View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Toulouse Capitole Publications

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

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

July 18, 2017

Abstract

We study how net neutrality regulations affect a high-bandwidth content provider (CP)'s investment incentives to enhance its quality of services (QoS) in content delivery to end users. We find that the effects crucially depend on whether the CP's entry decision is constrained by the Internet service provider (ISP)'s network capacity. If capacity is relatively large, prioritized services reduce the QoS investment as they become substitutes, but improve traffic management. With limited capacity, by contrast, 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. Our analysis suggests that the optimal policy may call for potentially asymmetric regulations across mobile and fixed networks.

Key words: Net neutrality, network capacity, entry, quality of services (QoS), externalities, investments

[*] We thank Marc Bourreau, Jane Choi, Ben Hermalin, Jeroen Hinloopen, Joshua Gans, Vahagn Jerbashian, Bruno Jullien, Martin Peitz, Wilfried Sand-Zantman, Glenn Woroch, and seminar participants at various conferences and university seminars for helpful comments. We gratefully acknowledge financial support from the NET Institute (www.NETinst.org). An earlier version of this paper was circulated as NET Institute Working Paper #13-24 under the title "Asymmetric Neutrality Regulation and Innovation at the Edges: Fixed vs. Mobile Networks." Jay Pil Choi's research was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2016S1A5A2A01022389).

[†] Michigan State University, 220A Marshall-Adams Hall, East Lansing, MI 48824-1038 and School of Economics, Yonsei University, Seoul, Korea. E-mail: choijay@msu.edu.

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

[§] Corresponding author. Department of Economics, Finance and Legal Studies, University of Alabama, 265 Alston Hall, Tuscaloosa, AL 35487. E-mail: byung-cheol.kim@ua.edu.

1 Introduction

Net neutrality is the principle that all packets on the Internet must be treated equally in their delivery regardless of their content source, destination, and type, and Internet service providers (ISPs) cannot charge content providers for the provision of best efforts access. The "open Internet" order in 2010 represents the U.S. Federal Communication Commission (FCC)'s initial attempt at securing net neutrality, and has served as a focal guideline for neutrality regulations.[1] However, the FCC's order has faced legal challenges by major ISPs such as Comcast and Verizon Communications. The United States Court of Appeals for the District of Columbia Circuit concurred with the ISPs and ruled that the FCC overstepped its authority.[2] In response, on February 26, 2015, the FCC adopted new rules for broadband Internet service by reclassifying high-speed Internet service as a "telecommunications service" rather than an "information service."[3] This seemingly technical maneuver allows the FCC to circumvent the legal issue of its authority over Internet service and treat the service as a public utility under Title II of the Telecommunications Act. Several lawsuits were filed immediately after the FCC's new rules; the court upheld the reclassification and thus the net neutrality was legally protected. In this process, the issue of net neutrality has emerged as the most important and controversial regulatory agenda since the inception of the Internet.[4] More recently, the new FCC chairman Ajit Pai outlined that high-speed internet service should no longer be treated like a public utility with strict rules.[5] Even so, since broadband providers have conflicts of interests against content service providers and vice versa, the controversy appears to be unabated continuously.

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

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

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

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

[5] See "F.C.C. Head Plots Course To Ease Rules On Internet" by Cecilia Kang in the *The New York Times* on April 27, 2017.

The extant literature on network neutrality has mainly focused on the expansion of ISPs' network capacity as innovation at the "core." In this paper we focus on potential implications of neutrality regulation on innovation incentives at the "edges" to reflect the growing importance of content providers (CPs)' investments in the modern Internet ecosystem.[6] More specifically, the ISPs' capacity expansion making bigger "pipes" is not the only solution to resolving the congestion problem. 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. 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] Real-world anecdotal evidences regarding the importance of CDN technology can be easily found. For example, Spotify (a Swedish commercial music streaming platform) and Dailymotion (a French video site) have adopted CDN technology to make their content business more reliable (Economist, 2014). Malone, Jacob, and Nevo (2015) offer some statistics regarding CDN using detailed residential broadband usage data: they report that CDN applications are far more popular than file sharing protocols such as BitTorrent and FTP but slightly less than online gaming such as Xbox Live and Clash of Clans or music streaming services such as Spotify and Pandora.

From an end user's perspective, the fundamental goal is to enjoy the highest quality of service at a minimum fee; the channel through which this is achieved, either through ISPs' capacity investments or CPs' CDN investments, is of little interest. In fact, the use of CDNs is well acknowledged as one of the best ways to improve page load time which

[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 requests of such content to the closest 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] Greenstein, Peitz, and Valleti (2016) offers an excellent review on the network neutrality debate and they provide why innovative investments by content providers are important and what kinds of technologies are for these investments. Martin and Schuett (2016) also discuss innovative actions taken by content providers and their implications for the network neutrality debate.

critically affects any web experience. Unfortunately, researchers have seldom studied how these new technological changes relate to regulatory decisions while regulators and policy-makers need to understand how the network regulations would affect the content providers' investments in alternative technology solutions (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 depend on the relative size of the ISPs' network capacity vis-à-vis major content providers' bandwidth usage.

The intuition is as follows. If the network capacity is large enough, prioritized delivery and QoS investment become substitutes. Consider a high network capacity case in which the entry of new content is always warranted even without the prioritized service. 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." However, because the marginal benefit of the QoS investment increases with the severity of network congestion, the MCP will invest less in a non-neutral network compared to in a neutral network. In other words, the availability of the prioritized service may dampen content providers' incentives to invest in QoS. This logic is combined with the standard argument that the ISP's investment is suboptimal because it does not fully incorporate the impact of its investment on the NCPs, which means non-internalized externality. We refer to this force as the "QoS investment effect."[9] The social welfare depends on the relative magnitude of these two forces, and we consider it more applicable to the fixed network where the entry of content providers has not been treated as a serious concern, relative to the mobile network where the network capacity can be a constraint on the entry decision.

[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.

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

As several comprehensive reviews of the literature on net neutrality are available (e.g., Lee and Wu (2009), Schuett (2010), Lee and Hwang (2011), and Krämer, Wiewiorra, and Weinhardt (2013)) including the most recent article by Greenstein, Peitz, and Valletti (2016), we briefly provide a selective review of notable works in relation to this paper.

The main focus of the extant studies has been investment incentives for the ISPs on their "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 "pipes." Academic research on this issue includes Musacchio, Schwartz, and Walrand (2009), Choi and Kim (2010), Cheng, Bandyopadhyay and Guo (2011), Economides and Hermalin (2012), Krämer and Wiewiorra (2012), and Njoroge et al. (2013). A related issue is the content providers' hold-up concern that may result in no entry or less investment in content. This concern arises because investments by high-value content providers may be expropriated *ex post* by ISPs who can act as gatekeepers with paid prioritization services. For studies along this avenue, we refer to Bandyopadhyay, Guo, and Cheng (2012), Choi and Kim (2010), Grafenhofer (2010), Reggiani and Valletti (2016), and Bourreau, Kourandi, and Valletti (2015). We depart from this literature by exploring new, but highly important, innovation channels adopted by major content providers such as Google, Amazon, Microsoft, and Netflix.[10]

In this regard, Peitz and Schuett (2016) is closely related to our paper. They consider so-called *congestion control techniques* that decrease packet losses during delivery to users with an “inflation of traffic” by sending multiple redundant packets. This practice may be privately optimal but aggravates the congestion problem on the network. They introduce the tragedy of common property resources into the net neutrality discussion and show that net neutrality regulation may lead to socially inefficient inflation of traffic whereas the socially optimal allocation can be achieved with tiered pricing. They also look at compression in an extension. Differentiated from them, in this article we study a high-bandwidth CP’s QoS investments in alleviating network congestion focusing on various trade-offs associated with such decision. Economides and Hermalin (2015) is related to our paper in that they provide a new explanation for why ISPs offer plans with download caps by showing that congestion externality can induce ISPs to use download limits as a mechanism to restrict the aggregate bandwidth usage.[11] Both papers deal with quality decisions by CPs and address how congestion externalities affect social welfare, but through different channels.

As a policy implication, our findings suggest potential benefits of an asymmetric regulatory approach to mobile and fixed networks, which was adopted by the FCC’s 2010 Order, but subsequently discarded with its reclassification of Internet service. The FCC justified its discriminatory treatment of mobile networks by stating that “Mobile broadband is an earlier-stage platform than fixed broadband, and ... [m]obile 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

[10] There is a consensus on the basic premise that end-users’ quality of service must be the primary goal of a desirable network ecosystem (Xiao (2008), Altman et al. (2012), and Guo, Cheng, and Bandyopadhyay (2013)). The ISP’s capacity investments and CPs’ CDN investments can be alternative means to achieve this same goal.

[11] Recently, direct payments from consumers to content providers have received more attention by researchers. Gans (2015) and Economides and Hermalin (2015) consider a setting in which consumers need to pay content-specific prices to content providers, whereas Choi, Jeon, and Kim (2015) consider micropayments in a reduced form represented as CPs’ business models. Jullien and Sand-Zantman (2015) examine the net neutrality issues in the context of information transmission such as signaling and screening.

not typically encounter."[12] With a limited mobile network capacity, the paid prioritization can facilitate the entry of a congestion-sensitive content provider while the entry is not made under neutral networks because the content provider may find it too costly to invest up to its desired QoS. For this case, the prioritization complements innovation at the edges, and a lenient non-neutral treatment may facilitate the availability of innovative content and applications in the early-stage mobile network.

The remainder of our paper is organized as follows. We present our model in Section 2. In particular, a generalized queuing model is introduced to allow for a CP's investment. In Section 3, we first characterize the first-best outcome and then analyze the QoS investment decisions by the major content provider under neutral and non-neutral network regimes. In Section 4, we examine the "intensive margin case" in which a network capacity is large enough and thus a major content provider's entry is warranted without a costly QoS investment. In Section 5, we study the "extensive margin case" in which the network capacity is limited such that a content provider's entry becomes a critical issue. In Section 6, we extend our model to allow for the ISP's investment in network capacity prior to the entry of the major CP. We conclude in Section 7. Mathematical proofs are relegated to the Appendix.

2 The Model

2.1 ISP, CPs, and Consumers

We consider a monopolistic broadband ISP who is in charge of last mile delivery of online content to end-users. 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 an MCP's investment decision to improve QoS for a successful content

[12] FCC 10-201, *In the Matter of Preserving the Open Internet, Broadband Industry Practices* (the 2010 "FCC Order"), published in Fed. Reg. Vol. 76, No. 185, Sept. 23, 2011.

business; the MCP's relatively large scale of operation justifies the costly investment.

There is a continuum of homogeneous consumers whose mass is normalized to one. Let v and V denote the consumer's intrinsic utility from consuming the MCP's content and NCPs' respectively. Each consumer experiences some disutility from delays of content delivery due to network congestion. We adopt an additive utility specification in which the net surplus decreases in the average waiting time for both types of content, respectively denoted by w and W. Consumers are heterogeneous with respect to the sensitivity to delays across the content they consume; some applications such as email are not sensitive as much as some other such as streaming services. Each consumer earns the gross utility from each type of content:

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

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

For NCPs' content, we introduce a parameter $\beta \in [0, 1]$ to denote the ISP's share of the total surplus generated by NCPs' content delivery. In other words, the ISP receives $\beta U(W)$ from providing delivery services for NCPs' content; the rest of the surplus, $(1 - \beta)U(W)$, is shared among NCPs and end users.[13] One may regard β as a measure of the extent to which the ISP internalizes any externality inflicted on NCPs and end users by its decisions. If $\beta = 0$, the ISP will not take into account any potential effects on NCPs' content traffic when the ISP deals with 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 $\beta = 1$. However, for any $\beta < 1$, there may be a discrepancy between the ISP's optimal decision and the social planner's, with the potential for discrepancy more pronounced with a lower β.

Similarly, for the MCP's content, let $\alpha \in [0, 1]$ denote the ISP's share of the total net

[13] Since NCPs and consumers are "passive players" in our model, this simplification does not affect any qualitative results in this article.

surplus $u(w)$ generated by the MCP's content delivery. We assume that the MCP receives the remainder of the surplus from its content delivery, $(1-\alpha)u(w)$, and that consumers receive zero surplus in both network regimes.[14]

2.2 Network Congestion, MCP'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.[15]

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 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 NCPs is one, the overall demand parameter (i.e., the total volume of traffic) for NCPs' content is also normalized to one. By contrast, we envision 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. 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 NCPs' aggregate traffic volume.[16] The total traffic volume for the ISP thus amounts to $1+\lambda$ and we need the condition of $\mu > 1+\lambda$ for a meaningful analysis of network congestion; otherwise, the waiting time becomes infinity.

The MCP can make an investment of $h \geq 0$ to enhance the quality of service in its content delivery. As discussed earlier, the investment can take various forms, such as compression technology to reduce packet size or content delivery networks (CDN) that shorten

[14] More generally, we can let the share of the consumers denoted by α_C and the ISP's share by α_I. Then, the MCP's share will be $\alpha_M = 1-\alpha_C-\alpha_I$. Our assumption means $\alpha_C = 0$. If $\alpha_C \neq 0$, the joint payoff of the ISP and the MCP in a non-neutral network will depend on α_C. However, none of our qualitative results will change with this consideration.

[15] Choi and Kim (2010), Cheong et al. (2011), Bourreau et al. (2015), Krämer and Wiewiorra (2012), inter alia, adopt the M/M/1 queuing model to analyze network congestion.

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

the delivery distance by installing content servers at local data centers so that end-users' demands are served by the closest data center.[17] The common objective of all such investments is to speed up content delivery to enhance the user experience. We thus model them simply as an investment in a compression technology that would reduce the traffic volume of the MCP'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. We assume the investment cost is increasing and convex in the investment level, i.e., $c'(h) > 0$ and $c''(h) > 0$, and satisfies the Inada condition of $c(0) = 0$ and $c'(0) = 0$.

We consider two network regimes: neutral and non-neutral networks where subscript n stands for neutral networks and d for non-neutral (discriminatory) 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. We assume that there is no charge for best efforts under either regime. We focus on the last mile segment of the Internet where multiple ISPs make interconnection agreements such as peering and it is typical that the ISPs facing uploading CPs are different from the ISPs facing end-users. We do not question the interconnection agreements among ISPs and simply allow for the possibility that if a uploading CP wants its content delivered with prioritization at the last mile, then the ISP controlling the last mile delivery can charge for the prioritized delivery.

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

[17] 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.

total waiting time for the MCP's content:

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

The total volume of traffic (packet size) amounts to $1+a\lambda$ (one for NCPs' content and $a\lambda$ for the MCP's content with compression[18]), 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 MCP's content, the total waiting time is computed as (2). With no investment in the compression technology ($h=0$, or $a=1$), the average waiting time reduces to $\frac{1}{\mu-(1+\lambda)}$ as in the standard M/M/1 queuing system. For the NCP's content, the total waiting is

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

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

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

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

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

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

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

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 (2)-(5), 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 one 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 extends the waiting cost equivalence characterized in Choi and Kim (2010), Bourreau et al. (2015), Krämer and Wiewiorra (2012) to 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 regardless of whether or not a subset of the packets is prioritized.

Property 4 For a given pair of (a, μ), prioritizing the MCP'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 MCP'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, an MCP's decisions are straightforward since they do not involve a bargaining situation with the ISP. The ISP cannot charge content providers for the provision of best efforts access.

N-1. For a given ISP's network capacity μ, the MCP 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 MCP and the ISP bargain over the price of the prioritized service.[19]

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.

[19] This seems to be a reasonable assumption considering the MCP's market power, as illustrated by the recent deal between Netflix and Comcast. See "The Inside Story Of How Netflix Came To Pay Comcast For Internet Traffic," available at http://qz.com/256586/the-inside-story-of-how-netflix-came-to-pay-comcast-for-internet-traffic.

We assume 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: The 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. Note that 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: for a given μ

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

Then, we can establish the following intuitive result.

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

Proof. *See the Appendix.* ■

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

3.2 Neutral Networks

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

$$\max_{h \geq 0} \pi_n = (1-\alpha)[v - kw_n(h,\mu)] - c(h),$$

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

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

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

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

This relationship implies that a threshold network capacity $\underline{\mu}_n$ exists such that $\pi_n^*(\mu) \geq 0$ if and only if $\mu \geq \underline{\mu}_n$. 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 entry (and thus 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: $h_n^* = 0$

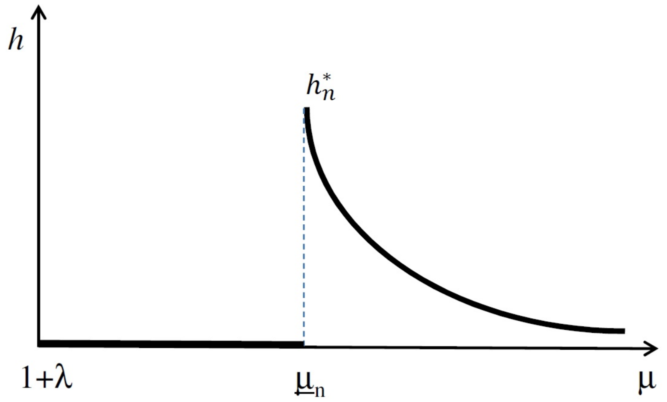

Figure 1: Optimal QoS Investment in the Neutral Network

for $\mu < \underline{\mu}_n$ and then $h_n^* > 0$ and $\frac{\partial h_n^*}{\partial \mu} < 0$ for $\mu \geq \underline{\mu}_n$.

Note that the threshold level of network capacity $\underline{\mu}_n$ depends on α and k. As either α or k increases, the MCP's optimized profit decreases and thus $\underline{\mu}_n$ increases. For notational brevity, we will use the simple notation $\underline{\mu}_n$ unless this concise notation causes any confusion.

3.3 Non-neutral Networks

Now consider a non-neutral network in which the MCP has an option to buy a 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 the 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 the terms of the prioritized service.[20] We define the MCP's profit gross of any payout for the priority as

$$\pi_d \equiv (1-\alpha)\left[v - kw_d(h,\mu)\right] - c(h), \tag{9}$$

[20] We assume that there is efficient bargaining between the ISP and the MCP. See section 4.1 for a discussion of alternative pricing schemes in which the ISP unilaterally sets a price for prioritized delivery of the MCP's traffic and the price depends on the amount of data carried.

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:

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

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

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

and the optimal investment decreases in the capacity, $\frac{\partial h_d^*}{\partial \mu} < 0$.[21]

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. In particular, if $\beta = 0$, the ISP and the MCP will agree on some price of prioritization whenever the surplus from the entry is non-negative, i.e., $v - kw_d(h_d^*(\mu),\mu) - c(h_d^*(\mu)) \geq 0$. 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 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 $\underline{\mu}_n > \underline{\mu}_d$.[22]

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

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

[22] Again we note that the threshold level of network capacity in a non-neutral network, $\underline{\mu}_d$, depends not only on α and k as in the neutral regime but also on β in the non-neutral one. As we discuss in the next subsection, this result holds for a small β. If β is sufficiently large and close to 1, we cannot rule out the possibility that this inequality is reversed.

increases with the severity of congestion in the network, as shown below.

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

Consequently, we establish the following lemma:

Lemma 2 *The MCP 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 Network Capacity and QoS Investments

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

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

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

(ii) For a larger capacity $\mu > \underline{\mu}_n$, prioritization and MCP's investment are "substitutes" in that purchasing prioritization reduces the MCP'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. For the range of $\underline{\mu}_d < \mu < \underline{\mu}_n$, there is the greater QoS investment under the non-neutral network compared to the neutral network. The reversal relationship occurs when $\mu > \underline{\mu}_n$. Intuitively, the prioritization reduces the QoS investment incentives because it provides an alternative technological solution to achieve the desired level of QoS.

Our analysis shows that the MCP's entry crucially depends on the ISP's network capacity. In the remainder of the paper, we refer to the limited capacity case of $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$ as "extensive margin case" and the high capacity case of $\mu > \underline{\mu}_n$ as "intensive margin case" in the sense that the MCP's entry is a focal issue in the former but not in the latter.

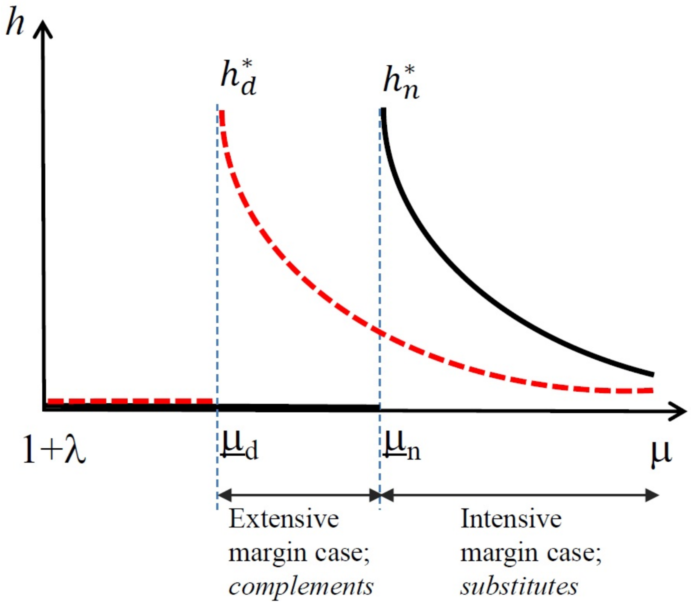

Figure 2: Optimal QoS investments, Net Neutrality, and Network Capacity

Remark 1 For any (α, β), we may compare the relative levels of the two thresholds, $\underline{\mu}_n(\alpha)$ and $\underline{\mu}_d(\alpha, \beta)$. Let $\underline{\mu}_d(\alpha, \beta)$ denote the network capacity at which the MCP's entry with prioritization yields the same joint payoff of the two bargaining parties with no entry, i.e., $\underline{\mu}_d(\alpha, \beta)$ is defined as

$$v - kw_d^*(\mu) - c(h_d^*(\mu)) - \beta\,[W_d^*(\mu) - W_n(\phi, \mu)] = 0$$

where ϕ stands for no entry of the MCP. We find that $\underline{\mu}_d(\alpha, \beta)$ increases with β if k is large enough and $\underline{\mu}_d(\alpha, 0) < \underline{\mu}_n(\alpha)$ for any $\alpha \in [0, 1]$.[23] However, the socially optimal entry threshold $\underline{\mu}_d(\alpha, 1)$ may or may not be higher than $\underline{\mu}_n(\alpha)$. If we have $\underline{\mu}_d(\alpha, 1) < \underline{\mu}_n(\alpha)$, the above distinction of the extensive and intensive margin cases is preserved and the entry becomes socially desirable. By contrast, if we have $\underline{\mu}_d(\alpha, 1) > \underline{\mu}_n(\alpha)$, then there will be a cutoff level of β denoted by $\widetilde{\beta}(< 1)$ such that $\underline{\mu}_d(\alpha, \widetilde{\beta}) \leq \underline{\mu}_n(\alpha)$ if and only if $\beta \leq \widetilde{\beta}$. In what follows, we restrict our attention to the case where $\underline{\mu}_d(\alpha, \beta) < \underline{\mu}_n(\alpha)$ holds.[24]

[23] The proof is available from the authors upon request.

[24] If $\underline{\mu}_d(\alpha, \beta) > \underline{\mu}_n(\alpha)$, the extensive margin case we discuss in section 5 does not arise.

4 The Intensive Margin Case

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

$$\Pi_r(h, \mu, \beta) = v - kw_r(h, \mu) - c(h) + \beta[V - W_r(h, \mu)]. \tag{11}$$

Hence, the prioritization will be adopted if the two parties find the non-neutral regime better than the neutral treatment: i.e.,

$$\Delta\Pi^I(\mu, \beta) = \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta) > 0, \tag{12}$$

where the superscript I indicates that we are considering the *intensive margin* case and $\Delta\Pi^I(\mu, \beta) > 0$ means a higher joint payoff under the non-neutral network. For the intensive margin case, we find that a non-neutral network may generate a trade-off between superior traffic management and more severe under-investment problem.

4.1 Traffic Management vs. Under-investment

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

$$\Delta\Pi^I(\mu, \beta) = \underbrace{[\Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta)]}_{\text{Traffic Management Effect } (+)} + \underbrace{[\Pi_n(h_d^*(\mu), \mu, \beta) - \Pi_n(h_n^*(\mu), \mu, \beta)]}_{\text{QoS Investment Effect } (-)}. \tag{13}$$

The first term in (13) is always positive and we refer to it as the "traffic management effect": for a given QoS investment level h, prioritizing the MCP's traffic reduces the total delay cost because the MCP's content is more sensitive to congestion ($k > 1$). Precisely, we have

$$\begin{aligned} \text{Traffic Management Effect} &= \Pi_d(h_d^*(\mu), \mu, \beta) - \Pi_n(h_d^*(\mu), \mu, \beta) \qquad (14) \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] + \beta\,[W_n(h_d^*(\mu)) - W_d(h_d^*(\mu))] \\ &= k[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] - \beta\,[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] \\ &= (k - \beta)[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))] \geq 0 \end{aligned}$$

where the third equality in (14) is obtained from Property 3, $w_n(h_d^*) + W_n(h_d^*) = w_d(h_d^*) + W_d(h_d^*)$.[25]

We refer to the second square bracket in (13) as the "QoS investment effect": the availability of a prioritized service will decrease the MCP's investment level from $h_n^*(\mu)$ to $h_d^*(\mu)$ (Lemma 2), which in turn affects the resulting joint payoff. To determine the sign of this term, let $h_n^J(\mu, \beta)$ denote the MCP's investment choice that maximizes the joint profit of the two parties in the neutral regime, i.e.,

$$h_n^J(\beta) = \arg\max_h \Pi_n(h, \mu, \beta) \ = v - kw_n(h, \mu) - c(h) + \beta[V - W_n(h, \mu)] \qquad (15)$$

which is alternatively defined as

$$h_n^J(\beta) = \arg\min_h \ kw_n(h, \mu) + c(h) + \beta W_n(h, \mu). \qquad (16)$$

From the profit maximization problem (15), we can see that the MCP's private optimal choice, $h_n^*(\mu)$ that maximizes $v - kw_n(h, \mu) - c(h)$, fails to incorporate its positive externality on the NCPs' content. The joint decision on the QoS investment internalizes only part of such externality, $\beta W_n(h)$. Thus, the joint decision yields an under-investment problem. This logic can be seen from the cost minimization problem (16) in which the MCP chooses $h_n^*(\mu)$ to minimize $kw_n(h) + c(h)$, but ignores the additional cost $\beta W_n(h)$. In short, in either interpretation, we find that an under-investment problem persists in that $h_n^*(\mu) < h_n^J(\mu, \beta)$

[25] All the derivations in (14) hold for any h, not just for $h_d^*(\mu)$. We evaluated the inequality at the optimal QoS choice.

for any $\beta > 0$. This result combined with Lemma 2, $h_d^*(\mu) < h_n^*(\mu)$, [26] proves that the QoS investment effect must be negative:

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

We should mention that the result on the MCP's QoS investment depends on our assumption that there is efficient bargaining between the ISP and the MCP. We discuss the robustness of our results to alternative bargaining assumptions in section 4.4.

4.2 Effects of Prioritization on Social Welfare

We now analyze a social planner's incentive to introduce the paid prioritization and compare it with the private incentive. We consider the constrained (second-best) social optimum in which the social planner can choose the network regime only, but the QoS investment is left to the MCP's decision. The social welfare in each regime coincides with the joint payoff of the ISP and the MCP when $\beta = 1$ for (11), which is given by

$$S_r(\mu) = \Pi_r(h_r^*(\mu), \mu, \beta = 1) = v - kw_r(h_r^*(\mu), \mu) - c(h_r^*(\mu)) + [V - W_r(h_r^*(\mu), \mu)], \tag{18}$$

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

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

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

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

NCPs' content which is represented by

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

We can further decompose the externality term in (20) and show $\Delta S(\mu) < \Delta\Pi^I(\mu, \beta)$ as follows:

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

The first term of (21) has a negative sign because of Lemma 2 ($h_n^* > h_d^*$), and the second term also takes a negative value following Property 2. Lastly, we notice that the discrepancy between the social incentives and the private incentives is inversely related to β. Interestingly, we find that if the discrepancy reaches its maximum ($\beta = 0$), the ISP and 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:

$$\begin{aligned}
\Delta\Pi^I(\mu, \beta &= 0) = [v - kw_d(h_d^*(\mu), \mu) - c(h_d^*(\mu))] - [v - kw_n(h_n^*(\mu), \mu) - c(h_n^*(\mu))] \qquad (22)\\
&\geq [v - kw_d(h_n^*(\mu), \mu) - c(h_n^*(\mu))] - [v - kw_n(h_n^*(\mu), \mu) - c(h_n^*(\mu))]\\
&= k[w_n(h_n^*(\mu), \mu) - w_d(h_n^*(\mu), \mu)] \geq 0,
\end{aligned}$$

where the first (weak) inequality comes from the revealed preference argument and the last inequality from Property 2. Proposition 3 summarizes findings thus far for the intensive margin case.

Proposition 3 *Consider the intensive margin case with network capacity $\mu > \underline{\mu}_n$ in which the MCP always enters. Then, we find*

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

(ii) In general, there exist socially excessive incentives to adopt a prioritization service unless $\beta = 1$.

4.3 Net Neutrality as a Second-Best Policy

One interesting policy implication from our study is that net neutrality regulation can be an optimal second-best policy. For the second-best comparison, let us focus on the trade-off between the traffic management and QoS investment effects. Note that the traffic management effect increases with k at the rate of $[w_n(h_d^*(\mu)) - w_d(h_d^*(\mu))]$ as is seen in (14). On the other hand, the QoS investment effect increases with k at the rate of $[w_n(h_n^*(\mu)) - w_n(h_d^*(\mu))]$ which can be confirmed in (17). In order to compare these two slopes, note that

$$w_n(h_d^*(\mu)) + w_n(h_d^*(\mu)) > w_n(h_n^*(\mu)) + w_d(h_d^*(\mu)).$$

because $h_d^*(\mu) < h_n^*(\mu)$ implies $w_n(h_d^*(\mu)) > w_n(h_n^*(\mu))$ and Property 2 states $w_n(h_d^*(\mu)) > w_d(h_d^*(\mu))$. The condition above is equivalent to

$$w_n(h_d^*(\mu)) - w_d(h_d^*(\mu)) > w_n(h_n^*(\mu)) - w_n(h_d^*(\mu)),$$

that is, the traffic management effect is more sensitively responding to k than the QoS investment effect. This implies that there exists k^* (> 1) below which net neutrality can be justified as the second-best policy because net neutrality is always preferred for the case of $k = 1$ as explained below.

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

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

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 MCP reduces its investment because the quality of service can be enhanced through prioritization.

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

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

4.4 Robustness of the Results to Alternative Priority Pricing Schemes

We point out that the result on the MCP's QoS investment depends on our assumption that there is efficient bargaining between the ISP and the MCP.[27] This implies that the MCP's investment does not depend on the price of priority. If we consider alternative pricing schemes in which (a) the ISP unilaterally sets a price for prioritized delivery of the MCP's traffic and (b) the price depends on the amount of data carried, our results can be modified in the following way for $\alpha = 0$.

Suppose, for instance, the ISP sets a linear tariff with a constant price per unit of traffic. In such a case, we have actually overinvestment compared to social optimum if β is very large and close to 1. The reason is that the ISP will generally set a price that exceeds the socially efficient level in an attempt to extract rents from the MCP, which in turn would increase the MCP's incentives to avoid such charges by investing in QoS. To be more precise, let p denote per unit price of prioritized traffic. The MCP then chooses its QoS investment level $a(=\frac{1}{1+h})$ to maximize $\pi_d \equiv [v - kw_d(a)] - c(a) - p\lambda a$ with associated first order condition $-kw_d'(a) - c'(a) = p\lambda$, where $c(a) = c(h^{-1}(a))$. The first order condition implicitly defines the inverse demand $p(a)$ facing the ISP. The ISP's problem is

$$\max_a \beta(V - W_d(a)) + p(a)\lambda a = \beta(V - W_d(a)) - a(kw_d'(a) + c'(a))$$

leading to the first order condition

$$-W_d'(a) - kw_d'(a) - c'(a) + (1-\beta)W_d'(a) - a(kw_d^{''}(a) + c^{''}(a)) = 0 \tag{23}$$

The social planner's first order condition is given by

$$-W_d'(a) - kw_d'(a) - c'(a) = 0 \tag{24}$$

A comparison of the two first order conditions (23) and (24) reveals that there are two

[27] We thank an anonymous referee for pointing this one out and kindly providing us with some analysis on this, which we replicate below.

types of distortions in the market equilibrium under a discriminatory network, represented by $(1-\beta)W_d'(a)$ and $-a(kw_d^{''}(a)+c^{''}(a))$. The first one, $(1-\beta)W_d'(a)>0$, is due to the failure to fully internalize the effect of a on NCPs' waiting time, which leads to insufficient QoS investment by the MCP. The second one, $-a(kw_d^{''}(a)+c^{''}(a))(<0)$, is due to the ISP's monopoly distortion in the absence of efficient bargaining and mitigates the first effect as it goes in the opposite direction from the first effect. In particular, if β is sufficiently large and close to 1, the second effect dominates the first one and we can have overinvestment by the MCP compared to the social optimum under a linear tariff pricing scheme.

However, if we consider a more general scheme of two-part tariff, we can restore the underinvestment result. To see this, suppose that the ISP offers a two-part tariff of (p,F), where p is per unit price of prioritized traffic and F is a fixed fee. When the ISP sets a per-unit traffic price of p to induce a traffic of a, the following individual rationality condition should be satisfied:

$$[v-kw_d(a)]-c(a)-p\lambda a-F\geq \underline{u},$$

where $\underline{u}$ is the payoff of the MCP when it rejects an offer by the ISP and its traffic is treated equally with NCPs' content. Thus, the highest fixed fee that can be charged is given by $F(a)=[v-kw_d(a)]-c(a)-p(a)\lambda a-\underline{u}$. The MCP's problem with a two-part tariff can be written as

$$\max_a \beta(V-W_d(a))+p(a)\lambda a+F(a)=\beta(V-W_d(a))+[v-kw_d(a)]-c(a)-\underline{u}$$

The first order condition is given by

$$-\beta W_d'(a)-kw_d'(a)-c'(a)=0 \tag{25}$$

It immediately follows that the MCP has less incentives to invest in QoS compared to the social optimum by comparing (24) and (25). The MCP has the same incentives as the social planner only when $\beta=1$, that is, only when the ISP fully extracts all surplus and internalizes any externality caused by the prioritized service. Otherwise, we restore the underinvestment result compared to the social optimum.

The comparison of MCP's investment levels across the discriminatory and neutral regimes, however, is more complicated. With a two part tariff, we can show that our result is robust if β is small and close to zero whereas we have more investment in the discriminatory regime if $\beta = 1$. This implies that there is a critical level of β below which our current result holds. To be more specific, the first order condition for the MCP's investment under a neutral regime is given by

$$-kw_n'(a) - c'(a) = 0 \tag{26}$$

A comparison of (25) and (26) indicates that we have more QoS investment under the neutral regime if either β is small or k is large enough because $w_n'(a) > w_d'(a)$. However, we cannot rule out the possibility of more investment in the discriminatory regime if $kw_n'(a) > W_d'(a) + kw_d'(a)$. This condition implies that there is more investment in the discriminatory when $\beta = 1$. Note that the MCP's QoS investment under the neutral system is independent of β whereas its investment under the discriminatory regime is increasing in β. Therefore, if the condition holds, there is a critical level of $\beta \in (0, 1)$, β^*, such that the MCP's QoS investment is lower under the discriminatory system with a two-part tariff if $\beta < \beta^*$. The intuition for this possibility can be explained as follows. As β increases, the ISP takes into account the effect of congestion on NCPs. One instrument to mitigate congestion from the perspective of the ISP is to raise the unit price of prioritized traffic (as in congestion pricing), which induces the MCP to invest more in QoS. It thus implies that the efficiency rationale for prioritization is stronger under a two-part tariff than under efficient bargaining if β is sufficiently large.

Lastly, we discuss what happens if we modify our model such that the ISP can charge both NCPs and MCP even for the best-effort basic service under non-neutrality. This change implies that β becomes endogenous at $\beta = 1$ because NCPs are homogeneous in our model and the ISP will extract all their surplus. In the bargaining between the MCP and the ISP, they will choose the outcome that maximizes their joint surplus among the three following options: (i) no entry, (ii) entry with a prioritized service, and (iii) entry without a prioritized service. The default option for the MCP is no entry. They will share the surplus from entry regardless of whether it is done with or without prioritization. We thus find that the full surplus extraction from the NCPs (by charging a price for non-prioritized service) together with efficient bargaining with the MCP leads to a socially optimal outcome both in terms of entry decision and prioritization decision.[28] One important difference from the main analysis is that the MCP's entry with a non-prioritized service is also on the table for negotiation. As a result, we cannot claim that entry always occur with prioritization as long as entry occurs without prioritization. The entry with prioritized service, compared to the entry without prioritized service, has the benefit of improving the traffic management at the cost of reducing the MCP's investment.

5 The Extensive Margin Case

In this section let us analyze the effects of net neutrality regulation on various participants when the MCP makes no entry under the neutral regime because $\pi_n^*(\mu) < 0$ for $\mu < \underline{\mu}_n$ but the entry becomes possible in the non-neutral network. We consider a general situation with $\beta \in [0, 1]$.

Here we attempt to make two points. First, we find that the paid prioritization can facilitate the entry of a congestion-sensitive content provider while the entry may not be made under neutral networks because the content provider may find it too costly to invest up to its desired QoS. For this case, the prioritization complements innovation at the edges. Second, the newly available content would generate additional value to the network, but the entry of new content not necessarily results in a higher social welfare. This is because the new content can consume a substantial portion of the existing network capacity, which increases the congestion for other content. Indeed, the surplus from new content can be outweighed by the efficiency loss from the elevated congestion for other content. Below we provide a simple analysis for these two results.

Under a non-neutral network, an MCP's entry has two countervailing effects. On the one hand, the new content generates a positive surplus $v - kw_d(h_d^*(\mu), \mu)$, which can be shared by the MCP and the ISP with Nash bargarining. On the other hand, the entry exacerbates the network congestion through the following two channels: the additional

[28] A detailed analysis is available from the authors upon request.

bandwidth taken by the MCP's new content means more congestion for a given network capacity; additionally the prioritized delivery of the MCP's content means a slower delivery for NCPs' content. Formally, we examine the difference in waiting time for the NCPs' content with the introduction of a two-tiered service, ΔW, that can be decomposed into two parts.

$$\begin{aligned}\Delta W &\equiv W_d(h_d^*(\mu),\mu) - W_n(\phi,\mu) \\ &= \underbrace{[W_n(h_d^*(\mu),\mu) - W_n(\phi,\mu)]}_{(+)\text{ due to new content entry}} + \underbrace{[W_d(h_d^*(\mu),\mu) - W_n(h_d^*(\mu),\mu)]}_{(+)\text{ due to different priority classes}}, \qquad (27)\end{aligned}$$

where we remind that ϕ stands for no entry of the MCP. The first bracketed term in (27) measures the increase in delivery time even in the absence of prioritization due to increased traffic volume with the MCP's entry. The second one captures the NCP 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

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

where $\mu > 1 + \lambda$ is assumed in the model.

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 $\Pi_n(\phi,\mu) = \beta\left[V - W_n(\phi,\mu)\right]$ because there is no entry in the neutral network. With a priority service in the non-neutral network, their joint profits are given by $\Pi_d(h_d^*(\mu),\mu) = v - kw_d(h_d^*(\mu),\mu) - c(h_d^*(\mu)) + \beta[V - W_d(h_d^*(\mu),\mu)]$. The change in joint profits due to introduction of the prioritization can be written as follows:[29]

$$\Delta\Pi^E(\mu,\beta) \equiv \Pi_d(h_d^*(\mu),\mu,\beta) - \Pi_n(\phi,\mu,\beta) = v - kw_d(h_d^*(\mu),\mu) - c(h_d^*(\mu)) - \beta\Delta W, \quad (28)$$

where the superscript E in $\Delta\Pi^E$ denotes that we consider the extensive margin case. As

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

(28) clearly shows, the MCP's entry generates the additional value of $v - kw_d(h_d^*(\mu), \mu)$ net of the MCP's investment cost $c(h_d^*(\mu))$ whereas the ISP must bear the loss of $\beta \Delta W$ due to its negative effects on NCPs. Recalling $\Delta \Pi^E(\mu, \beta = 1)$ is equal to the change in social welfare from the MCP's entry, one can immediately see that any private incentives to introduce a prioritized service under $\beta < 1$ exceeds the socially optimal incentives. Specifically, the discrepancy between the private and social incentives is measured by $(1 - \beta)\Delta W\ (\mu)$, which is inversely related to β. If $\beta = 1$, the ISP completely internalizes the effects on consumers and NCPs, with the private and social incentives coinciding. Proposition 4 summarizes our findings in this section.

Proposition 4 *Consider the extensive margin case. Then, the paid prioritization can facilitate the entry of a congestion-sensitive content. This newly available content generates a positive surplus to the network. The ISP's private incentives to introduce a prioritized service under $\beta < 1$ exceeds the socially optimal incentives.*

The MCP's entry can be either welfare-increasing or welfare-decreasing. If we want to discuss this point rigorously, we need somewhat lengthy mathematical derivations but here let us briefly deliver the main insight.[30] First, $\Delta \Pi^E(\mu, \beta)$ always increases with μ for an arbitrary $\beta \in [0, 1]$ if the delay sensitivity parameter k is sufficiently large. This is because the social benefit of assigning a fast lane to the MCP's content increases with k, while the negative effect of the entry on NCPs' content remains constant for a given β. Second, the MCP's entry is less likely to occur for a greater β as the ISP's loss from the higher level of network congestion increases with β. Then, we find that for a small enough β the MCP decides to enter with prioritization but such entry is not socially efficient.

However, it is also possible that the MCP decides not to enter with prioritization even when the entry is socially desirable once we consider a different environment than the one on which Proposition 4 is based. For example, suppose that consumers are heterogeneous and some consumers enjoy a positive surplus from the MCP's content. As the ISP does not take such surplus into account when it decides whether to adopt the prioritization or not, it is possible that socially desirable prioritization may not be adopted by the ISP.

[30] Rigorous mathematical derivations are available from the authors upon request.

6 ISP's Capacity Choice

Thus far we focused on the MCP's QoS investment for a *given* ISP's network capacity. To study potential interplay between the ISP's capacity choice and the MCP's entry/investment decisions, we augment our model by letting the ISP choose its network capacity μ before all subsequent plays ensue. Let $C(\mu)$ denote the investment cost of capacity $\mu$with $C' > 0$ and $C'' > 0$.

Let us begin with a summary of our findings with intuition. Overall, we find that the extension of the ISP's capacity choice generates results that are consistent with those obtained from our baseline model with exogenous capacity. When the MCP's entry is warranted, the net neutrality can give the ISP a higher investment incentive than the non-neutrality does. This is because the ISP is willing to invest less in a discriminatory network in order to enhance its bargaining position to such that the MCP needs to purchase a prioritized serve for its profitable entry. Such a strategic reason does not exist under neutral networks. By contrast, when the ISP's network capacity is limited and hence the entry of the MCP becomes a binding issue, the non-neutrality gives a higher investment incentive compared to the neutrality. This is because in a non-neutral network the ISP can internalize more surplus generated by the MCP's entry via bargaining with the MCP, but such a rent extraction channel is not available in a neutral network.

6.1 The intensive margin case

Consider first the intensive margin case in which the MCP enters *even without* prioritization, which is relevant only if $\mu \geq \underline{\mu}_n(\alpha)$.[31] Note that the waiting time depends on the capacity μ not only directly but also indirectly through the MCP's investment $h_r^*(\mu)$. For concise notation, we define $w_r^*(\mu) \equiv w_r(h_r^*(\mu), \mu)$ and $W_r^*(\mu) \equiv W_r(h_r^*(\mu), \mu)$ where $h_r^*(\mu)$ denotes the MCP's optimal investment for a given μ in the network regime $r = n, d$. When the ISP increases its capacity μ, the following two effects arise. The direct effect $\frac{\partial w_n^*(\mu)}{\partial \mu}$ makes the waiting time decreasing, whereas the indirect effect measured by $\frac{\partial w_r^*(\mu)}{\partial h_r} \times \frac{\partial h_r^*}{\partial \mu}$ rather

[31] Recall that even though $\underline{\mu}_n$ depends on α, we simply write $\underline{\mu}_n$ instead of $\underline{\mu}_n(\alpha)$.

increases the waiting time because the investment $h_r^*(\mu)$ decreases with μ. While these two effects go in opposite direction and thus the overall effect is not determined per se, in this subsection we assume that the direct effect dominates the indirect effect:

$$\frac{dw_r^*(\mu)}{d\mu} < 0, \qquad \frac{dw_d^*(\mu)}{d\mu} < 0.$$

Under a non-neutral network, the ISP's objective is given by

$$\begin{aligned} &\frac{[v - kw_d^*(\mu) - c(h_d^*(\mu)) + \beta W_d^*(\mu)] - [v - kw_n^*(\mu) - c(h_n^*(\mu)) + \beta W_n^*(\mu)]}{2} \\ &+ \{\alpha\,[v - kw_n^*(\mu)] + \beta\,[V - W_n^*(\mu)]\} - C(\mu). \end{aligned} \tag{29}$$

The first term measures a half of the surplus created by the prioritization. The braced second term represents the default payoff that the ISP obtains without prioritization. The first-order condition[32] with respect to μ is given by

$$-k\left[\frac{1}{2}\frac{dw_d^*(\mu)}{d\mu} - \left(\frac{1}{2} - \alpha\right)\frac{dw_n^*(\mu)}{d\mu}\right] + \frac{1}{2}\left[c'\frac{dh_n^*(\mu)}{d\mu} - c'\frac{dh_d^*(\mu)}{d\mu}\right] - \frac{\beta}{2}\left[\frac{dW_d^*(\mu)}{d\mu} - \frac{dW_n^*(\mu)}{d\mu}\right] = C'(\mu). \tag{30}$$

Although the LHS of (30) depends on the parameters (k, α, β), in order to pin down key forces, let us focus on the first bracketed term and only on the direct effect of μ on waiting time, assuming $\alpha = 0$. Then, the bracketed first term is given by

$$\frac{\partial w_d(h_d^*(\mu), \mu)}{\partial \mu} - \frac{\partial w_n(h_n^*(\mu), \mu)}{\partial \mu} = -\frac{a_d^*(\mu)\lambda}{(\mu - a_d^*(\mu)\lambda)^2} + \frac{a_n^*(\mu)\lambda}{(\mu - 1 - a_n^*(\mu)\lambda)^2} \tag{31}$$

where $a_r^*(\mu) = \frac{1}{1+h_r^*(\mu)}$ and $a_d^*(\mu) > a_n^*(\mu)$. The expression in (31) is *strictly positive* as long as $a_d^*(\mu) - a_n^*(\mu)$ is small enough, which implies that the marginal benefit of investment is strictly negative. Therefore, in this case, the ISP has no incentive to invest in capacity beyond $\mu \geq \underline{\mu}_n(\alpha)$. This effect does not disappear if the investment cost function $c(h)$ is

[32] A unique capacity choice requires that the LHS of the following equation, which is the marginal benefit of the capacity expansion, is decreasing with μ. The subsequent analysis is only valid under the set of parameters that this reasonable presumption is not violated.

convex enough.[33] The intuition behind this result is simple. A marginal capacity investment is more effective in reducing the waiting time in the neutral network, w_n, than that of the prioritized line in the non-neutral network, w_d, because the congestion problem is more severe in the former than in the latter. Hence, the surplus created by prioritization decreases in μ.[34] A similar effect was obtained by Choi and Kim (2010).

In contrast, under a neutral network, the ISP chooses its capacity to maximize $\alpha[v - kw_n^*(\mu)] + \beta[V - W_n^*(\mu)]$. Thus, the optimal capacity is determined by the following equation:

$$-\alpha k \left[\frac{dw_n^*(\mu)}{d\mu}\right] - \beta \left[\frac{dW_n^*(\mu)}{d\mu}\right] = C'(\mu), \tag{32}$$

where the first bracketed term represents the waiting time change for the MCP's content and the second the one for the NCPs. As long as the direct effect dominates the indirect effect, the marginal benefit from investment (i.e., RHS of (32)) is strictly positive no matter the level of $\mu(\geq \underline{\mu}_n)$ unless $\alpha = \beta = 0$. This is quite in contrast with what happens under non-neutral networks in which the marginal benefit from investment can be negative.

6.2 The extensive margin case

Consider now the extensive margin case in which the MCP may enter or not depending on the ISP's capacity choice.

We start by analyzing a benchmark in which the MCP is assumed to never enter. The ISP's profit obtains from the NCPs' content only and is equal to $\beta[V - W_n(\phi, \mu)] - C(\mu)$ where ϕ denotes the MCP's absence. Hence, the optimal capacity is characterized by the first-order condition:

$$-\beta \frac{\partial W_n(\phi, \mu)}{\partial \mu} = C'(\mu). \tag{33}$$

The LHS of (33) measures the marginal revenue increase by extracting more surplus from

[33] More precisely, assume $c(h) = \frac{ch^2}{2}$ where c is a positive constant. We assume that c is large enough relative to all other parameters including k such that the change in the MCP's investment is negligible.

[34] By contrast, if the ISP's rent extraction is large enough such that $\alpha \geq 1/2$, the first bracketed term in (30) will be always negative instead of being positive and hence the ISP will have an incentive to invest in capacity.

consumers who end up experiencing less congestion in accessing the NCPs' content; the RHS is the marginal cost associated with the capacity expansion. A unique solution to (33), denoted by $\mu_n^{\phi}(\beta)$, can be derived because the marginal revenue decreases with μ while $C'(\mu)$ increases with μ.[35] We assume $\mu_n^{\phi} < \underline{\mu}_r$ for $r = n, d$. This implies that whenever the ISP induces the MCP to enter, it invests more than μ_n^{ϕ}. In this sense, the ISP's capacity choice is closely related to its incentive to induce the MCP's entry. Hence, in what follows, we analyze the ISP's incentive to induce the entry. Let $\mu_r^e(\geq \underline{\mu}_r)$ denote the ISP's optimal capacity choice conditional on inducing the entry for $r = n, d$; in the Appendix, we characterize $\mu_r^e(\alpha, \beta)$.

Note first that regardless of the network regime, the ISP's payoff conditional on inducing no entry is given by

$$\beta[V - W_n(\phi, \mu_n^{\phi})] - C(\mu_n^{\phi}). \tag{34}$$

Under the non-neutral network, the ISP would induce entry if and only if (34) is smaller than

$$\frac{v - kw_d^*(\mu_d^e) - c(h_d^*(\mu_d^e)) - \beta\left[W_d^*(\mu_d^e) - W_n(\phi, \mu_d^e)\right]}{2} + \beta\left[V - W_n(\phi, \mu_d^e)\right] - C(\mu_d^e). \tag{35}$$

Under the neutral network, the ISP would induce entry if and only if (34) is smaller than

$$\alpha\left[v - kw_n^*(\mu_r^e)\right] + \beta\left[V - W_n^*(\mu_r^e)\right] - C(\mu_r^e). \tag{36}$$

The main difference between (35) and (36) occurs due to the difference in how much the ISP captures from the MCP after enabling its entry, which is captured by each first term in (35) and (36). Under the non-neutral network, the ISP can capture a half of the surplus generated by the MCP's entry with prioritization. In contrast, under the neutral network, it can capture α fraction of the surplus generated without prioritization. Hence, as long as α is small enough, the ISP is more likely to induce the MCP's entry (and hence to invest beyond μ_n^{ϕ}) under the non-neutral network compared to the ISP under the neutral network.

[35] Although μ_n^{ϕ} depends on β, we use μ_n^{ϕ} instead of $\mu_n^{\phi}(\beta)$ for simplicity.

7 Conclusion

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

In this paper, we develop a theoretical model that characterizes the relative size of network capacity as a distinguishing feature to allow the entry of a congestion-sensitive content provider, 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 networks with greater QoS investments, but this comes at the cost of increasing total traffic volume. When the entry is not constrained by the network capacity, the prioritization relieves content providers of their burden of QoS investments and improves efficiency by allocating the higher speed lane to more congestion-sensitive content. However, smaller QoS investments may be detrimental to social welfare. Our insight is consistent even when we consider the ISP's incentive to invest in capacity.

Our paper thus can inform policy-makers of important factors that should be considered in the formulation of net neutrality regulation as a second-best policy. For instance, it can shed light on the FCC's 2010 Order that treated mobile network operators more leniently than fixed wireline network operators. Specifically, its first two rules, namely, (i) 'transparency' and (ii) 'no blocking' were commonly applied to both types of network operators, but the third rule (iii) 'no unreasonable discrimination' appertained only to fixed line operators (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." In addition, this asymmetric regulatory approach was in sharp contrast to the European approach that does not allow for any differential treatment of fixed and mobile networks.[36] The rationale given by the FCC for its differential treatment between fixed and mobile networks was that its lenient non-neutral treatment may facilitate the availability of innovative content and applications in the early-stage mobile network. Our analysis indicates that such a policy can make sense unless the entry of new content consume a substantial portion of the limited mobile network capacity and the surplus from new content is outweighed by the efficiency loss from the elevated congestion for other content. For fixed networks with large capacity, however, net neutrality regulation can be the optimal second-best policy if the concern for weakened *dynamic* incentives for QoS investment with the prioritized service looms large compared to potential *static* efficiency gains from better traffic management.

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. 2012. "Net Neutrality, Broadband Market Coverage and Innovation at the Edge." *Decision Sciences* 43(1):141-172.

[3] Bourreau, Marc; Frago Kourandi, and Tommaso Valletti. 2015. "Net Neutrality with Competing Internet Platforms." *Journal of Industrial Economics* 63(1): 30-73.

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

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

[6] Choi, Jay Pil; Doh-Shin Jeon and Byung-Cheol Kim. 2015. "Net Neutrality, Business Models, and Internet Interconnection." *American Economic Journal: Microeconomics* 7(3): 104-141.

[36] 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.

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

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

[9] Economist. Video in Demand, The Economist Technology Quarterly, December 6, 2014.

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

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

[12] Greenstein, Shane; Martin Peitz and Tommaso Valletti. 2016. "Net Neutrality: A Fast Lane to Understanding the Trade-Offs." *Journal of Economic Perspectives* 30(2): 127-50.

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

[14] Malone, Jacob; Aviv Nevo and Jonathan Williams. 2015. "A Snapshot of the Current State of Residential Broadband Networks." NET Institute Working Paper #15-06.

[15] Jullien, Bruno and Wilfried Sand-Zantman. 2015. "Internet Regulation, Two-Sided Pricing, and Sponsored Data." mimeo.

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

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

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

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

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

[21] 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.

[22] 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.

[23] Peitz, Martin and Florian Schuett. 2016. "Net Neutrality and Inflation of Traffic." *International Journal of Industrial Organization* 46:16-62.

[24] Reggiani, Carlo and Tommaso Valletti. 2016. "Net Neutrality and Innovation at the Core and at the Edge." *International Journal of Industrial Organization* 45: 16-27.

[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 A: Mathematical Proofs

Proof of Proposition 1

The proof is straightforward as the following inequalities establish:

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

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

Proof of Lemma 1

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

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

One can easily determine the signs of the denominator and the numerator of $\frac{\partial h_n}{\partial \mu}\Big|_{h_n=h_n^*}$:

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

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

which proves Lemma 1. ■

Characterization of $\mu_r^e(\alpha,\beta)$

- Under neutral networks, $\mu_n^e(\alpha,\beta)$ is determined by the maximum between the solution of (32) and $\underline{\mu}_n$.
- Under non-neutral networks, suppose first $\mu \in [\underline{\mu}_d, \underline{\mu}_n)$. Then, the MCP enters only with prioritization and hence the ISP chooses its capacity to maximize the following objective:

  $$\frac{1}{2}\left\{v - kw_d^*(\mu) - c(h_d^*(\mu)) + \beta\left[V - W_d^*(\mu)\right] - \beta[V - W_n(\phi,\mu)]\right\} + \beta\left[V - W_n(\phi,\mu)\right] - C(\mu).$$

  The first braced term represents the surplus created by prioritization. The second term, $\beta\left[V - W_n(\phi,\mu)\right]$, is the ISP's default payoff. The first-order condition with respect to μ is given by

  $$-\frac{k}{2}\frac{dw_d^*(\mu)}{d\mu} - \frac{1}{2}c'(h_d^*(\mu))\frac{dh_d^*(\mu)}{d\mu} - \frac{\beta}{2}\frac{dW_d^*(\mu)}{d\mu} - \frac{\beta}{2}\frac{dW_n(\phi,\mu)}{d\mu} = C'(\mu). \qquad (37)$$

  Let $\widehat{\mu}(\alpha,\beta)$ denote the solution to Equation (37). If $\widehat{\mu}(\alpha,\beta) \geq \underline{\mu}_n$, $\mu_d^e(\alpha,\beta)$ is equal to the solution of (30). Otherwise, $\mu_d^e(\alpha,\beta)$ is determined by the one generating the highest profit between $\min\left\{\widehat{\mu}(\alpha,\beta), \underline{\mu}_d\right\}$ and the solution of (30).