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

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

October 14, 2013

Abstract

We study how net neutrality regulations affect high-bandwidth content providers investment incentives in quality of services (QoS). We find that the effects crucially depend on network capacity levels. With a limited network capacity, the prioritized delivery services are complements to content providers' investments and can facilitate entry of high-bandwidth content. By contrast, if the network capacity is large enough, the prioritized delivery and QoS investment are substitutes. In either case, the social welfare effects of the prioritized service is ambiguous. In the limited capacity case, the beneficial effects of entry by high-band width content should be weighed against the cost of increasing congestion for other existing content. In the high capacity case, the negative impact of reduced investment incentives can be counterbalanced by the benefit of improved traffic management. Our findings have important implications for the contrasting neutrality regulations across the Atlantic: US FCC treats mobile networks more leniently than fixed networks, while the EU treats them equally.

JEL Codes: L1, L5, O3

Key Words: Net neutrality, asymmetric regulation, quality of service, investment incentives, queuing, congestion, mobile/fixed Networks

---

*We gratefully acknowledge the financial support from the NET Institute (www.NETinst.org), 2013 summer grant. We thank seminar audiences at Fall 2013 Midwest Economic Theory Conference and Georgia Institute of Technology for helpful comments.

†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

The U.S. Federal Communication Commission (FCC) adopted an "open Internet" order in 2010.[1] While one highlighted controversy surrounding this order has been whether the FCC has legitimate authority to impose any regulatory obligations over the Internet,[2] another altercation—often it appears to have been ignored—is that the order treated mobile network operators more leniently than fixed wire-line network operators. More specifically, its first two rules, namely, (i) 'transparency' and (ii) 'no blocking' are commonly applied to both network operators, but the third rule (iii) 'no unreasonable discrimination' appertains only to fixed line operators (emphasis is ours):

> 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)

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." More important, this asymmetric regulatory approach has made a sharp contrast to the European approach to the same issue: the European regulatory standards, 2002 EC Directives on electronic communications and its revisions in 2009,[3] have no such a distinction between fixed and mobile networks. More generally, the uniform treatment reflects one of the European regulatory principles, "technological neutrality," which allows no differential treatment in all kinds of networks such as cable networks, mobile networks and fixed wire-line networks. Nevertheless, to the best of our knowledge, we have little rigorous analysis on how we can understand

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

[2] While the FCC has legitimate regulatory authority over telecommunication services under Title II of the Communications Act regarding "common-carriers," the Internet access service is now categorized as *information services,* thereby considered as a non-common carrier; the FCC's powers are considerably limited in the information services governed by Title I of the Act. As a result, some Internet broadband access providers such as Verizon Communications and Metro PCS are in court to challenge the legality of the FCC's order.

[3] 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")

such a contrast in regulatory approaches across the Atlantic; our study attempts to fill this void. We study when the substantial capacity limit in the mobile network (compared to the fixed network) calls for the asymmetric regulation like in US and when the uniform treatment is justified.

Another important motivation for our study lies in the need to understand the effects of net neutrality regulation on innovation incentives at the "edges," which is essentially inseparable from plausible rationale for asymmetric regulation. Though the extant literature of network neutrality has focused on the expansion of ISPs' network capacity as innovation at the "core,"[4] the ISPs' capacity expansion is not the unique way to resolve the congestion problem in the modern Internet ecosystem. In fact, major content providers such as Google, Netflix, Amazon have explored to improve the quality of service for their content and applications independently—for a given network infrastructure. For example, they have pursued alternative technological solutions such as content distribution (or delivery) network (CDN) or advanced compression technology to ensure a sufficient quality of service, without asking for preferential treatment of their own content (Xiao, 2008).[5] Yet, researchers have little studied how these new technological changes relates to the regulation decision, though, as Maxwell and Brenner(2012) argue, it has become important for regulators and policy-makers to understand how the network regulations would affect the content providers' reaction through their investments in alternative technology solutions to ensure their quality of services, independent of the ISPs.[6]

Motivated by both the different regulatory approaches and new technological progresses 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. Consistent with the FCC's interpretation,

[4]Some have asserted rationale for asymmetric treatment is the notion that fixed broadband networks are uniquely at the "core" of the Internet while content, applications and devices are at the "edge." From this metaphor, we consider the innovation at the edge. See Reggiani and Valletti (2012) for more discussion on this.

[5]It is well known that the innovative video compression technologies have contributed to a 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 level of QoS for content providers.

[6]From an end-user's perspective, the fundamental goal is to enjoy highest quality of service at a minimum fee; the channel may not be their interest, either though ISPs' capacity investment or CPs' CDN investment. "CDN is to cache frequently accessed content in various geographical locations, and redirect access request of such content to the closer place. ... by moving content closer to end users, CDN can dramatically reduce delay, delay variation, and placket loss ratio for users' applications and thus their perception of network QoS" (Xiao, 2008 p.117).

we characterize the network neutrality regulation as no prioritization: with the regulation, the ISPs cannot allocate some traffic into a prioritized lane for charge. In this setting, we find that the effects of net neutrality regulation substantially depends on the relative size of the ISPs' network capacity over major content providers' bandwidth usage.

We show that the prioritization with a limited network capacity like a mobile network can facilitate the entry of major content providers' high-bandwidth content when the new content would not be available due to a too high cost of reaching the required QoS, without the paid prioritization. Obviously, the additionally available content would generate a new value to the network; indeed, this reasoning of entry facilitation reminds of the following two parts of the FCC Order from which we can infer its rationale for the different treatments between fixed networks and mobile networks:

> Mobile broadband is an earlier-stage platform than fixed broadband, and it is rapidly evolving. (FCC Order, par. 94)
>
> 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.95)

The US FCC appears to believe that its lenient non-neutral treatment helps innovative content and applications more available in the early-stage mobile network; however, we should be alert that the new content entry may not necessarily result in a higher overall welfare. This is because the new content will consume a substantial portion of the existing network capacity, thus it also increases the congestion for other content in the network (for a given network capacity). Intuitively, the positive surplus from new content cannot outweigh the efficiency loss from elevated congestion in other content if the network capacity is too small, as the negative externality of congestion becomes more pronounced in the smaller capacity.

Now consider a relatively high network capacity such that the entry of new content is no longer a focal issue: regardless of the prioritization service, the high-bandwidth content providers enter the network. Even in this case, we find another interesting trade-off generated by the prioritization in the non-neutral network. On the one hand, the prioritization results in a more efficient traffic management by assigning the faster delivery service to the more delay-sensitive content. Thus, the prioritization increases *static* efficiency; however, another externality problem must not go

unnoticed. First, the content provider's private incentive to enhance its own quality of service falls short of the social incentive as long as such a QoS investment generates a positive spillover toward other content's delivery by reducing its own bandwidth usage. This under-investment problem gets even worse if the content provider relies on the alternative means to reach its QoS goal through the prioritized delivery service. As a result, the prioritization yields a negative effect on the social welfare by weakening *dynamic* incentives for the QoS investment. Consistent with this insight, Xiao (2008) claims that major content providers have increased their pursuit of quality of service through technological solutions, not by the prioritization, after the FCC's intensive efforts to apply network neutrality regulations. No matter regulatory obligations originally intended, our model exhibits that the effects of net neutrality regulation on dynamic investment incentives at the edges must not be overlooked.

The remainder of our paper is organized as follows. In the rest of this section we discuss some of the related literature. In Section 2, we describe our model in which we introduce a more general queuing framework and its properties. In Section 3, we analyze the QoS investment decision by the major content providers for the neutral and non-neutral networks. By doing so, we show how mobile networks and fixed networks can be differentiated depending on the network capacity. Then, we provide our main analysis for mobile networks in Section 4 and for fixed networks in Section 5. Section 6 include several extensions of our model and discussion points. We close the paper in Section 7. We put mathematical proofs, not covered in the main text, in the Appendix.

1.1 Related Literature

This paper makes novel contributions to the literature on net neutrality. Most extant studies have at large focused on innovations at the core and/or innovations at the edges. The former concerns Internet access service providers' investment incentives on its "last mile" network capacity. This is never surprising because the proponents and opponents of the regulation collided head-to-head in whether the content providers' alleged free-riding would have a chilling effect on the ISPs' incentives to upgrade their "pipelines." As prior studies in this group, we can refer to 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. (2012). The latter category encompasses quite a few research works that have centered on the content providers' hold-up concern that may result in no entry or less investment on content, because high-value content providers may be expropriated ex post by Internet service providers through the paid prioritization as gatekeepers.

Studies along this avenue include Bandyopadhyay, Guo, and Cheng (2009), Choi and Kim (2010), Grafenhofer (2010), Reggiani and Valletti (2011), Bourreau, Kourandi, and Valletti (2012).

Beyond the investment incentives, economists have studied how the network neutrality would affect consumers and social welfare from various perspectives. Earlier, Hermalin and Katz (2007) analyzed the network neutrality in the context of a product-line restriction over the vertical differentiation model. Economides and Tåg (2012) regarded the neutrality regulation as a zero-pricing regulation on the content side in a two-sided market. More recently, we have seen interesting, new approaches to the debate. For example, Mialon and Banerjee (2013) consider content network platforms and study their effects on Internet access (or subscription) price on the consumer side and social welfare. 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. Another interesting study is found in Jullien and Sand-Zantman (2013) who examine the net neutrality issues from the perspective of information transmission such as signaling and screening.

Several empirical studies have also enriched the literature. For example, Nurski (2012) studied the concern about ISPs' foreclosure of the content in competing with their own content, using UK survey data on household content usage and the data of ISP choices. She finds a prioritized service would unlikely end up the foreclosure scenario. Lee and Hwang (2011) estimated the efficiency differences of Internet application groups using meta-frontier methodology, and they find that the ISPs' discrimination against content providers would not decrease the efficiency to a significant degree.

In fact, reflecting the burgeoning stage of the literature, several papers have attempted to survey extant studies. Lee and Wu (2009), Schuett (2010), Lee and Hwang (2011), and Krämer, Wiewiorra, and Weinhardt (2012) are among those.

Although proponents or opponents of the regulation have been polarized in their opinions, they all have kept one basic premise in common throughout the debate: End-users' quality of service must be the primary goal of a desirable network ecosystem. Xiao (2008), Altman et al. (2012), and Guo, Cheng, and Bandyopadhyay (2013) are several examples among many to explicitly make such a point. Sharing this spirit, to the best of our knowledge, we are first to reflect practically emerging technology solutions into the network neutrality debate to enhance the quality of service such as content delivery networks or compression technology. More important, we offer a first model that explicitly capture the differences between the fixed- and mobile networks and address the FCC's

asymmetric regulation between the two networks in contrast to the EU's uniform treatment.[7]

2 The Model

2.1 ISP, CPs, and Consumers

We consider a monopolistic broadband Internet service provider (ISP) who is in charge of the "last mile" delivery of online content to end-users.[8] Since we are primarily interested in major content providers' independent investments to improve the quality of service (for a given network capacity), we consider two types of content providers: one major content provider such as Netflix, Xfinity TV, or Amazon Instant Video and a continuum of other (non-major) content providers whose mass is normalized to one. This distinction allows us to focus our analysis on the major content provider's investment decision to improve the QoS for successful content business; the major content provider's relatively large scale of operation justifies the costly investment.

There is a continuum of homogeneous consumers whose mass is normalized to one. Each consumer demands both the major content provider's content and the non-major CPs'. When a consumer receives the major CP's content with an average waiting time of w, his or her utility is given by

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

; for the non-major CPs' content,

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

for an average waiting time of W. Parameters u and U represent the consumer's utilities from receiving the major CP's and non-major's CPs' (aggregate) content, respectively, without any congestion. Obviously, each consumer faces the decreasing utilities in delays of content delivery. 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 how much more sensitive each consumer is to a delay in the major CP's content relative to a delay in the non-major CPs' content. Finally, recall that the mass of consumers is normalized to one, so $u(w)$ and $U(W)$ represent the entire surplus from the major CP's content and non-major CPs' as well, respectively.

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

[8]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 a complete delivery of content with several interconnected network providers being involved along a transit route.

We assume that the major content provider can extract the entire surplus $u(w)$ in the absence of priority service under net neutrality, but it negotiates with the ISP over the price of the priority service in the non-neutral network. For the non-major CPs' content, we introduce a parameter $\beta \in [0,1]$ to denote the ISP's share of the total surplus $\beta U(W)$ generated by the non-major CPs' content delivery. In other words, the ISP receives $\beta U(W)$ from providing delivery services for the non-major CPs' content; the rest of the surplus, $(1-\beta)U(W)$, is shared among the content providers and end users. In this setting, the parameter β represents the ISP's ability to extract rent from non-major CPs and end users via connection fees; it also measures how much of the congestion externalities are internalized by the ISP. If $\beta = 0$, the ISP is not concerned with any potential effects on congestion in the non-major CPs' content traffic when it deals with the major CP. In contrast, if $\beta = 1$, the ISP fully internalizes the congestion effect on non-major CPs' content. As will be shown later, this parameter plays an important role in assessing the welfare effects of net neutrality of regulations. In particular, when β is small, the introduction of a paid prioritization to charge the major CP for"speed" may generate social inefficiency because the ISP ignores the negative externality inflicted on the non-major CPs' content delivery.

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. To pin down a micro-foundation for a network congestion, we adopt the standard M/M/1 queuing system which has been adopted as a good approximation to the congestion issues of real computer networks: e.g., Choi and Kim (2010), Cheong et al. (2011), Bourreau et al. (2012), Krämer and Wiewiorra (2012).

Let $\mu \in R_+$ denote the ISP's network capacity. Each consumer demands a wide range of content from both the major and non-major CPs. The content request rate follows a Poisson process, which represents the intensity of content demand. For the non-major CPs' content, we normalize the hazard rate of the Poisson distribution and the size of packets for each content to one. Since the mass of the non-major CP is one, the overall demand parameter (i.e., the total volume of traffic) for the non-major CP's content is also normalized to one. By contrast, we envision a major CP as a discrete player operating a content network platform that provides a continuum of content whose mass is ξ and the packet size for each content is $m(>1)$ with $\xi \cdot m = \lambda$. Then, we can interpret λ as the relative size of packets that constitute the major CP's content or the sheer volume of content available from the major CP. With either interpretation, it measures the relative

traffic volume of the major CP's content vis-a-vis the non-major CPs' aggregate traffic volume. The total traffic volume for the ISP amounts to $1+\lambda$, which asks for $\mu > 1+\lambda$ not to make the waiting time reach infinity.

We consider the major content provider who can make an investment of $h \geq 0$ to enhance the quality of service in its content delivery. As we mentioned in Introduction, the investment can take various forms such as a compression technology to reduce packet-size to deliver the same content or content delivery networks to shorten the delivery distance by putting content servers in local data centers so that end-users' demand are served by the closest data center.[9] Because one common, main effect of such investments is to speed up the content delivery, we can model it simply as a "compression technology" that reduces the traffic volume of the major CP's content from λ to $\alpha\lambda$, where $\alpha = \frac{1}{1+h} \in (0,1]$. That is, the greater investment leads to the smaller packet size for the major CP's content; therefore, its delivery speed increases even without the ISP's capacity expansion. No investment ($h = 0$) corresponds to $\alpha = 1$, with the traffic volume remains at the level of λ. We assume that the investment cost increases 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 $h > 0$.

There are two network regimes: neutral network and non-neutral network. Consistent with the literature and regulatory obligations, we define the neutrality regulation as no paid prioritization: all traffics are equally (every packet is served according to the *best-effort* principle) treated without classes of service. In such a neutral network, each user in M/M/1 queuing system faces the following total waiting cost:

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

for the major CP's content. The total volume of traffic (packet size) amounts to $1+\alpha\lambda$ (1 for the non-major CPs' content and $\alpha\lambda$ for the major CP's content with compression), and thus the average waiting time per packet is given by $\frac{1}{\mu-(1+\alpha\lambda)}$. Thus, for the total packet size of $\alpha\lambda$, the

[9] According to Xiao (2008), there are at large three different types of delays that account for a total delay from a one end of the network to an other end: (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 the caching or content delivery network (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.

total waiting cost is computed as (3).[10] Similarly, for the non-major CP's content, we can derive the total waiting time of

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

Without the neutrality obligations, the ISP is allowed to adopt a paid prioritization in which the major CP can buy the premium service at some price to send its content ahead of the non-major CPs' packets in queue so that the waiting time for the prioritized packet is given by

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

Remarkably, such a gain in speed is realized by making other non-prioritized content getting slow down not only relative to the prioritized content but also relative to the neutral network speed. Precisely, the waiting time for the "basic" service in the non-neutral network is

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

In what follows, we often use interchangeably $w_r(h, \mu) = w_r(\alpha(h), \mu)$ and $W_r(h, \mu) = W_r(\alpha(h), \mu)$ for $r = n, d$ if the notations make no confusion.

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 microfoundation for our analysis.

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

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

This result is intuitive because one content's less use of bandwidth means more network capacity for other content in a fixed network capacity.

Property 2 For a given network capacity and a major content provider's investment, the waiting time for the prioritized lane is shorter than that under the neutral network and the waiting time for the non-major content in the non-neutral network is longer than that in the neutral

[10] Notice that with no investment investment in the compression technology ($h = 0 \Leftrightarrow \alpha = 1$), the average waiting time reduces to the formula of $\frac{1}{\mu - (1+\lambda)}$ as in the standard standard M/M/1 queuing system.

network: i.e.,

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

Property 3 The total waiting time is equal regardless of the network regimes: i.e.,

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

This result is an extended version of the waiting cost equivalence characterized in Choi and Kim (2010), Bourreau et al. (2012), Kramer and Wiewiorra (2012) into a more generalized queueing system that allows for the content provider's investment and its spillover.

2.4 Decision and Bargaining Timing

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

1. Given an ISP's network capacity μ, the major CP makes a decision on whether to enter the market. If it enters, it chooses its investment level h.
2. Given the network infrastructure (μ, h), content is delivered to consumers and the payoffs are accordingly realized.

One notable difference arises in the non-neutral network, for the major CP and the ISP need to bargain over a price of the prioritized service.

1. Given an ISP's network capacity μ, the CP and the ISP bargain over the price of the prioritized service.
2. With an agreement on the price of the prioritized service, the major content provider makes its entry and investment decisions taking the prioritized service into account. Without a mutual agreement on the price of the prioritized service, the situation becomes the same as in the neutral regime: all traffics are delivered uniformly by the best effort principle. The major content provider's entry and investment decision remain the same as in the neutral regime.
3. Given the network infrastructure and prioritization, content is delivered to consumers and the payoffs are realized.

Notably, we assume here that the investment by the major CP is not contractible, but it is solely determined by the content provider.

3 Optimal QoS Investment and Network Regimes

3.1 Neutral Networks

Let us consider a neutral network in which all packets are equally treated based on the first-come-first-served principle. We first analyze the major CP's optimal investment decision assuming that the major CP enters the market. We then consider the major CP's entry decision with the possibility of a corner solution, i.e., no entry without any investment. Assuming entry, the content provider's optimal choice of h in the neutral network is to maximize its profit:

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

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

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

for an interior solution h_n^*. The marginal benefit of the investment decreases in the capacity, which is easily confirmed by the cross-partial derivative:

$$\frac{\partial^2 \pi_n}{\partial \mu \partial h} = k\lambda \frac{-\lambda-(\mu-1)(1+h)}{[(\mu-1)(1+h)-\lambda]^3} < 0. \tag{8}$$

Let $\pi_n^*(\mu) \equiv \pi_n(h_n^*(\mu),\mu)$ denote the maximized CP profit at the optimal investment level $h_n^*(\mu)$ for a given network capacity μ. By the Envelope Theorem, we find that the major content provider obtains the higher profit for the larger network capacity:

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

This relationship tells 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 major CP makes an investment only when the ISP's capacity is above this threshold level; the investment cost is too high to justify entry into the content service market for $\mu < \underline{\mu}_n$. Hence, there will be a discontinuity in the major CP'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 can see how the (interior) optimal investment h_n^* changes over the capacity level for $\mu > \underline{\mu}_n$. For this comparative statics, let us define an implicit function $G(h_n^*;\mu,k,\lambda) \equiv \frac{k\lambda(\mu-1)}{[(\mu-1)(1+h_n^*)-\lambda]^2} - c'(h_n^*) = 0$ from (7). Then, we can apply the Implicit Function Theorem as

follows:

$$\frac{\partial h_n^*}{\partial \mu} = -\frac{\partial G/\partial \mu}{\partial G/\partial h_n^*}$$

where the denominatior has the negative sign as is seen from

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

and thus the overall sign is determined by the sign of the numerator. From the partial derivative of G with respect to μ, we find

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

Thus, we can establish the following result as Lemma 1:

Lemma 1 *The investment for quality of service decreases in network capacity* (μ)*, i.e.,* $\frac{\partial h_n^*}{\partial \mu} < 0$ *for* $\mu \geq \underline{\mu}_n$.

In summary, 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 $\mu \geq \underline{\mu}_n$.

3.2 Non-neutral Networks

Now let us consider the non-neutral network in which the major content provider has an option to buy the prioritized delivery service at a negotiated price. One benefit of such an arrangement is that the content provider may lower its investment for the same quality of service owing to a preferential treatment of its content delivery. The analysis for the non-neutral network is not much different from that of the neutral network. We start to define the major content provider's profit before any payout for the priority as

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

where $w_d(h) = \frac{\lambda}{\mu(1+h)-\lambda}$. Then, the content provider's optimal investment choice of h is to maximize its profit net of the prioritization price p:

$$\max_{h \geq 0} \pi_d - p.$$

The first order condition yields the following equation:

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

for an interior solution h_d^*. 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 capacity, i.e.,

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

and the investment decreases in the capacity, $\frac{\partial h_d^*}{\partial \mu} < 0$.[11]

Though both h_n^* and h_d^* are independent of β, one notable difference in the non-neutral network is that the price of prioritization will be affected by β. This is because the paid prioritization will make the ISP earn less from non-major content providers due to their higher congestion in the basic lane, for which the major content provider needs to compensate via the higher priority price. Note that $\beta = 0$ results in the lowest capacity level that induces the entry (and investment) of the major content provider. The reason is following: the ISP has no opportunity cost of selling prioritization for $\beta = 0$ that the ISP and the major content provider will agree on some price of prioritization whenever $\pi_d^*(\mu) > 0$ if $\beta = 0$; the joint surplus generated by introducing the prioritization becomes smaller for $\beta > 0$ compared to when $\beta = 0$. Therefore, we here consider the extreme case of $\beta = 0$ and relegate the analysis of $\beta > 0$ to the next section.[12]

Back to the major content provider's optimal QoS investment choice, we know that for a given β, its profit $\pi_d^*(\mu)$ strictly increases with μ as in the neutral network. So, there is a critical level of capacity $\underline{\mu}_d$ such that $\pi_d^*(\mu) \geq 0$ if and only if $\mu \geq \underline{\mu}_d$. As in the neutral network, the major CP's investment jumps up at the threshold $\underline{\mu}_d$, then decreases over μ for $\mu > \underline{\mu}_d$. Because $\pi_d^*(\mu) > \pi_n^*(\mu)$ and $\pi_d^*(\mu)$ increases in μ, we must have $\underline{\mu}_n > \underline{\mu}_d$. Finally, to compare h_n^* and h_d^*, we verify that

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

which leads to the following lemma:

Lemma 2 *As long as the major CP makes positive QoS investment, prioritization reduces the investment, i.e.,* $h_n^*(\mu) > h_d^*(\mu)$.

[11]The proof is omitted as it is similar to the process leading to Lemma 1 in Section 3.1.

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

Summing up thus far analyses, we can categorize the investment decisions by the major content provider, depending on the size of network capacity, as follows in words and an illustration in Figure 1:

Proposition 1 *Consider $\beta = 0$ in the non-neutral network.*

(i) For a sufficiently large capacity $\mu > \underline{\mu}_n$, a prioritization and the major CP's investment are "substitutes" in that purchasing prioritization reduces the major CP's QoS investment.

(ii) For a small capacity $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$, a prioritization and the major CP's investment are "complements" in that the major CP enters and makes positive QoS investment in the non-neutral network though there is no entry in the neutral network.

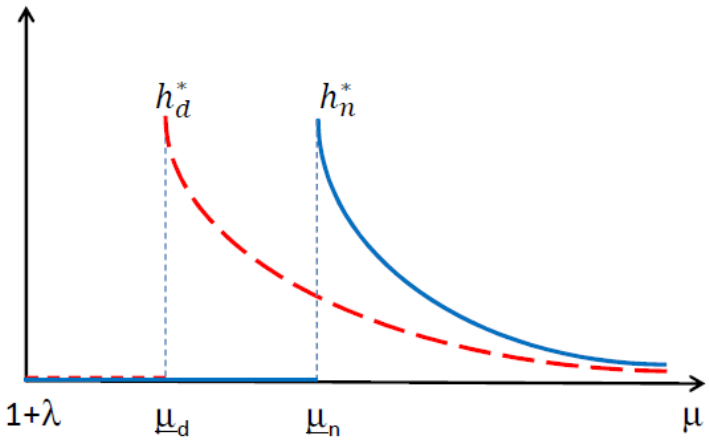

Figure 1: Investment by Content Providers

Based on thus far analysis for both networks, now we are ready to link the content provider's entry and investment decision to the FCC's reasoning for the distinction between fixed and mobile operators. Given the FCC's statement, "Mobile broadband speeds, capacity, and penetration are typically much lower than for fixed broadband," it is natural to consider that the smaller capacity is aligned to the mobile network, specifically to the case of $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$; the case of $\mu > \underline{\mu}_n$ to the fixed network. For the mobile network, there is a higher concern for the entry of high-bandwidth content and complex applications. By contrast, even high-bandwidth content is more likely to prevail when capacity is large regardless of the neutrality obligations.

4 Mobile Networks

For the mobile network such that $\mu \in (\underline{\mu}_d, \underline{\mu}_n)$, recall that the major content provider makes no entry under the neutral network because of $\pi_n^*(\mu < \underline{\mu}_n) < 0$, whereas it enters with a paid prioritization in the non-neutral network, at least for $\beta = 0$. The entry has two effects. One the one hand, the new content generates the surplus $\pi_d^*(\mu)(> 0)$, which the content provider and the ISP shares according to bargaining powers. On the other hand, the new content's entry increases the congestion against the non-major content on the following two channels: (1) one content's prioritized delivery means slow delivery for other content in the basic lane and (2) additional bandwidth taken by the new content means the more congestion for a given network capacity. Precisely, the difference in waiting cost for the non-major CPs' content, ΔW, can be decomposed into the following two parts:

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

where ϕ stands for 'no entry' of the major content provider. The first part captures the non-major content's waiting time increase due to the prioritization for a given QoS investment h_d. The second part measures the increasing delivery time even in the absence of prioritization just because now the major content provider's content occupies some bandwidth in the non-neutral network—the whole bandwidth would have been occupied by the non-major CPs' content in the neutral network. On both accounts, we must have $\Delta W > 0$, which is mathematically clear because, for any $\alpha_d^* > 0$, we have

$$\Delta W = \frac{\alpha_d^* \lambda \left(2\mu - \alpha_d^* \lambda - 1\right)}{\left[\mu - (1 + \alpha_d^* \lambda)\right] \left(\mu - \alpha_d^* \lambda\right) (\mu - 1)} > 0.$$

Given this finding, let us further examine how the prioritization would affect the overall welfare. We define the joint profit of the ISP and the major content provider as follows: for the neutral network, the joint profit will be given by $\Pi_n(\phi, \mu, \beta) = \beta[v - W_n(\phi, \mu)]$; for the non-neutral network, $\Pi_d(h, \mu, \beta) = \pi_d(h, \mu) + \beta[v - W_d(h, \mu)]$. Then, we can compute the change in the joint profit by introducing the prioritization as follows:[13]

$$\begin{aligned} \Delta\Pi^m(\mu, \beta) &= \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(\phi, \mu, \beta) \\ &= \underbrace{\pi_d^*(\mu)}_{(+) \text{ major CP's surplus from entry}} \quad \underbrace{-\beta \Delta W(\mu)}_{(-) \text{ more congestion for non-major content}}, \end{aligned}$$

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

where the superscript m in $\Delta\Pi^m$ stands for the *mobile* network. Recognizing that the welfare crucially depends on the network capacity (μ), we need to examine how $\Delta\Pi^m(\mu,\beta)$ changes over μ for a given β; note that $\Delta\Pi^m(\mu,1)$ captures the change in social welfare from the major CP's entry under non-neutral networks. Intuitively, the capacity affects the network congestion in various manners. First, the higher capacity has a direct and positive effect on the quality of service for the prioritized content. However, because the major content provider would reduce its investment for the quality of service in a larger network capacity, there is also an indirect and negative effect. In math, the total derivative of w_d with respect to μ shows these subtle interplays as follows:

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

Then, we establish Lemma 3(i) that the positive direct-effect dominates the negative indirect-effect by showing $\left|\frac{\partial h_d^*}{\partial \mu}\right|$ turning out not that large; similarly, for the non-major content delivery as in Lemma 3(ii).

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 proof in the Appendix.* ■

The welfare gain through reduced congestion in the major content becomes larger for the higher k as the users' benefit from the lower congestion is proportional to k. In contrast to this gain side, the larger capacity reduces the negative externality of the major CP's entry onto the non-major content. More precisely, a capacity increase reduces W_n, without increasing W_d (Lemma 3(ii)). This means that the upper-bound of the loss is the reduced congestion in W_n, which is proportional to β from the ISP and the major content provider's viewpoint. Hence, if k is large enough, the gain would exceed the loss. Using Lemma 3, we prove that $\Delta\Pi^m(\mu,\beta)$ strictly increases in μ if k is large enough:

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

Proof. *See the proof in the Appendix.* ■

Indeed, if β is small enough, the ISP and the major CP are not much affected adversely by the prioritization; the major CP's subsequent entry may occur even for the case that k is close to 1.

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

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

Interestingly, Lemma 5 tells us that any entry under $\mu < \underline{\mu}_d(1)$ is socially harmful. Such an inefficient entry arises because the ISP and the major CP's coalition does not fully internalize the negative externality of increased congestion onto the non-major content.

Proposition 2 (Mobile Networks) *Consider the mobile network with a limited network capacity of $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$. Then, we find*

(i) Under neutral networks, the major content provider's congestion-sentitive content is not served due to lack of the quality of service, $\pi_n^*(\mu) < 0$.

(ii) Under non-neutral networks, given β, the two-tiered service with traffic prioritization makes the congestion-sensitive content served as long as $\mu \geq \underline{\mu}_d(\beta)$ where $\underline{\mu}_d(\beta)$ strictly increases with β and $\underline{\mu}_d(0) = \underline{\mu}_d$.

(iii) If $\underline{\mu}_d(1) < \underline{\mu}_n$, then the congestion-sentitive content enters the network due to the prioritization; the entry is not socially desirable for $\mu \in (\underline{\mu}_d(\beta), \underline{\mu}_d(1))$, but it is socially efficient for $\mu \in (\underline{\mu}_d(1), \underline{\mu}_n)$. By contrast, if $\underline{\mu}_n < \underline{\mu}_d(1)$, there is no socially efficient entry due to the prioritization.

Up to now, we have analyzed the mobile network in which a limited network capacity can make a high-bandwidth content unavailable in the neutral network, but such a barrier-to-entry becomes loose through a paid prioritization in non-neutral networks. Though our analysis can stand by itself, here we offer a complementary numerical example for expositional purpose. We set the values of parameters: $k = 2$, $\lambda = 2$, $u = 5$, $U = 3$, $F = 2$, $\beta = 1$, and $\mu \in [2.5, 3.0]$. The cost function is

Table 1: Private Incentives to Entry and Social Welfare in Mobile Networks

μ	h_n^e	h_d^*	π_n^e	π_d^*	ΔW^*	$\Delta\Pi^m(\mu,1)$
2.5	1.196	0.683	-1.522	0.772	5.442	-4.721
2.6	1.115	0.654	-1.134	0.834	4.153	-3.320
2.7	1.043	0.625	-0.803	0.934	3.324	-2.390
2.8	0.980	0.597	-0.518	1.025	2.751	-1.726
2.9	0.924	0.568	-0.270	1.107	2.335	-1.228
3.0	0.874	0.539	-0.053	1.181	2.020	-0.840

set as $c(h) = h^2$. The parameter values are set to describe a situation that the network capacity is relatively small to overall traffic volume so that the capacity affects congestion costs substantially.

Columns 2 and 3 in Table 1 record the major content provider's QoS investments. Note that we use the superscript 'e' for the QoS investment h_n^e in the neutral network because it is expected ex post payoff upon entry. As Column 4 clearly shows, the major content provider will not enter the network because the QoS investment is too costly in the neutral network; by contrast, the entry is ensured ($\pi_d^* > 0$ in Column 5). We also computed the waiting time difference due to the major content provider's entry in ΔW^*. The last column, $\Delta\Pi^m(\mu, 1)$ measures the total welfare change driven by the major CP's entry; for this specific numerical example, it shows that the entry is inefficient (but we would have other results for different sets of parameters).

5 Fixed Networks

In this section, we consider the 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 private, joint incentives to introduce the prioritization service. Their joint payoffs in the network regime $r = n, d$ are 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

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

which means a higher joint payoff under the non-neutral network. Interestingly and importantly, 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.

$$\Delta\Pi^f(\mu,\beta) = \underbrace{\Pi_d(h_d^*(\mu),\mu,\beta) - \Pi_n(h_d^*(\mu),\mu,\beta)}_{\text{Efficient Traffic Management Effect } (+)} + \underbrace{\Pi_n(h_d^*(\mu),\mu,\beta) - \Pi_n(h_n^*(\mu),\mu,\beta)}_{\text{QoS Investment Effect } (-)} \tag{10}$$

where the superscript f stands for the *fixed* network.

The first term in (10) is always positive and represents the *static effect from efficient traffic management* (TM): 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 can show that

$$\begin{aligned} TM &= \Pi_d(h_d^*(\mu),\mu,\beta) - \Pi_n(h_d^*(\mu),\mu,\beta) \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] + \beta\left[W_n(h_d^*(\mu)) - W_d(h_d^*(\mu))\right] \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] - \beta\left[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))\right] \\ &= (k-\beta)[w_n(h_d^*) - w_d(h_d^*)] > 0 \end{aligned} \tag{11}$$

where the third equality in (11) is obtained from Property 3 $w_n(h_d^*) + W_n(h_d^*) = w_d(h_d^*) + W_d(h_d^*)$.

The second term in (10) represents the *effect from reduced QoS investment* (IE)*:* the availability of the prioritized lane decreases the major CP's investment incentives from $h_n^*(\mu)$ to $h_d^*(\mu)$, which affects the joint payoff. To verify the sign of this term 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.,

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

The privately optimal choice by the major CP, $h_n^*(\mu)$, maximizes $\pi_n(h,\mu)$ and does not take into account the positive effect of its investment on $\beta W_n(h)$. This implies there is under-investment from the perspective of joint profit maximization, i.e., $h_n^*(\mu) < h_n^J(\mu,\beta)$ (unless $\beta = 0$). The objective function $[kw_n(h) + \beta W_n(h)] + c(h)$ is convex as each of $w_n(h)$, $W_n(h)$ and $c(h)$ is convex in h. Because we have $h_d^*(\mu) < h_n^*(\mu)$, this implies that the QoS investment effect must be negative:

$$IE = \Pi_n(h_d^*(\mu),\mu,\beta) - \Pi_n(h_n^*(\mu),\mu,\beta) < 0.$$

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 major CP. Note that social welfare in each regime coincides with the joint payoff of the ISP and the major CP when $\beta = 1$. It thus can be written as

$$S_r(\mu) = \Pi_r(h_r^*(\mu), \mu, \beta = 1) = \pi_r(h_r^*(\mu), \mu) + [V - W_r(h, \mu)],$$

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

$$\begin{aligned} \Delta S(\mu) &= S_d(\mu) - S_n(\mu) \\ &= \Delta\Pi^f(\mu, \beta) + (1 - \beta)\left[W_n(h_n^*) - W_d(h_d^*)\right] \end{aligned} \tag{12}$$

When $\beta = 1$, the private incentive to use the prioritization service is perfectly aligned with the social incentive with the ISP fully internalizing any effect of providing the prioritized service on end consumers and non-major CPs ($\Delta S(\mu) = \Delta\Pi^f(\mu, 1)$). 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 non-major CPs' content due to the prioritized service. More precisely, we have

$$\Delta S(\mu) - \Delta\Pi^f(\mu, \beta) = (1 - \beta) \underbrace{\left[W_n(h_n^*) - W_d(h_d^*)\right]}_{\text{externality on existing content } (-)} \tag{13}$$

The externality term in (13) can be decomposed as follows.

$$W_n(h_n^*) - W_d(h_d^*) = [W_n(h_n^*) - W_n(h_d^*)] + [W_n(h_d^*) - W_d(h_d^*)] < 0$$

The expression in the first square bracket is negative because $h_n^* > h_d^*$ (Lemma 2). The expression in the second square bracket is also negative by Property 2. Therefore, the externality term is negative.

The discrepancy between the social incentives and the private incentives is inversely related to β. When $\beta = 0$, the discrepancy reaches its maximum; the ISP and the major CP will always find it profitable to adopt the prioritization. For instance, suppose $k = 1$ so that there is no social gain from the efficient traffic management. Even so, they will introduce the prioritized service at the expense of the end consumers and non-major content providers though the neutrality regulation

gives a higher social welfare. To see this, we can check out the following:

$$\begin{aligned} \Delta\Pi^f(\mu, \beta = 0) &= \pi_d(h_d^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &\geq \pi_d(h_n^*(\mu), \mu) - \pi_n(h_n^*(\mu), \mu) \\ &= w_n(h_n^*(\mu), \mu) - w_d(h_n^*(\mu), \mu) > 0, \end{aligned} \tag{14}$$

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 networks as follows.

Proposition 3 (Fixed network) *Consider the fixed network with a sufficient network capacity $\mu > \underline{\mu}_n$ that the major content provider's entry is always ensured. Then, we find*

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

(ii) For any $\beta < 1$, there is socially excessive incentive to adopt the prioritization service as the ISP and the major CP do not fully internalize the effect of increased delay on non-major CPs' content due to the prioritized service .

As before, we wrap up this section with a simple numerical example that clearly shows the traffic management effect and the QoS investment effect in Table 2. We set the parameters such that $\mu = 3$, $\lambda = 2$, $u = 5$, and $U = 3$ with $k \in \{1, 2, 3\}$.

Table 2: Fixed Networks: TM vs. IE

k	h_n^*	h_d^*	TM	IE	ΔS
1	0.693	0.406	0.000	−1.213	−1.213
2	0.874	0.559	1.043	−1.162	−0.120
3	1.000	0.667	1.667	−1.194	0.472

For $k = 1$ where there is no efficiency gain through traffic management due to the equal per unit congestion cost. But, regardless of the value of k, it is clear that the non-neutral network suffers more from the under-investment problem compared to the neutral network. As k increases, the traffic management effect increases because of the larger gain from assigning the prioritized lane to more congestion-sensitive content. For $k = 3$, such gain exceeds the cost of welfare loss from a suboptimal investment; the social welfare ΔS is higher in the non-neutral network.

6 Discussion

6.1 Consumer Surplus

In our main analysis, we have considered homogeneous consumers and the major content provider's full rent extraction for its content.[14] In such a setting, consumers always suffer from the entry of major content provider's high-bandwidth content due to the negative externality into existing content, unless the ISP is able to extract all consumer surplus from non-major CPs' content (i.e., $\beta = 1$). Of course, this simplification is innocuous to the results that we have derived, as long as we focus on the social welfare. To the contrary, if one wanted to say about consumer welfare analysis and the ISP's investment, we should extend our model into a direction that each consumer enjoys some positive surplus from the major content provider's successful content delivery.

To model potential consumers' benefit from the major content, let us introduce consumer heterogeneity in their valuation: the simplest way is to consider two types of consumers, H with proportion $\theta \in (0,1)$ and L with $1-\theta$. Then, the utility level type i consumer derives from the major CP's content in the network regime r is given by

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

where $i \in \{H, L\}$ and $r \in \{n, d\}$ with $\Delta = u_H - u_L > 0$.

Suppose that the major content provider prefers to serve all consumers than to serve the high type only.[15] This would be the case if u_L is sufficiently large compare to Δ or θ is relatively small. In such a case, the low type consumers once again always suffer from the major CP's entry because their surplus is fully extracted and the only effect from the entry is more congestion on existing content. The high type consumers now receive a rent of Δ from the new content, despite they suffer from the same negative externality for the existing content. Then, the social welfare comparison in the non-neutral mobile network will be determined by the trade-off between $\pi_d^* + \theta\Delta$ and $\Delta W = [W_d(h_d^*(\mu);\mu) - W_n(\phi;\mu)]$. The difference from our analysis in Section 4 is that now high type consumers receive a positive rent. Regarding $\pi_d^* + \theta\Delta$ as a new level of profit π_d^{**}, our qualitative results do not change.

[14] 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.

[15] Choi, Jeon and Kim (2013) study an ISP's second-degree price discrimination against two types of content providers and show how the neutrality regulation affects the ISP's incentive to exclude a low type of content provider. Here the major content provider—not the ISP—is discriminating against consumers.

6.2 Discrete QoS in Congestion

We have considered a continuous utility function in its congestion level; in real-world applications, the utility may show some discontinuity over the quality of service. In other words, depending on content/application types, many users would perceive a content delivery as "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 the service when he or she finds the content delivery unsatisfactory due to a frequent buffering or a blurry screen. A user would not value a Voice over Internet Protocol (VoIP) when call drops are too often or the call quality is poorer below a certain level. Of course, users may not always gain increasingly from the higher quality of service: in most cases, users may be indifferent over the delivery speed oncee it is above some required level. In this spirit, we can consider the utility function as the following step function:

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

Under this kind of discrete quality of service, one can easily derive explicit solutions for QoS investments in both network regimes. In the neutral network with a sufficiently large capacity μ, there will be no need for any investment from the major content provider 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:

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

For any capacity $\mu > \overline{\mu}_n$, there is no investment in the neutral network. The major CP's optimal investment to ensure the required QoS, denoted by h_n, is derived from $w_n(h_n) = \frac{\lambda}{(\mu-1)(1+h_n)-\lambda} = w_o$:

$$h_n^*(\mu) = \frac{1+w_o}{w_o}\frac{\lambda}{\mu-1} - 1 \quad \text{for} \;\; \mu < \overline{\mu}_n. \tag{15}$$

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

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

So, there will be two cases depending on the range of network capacity. For $\overline{\mu}_d < \mu < \overline{\mu}_n$, 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

$$h_d^*(\mu) = \frac{1+w_o}{w_o}\frac{\lambda}{\mu} - 1 \quad \text{for} \quad \mu < \overline{\mu}_d. \tag{16}$$

From the optimal QoS investment derived in (15) and (16), we can confirm the results in Proposition 1-3 with explicit solutions.

Two differences are noteworthy when we use this specification of a discrete quality of service. First, the purchase of the prioritized delivery class can be a complete substitute for the QoS investment when $\mu > \overline{\mu}_d$. That is, the major CP needs no investment with the prioritized delivery service; this is not the case for a continous utility function in quality of service. Second, now that the ex post waiting cost for the major CP's content is given as a constant w_o regardless of network regimes, the efficient traffic management effect cannot be captured. In this aspect, the continuous utility function (1) is better to capture the advantage of a different treatment of content delivery based on its congestion sensitivity.

6.3 First-Best vs. Second-Best

One interesting issue arises when we think of the first-best investment that is determined by the social planner who can choose both the QoS investment level and the network regime. Let us compare this outcome with the second-best outcome that the social planner can only choose the network regime, either neutral or non-neutral network, which is studied in Section 4. Then, it would be interesting to see whether there is such a case that the first-best discriminatory network is better than the first-best neutral network, while the second-best discriminatory network is worse than the second-best neutral network. If so, it would imply that the inability of ordering the private CP to choose the socially optimal QoS makes the neutral network yield a better social outcome.

Let us denote the socially desirable QoS investment level, h_n^o, in the neutral network; similarly h_d^o for the non-neutral network. For discussion's sake, let us continue the same numerical example for Table 2 ($\mu = 3, \lambda = 2$, $u = 5$, $U = 3$, and $F = 0$ with $k \in \{1, 2, 3\}$); Table 3 show the contrast between the first-best outcome and the second-best outcome.

The comparison among the optimal QoS investments h_n^*, h_d^*, and h_n^o clearly shows that the under-investment problem occurs in both network regimes; it is more severe in the non-neutral network where the major CP reduces its investment because the quality of service can be enhanced

Table 3: First-Best vs. Second-Best

k	h_n^*	h_d^*	h_n^o	h_d^o	$S_d^* - S_n^*$	$S_d^{FB} - S_n^{FB}$
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

through the prioritization. Let us start considering the case of symmetric waiting cost, i.e., $k = 1$: then, the neutral network performs better in terms of social welfare because of the less severe under-investment problem in the neutral network and zero efficiency gain from traffic management by prioritization. For the first-best outcome, h_n^o and h_d^o will be the same if $k = 1$ because of the total waiting cost equivalence (Property 3), which means the QoS investment effect is also zero. The social planner will be indifferent between the two network regimes for $k = 1$.

However, as the numerical example shows, the non-neutral network outperforms the neutral network for $k = 2$: now the static efficiency gain is earned in the discriminatory treatment of congestion-sensitive content and the social planner is not suffering from a suboptimal QoS investment choice in either network. Hence, we can see the conflicts of interest in that the social planner would prefer the discriminatory regime in the first-best hypothetical situation whereas she will prefer the neutral network when she must let the private ISP choose the investment to warrant its desirable QoS level. Interestingly, if k is sufficiently large (here, $k = 3$), there will be no further such conflict. Regardless of the inability of ordering the private CP to choose the socially optimal QoS investment levels, the non-neutral network will yield a higher social welfare simply because of a sufficiently large traffic management effect. Our discussion suggests that as the regulator is unable to dictate the socially desirable level of QoS in the real world the neutrality regulation can be preferred by the regulator.

6.4 ISP's Capacity Choice

We have performed our analysis for an exogenously given capacity; now let us think of how the possibility of the major CP's QoS investment can affect the ISP's incentive to invest in capacity. For this purpose, we insert a new stage before the major CP makes its investment decision—the ISP unilaterally chooses its capacity μ. For the neutral network this means the following sequence of choices:

1. The ISP chooses its network capacity μ.

2. Given the ISP's network capacity μ, the major CP decides its investment level h.

3. Given the network infrastructure (μ, h), content is delivered to consumers and the payoffs are accordingly realized.

Similarly, in the non-neutral network, the ISP chooses its capacity before the ISP and the CP bargain over the price of prioritization.

The extant literature has studied how the net neutrality regulation affects the ISP's capacity choice in the absence of the major CP's investment. For instance, Choi and Kim (2010) find that non-neutrality may reduce the ISP's incentive to invest in capacity because a large capacity reduces the price of prioritization. Instead of the cross-regime comparison, here we analyze how the anticipated reaction of the major CP in terms of its investment choice affects the ISP's incentive to invest in capacity within a given regime.

It turns out that the result crucially depends on whether the benchmark capacity chosen in the absence of the major CP's investment is large enough to induce the entry of the major CP. Given a regulatory regime, if the capacity choice in the absence of the CP's investment is large enough to allow the major CP to enter (as in the fixed network), the possibility for the CP to make QoS investment reduces the ISP's incentive to invest in capacity. By contrast, if the benchmark capacity is too small to induce the entry of the major CP (as in the mobile network), the possibility of CP's investment is likely to increase the ISP's incentive to invest in capacity. These effects are qualitatively the same regardless of whether neutrality regulation is imposed. So, below we try to disuss in more detils only for the neutral network for brevity.

Recall from Section 3 that we denote by $\underline{\mu}_n$ the cut-off network capacity for the major content provider's entry. Let us define a new threshold capacity $\underline{\mu}_n^N$ in the benchmark when the major content provider decides to enter or not, but it is unable to make any QoS investment. The superscript N means no possible investment. Then, obviously we have $\underline{\mu}_n < \underline{\mu}_n^N$. Recall the notations that 'ϕ' stands for no entry of the major content provider and '0' represents the situation in which the major CP entered but made zero investment. For the investment choices, let μ_n^* denote the ISP's network capacity choice when the major CP cannot invest in the compression technology; μ_n^{**} when the major CP can invest in it.

Consider first the case of $\mu_n^* \geq \underline{\mu}_n^N$: the major content provide chooses to enter even without its possible QoS investment. Then, μ_n^* is determined by the equality between the ISP's marginal

benefit from reducing the waiting time for the non-major content and its marginal cost:

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

where $C(\mu)$ represents the cost of investing μ. Suppose now that the major CP can make investment. Then, the major CP still enters and the capacity chosen by the ISP, denoted by μ_n^{**}, is determined by

$$-\beta\left[\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial \mu} + \frac{\partial W_n(h_n^*(\mu),\mu)}{\partial h}\frac{\partial h_n^*}{\partial \mu}\right] = C'(\mu). \tag{17}$$

Since $\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial h}\frac{\partial h_n^*}{\partial \mu} > 0$, we have $\mu_n^{**} < \mu_n^*$: the major CP's investment possibility reduces the ISP's capacity choice. Since, for high level of capacity, the two investments are substitutes, the first mover prefers to free ride on the follower by reducing its own investment.

Consider now the case of $\mu_n^* \leq \underline{\mu}_n$. Hence, in the absence of QoS investment opportunity, the major content provider does not enter and the ISP's capacity choice μ_n^* is determined by

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

Suppose now that the major CP can make investment. Then, the ISP has two choices: (1) it chooses $\mu^{**} = \mu_n^*$ and the major CP does not enter, or (2) it chooses $\widehat{\mu} \geq \underline{\mu}_n$ to induce the major CP's entry (and investment) so that the ISP chooses its capacity $\widehat{\mu}$ is determined by (17). To sum up, if $\mu_n^* \leq \underline{\mu}_n$, we have either $\mu_n^{**} = \mu_n^*$ or $\mu_n^{**} = \widehat{\mu}$; in the latter case, the major CP's investment possibility increases the ISP's investment. Since, for low level of capacity, the two investments are complements, the first mover increases its investment to promote the entry and the investment of the follower.

Last, consider now the case of $\mu_n^* \in (\underline{\mu}_n, \underline{\mu}_n^N)$. In this case, no entry occurs in the default situation. But, maintaining $\mu = \mu_n^*$ allows the major content provider to enter with its QoS investment. So, the ISP chooses $\mu_n^{**} = \widehat{\mu}$ determined by (17). Since we have

$$\left|\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial \mu}\right| = \frac{1}{\left[\mu - (1 + \frac{\lambda}{1+h_n^*(\mu)})\right]^2} > \left|\frac{\partial W_n(\phi,\mu)}{\partial \mu}\right| = \frac{1}{(\mu-1)^2}$$

and $\frac{\partial W_n(h_n^*(\mu),\mu)}{\partial h}\frac{\partial h_n^*}{\partial \mu} > 0$, both $\mu_n^{**} < \mu_n^*$ and $\mu_n^{**} > \mu_n^*$ can happen.

Summarizing, we have

(i) For small default capacity (i.e., $\mu_n^* \leq \underline{\mu}_n$), the possibility for the major CP to make QoS investment induces the ISP to expand its capacity.

(ii) For intermediate default capacity (i.e., $\mu_n^* \in (\underline{\mu}_n, \underline{\mu}_n^N)$), the possibility for the major CP to make QoS investment can induce the ISP to expand or to reduce its capacity.

(iii) For large default capacity (i.e., $\mu_n^* \geq \underline{\mu}_n^N$), the possibility for the major CP to make QoS investment induces the ISP to reduce its capacity.

7 Conclusion

In recent years, there has been an explosive growth in the mobile traffic. According to the Cisco's report (2013),[16] the 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 trend and local network needs, the US and the European Union have taken a quite contrasting approach in their view on network regulation. While US FCC imposes the critical rule of "no unreasonable discrimination" only on fixed operators and gives exemption of such restriction to mobile operators, the European Commission treats all types of networks—both fixed and mobile—in a uniform fashion under the principle of technological neutrality. While the FCC's asymmetric regulation—ex ante regulations on fixed broadband service providers, but ex post intervention on mobile broadband service providers—has been a controversial issue (Eisenach 2012, Maxwell and Brenner 2012), there has been little academic research on this aspect of net neutrality regulations.

To address this issue, we develop a theoretical model that characterizes the relative size of network capacity as a distinguishing feature between mobile networks and fixed networks; mobile network capacity is inherently limited by the availability of radio spectrum and the laws of physics. In such a framework, we investigate major content providers' incentives to invest in the quality of service, and identify conditions under which the paid prioritization service constitutes complements or substitutes to the QoS investments. Our analysis has implications for asymmetric regulation between fixed and mobile networks. We find that net neutrality regulations in mobile networks may pose more serious concerns than in fixed networks. In a mobile network with highly limited capacity,

[16]The reference is "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)

the regulation may hinder the desirable entry of delay-sensitive content without the availability of the prioritized service. By contrast, net neutrality regulations may pose less serious concerns in fixed networks. With a relatively more abundant capacity, the entry is not an issue and innovations at the edges can substitute for the lack of the prioritized service.

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, M., F. Kourandi, and T. 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." The 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." mimeo.

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

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

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

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

[11] Guo, Hong; Hsing Kenneth Cheng and Subhajyoti Bandyopadhyay. 2013. "Broadband Network Management and the Net Neutrality Debate." Production and Operations Management. forthcoming.

[12] 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.

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

[14] 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.

[15] Krämer, Jan; Lukas Wiewiorra and Christof Weinhardt. 2012. "Net Neutrality: A Progress Report." Telecommunications Policy.

[16] Lee, Daeho, and Junseok Hwang, 2011, "Network Neutrality and Difference in Efficiency Among Internet Application Service Providers: A Meta-frontier Analysis," Telecommunications Policy, 35(8): 764-772.

[17] 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).

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

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

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

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

[22] 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.

[23] Njoroge, P., A.E. Ozdaglar, N.E. Stier-Moses, , G.Y. Weintraub. 2012. "Investment in Two Sided Markets and the Net Neutrality Debate." Columbia Business School DRO (Decision, Risk and Operations) Working Paper No. 2010-05.

[24] Nurski, Laura. 2012. "Net Neutrality, Foreclosure and the Fast Lane: An Empirical Study of the UK." Foreclosure and the Fast Lane: An Empirical Study of the UK (October 1, 2012) .NET Institute Working Paper 12-13.

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

[26] Reggiani, Carlo and Tommaso Valletti. 2011. "Net Neutrality and Innovation at the Core and at the Edge." mimeo.

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

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

Appendix: Mathematical Proofs

Proof of Lemma 3

Proof of (i). The proof is based on a cotradiction argument. Let μ' be the initial capacity and $\mu^{''}(>\mu')$ the new capacty. Suppose not: $w_d(h_d^*(\mu'),\mu') < w_d(h_d^*(\mu''),\mu'')$. Then, let h'' be defined as

$$w_d(h_d^*(\mu'),\mu') = w_d(h'',\mu''), \tag{18}$$

which is equivalent to

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

$\mu^{''} > \mu'$ and (18) imply $h'' < h_d^*(\mu')$. In addition, $w_d(h_d^*(\mu'),\mu') < w_d(h_d^*(\mu''),\mu'')$ implies that the major CP wants to invest less than h'' when $\mu = \mu''$. By definition of $h_d^*(\cdot)$, we have:

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

The marginal gain of investment for the major CP at $h = h''$ and $\mu = \mu''$ is given by

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

and we have

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

Therefore, at $h = h''$ and $\mu = \mu''$, the marginal gain is larger than the marginal cost, which contradicts $w_d(h_d^*(\mu'),\mu') < w_d(h_d^*(\mu''),\mu'')$.

Proof of (ii). From (i), we have

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

The last inequality is equivalent to

$$\frac{\partial\alpha_d^*}{\partial\mu} < \frac{\alpha_d^*}{\mu} \tag{19}$$

We have

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

We show below

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

Regarding the first term, we have

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

This is negative if

$$\frac{\partial\alpha_d^*}{\partial\mu} < \frac{1+\alpha\lambda}{\lambda\mu}.$$

This holds from (19) as $\frac{\alpha}{\mu} < \frac{1+\alpha\lambda}{\lambda\mu}$.

Regarding the second term, we have

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

This is negative if

$$\frac{\partial\alpha_d^*}{\partial\mu} < \frac{1}{\lambda}.$$

This holds from (19) as $\mu > \alpha\lambda$.

Proof of Lemma 4

Note that the result trivially holds for $\beta = 0$ for $k > 1$. By continuity, the result holds for β small enough. Consider β not small enough. We have

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

where in the first inequality we use Lemma 3(ii) ($\frac{dW_d}{d\mu} < 0$). Then, we have

$$\overline{k}(\mu) = \frac{\beta}{\frac{\lambda}{1+h_d^*(\mu)}}\left(\frac{\mu-\frac{\lambda}{1+h_d^*(\mu)}}{\mu-(1+\lambda)}\right)^2.$$