Asymmetric Neutrality Regulation and Innovation at the Edges: Fixed vs. Mobile Networks*

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

August 31, 2014

Abstract

We study how net neutrality regulations affect a high-bandwidth content provider's (CP) investment incentives in quality of services (QoS). We find that the effects crucially depend on network capacity levels. With limited capacity, as in mobile networks, 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, as in fixed-line networks, prioritized services reduce QoS investment as they become substitutes, but improves 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, asymmetric regulation, quality of service, investment incentives, queuing, congestion, mobile/fixed networks

---

*We thank Marc Bourreau, Jane Choi, Jeroen Hinloopen, Bruno Jullien, Martin Peitz, Wilfried Sand-Zantman, Glenn Woroch, and seminar participants at 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. The usual disclaimer applies.

†School of Economics, University of New South Wales, Sydney, NSW 2052, Australia and Department of Economics; 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.

§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 debate on net neutrality has been the most important and controversial regulatory agenda since the inception of the Internet. The "open Internet" order in 2010[1] adopted by the U.S. Federal Communication Commission (FCC) has played as a focal guideline for neutrality regulations. One well-known controversy surrounding this order has been whether the FCC has legitimate statutory authority to impose any regulatory obligations over the Internet. While the FCC has legitimate regulatory authority over telecommunication services under Title II of the Communications Act regarding "common-carriers," the Internet is currently categorized as *information services*, and thereby considered a non-common carrier; the FCC's powers are considerably limited in the information services governed by Title I of the Act.[2]

As a result, some Internet broadband access providers such as Comcast, Verizon Communications, and Metro PCS have challenged the legality of the FCC's order. The United States Court of Appeals for the District of Columbia Circuit ruled on the case *Comcast Corp. v. FCC* (600 F.3d 642) on April 6, 2010 that the FCC overstepped its authority when it imposed anti-discrimination rules on Comcast, which blocked BitTorrent applications in the summer of 2008. For the case *Verizon v. FCC* (740 F.3d 623), the same D.C. Circuit found on Jan. 14, 2014 that the FCC's Order that prevented deals between Verizon and content providers for faster delivery is not legitimate, while at the same time stating that the Commission *does* have some basic authority "to promulgate rules governing broadband providers' treatment of Internet traffic." After all, the verdicts were mixed.

The current regulatory stance on net neutrality is also fluid with mixed messages. The FCC recently announced that it would propose new rules that allow major content providers like Google, Netflix, and Disney to pay Internet service providers for preferential treatment of their content[3] whereas the FCC Chairman, Tom Wheeler, made a strong statement in his speech (April 30, 2014) at the National Cable & Telecommunications Association that he would consider reclassifying the

[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]For more details on reclassification and related issues, we refer to "Net neutrality is on trial in Washington. Here's what you need to know" by Timothy B. Lee in *The Washington Post* on Sept. 10, 2013.

[3]Netflix recently struck such a deal with Comcast. For a related newspaper article, see "F.C.C. in a Shift, Backs Fast Lanes for Web Traffic" by Edward Wyatt in *The New York Times* on April, 24, 2014.

Internet as a telecommunication service to enable the regulation of the Internet under Title II. With a split vote of 3-2, the FCC decided to open up for public debate regarding new rules for the open Internet (May 15, 2014). Net neutrality thus still remains a contentious regulatory issue.

Another issue of importance, which appears to have been somewhat ignored, is that the FCC's Order treated mobile network operators more leniently than fixed wireline network operators. More specifically, its first two rules, namely, (i) 'transparency' and (ii) 'no blocking' are commonly applied to both types of network operators, but the third rule (iii) 'no unreasonable discrimination' appertains only to fixed line operators:

> A person engaged in the provision of *fixed* broadband Internet access service, insofar as such person is so engaged, shall not unreasonably discriminate in transmitting lawful network traffic over a consumer's broadband Internet access service. Reasonable network management shall not constitute unreasonable discrimination. (47 of CFR §8.7, italics added)

Maxwell and Brenner (2012) described such asymmetric treatment of fixed and mobile networks as "by far the most controversial aspect of the FCC's order insofar as it is designed to prohibit paid prioritization arrangements between an Internet access provider and upstream content, application or service providers." Importantly and interestingly, this asymmetric regulatory approach is in sharp contrast to the European approach to the same issue; European regulatory standards, the 2002 EC Directives on electronic communications and its revisions in 2009,[4] have no such distinction between fixed and mobile networks. The uniform treatment reflects one of the European regulatory principles, "technological neutrality,"which allows no differential treatment across all types of networks including cable, mobile, and fixed wireline networks.[5]

Nevertheless, to our knowledge, no rigorous analysis has been done on the sharp contrast in regulatory approaches between the US and EU; our study fills this void. We study when mobile networks would call for asymmetric regulation and when uniform treatment may be justified.

Our study is not only motivated by regulatory differences, but also by the implications of

[4]Directive 2002/22/EC of the European Parliament and of the Council ("Universal Service Directive") and Directive 2002/21/EC ("Framework Directive"); amendments were made under 2009/1140/EC (the "Better Regulation Directive") and Directive 2009/136/EC (the "Consumer Rights Directive").

[5]For a specific example, the Netherlands enacted net neutrality law in 2011 that prohibited *mobile network operators* from charging extra fees to customers on certain applications, which is opposite to the US FCC's rather lenient treatment of mobile network operators. Krämer, Wiewiorra, and Weinhardt (2013) offer a comprehensive literature review on recent progress of net neutrality issues.

neutrality regulation on innovation incentives at the "edges." The extant literature on network neutrality has mainly focused on the expansion of Internet service providers (ISPs)' network capacity as innovation at the "core."[6] However, the ISPs' capacity expansion making bigger "pipelines"is not the only solution to resolving the congestion problem in the modern Internet ecosystem. In fact, major content providers such as Google, Netflix, and Amazon have 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 is not made under neutral

[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.

[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.

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, which resonates with the rationale given by the FCC for its differential treatment between fixed and mobile networks:

> Mobile broadband is an earlier-stage platform than fixed broadband, and it is rapidly evolving. Mobile broadband speeds, capacity, and penetration are typically much lower than for fixed broadband. (...) In addition, existing mobile networks present operational constraints that fixed broadband networks do not typically encounter. (FCC Order, par. 94-95)

The FCC appears to believe that its lenient non-neutral treatment facilitates the availability of innovative content and applications in the early-stage mobile network. However, the entry of new content does not necessarily result in higher welfare. This is because the new content will consume a substantial portion of the existing network capacity, which increases the congestion for other content. Such a negative externality of congestion becomes more pronounced with a limited capacity network such as mobile. Indeed, the surplus from new content can be outweighed by the efficiency loss from the elevated congestion for other content when the negative externality is not internalized in the content provider's entry decision.

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 type of 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 applicable

[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 FCC's intensive efforts to apply network neutrality regulations.

to the fixed network where the entry of content providers has not been treated as a serious concern.

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 incentive does not depend on the surplus created by the major CP than under non-neutral networks in which the ISP partially internalizes the surplus. Provided that the entry occurs, 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 for mobile networks neutrality regulation can be adverse to the entry of major CPs, whereas for fixed networks non-neutrality may reduce the ISP's incentive to invest in capacity.

Our study makes two primary contributions to the debate of net neutrality regulations. First, we provide a novel theoretical model of major content providers' QoS investment. Considering the importance of innovations at the edges of the Internet, we think it critical to offer a formal theory to understand innovation incentives and associated externalities across different network capacities. Second, our model provides a useful framework through which one can comprehend the contrasting neutrality regulations between the US FCC, which treats mobile networks more leniently than fixed networks, and the EU which treats both networks uniformly.

The remainder of our paper is organized as follows. After reviewing related literature, we present our model in Section 2 including a generalized queuing model which describes how prioritization and QoS 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 neutrality regulation can be justified only as a second-best policy when a social planner cannot directly control content providers' entry and investment decisions. After the first-best, we analyze the QoS investment decisions by the major content providers under neutral and non-neutral network regimes. We show how mobile networks and fixed networks can be differentiated depending on network capacity. In Sections 4 and 5, we provide our main analysis for mobile and fixed networks, respectively. In Section 6, we analyze the ISP's capacity choice. Section 7 presents two extensions of our model with consumer heterogeneity and discrete QoS. We wrap up with concluding remarks in Section 8. Lengthy mathematical proofs are relegated to the Appendix.

1.1 Related Literature

Several survey articles such as Lee and Wu (2009), Schuett (2010), Lee and Hwang (2011), and Krämer, Wiewiorra, and Weinhardt (2012) have offered comprehensive reviews on the literature of net neutrality. So, we here briefly mention notable works in relation to this paper.

The main focus of the extant studies has been investment incentives for the ISPs (at the core) and content providers (at the edges). One major issue in the net neutrality debate is Internet access service providers' investment incentives 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." Economic 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 can refer to Bandyopadhyay, Guo, and Cheng (2009), Choi and Kim (2010), Grafenhofer (2010), Reggiani and Valletti (2012), and Bourreau, Kourandi, and Valletti (2012).

Beyond investment incentives, economists have studied how network neutrality would affect consumer and social welfare from various perspectives. Hermalin and Katz (2007) analyze network neutrality from the perspective of product line restrictions in a vertical differentiation model. Economides and Tåg (2012) regard neutrality regulation as a zero-pricing regulation on the content side in a two-sided market. Mialon and Banerjee (2013) study how the effects of net neutrality on Internet access (or subscription) price and social welfare crucially depends on the market structure of the content side. Choi, Jeon, and Kim (2013) 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. Jullien and Sand-Zantman (2013) examine the net neutrality issues in the context of information transmission such as signaling and screening.

We find Peitz and Schuett (2014) more closely related to our paper though they consider a different type of externality to the network derived from content providers. 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.

Our paper departs from the earlier literature in several respects. We focus on content providers' incentives to invest in alternative ways of reducing congestion beyond the ISPs' network capacity. This is in line with the basic premise in the debate that end-users' quality of service must be the primary goal of a desirable network ecosystem (See Xiao (2008), Altman et al. (2012), and Guo, Cheng, and Bandyopadhyay (2013)). We show how these alternative mechanisms can be complements or substitutes to network capacity depending on the ISP's capacity limit. Our analysis captures the differences between fixed and mobile networks because mobile networks encounter technical and physical constraints in expanding capacity due to the limited availability of spectrum. It highlights the FCC's asymmetric regulation between the two networks in contrast to the EU's uniform treatment.[10]

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.[11] 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 the 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. Each

[10]See Read (2012) and Hairong and Reggiani (2011) for the EU's regulatory framework.

[11]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, Jeon, and Kim (2013) addresses the equivalence in network quality choices between interconnected ISPs and a monopoly ISP.

consumer demands both the MCP's and NCPs' content. When a consumer receives the MCP's content with average waiting time of w, the consumer earns utility

$$u(w) = v - kw. \tag{1}$$

where parameter v represents the consumer's intrinsic utility from receiving the MCP's content. The corresponding utility from the NCPs' content with an average waiting time of W is given by

$$U(W) = V - W. \tag{2}$$

where V represents the consumer's intrinsic utility from receiving the NCPs' aggregate content. 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. The parameter $k \geq 1$ measures the relative sensitivity of the MCP's content to delays compared to the NCPs'. Since we assume that the mass of consumers is normalized to one, $u(w)$ and $U(W)$ respectively represent the entire surplus from the MCP's content and the NCPs'.

We assume that the MCP can extract the entire surplus $u(w)$ 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.[12] For the NCPs' content, we introduce a parameter $\beta \in [0, 1]$ to denote the ISP's share of the total surplus generated by the NCPs' content delivery. In other words, the ISP receives $\beta U(W)$ from providing delivery services for the 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 the NCPs and end users by its decisions. If $\beta = 0$, the ISP will not take into account any potential effects on the NCPs' content traffic when the ISP deals with the 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 the 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

[12] 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.

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.[13]

Let μ denote the ISP's network capacity. Each consumer demands a wide range of content from both the MCP and NCPs. The content request rate follows a Poisson process, which represents the intensity of content demand. For the 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 the NCPs' content is also normalized to one. By contrast, we envision the MCP as one discrete player operating a content network platform that provides a continuum of content whose aggregate packet size is given by λ.[14] Then, we can interpret λ as the sheer volume of the MCP's content or a measure of the relative traffic volume of the MCP's content vis-à-vis the NCPs' aggregate traffic volume. The total traffic volume for the ISP thus amounts to $1 + \lambda$. Note that 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.[15] 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

[13]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.

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

[15]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.

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 the 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 that 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 investment $F(\geq 0)$ for any positive investment $h > 0$.

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 the 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 for the major CP's content:

$$w_n(a, \mu) = \underbrace{\frac{1}{\mu - (1 + a\lambda)}}_{\text{waiting time per packet}} \times \underbrace{a\lambda}_{\text{total packet size}}. \tag{3}$$

The total volume of traffic (packet size) amounts to $1 + a\lambda$ (one for the NCPs' content and $a\lambda$ for the MCP's content with compression), 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 (3). 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

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

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

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

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

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

$$W_d(a,\mu) = \frac{\mu}{\mu - (1+a\lambda)} \frac{1}{\mu - a\lambda} \times 1. \tag{6}$$

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 $r = n, d$.

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, the 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 the 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, the 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 that the 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 the MCP.

3 Optimal QoS Investment and Network Regimes

3.1 Benchmark: First-best

We first characterize the first-best outcome (given a network capacity μ) in which the social planner can control the 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 the 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 the 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:

$$h_r^{FB} = \arg\min_{h_r} \Psi_r(h) = kw_r(h) + W_r(h) + c(h). \tag{7}$$

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.

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

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

Proposition 1 tells us that the first-best outcome always entails a non-neutral network when the MCP's entry is socially desirable because it allows more efficient traffic management (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 the second-best neutral network can offer higher welfare than the second-best non-neutral network.

3.2 Neutral Networks

Let us 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 the MCP has entered. Assuming the MCP's entry, the content provider's optimal choice of h is to maximize its profit:

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

where $w_n(h,\mu) = \frac{\lambda}{(\mu-1)(1+h)-\lambda}$ from (3). The first order condition with respect to h becomes

$$\left.\frac{\partial \pi_n}{\partial h}\right|_{h_n^*} = \frac{k\lambda(\mu-1)}{[(\mu-1)(1+h)-\lambda]^2} - c'(h) = 0, \tag{8}$$

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 the MCP at the optimal investment level $h_n^*(\mu)$ for a given network capacity μ. By the Envelope Theorem, we find that the MCP obtains a higher profit as the network capacity increases:

$$\frac{d\pi_n^*}{d\mu} = \frac{\partial \pi_n}{\partial \mu} = -k\frac{\partial w_n(h_n^*,\mu)}{\partial \mu} = k\frac{\lambda(1+h_n^*)}{[(\mu-1)(1+h_n^*)-\lambda]^2} > 0. \tag{9}$$

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, the MCP makes an investment only when the ISP's capacity is above 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 the 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 let us consider the non-neutral network in which the MCP has an option to buy the prioritized delivery service at a negotiated price. One benefit of such an arrangement is that the 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 the MCP and the ISP agree on a price of the prioritized service. We define the MCP's profit gross of any payout for the priority as

$$\pi_d \equiv u - kw_d(h,\mu) - c(h) - F,$$

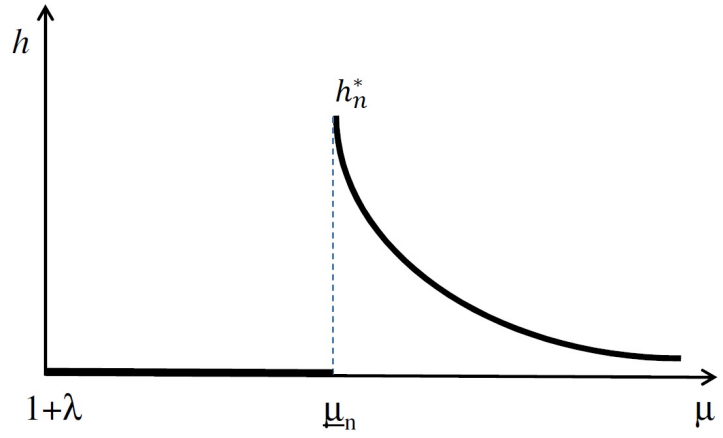

Figure 1: Optimal QoS Investment in the Neutral Network

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

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

As in the neutral network case, by 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, $\frac{\partial h_d^*}{\partial \mu} < 0$.[16]

Note that 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 the 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 the MCP's entry is facilitated to the maximum extent, and relegate the analysis of $\beta > 0$ to the next section.[17] 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 the 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

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

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

if $\mu \geq \underline{\mu}_d$. Again, the 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 $\mu > \underline{\mu}_n$.*

3.4 Mobile/Fixed Networks 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 the MCP's investment are "complements" in that prioritization induces the 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 the 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.

Our analysis on the content provider's entry and investment decisions offer a framework to assess the US FCC's asymmetric regulation across fixed and mobile networks. For mobile networks,

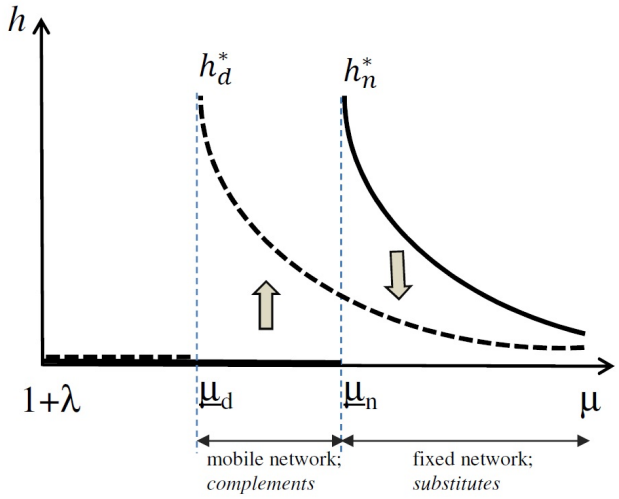

Figure 2: The QoS Investments and Mobile/Fixed Networks

there is a heightened concern for congestion and the entry of high-bandwidth content and complex applications can be facilitated with the availability of the prioritized service. The mobile networks encounter technical and physical constraints in expanding capacity due to the limited availability of spectrum whereas such constraints are not much of a restriction in fixed networks. In the remainder of the paper, we thus take the limited capacity case of $\mu \in (\underline{\mu}_d, \underline{\mu}_n)$ as a representation of mobile networks and the high capacity case of $\mu > \underline{\mu}_n$ as a representation of fixed networks.

4 Mobile Networks

In this section, we analyze the effects of net neutrality regulation on various participants in the mobile network with $\mu \in (\underline{\mu}_d, \underline{\mu}_n)$,with a particular focus on social welfare. As in the previous section, we characterize the mobile network as a limited capacity case in which the MCP makes no entry under the neutral regime because $\pi_n^*(\mu) < 0$ for $\mu < \underline{\mu}_n$ but is able to enter with paid prioritization in the non-neutral network (at least for $\beta = 0$).

4.1 Effects of MCP's Entry on NCPs

Under non-neutral network, the MCP's entry has two countervailing effects. On one hand, the new content generates a surplus of $\pi_d^*(\mu) > 0$, 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 congestion in the existing non-major content traffic through the 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 the MCP's content means a slower delivery for NCPs' content that is now relegated to the non-prioritized, slow lane.

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' by 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

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

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 prioritization. The prioritized service will be provided to the MCP and its price will be agreed upon between the ISP and the MCP if their joint profits increase with the service. The joint profits under the neutral regime will be given by $\Pi_n(\phi,\mu,\beta) = \beta[V - W_n(\phi,\mu)]$ in the mobile network because there is no entry by the MCP. With the 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:[18]

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

where the superscript m in $\Delta\Pi^m$ stands for the *mobile* network. Expression (12) clearly shows the trade-off associated with the prioritization in the mobile network. The MCP's entry generates the value of $\pi_d^*(\mu)$ but the ISP must bear the loss of $\beta\Delta W\left(\mu\right)$ due to the negative effects on the NCPs.

Note that $\Delta\Pi^m(\mu,1)$ captures the change in social welfare from the major CP's entry under the non-neutral network. One immediately sees that private incentives to introduce prioritized service

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

are thus excessive from a social planner's point of view. The discrepancy between the private and social incentives can be represented 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 that the welfare effects of prioritization and private incentives to provide prioritization crucially depend on the network capacity (μ), we need to examine how $\Delta\Pi^m(\mu,\beta)$ changes with μ for a given β. First, we analyze the effects of a higher network capacity on congestion. Consider the effect of network capacity (μ) on the waiting time for the MCP's content in the non-neutral network (w_d).

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

A higher network capacity has direct positive effects on the quality of service, which is represented by the first term on the RHS in (13). However, there are also indirect negative effects which counteract the direct effects on the waiting cost: the major CP responds to a higher network capacity by reducing its QoS investment. Nonetheless, we establish in Lemma 3 that the positive direct effects dominate the negative indirect effects.

Lemma 3 *The waiting time in the non-neutral network decreases as the network capacity increases regardless of the priority class:*

(i) $\frac{dw_d(h_d^*(\mu),\,\mu)}{d\mu} < 0$;

(ii) $\frac{dW_d(h_d^*(\mu),\,\mu)}{d\mu} < 0$.

Proof. See the Appendix. ■

We now analyze the private incentives to introduce prioritized service in the non-neutral network. From (12) and Lemma 3(i), it is clear that $\Delta\Pi^m(\mu,\beta)$ strictly increases in μ for $\beta = 0$ for any $k \geq 1$, and by continuity this result holds for small enough β. Even if β is not sufficiently small, we can still establish that the private incentives to introduce prioritized service increase with network capacity if k, the delay sensitivity parameter for the MCP's content, 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 basic intuition for Lemma 4 is that as k increases, the ability to access the fast lane by the MCP's content offers a greater benefit of allocating more congestion-sensitive content to the faster

lane while the negative externality on NCP's content is independent of k. In addition, because only a proportion β of the externality is internalized by the ISP, if β is small enough, the ISP and the major CP are not much affected adversely by the prioritization. In such a case, the prioritized service will be provided and the MCP will enter even when k is close to one.

Now let us define $\underline{\mu}_d(\beta)$ as the cutoff capacity above which the MCP enters and below which there is no entry: the cutoff capacity defined for $\beta = 0$ in Section 4 is denoted by $\underline{\mu}_d(0) = \underline{\mu}_d$. The MCP's investment $h_d^*(\mu)$ does not depend on β, given its entry. Because $[W_d(h_d^*(\mu);\mu) - W_n(\phi;\mu)] < 0$ is a constant (independent of β), the joint surplus conditional on the MCP's entry strictly decreases with β for a given μ, which yields the following:

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, and Lemma 5 tells us that for $\beta < 1$, such excessive entry can occur. This is because the coalition of the ISP and the major CP 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 (Mobile Networks) *Consider the mobile network with $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$. Suppose $k \geq \overline{k}(\mu,\beta)$.*

(i) Given β, the paid prioritization with two-tiered service induces the MCP's entry with congestion-sensitive content as long as $\mu \geq \underline{\mu}_d(\beta)$, where $\underline{\mu}_d(\beta)$ strictly increases with β and $\underline{\mu}_d(0) = \underline{\mu}_d$.

(ii) If $\underline{\mu}_d(1) < \underline{\mu}_n$, the MCP enters due to the prioritization though the entry is not socially desirable for $\mu \in (\underline{\mu}_d(\beta), \underline{\mu}_d(1))$. But, the entry is socially efficient for $\mu \in (\underline{\mu}_d(1), \underline{\mu}_n)$.[19]

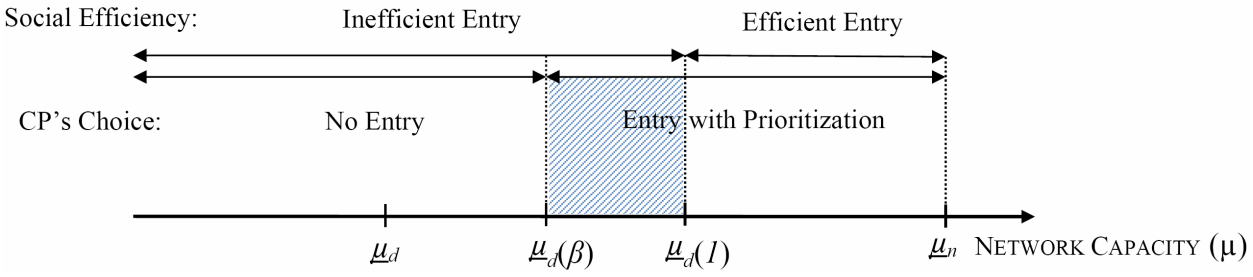

Figure 3: Social Efficiency and Private Entry Decision

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

Figure 3 illustrates Proposition 2-(ii): when $\underline{\mu}_d(1) < \underline{\mu}_n$, the MCP makes a socially inefficient entry due to the paid prioritization for the shaded range of μ. However, we note that socially excessive entry is not an essential result. Later in Section 7.1, where we consider consumer heterogeneity and/or competition between ISPs, we discuss when this result is relaxed and how an insufficient entry can occur.

5 Fixed Networks

In this section, we consider a fixed network in which the network capacity is large enough to induce the major content provider's entry regardless of the network regimes, i.e., $\mu \geq \underline{\mu}_n$. We first analyze the ISP and the major CP's joint private incentives to introduce the prioritization service. Their joint payoff in the network regime $r = n, d$ is given by

$$\Pi_r(h, \mu, \beta) = \pi_r(h, \mu) + \beta[V - W_r(h, \mu)].$$

As a result, the prioritization will be adopted if and only if

$$\Delta\Pi^f(\mu, \beta) = \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta) > 0,$$

where the superscript f stands for the *fixed* network and $\Delta\Pi^f(\mu, \beta) > 0$ means a higher joint payoff under the non-neutral network.

5.1 Traffic Management and QoS Investment Effects

We can decompose the effects of the prioritization on the joint payoff into the following: (1) the static traffic management effect and (2) the dynamic investment effect.

$$\Delta\Pi^f(\mu, \beta) = \underbrace{[\Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta)]}_{\text{Traffic Management Effect } (+)} + \underbrace{[\Pi_n(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta)]}_{\text{QoS Investment Effect } (-)}. \quad (14)$$

The first term in (14) is always positive and represents the *traffic management effect*: for any given QoS investment level h, prioritizing the major CP's traffic reduces the total delay cost because

the major CP's content is assumed to be more sensitive to congestion ($k > 1$). Precisely, we have

$$\begin{aligned} \text{traffic management effect} &= \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta) \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] + \beta\left[W_n(h_d^*(\mu)) - W_d(h_d^*(\mu))\right] \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] - \beta\left[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))\right] \\ &= (k-\beta)[w_n(h_d^*) - w_d(h_d^*)] > 0 \end{aligned} \tag{15}$$

where the third equality in (15) is obtained from Property 3: $w_n(h_d^*)+W_n(h_d^*) = w_d(h_d^*)+W_d(h_d^*)$.[20]

The expression in the second square bracket in (14) represents the *QoS investment effect*: the availability of the prioritized lane decreases the major CP's investment from $h_n^*(\mu)$ to $h_d^*(\mu)$, which affects the joint payoff. To determine the sign of this term, let $h_n^J(\mu, \beta)$ be the collectively optimal level of QoS investment which maximizes the joint profit of the two parties in the neutral regime, i.e.,

$$\begin{aligned} h_n^J(\beta) &= \arg\max_h \Pi_n(h, \mu, \beta) \ (= \pi_n(h, \mu) + \beta[U - W_n(h, \mu)]) \\ &= \arg\min_h \ [kw_n(h) + \beta W_n(h)] + c(h). \end{aligned} \tag{16}$$

From the profit maximization problem, the privately optimal choice by the MCP, $h_n^*(\mu)$, maximizes $\pi_n(h, \mu)$ and does not take into account the positive effect of its investment on $\beta W_n(h)$. Alternatively, from the cost minimization problem, the MCP chooses $h_n^*(\mu)$ to minimize $kw_n(h) + c(h)$, but ignores the negative externality $\beta W_n(h)$. In either way, such non-internalization of externalities implies that under-investment occurs from the perspective of joint profit maximization, i.e., $h_n^*(\mu) < h_n^J(\mu, \beta)$ unless $\beta = 0$.

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. Because we derived $h_d^*(\mu) < h_n^*(\mu)$ in Lemma 2 and just verified $h_n^*(\mu) < h_n^J(\mu, \beta)$, we find $\Pi_n(h_d^*(\mu), \mu, \beta) < \Pi_n(h_n^*(\mu), \mu, \beta)$, which implies that the QoS investment effect must be negative:

$$\text{QoS investment effect} = \Pi_n(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta) < 0. \tag{17}$$

[20]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^*.

5.2 Effects of Prioritization on Social Welfare

We now analyze a social planner's incentives to introduce the prioritization service and compare them to the private incentives. We consider a constrained (second-best) social optimum in which the social planner can only choose the network regime while the investment decision is left to the MCP. Note that social welfare in each regime coincides with the joint payoff of the ISP and the MCP when $\beta = 1$, and is given by

$$S_r(\mu) = \Pi_r(h_r^*(\mu), \mu, \beta = 1) = \pi_r(h_r^*(\mu), \mu) + [U - W_r(h, \mu)], \tag{18}$$

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

$$\begin{aligned} \Delta S(\mu) &= S_d(\mu) - S_n(\mu) \\ &= \Delta\Pi^f(\mu, \beta) + (1 - \beta)\left[W_n(h_n^*) - W_d(h_d^*)\right] \end{aligned} \tag{19}$$

When $\beta = 1$, the private incentive to use the prioritization service is perfectly aligned with the social incentive since the ISP fully internalizes any effect of providing the prioritized service on end consumers and non-major CPs (i.e., $\Delta S(\mu) = \Delta\Pi^f(\mu, 1)$). For any $\beta < 1$, however, the two parties have socially excessive incentive to adopt the prioritization service as they do not fully internalize the effect of increased delay on NCPs' content due to the prioritized service. More precisely, we have

$$\Delta S(\mu) - \Delta\Pi^f(\mu, \beta) = (1 - \beta) \underbrace{\left[W_n(h_n^*) - W_d(h_d^*)\right]}_{\text{externality on NCP's content}} \tag{20}$$

The externality term in (20) can be decomposed as follows.

$$W_n(h_n^*) - W_d(h_d^*) = [W_n(h_n^*) - W_n(h_d^*)] + [W_n(h_d^*) - W_d(h_d^*)] < 0 \tag{21}$$

The first square bracket of (21) has a negative sign because of Lemma 2 ($h_n^* > h_d^*$). The second term also takes a negative value by Property 2 in Section 2.3. Therefore, we verify the intuition that the externality term must be negative.

The discrepancy between the social incentives and the private incentives is inversely related to β. In particular, when the discrepancy reaches its maximum ($\beta = 0$), the ISP and the 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:

$$\begin{aligned} \Delta\Pi^f(\mu, \beta = 0) &= \pi_d(h_d^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &\geq \pi_d(h_n^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &= w_n(h_n^*(\mu), \mu) - w_d(h_n^*(\mu), \mu) > 0, \end{aligned} \tag{22}$$

where the first (weak) inequality comes from the revealed preference argument, and the inequality in the third line comes from Property 2.

Thus, we can summarize our findings for the fixed network as follows.

Proposition 4 (Fixed network) *Consider the fixed network with $\mu > \underline{\mu}_n$ in which the MCP always enters. Then, we find that*

(i) A prioritization service involves a trade-off between the positive efficient traffic management effect and the negative QoS investment effect; prioritization is adopted when the gain from the better traffic management due to the prioritization exceeds the loss from the diminished QoS investment.

(ii) There are socially excessive incentives to adopt a prioritization service, i.e., $\Delta S(\mu) \leq \Delta\Pi^f(\mu, \beta)$.

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 higher welfare than neutrality due to the traffic management effect. The example shows that in the second-best world, the reverse can hold because the under-investment problem is more severe in a non-neutral network than in a neutral network. We 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 the major CP 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 still outweighs the positive traffic management effect (Proposition 4). If k is sufficiently large ($k = 3$), such conflict disappears. 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, Jeon and Kim (2013) who show that it is possible for a discriminatory network to offer a lower social welfare than a neutral network due to its excessive quality distortion for the basic service. While the message sounds similar, the logic differs. Here, the concern about a non-neutral network arises from the negative effect of prioritization on innovation at edges from MCPs which generate positive externality onto other existing content.

6 ISP's Capacity Choice

In our baseline model, we focused on the MCP's QoS investment for a *given* ISP's network capacity μ. In this section, we allow the ISP to choose its capacity before the MCP entry decision, which helps us understand the interplay between the ISP's capacity choice and the MCP's entry with its ensuing investment. We add a new initial stage. In a neutral network, at stage **N-0**, the ISP chooses its network capacity before the **N-1** and **N-2** stages follow; similarly, **D-0** is added in a non-neutral network.

Note 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 $\partial h_r^*/\partial \mu$ provided that the MCP enters. In order to focus on the key effects, we conduct our analysis by assuming that the indirect effect is of second-order to the direct effect (which holds if $c(h)$ is convex enough). In addition, we assume that v and k are large enough to

capture a situation in which the QoS of the high-bandwidth content is crucial to consumer utility.

As a benchmark case, consider the ISP's optimal capacity investment in the absence of the MCP. Let $C(\mu)$ denote the investment cost of capacity μ with $C' > 0$ and $C'' > 0$. Then, the ISP's optimal capacity choice will be characterized by the equality of marginal benefit and marginal cost:

$$-\beta \frac{\partial W_n(\phi, \mu)}{\partial \mu} = C'(\mu). \tag{23}$$

Let $\mu_n^{\phi}(\beta)$ be the solution of Equation (23).

6.1 Neutral networks

The ISP's optimal capacity choice depends on whether the ISP induces the MCP to enter or not. If the ISP induces the MCP to enter in a neutral network, its optimal capacity is determined by

$$-\beta \left[\frac{\partial W_n(h_n^*(\mu), \mu)}{\partial \mu} + \frac{\partial W_n(h_n^*(\mu), \mu)}{\partial h} \frac{\partial h_n^*}{\partial \mu} \right] = C'(\mu). \tag{24}$$

Let $\mu_n^E(\beta)$ be the solution to Equation (24). If we consider a hypothetical situation where the MCP is unable to invest in QoS (i.e., $h = 0$), the ISP's optimal capacity would be $\widetilde{\mu}_n^E(\beta)$, where $\widetilde{\mu}_n^E(\beta)$ satisfies the following condition:

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

We can show that the MCP's QoS and the ISP's capacity investments constitute *substitutes* in that the ISP invests less when the MCP can make investment at the edge than when the MCP cannot (i.e., $h = 0$), that is, $\widetilde{\mu}_n^E(\beta) > \mu_n^E(\beta)$. To verify this, we note that the marginal benefit of capacity expansion with the MCP's non-negative investment is smaller than that with zero investment. That is, we have that

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

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 $\frac{\partial h_n^*}{\partial \mu} < 0$. Furthermore, because the ISP does not internalize the effect of its investment on $w_n(h_n^*(\mu), \mu)$ in a neutral network, the ISP always chooses a suboptimal investment level from a social viewpoint in a neutral network when it induces the MCP's entry.

Recall that the minimum capacity level $\underline{\mu}_n$ is required to induce the MCP's entry under a neutral network. Then, a necessary condition for the ISP to induce the MCP's entry is that $\mu_n^E(\beta) > \underline{\mu}_n$. Otherwise, the ISP invests $\underline{\mu}_n$ when it induces the MCP's entry. In this case, however, inducing

entry would give a lower profit to the ISP than inducing no entry by choosing $\underline{\mu}_n - \varepsilon$, where $\varepsilon > 0$ is infinitesimal. This is because the entry of the MCP would only increase congestion as is seen from $W_n(\phi, \underline{\mu}_n) < W_n(h_n^*(\underline{\mu}_n), \underline{\mu}_n)$.

Therefore, the ISP would induce entry if and only if it yields more profit than no entry, i.e.,

$$-\beta W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_n\}) - C(\min\{\mu_n^\phi, \underline{\mu}_n\}) \leq -\beta W_n(h_n^*(\mu_n^E), \mu_n^E) - C(\mu_n^E), \tag{25}$$

where the minimum operator appears in the LHS of (25) because, 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 $\mu_n^\phi(\beta) < \underline{\mu}_n$. Inequality (25) will hold 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 the MCP's entry if and only if

$$-W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_n\}) - C(\min\{\mu_n^\phi, \underline{\mu}_n\}) \leq \pi_n(h_n^*(\mu_n^E), \mu_n^E) - W_n(h_n^*(\mu_n^E), \mu_n^E) - C(\mu_n^E), \tag{26}$$

where $\pi_n(h_n^*(\mu), \mu) \equiv v - kw_n(h_n^*(\mu_n^E), \mu_n^E) - c(h_n^*(\mu_n^E)) > 0$ must hold since $\mu_n^E(\beta) > \underline{\mu}_n$. By comparing (25) and (26) 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 that the 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 the MCP's profit.

6.2 Non-neutral Networks

Consider a non-neutral network in which the ISP and the MCP bargain over the price of prioritization, so that without the neutrality regulation, the ISP's investment decision depends on its bargaining power against the MCP and its default payoff if the bargaining fails. We assume the Nash bargaining with equal bargaining power between the two parties. Because of the differences in default payoffs, we should distinguish two cases depending on whether or not the MCP enters without prioritization.

Consider the case in which the MCP does not enter without prioritization. Then, the ISP chooses its capacity to maximize the following objective:

$$\frac{\{\pi_d(h_d^*(\mu), \mu) - \beta\left[W_d(h_d^*(\mu), \mu) - W_n(\phi, \mu)\right]\}}{2} + \beta\left[U - W_n(\phi, \mu)\right] - C(\mu). \tag{27}$$

The first term in (27) is the half of the surplus created by prioritization. The second term is the ISP's default payoff in a neutral network without the MCP's entry. The first-order condition with

respect to μ is given by

$$-\frac{k}{2}\frac{dw_d(h_d^*(\mu),\mu)}{d\mu}-\frac{\beta}{2}\frac{dW_d(h_d^*(\mu),\mu)}{d\mu}-\frac{\beta}{2}\frac{dW_n(\phi,\mu)}{d\mu}=C'(\mu). \tag{28}$$

Let $\widehat{\mu}(\beta)$ denote the solution to Equation (28). In order not to induce the MCP's entry, the ISP's capacity cannot exceed $\underline{\mu}_n$. Thus, the optimal capacity conditional on no entry without prioritization (but entry with prioritization) is given by $\mu_d^E(\beta)=\min\left\{\widehat{\mu}(\beta),\underline{\mu}_n\right\}$.[21]

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

$$\frac{\{\pi_d(h_d^*(\mu),\mu)-\pi_n(h_n^*(\mu),\mu)-\beta\left[W_d(h_d^*(\mu),\mu)-W_n(h_n^*(\mu),\mu)\right]\}}{2}+\beta\left[U-W_n(h_n^*(\mu),\mu)\right]-C(\mu). \tag{29}$$

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

$$-\frac{k}{2}\left[\frac{dw_d(h_d^*(\mu),\mu)}{d\mu}-\frac{dw_n(h_n^*(\mu),\mu)}{d\mu}\right]-\frac{\beta}{2}\frac{dW_d(h_d^*(\mu),\mu)}{d\mu}-\frac{\beta}{2}\frac{dW_n(h_n^*(\mu),\mu)}{d\mu}=C'(\mu). \tag{30}$$

When k is large enough, the LHS of (30) is predominantly determined by the bracketed term. In addition, since we assume that the indirect effect through the change in $h_r^*(\cdot)$ is of 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 (30) takes on a negative value: the ISP has no incentive to invest. This implies that the constraint that the 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 entry does occur with prioritization, we can first show that the ISP never chooses a capacity level that allows MCP's entry without prioritization. Suppose to the contrary that the

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

ISP chooses $\underline{\mu}_n$. Then, comparing (27) and (29) reveals that the ISP's profit is higher when entry is not allowed in the absence of prioritization because $W_n(\phi, \mu) < W_n(h_n^*(\mu), \mu)$.[22] We break ties in favor of the ISP and assume that the MCP does not enter without prioritization given $\underline{\mu}_n$; e.g., 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.

We now examine the ISP's incentive to induce the MCP's entry with prioritization in a non-neutral network. We proceed as in the analysis of neutral networks. If the ISP does not induce entry, the ISP chooses $\min\{\mu_n^\phi, \underline{\mu}_d\}$. But if the ISP does induce entry, it chooses $\mu_d^E(\beta)$. The ISP induces entry if and only if

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

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

In the RHS of (31), the ISP internalizes half of the surplus created by entry, $\pi_d(h_d^*(\mu_d^E), \mu_d^E)$, while this term does not appear in the RHS of (25). Therefore, 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

$$-W_n(\phi, \min\{\mu_n^\phi, \underline{\mu}_d\}) - C(\min\{\mu_n^\phi, \underline{\mu}_d\}) \leq \pi_d(h_d^*(\mu_d^E), \mu_d^E) - W_d(h_d^*(\mu_d^E), \mu_d^E) - C(\mu_d^E), \tag{32}$$

To compare (31) with (32), the entry bias in the market can operate in both directions. The ISP can capture only half of the surplus created by entry, which induces insufficient entry. In contrast, the negative externality on NCPs' content arises due to the MCP's entry with priority in a non-neutral network, which is not fully internalized by the ISP, implies excessive entry. However, if we assume that $\pi_d(h_d^*(\mu_d^E), \mu_d^E)$ is sufficiently large compared to the effects on NCPs' content delivery time due to the 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 (31) holds but (32) does not.

We summarize our findings in the following proposition.

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

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

(i) (Incentives to induce entry of the MCP) Regardless of neutrality regulation, the ISP's incentive to induce entry of the MCP is suboptimally low. But, this problem is much more severe in a neutral network in which the ISP's incentive does not depend on (v, k) than in a non-neutral network in which the ISP partially internalizes the surplus created by the entry and thus its incentive to induce entry increases with v and k.

(ii) (Incentives to invest in capacity when the MCP enters in equilibrium)

- *Under neutral networks, 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. However, the anticipation of the MCP's QoS investment reduces the ISP's investment incentive.*
- *Under non-neutral networks, provided that the ISP induces entry of the MCP with prioritization, the ISP invests $\mu_d^E(\beta)$ where $\mu_d^E(\beta) \leq \underline{\mu}_n$ and induces the MCP not to enter without prioritization. The ISP's investment incentive is limited because the larger capacity decreases the surplus created by introducing the 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 larger "pipeline" means a lower value of the priority. 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 for mobile networks the neutrality regulation can be adverse to the entry of major content providers, whereas for fixed networks non-neutrality may reduce the ISP's incentive to invest in capacity. Therefore, we find that the extended model with the ISP's capacity investment provides consistent implications with those in our baseline model.

7 Extensions

7.1 Consumer Heterogeneity and Suboptimal Entry

In our baseline model, we have assumed homogeneous consumers and full surplus extraction by the MCP from its content delivery.[23] In such a setting, consumers always suffer from entry of the 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

[23] 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.

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 each consumer enjoys some positive surplus from the MCP's content delivery. To model such potential consumer benefits from the MCP's entry, let us introduce consumer valuation heterogeneity. The simplest way is to consider two types of consumers, H with proportion $\gamma \in (0,1)$ and L with $1-\gamma$.[24] The utility level that a type i consumer derives from the major CP's content in the network regime r is given by

$$u_i(w_r) = v_i - kw_r,$$

where $i \in \{H, L\}$ and $r \in \{n, d\}$ with $\delta = v_H - v_L > 0$.

Suppose that the MCP prefers to serve all consumers than to serve the high type only. This would be the case if v_L is sufficiently large compared to δ and/or γ is relatively small. In such a case, the L-type consumers once again always suffer from the MCP's entry because their surplus is fully extracted and the only effect from the entry is more congestion on existing content. The H-type consumers now receive a rent of δ from the new content, despite them suffering from the same negative externality for existing content. Then, the social welfare comparison in the non-neutral mobile 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 the MCP's profit from full surplus extraction.

Our previous analysis in Section 4 remains qualitatively intact when performed with π_d^{**} except that consumer heterogeneity can generate suboptimal entry. More specifically, prioritization is introduced only if

$$\Delta\Pi^m(\mu,\beta) \equiv \Pi_d(h_d^*(\mu),\mu,\beta) - \Pi_n(\phi,\mu,\beta) = \pi_d^*(\mu) - \gamma\delta - \beta\Delta W(\mu) > 0. \tag{33}$$

However, it is possible to have $\Delta\Pi^m(\mu,\beta)|_{\mu=\underline{\mu}_d(1)} < 0$. Precisely, 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 socially optimal entry may not take place. Therefore, when the MCP cannot extract the entire consumer surplus, the concern for socially excessive entry is mitigated and we may have insufficient entry.[25]

[24] 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.

[25] In the same spirit with consumer heterogeneity, suppose that a competition between MCPs plays a role of reducing the 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.

7.2 Discrete QoS in Congestion

For our main analysis, we have considered a continuous utility function in the congestion level; however, we also realize that utility may show some discontinuity over the quality of service for some real-world applications. In other words, depending on the content/application type, many users would perceive content delivery as a "failure" once the quality of service falls short of a certain level. 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 the 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 $w_n(h_n) = \frac{\lambda}{(\mu-1)(1+h_n)-\lambda} = w_o$:

$$h_n^*(\mu) = \frac{1+w_o}{w_o}\frac{\lambda}{\mu-1} - 1 \quad \text{for} \;\; \mu < \overline{\mu}_n. \tag{34}$$

In the non-neutral network, the 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 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

$$h_d^*(\mu) = \frac{1 + w_o}{w_o} \frac{\lambda}{\mu} - 1 \quad \text{for} \quad \mu < \overline{\mu}_d. \tag{35}$$

From the optimal QoS investments explicitly derived in (34) and (35), 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, the 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 the MCP's content has no effect on its *effective* waiting cost, but only increases the waiting cost for the NCPs' content, implying a negative traffic management effect.

8 Conclusion

Mobile network traffic has explosively grown in recent years. According to a report by Cisco,[26] global mobile data traffic grew 70 percent in 2012 alone, with mobile video traffic accounting for 51 percent of the total mobile traffic. These statistics imply that mobile operators have emerged as primary network access providers for many users, and a large portion of their usage involves high-bandwidth video content. Though regulatory agencies and market participants have agreed on these global trends and local network needs, regulatory agencies have taken different stances on how to address them. While the FCC imposes the critical rule of 'no unreasonable discrimination' only on fixed operators and gives exemption of such restrictions to mobile operators, the European Commission treats all types of networks in a uniform fashion under the principle of technological neutrality. Despite the FCC's controversial asymmetric regulation (Eisenach 2012, Maxwell and Brenner 2012), little rigorous analysis has been put forth on this aspect of net neutrality regulations.

[26] See "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017." The driving force behind this trend is widespread adoption of smart-phones. "In 2012, the typical smart-phone generated 50 times more mobile data traffic (342 MB per month) than the typical basic-feature cell phone (6.8 MB per month) of mobile data."(Id., p.2)

In this paper, we develop a theoretical model that characterizes the relative size of network capacity as a distinguishing feature between mobile networks and fixed networks, 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. In fixed networks, 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, arguably the most controversial regulatory agenda since the inception of the Internet, by providing a new perspective on major content providers' innovation incentives, which may have different implications across mobile and fixed networks.

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).

[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. 2013. "Net Neutrality, Business Models, and Internet Interconnection." forthcoming at *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] Eisenach, Jeffrey A. 2012. "Broadband Competition in the Internet Ecosystem." American Enterprise Institute Working Papers 35845.

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

[10] 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.

[11] 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.

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

[13] 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.

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

[15] 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).

[16] 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.

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

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

[19] 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.

[20] 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.

[21] 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.

[22] Peitz, Martin and Florian Schuett. 2014. "Net Neutrality and Inflation of Traffic." mimeo.

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

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

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

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

Appendix: Mathematical Proofs

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 (8) 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

$$w_d(h_d^*(\mu'),\mu') = w_d(h'',\mu''), \tag{36}$$

which is equivalent to

$$\frac{\lambda}{\mu'(1+h_d^*(\mu'))-\lambda} = \frac{\lambda}{\mu''(1+h'')-\lambda}. \tag{37}$$

Note that $\mu''>\mu'$ combined with (36) 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:

$$k\frac{\lambda\mu'}{\left[\mu'(1+h_d^*(\mu'))-\lambda\right]^2} = C'(h_d^*(\mu')). \tag{38}$$

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 (37), and the second one is from (38). 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

$$\frac{\partial a_d^*}{\partial \mu} < \frac{a_d^*}{\mu}. \tag{39}$$

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

$$\frac{\partial a_d^*}{\partial \mu} < \frac{1 + a_d^*\lambda}{\lambda\mu} = \frac{1}{\lambda\mu} + \frac{a_d^*}{\mu}. \tag{40}$$

Using (39), we show that inequality (40) 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 (39) 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

$$\begin{aligned}\frac{d\Delta\Pi^m(\mu,\beta)}{d\mu} &= \frac{d\pi_d^*(\mu)}{d\mu} - \beta\frac{d\left[W_d(h_d^*(\mu),\mu) - W_n(\phi,\mu)\right]}{d\mu} = k\left|\frac{\partial w_d}{\partial\mu}\right| - \beta\frac{dW_d}{d\mu} - \beta\left|\frac{\partial W_n}{\partial\mu}\right| \\ &> k\left|\frac{\partial w_d}{\partial\mu}\right| - \beta\left|\frac{\partial W_n}{\partial\mu}\right| = k\frac{a_d^*\lambda}{(\mu - a_d^*\lambda)^2} - \beta\frac{1}{(\mu-1)^2}, \end{aligned} \tag{41}$$

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

$$\overline{k}(\mu,\beta) = \frac{\beta}{a_d^*\lambda}\left(\frac{\mu - a_d^*\lambda}{\mu - 1}\right)^2, \tag{42}$$

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 (42) 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. ■