	0.559	1.357	1.242	-0.120	0.511
3	1.000	0.667	1.518	1.330	0.472	0.912

The comparison between optimal QoS investments shows that the under-investment problems occur in both network regimes (i.e., $h_n^{FB} > h_n^*$, $h_d^{FB} > h_d^*$), but the extent of the under-investment is larger in the non-neutral network ($h_d^{FB} - h_d^* > h_n^{FB} - h_n^*$) where MCP reduces its investment because the quality of service can be enhanced through prioritization.

When one considers the symmetric waiting cost, i.e., $k = 1$, the first-best outcomes are the same in both network regimes ($S_d^{FB} = S_n^{FB}$). For the second-best, the neutral network is better ($S_d^* < S_n^*$) because of the less severe under-investment problem in the neutral network and zero efficiency gains from traffic management by prioritization. For a modest asymmetry in the congestion costs ($k = 2$), the non-neutral network outperforms the neutral network for the first-best ($S_d^{FB} > S_n^{FB}$) because the efficiency gain via the better traffic management gives rise to a higher first-best welfare in the non-neutral network (Proposition 1). However, the opposite holds for the second-best ($S_d^* < S_n^*$): the more severe negative effect of the under-investment problem in the non-neutral network outweighs the positive traffic management effect (Proposition 4). If k is sufficiently large ($k = 3$), such conflict disappears. Now the non-neutral network starts to give higher social welfare both in the first-best and second-best sense because the traffic management effect dominates the QoS investment effect even in the second-best outcome.

The potential necessity of net neutrality regulations as a second-best policy is reminiscent of Choi *et al.*(2015). While our finding sound similar, the logic differs. In our earlier work, we show that a menu of multiple qualities may yield an excessive quality distortion for the basic service such that the resulting social welfare is even lower in the non-neutral network compared to in the neutral network. Here, this possibility comes from the substitution between the QoS investment and the prioritization available only under the non-neutral network.

6 ISP's Capacity Choice

In our baseline model, we focused on MCP's QoS investment for a *given* ISP's network capacity. Because it is very important to understand potential interplays between the ISP's capacity choice and the MCP's entry/investment decisions, here we augment our model to include a new initial

stage when the ISP decides μ prior to the MCP's actions. Precisely, at the new initial stage **N-0** or **D-0**, the ISP chooses its network capacity before all subsequent plays ensue. Let $C(\mu)$ denote the investment cost of capacity μ with $C' > 0$ and $C'' > 0$. For analytic simplicity and clarity, we adopt two assumptions for this section. First, although we know that the waiting time $w_r(h_r^*(\mu), \mu)$ and $W_r(h_r^*(\mu), \mu)$ depend on μ both directly and indirectly through $\partial h_r^*/\partial \mu$, we conduct our analysis by assuming that the indirect effect is of a second-order to the direct effect, which should hold if $c(h)$ is convex enough. Second, we assume v and k are large enough to capture a situation in which the QoS of the high-bandwidth content is crucial to consumer utility. Both assumptions are not restrictive for main results.

As a benchmark case, consider the ISP's optimal capacity choice in the absence of MCP, which is characterized by

$$-\beta \frac{\partial W_n(\phi, \mu)}{\partial \mu} = C'(\mu). \tag{26}$$

The LHS measures the marginal benefit from a small increase in the capacity, which is the reduced waiting time for the NCPs' content. the RHS captures the marginal cost. Let $\mu_n^{\phi}(\beta)$ be the solution to (26).

6.1 Neutral networks

The ISP's optimal capacity in a neutral network must depend on whether the ISP induces the MCP to enter or not. If the ISP induces the entry and MCP chooses $h_n^*(\mu)$, the optimal capacity is determined by

$$-\beta \left[\frac{\partial W_n(h_n^*(\mu), \mu)}{\partial \mu} + \frac{\partial W_n(h_n^*(\mu), \mu)}{\partial h} \frac{\partial h_n^*}{\partial \mu} \right] = C'(\mu). \tag{27}$$

Let $\mu_n^E(\beta)$ be the solution to (27). Let us consider a hypothetical situation where MCP enters but choose zero investment (i.e., $h = 0$), and denote the ISP's optimal capacity $\widetilde{\mu}_n^E(\beta)$ in this scenario. Then, $\widetilde{\mu}_n^E(\beta)$ will satisfy

$$-\beta \frac{\partial W_n(0, \mu)}{\partial \mu} = C'(\mu).$$

Intuitively, the MCP's QoS and the ISP's capacity investment become *substitutes* as the ISP would invest less when MCP can make a positive investment compared to when MCP cannot (i.e., $h = 0$). To verify this intuition, we compare the marginal benefit of MCP's capacity expansion under $h_n^*(\mu)$ with that with $h = 0$. Then, we can confirm $\mu_n^E(\beta) < \widetilde{\mu}_n^E(\beta)$ from the inequality of

$$-\beta \left[\frac{\partial W_n(h_n^*(\mu), \mu)}{\partial \mu} + \frac{\partial W_n(h_n^*(\mu), \mu)}{\partial h} \frac{\partial h_n^*}{\partial \mu} \right] < -\beta \frac{\partial W_n(0, \mu)}{\partial \mu},$$

because the direct effect $-\beta\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial \mu}$ decreases with h and the indirect effect is also negative from $\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial h} < 0$ and $\frac{\partial h_n^*}{\partial \mu} < 0$. Recall that $\underline{\mu}_n$ denotes the minimum level of the network capacity to induce the MCP's entry in a neutral network. Thus, a necessary condition for the ISP to induce MCP's entry requires $\underline{\mu}_n < \mu_n^E(\beta)$. Otherwise, the ISP only need to invest $\underline{\mu}_n$ to induce the MCP's entry, but this strategy would give a lower profit to the ISP compared to inducing no entry by choosing $\underline{\mu}_n - \varepsilon$, where $\varepsilon > 0$ is infinitesimal. This is because the entry of MCP would only increase congestion to the NCPs' content as is seen from $W_n(\phi, \underline{\mu}_n) < W_n(h_n^*(\underline{\mu}_n), \underline{\mu}_n)$. In general, no entry can occur with either $\mu = \underline{\mu}_n$ when $\underline{\mu}_n < \mu_n^\phi(\beta)$ or $\mu = \mu_n^\phi(\beta)$ when $\mu_n^\phi(\beta) < \underline{\mu}_n$. So, we use the minimum operator $\min\{\mu_n^\phi, \underline{\mu}_n\}$ for the capacity inducing no entry. Obviously, the ISP would induce entry if and only if it yields more profit than no entry, i.e.,

$$-\beta W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_n\}) - C(\min\{\mu_n^\phi, \underline{\mu}_n\}) \leq -\beta W_n(h_n^*(\mu_n^E), \mu_n^E) - C(\mu_n^E), \tag{28}$$

which is equivalently given as

$$C(\mu_n^E) - C(\min\{\mu_n^\phi, \underline{\mu}_n\}) \leq \beta[W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_n\}) - W_n(h_n^*(\mu_n^E), \mu_n^E)]. \tag{29}$$

Inequality (29) holds if the cost of capacity expansion is sufficiently cheap and β is sufficiently large. Given that the ISP chooses either $\min\{\mu_n^\phi, \underline{\mu}_n\}$ or μ_n^E, welfare is higher with MCP's entry if and only if

$$-W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_n\}) - C(\min\{\mu_n^\phi, \underline{\mu}_n\}) \leq \pi_n(h_n^*(\mu_n^E), \mu_n^E) - W_n(h_n^*(\mu_n^E), \mu_n^E) - C(\mu_n^E), \tag{30}$$

where $\pi_n(h_n^*(\mu), \mu) > 0$ always holds since $\mu_n^E(\beta) > \underline{\mu}_n$. By comparing (28) and (30) using $C(\mu_n^E) > C(\min\{\mu_n^\phi, \underline{\mu}_n\})$ from $\mu_n^E > \min\{\mu_n^\phi, \underline{\mu}_n\}$, we find MCP's entry is socially desirable whenever the ISP induces entry. However, the converse is not always true. A socially desirable entry can be blocked by the ISP since it does not take into account MCP's profit. Also, the ISP's capacity choice in a neutral network is always suboptimal when it induces MCP's entry, because the ISP ignores the effect of its investment on $w_n(h_n^*(\mu), \mu)$.

6.2 Non-neutral Networks

Now let us examine a non-neutral network in which the ISP and MCP bargain over the price of prioritization, so that without the neutrality regulation, the ISP's investment decision depends on its bargaining power against MCP and its default payoff if the bargaining fails. We assume the Nash

bargaining with equal bargaining power between the two. Because of differences in default payoffs, we should distinguish two cases depending on whether or not MCP enters without prioritization.

Consider the first case in which MCP does not enter without prioritization and the default payoffs follow no entry payoffs. Then, the ISP chooses its capacity to maximize the following objective:

$$\frac{\{\pi_d(h_d^*(\mu),\mu) - \beta\left[W_d(h_d^*(\mu),\mu) - W_n(\phi,\mu)\right]\}}{2} + \beta\left[U - W_n(\phi,\mu)\right] - C(\mu). \tag{31}$$

The first term in (31) is the half of the surplus created by prioritization, and the second term is the ISP's default payoff. The first-order condition with respect to μ is given by

$$-\frac{k}{2}\frac{dw_d(h_d^*(\mu),\mu)}{d\mu} - \frac{\beta}{2}\frac{dW_d(h_d^*(\mu),\mu)}{d\mu} - \frac{\beta}{2}\frac{dW_n(\phi,\mu)}{d\mu} = C'(\mu). \tag{32}$$

Let $\widehat{\mu}(\beta)$ denote the solution to Equation (32). For the case of no entry, the ISP's capacity must be constrained by the upper-bound $\underline{\mu}_n$, which means the optimal capacity (conditional on no entry without prioritization) is $\mu_d^E(\beta) = \min\left\{\widehat{\mu}(\beta), \underline{\mu}_n\right\}$.[25]

Consider the second case in which MCP enters even without prioritization. Then, the ISP's objective is given by

$$\frac{\{\pi_d(h_d^*(\mu),\mu) - \pi_n(h_n^*(\mu),\mu) - \beta\left[W_d(h_d^*(\mu),\mu) - W_n(h_n^*(\mu),\mu)\right]\}}{2} + \beta\left[U - W_n(h_n^*(\mu),\mu)\right] - C(\mu). \tag{33}$$

The first-order condition with respect to μ is given by

$$-\frac{k}{2}\left[\frac{dw_d(h_d^*(\mu),\mu)}{d\mu} - \frac{dw_n(h_n^*(\mu),\mu)}{d\mu}\right] - \frac{\beta}{2}\frac{dW_d(h_d^*(\mu),\mu)}{d\mu} - \frac{\beta}{2}\frac{dW_n(h_n^*(\mu),\mu)}{d\mu} = C'(\mu). \tag{34}$$

When k is large enough, the LHS of (34) is predominantly determined by the bracketed term. In addition, since we assume the indirect effect through the change in $h_r^*(\cdot)$ is of a second-order compared to the direct effect, the bracketed term is by and large determined by

$$\frac{\partial w_d(h_d^*(\mu),\mu)}{\partial\mu} - \frac{\partial w_n(h_n^*(\mu),\mu)}{\partial\mu} = -\frac{a_d^*(\mu)\lambda}{\left(\mu - a_d^*(\mu)\lambda\right)^2} + \frac{a_n^*(\mu)\lambda}{\left(\mu - 1 - a_n^*(\mu)\lambda\right)^2} > 0$$

where $a_r^*(\mu) = \frac{1}{1+h_r^*(\mu)}$. Hence, if k is large enough and $c(\cdot)$ is convex enough, the LHS of (34) takes on a negative value: the ISP has no incentive to invest. This implies that the constraint that

[25] A necessary condition for the ISP to induce MCP's entry with prioritization is $\mu_d^E(\beta) > \underline{\mu}_d(\beta)$.

MCP enters without prioritization (i.e., $\mu \geq \underline{\mu}_n$) binds. Therefore, our analysis shows that the ISP invests $\underline{\mu}_n$, conditional on inducing entry of the MCP. The ISP wants to maximize the surplus created by its prioritization service, which is mainly driven by the difference in the waiting time, $k[w_n(h_n^*(\mu),\mu) - w_d(h_d^*(\mu),\mu)]$ for a sufficiently large k. A marginal capacity investment is more effective in reducing w_n than w_d, which in turn decreases the surplus created by prioritization. This is why the ISP wants to minimize its investment; a similar effect was obtained by Choi and Kim (2010).

Given that MCP only enters with prioritization, we can show that the ISP never chooses a capacity level that allows MCP's entry without prioritization. Suppose to the contrary that the ISP chooses $\underline{\mu}_n$. Then, comparing (31) and (33) reveals that the ISP's profit is higher when entry is not allowed in the absence of prioritization because $W_n(\phi,\mu) < W_n(h_n^*(\mu),\mu)$.[26] We break ties in favor of the ISP and assume MCP does not enter without prioritization given $\underline{\mu}_n$.[27]

We now examine the ISP's incentive to induce MCP's entry with prioritization in a non-neutral network. Following similar analysis applied to the neutral network, we find that the ISP chooses $\min\{\mu_n^\phi, \underline{\mu}_d\}$ if the ISP does not induce entry, but $\mu_d^E(\beta)$ if the ISP induces entry. Thus, the ISP will induce entry if and only if

$$-\beta W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_d\}) - C(\min\{\mu_n^\phi, \underline{\mu}_d\}) \leq$$

$$\frac{\left\{\pi_d(h_d^*(\mu_d^E),\mu_d^E) - \beta\left[W_d(h_d^*(\mu_d^E),\mu_d^E) - W_n(\phi,\mu_d^E)\right]\right\}}{2} - \beta W_n(\phi,\mu_d^E) - C(\mu_d^E). \tag{35}$$

As is shown in the RHS of (35), the ISP internalizes a half of the surplus created by entry, $\pi_d(h_d^*(\mu_d^E),\mu_d^E)$ in the non-neutral network. By contrast, such a rent does not show in the RHS of (28) in the neutral network. This implies that non-neutrality incentivizes the ISP to make investments to facilitate the MCP's entry. Given that the ISP chooses either $\min\{\mu_n^\phi, \underline{\mu}_d\}$ or $\mu_d^E(\beta)$, welfare is higher with entry if and only if

$$-W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_d\}) - C(\min\{\mu_n^\phi, \underline{\mu}_d\}) \leq \pi_d(h_d^*(\mu_d^E),\mu_d^E) - W_d(h_d^*(\mu_d^E),\mu_d^E) - C(\mu_d^E), \tag{36}$$

Comparing (35) with (36), we can see that the entry bias can operate in either direction. The

[26] Here we use $\pi_n(h_n^*(\mu),\mu) = 0$ at $\mu = \underline{\mu}_n$.

[27] For example, the ISP can choose $\underline{\mu}_n - \varepsilon$ with an infinitesimal $\varepsilon > 0$ to induce no entry. This argument shows that when entry occurs with prioritization, the ISP chooses $\mu_d^E(\beta)$ and induces no entry without prioritization.

ISP can capture only a half of the surplus created by entry, which induces insufficient entry. In contrast, the negative externality on NCPs' content arises due to MCP's entry in a non-neutral network, which is not fully internalized by the ISP, implying an excessive entry. However, if we assume $\pi_d(h_d^*(\mu_d^E), \mu_d^E)$ is sufficiently large compared to the effects on NCPs' content delivery time due to MCP's entry (which trivially holds when k and v are large enough), then whenever the ISP does induce entry, it is also socially desirable. But, socially desirable entry is blocked by the ISP if (35) holds but (36) does not.

Proposition 5 summarize our findings in this section.

Proposition 5 *Consider the ISP's investment problem before the MCP's entry. Assume that v and k are large enough and that $c(\cdot)$ is convex enough.*

(i) (Incentives to induce the entry) Regardless of the neutrality regulation, the ISP's incentives to induce the entry of MCP is lower than the social incentives. This problem is more severe in a neutral network than in a non-neutral network in which the ISP partially internalizes the surplus created by the entry.

(ii) (Investment incentives when MCP enters in equilibrium)

- *In the neutral network, the ISP invests $\mu_n^E(\beta)$ where $\mu_n^E(\beta) > \underline{\mu}_n$. The ISP invests to reduce the waiting time for NCPs. The ISP's investment is lower by expecting the MCP's subsequent QoS investment compared to the MCP's commitment to zero QoS investment.*
- *In the non-neutral network, the ISP invests $\mu_d^E(\beta)$ where $\mu_d^E(\beta) \leq \underline{\mu}_n$ and induces MCP not to enter without prioritization. The ISP's investment is optimally limited because the larger capacity would decrease the ISP's' extraction of the surplus created by inducing the entry via prioritization.*

The result in Proposition 5-(ii) is reminiscent of Choi and Kim (2010): a discriminatory network may not warrant a higher investment of the ISP since a lower congestion would mean a declining value of a prioritized service. In a neutral network, however, there is another kind of concern: the incentives to induce entry is lower than in a non-neutral network. Overall, Proposition 5 suggests that when a network capacity is limited, the neutrality regulation can be adverse to the entry of major content providers, whereas when the entry is not an issue, non-neutrality may reduce the ISP's incentive to invest in capacity. Therefore, we find the extended model with the ISP's capacity investment provides consistent implications with those in our baseline model.

7 Extensions

7.1 Consumer Heterogeneity

In our baseline model, we considered homogeneous consumers and full surplus extraction by MCP from its content delivery.[28] In such a setting, consumers always suffer from entry of MCP's high-bandwidth content due to the negative externality on the existing NCPs' content for any $\beta < 1$. This simplification is innocuous to the results that we have derived, for we have focused on social welfare. However, for a more realistic analysis of consumer welfare and its implications for net neutrality regulations, we need to extend our model such that some consumers enjoy a certain positive surplus from MCP's content delivery. For this spirit, let us introduce consumer valuation heterogeneity in the simplest way: two types of consumers, H with proportion $\gamma \in (0,1)$ and L with $1-\gamma$.[29] The utility level that a type $i \in \{H, L\}$ consumer derives from MCP's content in the neutral network will equal to $u_i(w_n) = v_i - kw_n$ and in the non-neutral network to $u_i(w_d) = v_i - kw_d$. Let $\delta = v_H - v_L > 0$. Furthermore, assume that MCP prefers to serve both types than to serve only H-type.[30] In such a case, L-type consumers only suffer from MCP's entry, whereas now H-type consumers earns a surplus of δ from the new content though they still suffer from the elevated congestion for the existing NCPs' content. Then, the social welfare comparison in the non-neutral network will be determined by the trade-off between π_d^{**} and $\Delta W = [W_d(h_d^*(\mu);\mu) - W_n(\phi;\mu)]$, where $\pi_d^{**} = \pi_d^* - \gamma\delta$ and π_d^* is MCP's profit from full surplus extraction.

Our previous analysis in Section 4 remains qualitatively intact by replacing π_d^* with π_d^{**}. Remarkably, with consumer heterogeneity, we find that the MCP's entry can be socially insufficient. This result provides a nuanced contrast to Proposition 3-(ii) that we focused on the excessive MCP's entry. Specifically, a prioritization will be introduced only if

$$\Delta\Pi^E(\mu,\beta) \equiv \Pi_d(h_d^*(\mu),\mu,\beta) - \Pi_n(\phi,\mu,\beta) = \pi_d^*(\mu) - \gamma\delta - \beta\Delta W(\mu) > 0, \tag{37}$$

but the entry will be socially inefficient if $\Delta\Pi^E(\mu,\beta)|_{\mu=\underline{\mu}_d(1)} < 0$. If the condition $\gamma\delta + \beta\Delta W(\mu)|_{\mu=\underline{\mu}_d(1)} > \Delta W(\mu)|_{\mu=\underline{\mu}_d(1)}$ (i.e., $\Delta W(\underline{\mu}_d(1)) < \frac{\gamma\delta}{1-\beta}$) holds, then a socially optimal entry does not take place.

[28] We did not specify how the rent was extracted. One can think of micro-payments such as pay-per-view, membership fees, and/or various types of online advertising.

[29] One caveat is that one may introduce heterogeneity of consumers by assuming a uniform distribution over u_i and derive a linear demand. But, it would unnecessarily complicate the analysis without further insight to be gained.

[30] This would be the case if v_L is sufficiently large compared to δ and/or γ is relatively small.

This implies that the concern for socially excessive entry is mitigated and we may have insufficient entry when MCP cannot extract the entire consumer surplus.[31]

7.2 Discrete QoS in Congestion

We have considered a consumer's utility is a continuous function in the congestion level. However, it is not necessarily the case for some real-world applications. For example, a consumer who is watching a movie through a video streaming platform such as Netflix may stop subscribing to the service when he or she finds the content delivery unsatisfactory due to frequent buffering or a blurry screen. A user would not value a Voice over Internet Protocol (VoIP) when calls drop too often or the call quality is below a certain level, whereas the same user may feel indifferent once the QoS is above a certain level. Accordingly, we could consider the utility function as the following step function,

$$u(w) = \begin{cases} u & \text{for} \quad w \leq w_o \\ 0 & \quad\quad\;\; w > w_o \end{cases}$$

while the non-major content is assumed to have no discontinuity in the QoS. One advantage of working with a discrete QoS function is to be able to derive explicit solutions for QoS investments. In the neutral network with a sufficiently large capacity μ, there will be no need for any investment from MCP to warrant its minimum quality requirement. The upper-bound capacity, denoted by $\overline{\mu}_n$, can be derived from $w_n(h=0) = \frac{\lambda}{\mu-(1+\lambda)} = w_0$ as follows:

$$\underline{\mu}_n = 1 + \lambda + \frac{\lambda}{w_o}.$$

The MCP's optimal investment to ensure the required QoS, denoted by h_n, is derived from $w_n(h_n) = \frac{\lambda}{(\mu-1)(1+h_n)-\lambda} = w_o$:

$$h_n^*(\mu) = \frac{1+w_o}{w_o}\frac{\lambda}{\mu-1} - 1 \quad \text{for} \;\; \mu < \overline{\mu}_n. \tag{38}$$

In the non-neutral network, MCP can have an option to buy the prioritized delivery service at a certain price. The benefit of such an arrangement is that the investment level that ensures the required common QoS for the content can be lowered compared to in the neutral network. Solving $w_d(h=0) = \frac{\lambda}{\mu(1+h)-\lambda} = w_o$, we can derive the threshold capacity above which no investment is

[31] In the same spirit with consumer heterogeneity, suppose that a competition between MCPs plays a role of reducing MCPs' payoffs. Then, even without consumer heterogeneity, an insufficient entry of major content providers is possible. We leave explicit modeling MCPs' competition for further research.

required to ensure the required QoS in the non-neutral network:

$$\overline{\mu}_d = \lambda + \frac{\lambda}{w_o}.$$

There will thus be two cases depending on the range of network capacity. For $\overline{\mu}_d \leq \mu$, the purchase of the priority leads to no extra investment: that is, $h_d^* = 0$. By contrast, for $\mu < \overline{\mu}_d$ the major content provider would need an additional investment of

$$h_d^*(\mu) = \frac{1 + w_o}{w_o}\frac{\lambda}{\mu} - 1 \quad \text{for} \quad \mu < \overline{\mu}_d. \tag{39}$$

From the optimal QoS investments explicitly derived in (38) and (39), we can replicate most of qualitative results that we have thus far obtained.

However, two differences are noteworthy when we use this specification of a discrete quality of service. First, the purchase of the prioritized delivery class becomes a complete substitute for the QoS investment when $\mu > \overline{\mu}_d$. That is, MCP will make no investment with the prioritized delivery service, which is not the case for a continuous utility function in QoS as in (1). Second and more important, in a discrete QoS utility setting the traffic management effect is no longer guaranteed to be positive. For instance, consider (a, μ) such that $w_n(a, \mu) \geq w_o$. Then, prioritizing MCP's content has no effect on its *effective* waiting cost, but only increases the waiting cost for NCPs' content, implying a negative traffic management effect.

7.3 Multiple MCPs

One may wonder how much our model of a single MCP would restrict economic insights compared to multiple MCPs. For discussion's sake, let us consider two MCPs but all logic can be applied generally for any $N \geq 2$. With multiple MCPs, there are at least three new factors to be involved. First of all, heterogeneous MCPs may have different entry conditions so that each may or may not enter the network without the priority. Second, we need to distinguish the two different situations: (i) the ISP sells the prioritized service to one of the two MCPs exclusively (Choi and Kim, 2010) or (ii) the ISP sells it to both at the same time (Cheng *et al*, 2011). Third, the entry and resulting payoffs are affected by whether the two MCPs are competing entities (if so, how intensive the competition is) or not. One's exclusive priority can imply one's competitive advantage over the other when the two are competing, which is similar to an instrument of raising rival's cost which may block the potential competitor's entry. Thus, even with the two MCPs there are many sub-cases to be examined and three factors can interplay intricately. Here let us discuss economic

insights to be gained in this extended model, leaving detailed analysis a future research agenda.

Suppose the network capacity is sufficiently large so that two MCPs can provide their own content without entry concern. From social welfare perspective, the neutrality regulation involves the same trade-off characterized in Section 5 for the internal margin case between static traffic management and dynamic QoS investment. If the ISP allocates the fast-lane to only one MCP, the difference between the payoffs with and those without it will determine the winner and the transfer price will depend on bargaining powers. The competition between the MCPs will make the exclusive priority more valuable and thus yields more rent to the ISP. If the ISP delivers the priority to both MCPs, the relative value of the priority decreases compared to the exclusive priority. While these new factors would generate more nuanced results, we find that our key insight is robust to this extension.

Now suppose the capacity is small and each MCP needs a prioritized service (regardless of its exclusiveness) for effective content business. In this external margin case, from the viewpoint of the social welfare, we would again obtain the same trade-off featured in Section 4 between new content and congestion externality. With the exclusive priority, the new content effect is limited but congestion externality is also refrained. Again the value of the priority is expected to be higher with more intense competition among MCPs. Overall, we find our model does not lose the main economic forces.

However, some interesting issues may arise with multiple MCPs. For example, suppose that MCPs are heterogeneous in their costs of QoS investment and only highly efficient MCP(s) can enter without the prioritization, but less efficient MCPs can enter the network only with the prioritized service. Would the social welfare be higher in a non-neutral network than in a neutral network? We think that it crucially depends on how the efficient MCPs (whose entry is unrestricted) would respond to new entrants in a non-neutral network. The new entrants increases the overall traffic volume so that the incumbents need to invest more on their QoS investments to deliver the content successfully, which generate positive externality. Interestingly, the extent of the new entry will be endogenously affected by the incumbents' investments. Since non-major CPs will be also affected by the changes in QoS investments and the congestion externality, and social welfare varies with new content entry, more delicate analysis is required to sort out various interactions. So, we acknowledge a model of multiple MCPs offers more nuanced results as well as new research agenda, but our simple model maneuvers to deliver key economic insights with substantial tractability.

8 Conclusion

The debate on net neutrality has been the most important and controversial regulatory agenda since the inception of the Internet.[32] Not surprisingly, economists have extensively studied and helped to frame various issues, e.g., effects of network neutrality regulations on ISPs' investment incentives and on consumer surplus and social welfare, as well as on the entry/exit of content providers. Yet, the extant literature have not paid due attention to its effects on another crucial innovations taking place at the edges of the Internet in reality, although it is imperative for scholars, regulators and policy-makers to grasp how network regulations would affect these innovations at the edges (Maxwell and Brenner, 2012).

In this paper, we develop a theoretical model that characterizes the relative size of network capacity as a distinguishing feature to allow the entry of a congestion-sensitive content, and investigate major content providers' incentives to invest in QoS. Our analysis sheds new light on various trade-offs that net neutrality regulations bring forth to social welfare. The paid prioritization service can induce high-bandwidth content providers to enter the limited capacity mobile networks with greater QoS investments, but this comes at the cost of increasing total traffic volume. When the entry is not constrained by the network capacity, the prioritization relieves content providers of their burden of QoS investments and improves efficiency by allocating the higher speed lane to more congestion-sensitive content. However, smaller QoS investments may be detrimental to social welfare. Our insight is consistent or even strengthened when we consider the ISP's incentive to invest in capacity. We hope that our analysis benefits the on-going neutrality debate.

[32] As of May 6, 2015 'net neutrality' gives 1.28 million search results at news.google.com and 14.2 million Google web results, which indicates how net neutrality has become newsworthy in a short period.

References

[1] Altman, Eitan; Julio Rojas; Sulan Wong; Manjesh Kumar Hanawal and Yuedong Xu. 2012. "Net Neutrality and Quality of Service," Game Theory for Networks. Springer, 137-52.

[2] Bandyopadhyay, Subhajyoti; Hong Guo and Hsing Cheng. 2009. "Net Neutrality, Broadband Market Coverage and Innovation at the Edge." Broadband Market Coverage and Innovation at the Edge (May 15, 2009).

Becker, Gary S.; Dennis W. Carlton; Hal S. Sider. 2010. "Net Neutrality and Consumer Welfare." *Journal of Competition Law and Economics* 6(3): 497-519.

[3] Bourreau, Marc, Frago Kourandi, and Tommaso Valletti. 2012. "Net Neutrality with Competing Internet Platforms." mimeo.

[4] Cheng, Hsing Kenneth; Subhajyoti Bandyopadhyay and Hong Guo. 2011. "The Debate on Net Neutrality: A Policy Perspective." *Information Systems Research* 22(1): 60-82.

[5] Choi, Jay Pil and Byung-Cheol Kim. 2010. "Net Neutrality and Investment Incentives." *Rand Journal of Economics* 41(3): 446-71.

[6] Choi, Jay Pil; Doh-Shin Jeon and Byung-Cheol Kim. 2015. "Net Neutrality, Business Models, and Internet Interconnection." (in press) *American Economic Journal: Microeconomics.*

[7] Economides, Nicholas and Benjamin E Hermalin. 2012. "The Economics of Network Neutrality." *Rand Journal of Economics* 43(4): 602-629.

[8] Economides, Nicholas and Benjamin E Hermalin. 2015. "The Strategic Use of Download Limits by a Monopoly Platforms." *Rand Journal of Economics* 46(2): 297-327.

[9] Economides, Nicholas and Joacim Tåg. 2012. "Network Neutrality on the Internet: A Two-Sided Market Analysis." *Information Economics and Policy* 24(2): 91-104.

[10] Eisenach, Jeffrey A. 2012. "Broadband Competition in the Internet Ecosystem." American Enterprise Institute Working Papers 35845.

[11] Gans, Joshua. 2015. "Weak versus Strong Net Neutrality." *Journal of Regulatory Economics* 47(2): 183-200.

[12] Grafenhofer, Dominik. 2010. "Price Discrimination and the Hold–up Problem: A Contribution to the Net–Neutrality Debate." mimeo.

[13] Guo, Hong; Hsing Kenneth Cheng and Subhajyoti Bandyopadhyay. 2013. "Broadband Network Management and the Net Neutrality Debate." *Production and Operations Management* 22(5):1287-1298.

[14] Hermalin, Benjamin E and Michael L Katz. 2007. "The Economics of Product-Line Restrictions with an Application to the Network Neutrality Debate." *Information Economics and Policy* 19(2): 215-48.

[15] Jullien, Bruno and Wilfried Sand-Zantman. 2013. "Pricing Internet Traffic: Exclusion, Signalling, and Screening." mimeo.

[16] Krämer, Jan and Lukas Wiewiorra. 2012. "Network Neutrality and Congestion Sensitive Content Providers: Implications for Content Variety, Broadband Investment, and Regulation." *Information Systems Research* 23(4): 1303-21.

[17] Krämer, Jan; Lukas Wiewiorra and Christof Weinhardt. 2013. "Net Neutrality: A Progress Report." *Telecommunications Policy* 32: 794-813.

[18] Lee, Daeho, and Junseok Hwang. 2011. "The Effect of Network Neutrality on the Incentive to Discriminate, Invest and Innovate: A Literature Review." No. 201184. Seoul National University; Technology Management, Economics, and Policy Program (TEMEP).

[19] Lee, Robin S. and Tim Wu. 2009. "Subsidizing Creativity through Network Design: Zero-Pricing and Net Neutrality. " *Journal of Economics Perspective* 23(3): 61-76.

[20] Maxwell, Winston J. and Daniel L. Brenner. 2012. "Confronting the FCC Net Neutrality Order with European Regulatory Principles." *Journal of Regulation*, June.

[21] Mialon, Sue H and Samiran Banerjee. 2013. "Net Neutrality and Open Access Regulation on the Internet." mimeo.

[22] Mu, Hairong and Carlo Reggiani. 2011. "The Internet Sector and Network Neutrality: Where Does the EU Stand?" Indra Spiecker, Jan Krämer, editor(s). Network Neutrality and Open Access. Baden-Baden: Nomos Verlag, 115-151.

[23] Musacchio, John; Galina Schwartz and Jean Walrand. 2009. "A Two-Sided Market Analysis of Provider Investment Incentives with an Application to the Net-Neutrality Issue." *Review of Network Economics* 8(1): 1-18.

[24] Njoroge, Paul, Asuman Ozdaglar, Nicolás E. Stier-Moses, Gabriel Y. Weintraub. 2013. "Investment in Two Sided Markets and the Net Neutrality Debate." *Review of Network Economics* 12(4): 355-402.

[25] Peitz, Martin and Florian Schuett. 2015. "Net Neutrality and Inflation of Traffic." mimeo.

[26] Read, Darren. 2012. "Net Neutrality and the EU Electronic Communications Regulatory Framework." *International Journal of Law and Information Technology* 20(1): 48-72.

[27] Reggiani, Carlo and Tommaso Valletti. 2012. "Net Neutrality and Innovation at the Core and at the Edge." mimeo.

[28] Schuett, Florian. 2010. "Network Neutrality: A Survey of the Economic Literature." *Review of Network Economics* 9(2): Article 1.

[29] Xiao, XiPeng. 2008. Technical, Commercial and Regulatory Challenges of Qos: An Internet Service Model Perspective. Elsevier Science.

Appendix: Proofs

Proof of Proposition 1
The proof is straightforward as the following inequalities establish:

$$\begin{aligned}\Psi_d(h_d^{FB}) &= kw_d(h_d^{FB}) + W_d(h_d^{FB}) + c(h_d^{FB}) \leq kw_d(h_n^{FB}) + W_d(h_n^{FB}) + c(h_n^{FB}) \\ &< kw_n(h_n^{FB}) + W_n(h_n^{FB}) + c(h_n^{FB}) = \Psi_n(h_n^{FB})\end{aligned}$$

The first line of the above proof is by a revealed preference argument. The second inequality is based on Property 4.

Proof of Lemma 1
For the comparative statics, let us define an implicit function $G(h_n; \mu, k, \lambda) \equiv \frac{k\lambda(\mu-1)}{[(\mu-1)(1+h_n)-\lambda]^2} - c'(h_n) = 0$ from (7) around the point h_n^*. Then, we can apply the Implicit Function Theorem as follows:

$$\left.\frac{\partial h_n}{\partial \mu}\right|_{h_n=h_n^*} = -\frac{\frac{\partial G}{\partial \mu}(h_n^*)}{\frac{\partial G}{\partial h_n}(h_n^*)}.$$

Once can easily determine the signs of the denominator and the numerator of $\left.\frac{\partial h_n}{\partial \mu}\right|_{h_n=h_n^*}$:

$$\frac{\partial G}{\partial h_n}(h_n^*) = \frac{-2k\lambda(\mu-1)^2}{[(\mu-1)(1+h_n^*)-\lambda]^3} - c''(h_n^*) < 0;$$

$$\frac{\partial G}{\partial \mu}(h_n^*) = \frac{-k\lambda(\mu-1)(1+h_n^*) - k\lambda^2}{[(\mu-1)(1+h_n^*)-\lambda]^3} < 0,$$

which proves Lemma 1. ■

Proof of Lemma 3

Proof of Part (i)

Our reasoning follows proof by contradiction. Let μ' be an initial capacity and $\mu''(>\mu')$ a new capacity. Suppose, as a working hypothesis, in negation that $w_d(h_d^*(\mu'), \mu') < w_d(h_d^*(\mu''), \mu'')$. Then, let h'' be defined as

$$w_d(h_d^*(\mu'), \mu') = w_d(h'', \mu''), \tag{40}$$

which is equivalent to

$$\frac{\lambda}{\mu'(1+h_d^*(\mu'))-\lambda} = \frac{\lambda}{\mu''(1+h'')-\lambda}. \tag{41}$$

Note that $\mu'' > \mu'$ combined with (40) means $h'' < h_d^*(\mu')$. In addition, $w_d(h_d^*(\mu'), \mu') < w_d(h_d^*(\mu''), \mu'')$ implies that the major content provider would invest less than h'' when $\mu = \mu''$. From the first-order condition for $h_d^*(\cdot)$, we know $h_d^*(\mu')$ must satisfy the following condition:

$$k\frac{\lambda\mu'}{\left[\mu'(1+h_d^*(\mu'))-\lambda\right]^2} = C'(h_d^*(\mu')). \tag{42}$$

The marginal gain of investment for the major CP with $h = h''$ at $\mu = \mu''$ can be expressed as the following equivalent equations:

$$k\frac{\lambda\mu''}{\left[\mu''(1+h'')-\lambda\right]^2} = k\frac{\lambda\mu''}{\left[\mu'(1+h_d^*(\mu'))-\lambda\right]^2} = C'(h_d^*(\mu'))\frac{\mu''}{\mu'}.$$

The first equality holds because of (41), and the second one is from (42). From $h'' < h_d^*(\mu')$, however, we must have

$$C'(h_d^*(\mu'))\frac{\mu''}{\mu'} > C'(h'').$$

Hence, at the choice of $h = h''$ at $\mu = \mu''$, the marginal gain exceeds the marginal cost. This contradicts the working hypothesis of $w_d(h_d^*(\mu'), \mu') < w_d(h_d^*(\mu''), \mu'')$. ■

Proof of Part (ii)
Using the result of Part (i), we can state that

$$\begin{aligned}
\frac{dw_d(h_d^*(\mu),\mu)}{d\mu} &= \frac{\partial w_d(h_d^*(\mu),\mu)}{\partial \mu} + \frac{\partial w_d(h_d^*(\mu),\mu)}{\partial h}\frac{\partial h_d^*}{\partial \mu} \\
&= \frac{\partial w_d(a_d^*(\mu),\mu)}{\partial \mu} + \frac{\partial w_d(a_d^*(\mu),\mu)}{\partial h}\frac{\partial a_d^*}{\partial \mu} \\
&= -\frac{a_d^*\lambda}{\left(\mu - a_d^*\lambda\right)^2} + \left[\frac{\lambda}{\left(\mu - a_d^*\lambda\right)} + \frac{a_d^*\lambda^2}{\left(\mu - a_d^*\lambda\right)^2}\right]\frac{\partial a_d^*}{\partial \mu} < 0
\end{aligned}$$

The last inequality is equivalent to

$$\frac{\partial a_d^*}{\partial \mu} < \frac{a_d^*}{\mu}. \tag{43}$$

Now, we use the waiting cost for the non-major content of

$$W_d(a_d^*(\mu),\mu) = \frac{\mu}{\mu - (1 + a_d^*\lambda)}\frac{1}{\mu - a_d^*\lambda}$$

and show the following two inequalities:

$$\frac{d\left[\frac{\mu}{\mu-(1+a_d^*(\mu)\lambda)}\right]}{d\mu} < 0 \text{ and } \frac{d\left[\frac{1}{\mu-a_d^*(\mu)\lambda}\right]}{d\mu} < 0.$$

Regarding the first inequality, we have

$$\frac{d\left[\frac{\mu}{\mu-(1+a_d^*(\mu)\lambda)}\right]}{d\mu} = \frac{1}{\mu - (1 + a_d^*\lambda)} - \frac{\mu}{\left[\mu - (1 + a_d^*\lambda)\right]^2} + \lambda\frac{\mu}{\left[\mu - (1 + a_d^*\lambda)\right]^2}\frac{\partial a_d^*}{\partial \mu},$$

which becomes negative if

$$\frac{\partial a_d^*}{\partial \mu} < \frac{1 + a_d^*\lambda}{\lambda\mu} = \frac{1}{\lambda\mu} + \frac{a_d^*}{\mu}. \tag{44}$$

Using (43), we show that inequality (44) always holds.

Similarly, regarding the second inequality, we show that

$$\frac{d\left[\frac{1}{\mu-a_d^*(\mu)\lambda}\right]}{d\mu} = -\frac{1}{\left(\mu - a_d^*\lambda\right)^2} + \frac{\lambda}{\left(\mu - a_d^*\lambda\right)^2}\frac{\partial a_d^*}{\partial \mu} < 0$$

if

$$\frac{\partial a_d^*}{\partial \mu} < \frac{1}{\lambda},$$

which also holds from (43) as $\mu > a\lambda$.

Because both product terms in $W_d(a_d^*(\mu), \mu)$ decrease in μ, the proof of Part (ii) is completed. ■

Proof of Lemma 4

The total derivative of $\Delta\Pi^m(\mu, \beta)$ with respect to μ yields

$$\begin{aligned} \frac{d\Delta\Pi^m(\mu,\beta)}{d\mu} &= \frac{d\pi_d^*(\mu)}{d\mu} - \beta\frac{d\left[W_d(h_d^*(\mu),\mu) - W_n(\phi,\mu)\right]}{d\mu} = k\left|\frac{\partial w_d}{\partial\mu}\right| - \beta\frac{dW_d}{d\mu} - \beta\left|\frac{\partial W_n}{\partial\mu}\right| \\ &> k\left|\frac{\partial w_d}{\partial\mu}\right| - \beta\left|\frac{\partial W_n}{\partial\mu}\right| = k\frac{a_d^*\lambda}{(\mu - a_d^*\lambda)^2} - \beta\frac{1}{(\mu-1)^2}, \end{aligned} \tag{45}$$

where in the first inequality we use Lemma 3(ii), i.e., $\frac{dW_d}{d\mu} < 0$. A sufficient condition for $\Delta\Pi^m(\mu, \beta)$ to increase in μ can be characterized by $k \geq \overline{k}(\mu, \beta)$, where

$$\overline{k}(\mu,\beta) = \frac{\beta}{a_d^*\lambda}\left(\frac{\mu - a_d^*\lambda}{\mu - 1}\right)^2, \tag{46}$$

that is, $\left.\frac{d\Delta\Pi(\mu,\beta)}{d\mu}\right|_{k\geq\overline{k}} \geq 0$ and the equality holds at $k = \overline{k}$.

The right-hand side of inequality (46) is decreasing in $a_d^*\lambda$. In particular, if $a_d^*\lambda > 1$, the threshold $\overline{k}$ is smaller than one, regardless of $k \geq 1$ and $\beta \in [0, 1]$. Intuitively, if the traffic volume of the high-bandwidth content is so large ($a_d^*\lambda > 1$), the relative merit of the non-neutral treatment is always increasing in the capacity. ■