Peitz, Martin; Schuett, Florian

Working Paper

Net neutrality and inflation of traffic

Working Paper Series, No. 15-05

Provided in Cooperation with:
University of Mannheim, Department of Economics

Suggested Citation: Peitz, Martin; Schuett, Florian (2015) : Net neutrality and inflation of traffic, Working Paper Series, No. 15-05, University of Mannheim, Department of Economics, Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-375355

This Version is available at:
https://hdl.handle.net/10419/129587

UNIVERSITY OF MANNHEIM

University of Mannheim / Department of Economics

Working Paper Series

Net neutrality and inflation of traffic

Martin Peitz Florian Schuett

Working Paper 15-05

February 2015

Net neutrality and inflation of traffic*

Martin Peitz† Florian Schuett‡

First version: April 2013; this version: February 2015

Abstract

Under strict net neutrality Internet service providers (ISPs) are required to carry data without any differentiation and at no cost to the content provider. We provide a simple framework with a monopoly ISP to evaluate different net neutrality rules. Content differs in its sensitivity to delay. Content providers can use congestion control techniques to reduce delay for their content, but do not take into account the effect of their decisions on the aggregate volume of traffic. As a result, strict net neutrality often leads to socially inefficient traffic inflation. We show that piece-meal departures from net neutrality, such as transmission fees or prioritization based on sensitivity to delay, do not necessarily improve efficiency. However, allowing the ISP to introduce bandwidth tiering and charge for prioritized delivery can implement the efficient allocation.

Keywords: Net neutrality, network congestion, telecommunications, quality of service

JEL-classification: L12, L51, L86

*We thank Cédric Argenton, Jan Boone, Marc Bourreau, Joan Calzada, Dennis Gärtner, Axel Gautier, Dominik Grafenhofer, Martin Hellwig, Viktoria Kocsis, Jan Krämer, Jens Prüfer, Bert Willems, Gijsbert Zwart, seminar participants at the Max Planck Institute for Research on Collective Goods (Bonn), the University of Liège, and Tilburg University, as well as participants at the 2013 "Economics of ICT"-conference in Mannheim, the 2014 "Economics of ICT"-conference in Paris, the 2014 Workshop on "Economics of ICT" in Porto, the 2014 Florence School of Regulation Scientific Seminar on "Media and Telecommunications" in Florence, the 2014 International Industrial Organization Conference (IIOC) in Chicago, the 2014 "Jornadas de Economia Industrial" in Barcelona, and the Symposium in honor of Jean Tirole in The Hague for helpful comments. Martin Peitz gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (project PE 813/2-2).

†Department of Economics, University of Mannheim, D-68131 Mannheim, Germany. Email: martin.peitz@googlemail.com. Also affiliated with CEPR, CERRE, CESifo, MaCCI, and ZEW.

‡TILEC & CentER, Tilburg University. Postal address: Tilburg University, Department of Economics, PO Box 90153, 5000 LE Tilburg, Netherlands. Email: f.schuett@uvt.nl.

1 Introduction

The net neutrality debate has focused on the question whether users' ISPs are allowed to prioritize certain types of services, and to charge content providers for the delivery of traffic, possibly dependent on the type of content and the priority which is assigned to the data packets. The debate within economics has focused on allocative consequences of various net neutrality rules. Apart from vertical foreclosure concerns, possible inefficiencies in the regulated or unregulated market may be due to market power, external effects between content providers and users, as well as direct network externalities arising from congestion in the network. The present paper adds to this debate by considering the incentives of content providers to distort traffic volumes in a setting with a monopoly ISP. We show that, under some conditions, strict net neutrality leads to traffic inflation and a loss in social welfare compared to the first best, while the first best can be implemented in a regime with bandwidth tiering and prioritized delivery.

Our analysis is motivated by three observations. First, there are congestion issues on the Internet. The increase in high-bandwidth applications and content, combined with limited last-mile capacity, results in congestion during peak hours, leading to delay. This issue is of particular importance on mobile networks. Second, some content is more sensitive to delay than other content. Time-sensitive content includes voice and video telephony, online games, real-time video streaming, and certain cloud services; less time-sensitive content includes email, web browsing, and file sharing, where modest delays in transmission do not matter much. Third, and most importantly, certain techniques used to minimize delay – so called *congestion control techniques* – affect the volume of traffic on the network. Some of them work by creating additional traffic; these include forward-error-correction (FEC) schemes, used to protect video packets,[1] and multiple multicast trees to provide redundant paths. Roughly speaking, these techniques introduce redundancies which increase packet size but partially insure the sender against packet losses. Similarly, Google has been reported to have implemented a technique to preload YouTube video clips on a user's device before that user has pressed the play button, based on information it has about this user (see Economist, 2014). Since the user will not play all those preloaded clips, this tends to increase traffic. Other congestion control techniques reduce the traffic volume, for example by lowering the quality of the sender's product; alternatively, senders may use compression techniques. Several providers of over-the-top content (such as Netflix) are known to adjust the quality of their service to the risk of congestion.

From an economic point of view, the use of congestion control and compression techniques causes externalities in traffic generation. Congestion control techniques that create additional traffic reduce individual delay but increase aggregate congestion on the network. Techniques that reduce traffic volumes, including compression, reduce individual traffic (usually at a cost to the sender) but also decrease aggregate congestion. In either of these environments, private and social incentives may not be aligned. Inefficiencies may arise for two reasons: (1) misallocation of traffic and (2) traffic inflation. Under a strict version of net neutrality (best effort for all traffic, no prioritization, zero prices on the con-

[1] Skype has been reported to react to persistent packet losses by increasing packet size (De Cicco *et al.*, 2011).

tent side), the network essentially constitutes an unmanaged common property resource. Net neutrality therefore leads to excessive exploitation by CPs (traffic inflation). In addition, the symmetric treatment of time-sensitive and time-insensitive traffic is inefficient (misallocation of traffic). By charging for traffic and handling time-sensitive traffic with priority, the ISP can serve as the guardian of the common property resource. This possibly reduces redundancies and other sources of inflation and gives time-insensitive traffic lower priority, which increases the capacity effectively available for time-sensitive traffic.

In our formal framework, there may be one or two lanes of traffic. The speed with which traffic flows is endogenous and can be controlled by the ISP subject to the constraints imposed by the regulator. There are two types of content: time-sensitive content and time-insensitive content. Time-sensitive content must be delivered without delay for consumers to derive utility from it; for time-insensitive content, delay does not matter. The capacity (bandwidth) of the ISP's network is fixed and constitutes a bottleneck needed to reach consumers. We assume that the probability that a given packet arrives without delay depends on the ratio of bandwidth to total traffic. To obtain a simple, tractable setting, we postulate that content providers can enhance the likelihood of on-time delivery by sending packets more than once. This increases the probability that at least one packet arrives on time, but also increases total traffic, and hence network congestion. In an extension we also consider the use of congestion control techniques that reduce the individual volume of traffic.

The first-best allocation in this framework always involves prioritization of time-sensitive content, with the volume of traffic adjusted so as to avoid congestion. In a second-best world, where all content must be carried in a single transport class (best effort), some congestion arising from traffic inflation is generally optimal, as it increases delivery probabilities for time-sensitive content at the expense of time-insensitive content. We show that net neutrality regulation leads to an equilibrium level of traffic that generally exceeds the second-best level, as content providers fail to internalize the effect of their own traffic on the overall network congestion.

We consider several departures from the above net neutrality rules – namely, deep packet inspection, transmission fees, and bandwidth tiering – and show that they can increase efficiency. Deep packet inspection allows the ISP to distinguish different types of content and prioritize time-sensitive content. Although this can lead to efficient outcomes in some cases, there are other cases in which time-sensitive CPs dissipate the reductions in delay by increasing traffic, and overall delivery probabilities may even be lower than under net neutrality.

When the ISP can charge a uniform transmission fee but cannot prioritize traffic, it sets the fee so as to price out congestion. The second-best traffic volume generally does involve some congestion, however, implying that transmission fees tend to be excessive. A price cap can implement the second-best efficient level.

Better outcomes can be achieved under bandwidth tiering. If the ISP can route traffic through two tiers – a fast lane and a slow lane – and charge differentiated fees for these tiers, the fee structure that maximizes the ISP's profit also leads to efficiency, as it implements the first-best allocation.

Related Literature. Our paper draws on the old literature on common property

resources and on recent work on information congestion (Van Zandt, 2004, and, more closely related, Anderson and De Palma, 2009). It also links to work on gatekeepers on the Internet. Anderson and De Palma show, among other things, that a monopoly gatekeeper completely prices out congestion. In their setting, the gatekeeper sets a uniform price for all incoming traffic, which allows to restrict traffic to the capacity of consumers to process information. In our context, it is not the limited processing ability of consumers, but the limited capacity of the network or, more precisely, of switches and interconnection points, which limits the pass-on of information. In contrast to previous work on information congestion, in response to the regulatory intervention in telecommunications markets, we draw a richer picture of the instruments available to the ISP as the gatekeeper. We also show that monopoly pricing is efficient in some regimes but not in others.

The paper contributes to the literature on net neutrality (see, e.g., Hermalin and Katz, 2007; Economides and Tåg, 2012; Choi and Kim, 2010; Cheng *et al.*, 2011; Economides and Hermalin, 2012; Jullien and Sand-Zantman, 2013; Bourreau *et al.*, 2014; Kourandi *et al.* , 2014). We borrow from Economides and Hermalin (2012) the notion that delivery speed is related to the ratio of traffic to bandwidth. Like Choi and Kim (2010) and Krämer and Wiewiorra (2012), we provide a rationale for why prioritization and quality differentiation may be efficiency enhancing.[2]

Choi *et al.* (2014) consider heterogeneous content providers and allow for interconnection between competing ISPs. At an initial stage, ISPs agree on quality levels and interconnection fees. Then, absent net neutrality, competing ISPs set menus of delivery qualities and subscription prices on the content provider side and CPs make subscription decisions. Afterwards, prices on the consumer side are set and consumers make subscription decisions. In their model, competing ISPs agree on access charges and delivery qualities such that they behave like monopoly bottlenecks against CPs. Without net neutrality ISPs have more instruments to extract CPs' surplus because under net neutrality, they are forced to provide one level of quality for all CPs. In equilibrium without net neutrality regulation, ISPs may focus on extracting surplus on the content provider side, while they may focus on extracting consumer surplus under net neutrality. Welfare results are, however, less clear. More closely related, Choi *et al.* (2013) consider congestion externalities on the Internet. They investigate the interplay of prioritized delivery and quality of service (QoS) investments by content providers, such as improved compression technologies. They show that, given a small network capacity, prioritization can facilitate entry of high-bandwidth content with the negative side effect that congestion of other content increases. Given a large network capacity, entry is less of an issue and prioritization allows for a faster delivery of time-sensitive content which tends to increase welfare. However, content providers have less incentive to invest in quality of service. This suggests a differential treatment of traffic on mobile versus fixed networks.

Our paper can be seen as complementary to Choi *et al.* (2013). Unlike us, they model congestion using an M/M/1 queuing model. Furthermore, their setting is asymmetric in the sense that it features a single high-bandwidth CP who can invest in QoS improvements;

[2]Including network investments may overturn the result in the model by Choi and Kim (2010). As they show, a monopoly ISP may invest more in capacity under net neutrality because expanding capacity reduces the CPs' willingness to pay for prioritization of their services.

all other CPs have low-bandwidth content and cannot invest in QoS. Yet the other CPs' content is sensitive to congestion as well, albeit less so than the major CP's. By contrast, in our model there is a continuum of time-sensitive CPs who are all symmetric in both their ability to use congestion control techniques and the sensitivity of their content to delay. Time-insensitive CPs cannot use congestion control but are unaffected by delay. Moreover, we provide a rich picture of the short-term effects of various regulatory regimes that are part of the net neutrality debate, whereas Choi *et al.* (2013) focus on the effects of prioritization and how they differ depending on the type of network.

The remainder of the paper is organized as follows. Section 2 lays out the model, introduces congestion and considers two efficiency benchmarks. Section 3 considers equilibrium traffic volumes under net neutrality and various other regimes. Section 4 discusses some extensions. In particular, it is shown that our main insights are robust in alternative settings in which firms can reduce individual traffic volume at a cost (through the use of compression techniques or quality reduction). Section 5 concludes. Proofs are relegated to the Appendix.

2 The model and efficiency

2.1 The model

We consider a market for Internet services which is intermediated by a monopoly ISP delivering content from content providers to users. There are thus three types of actors: consumers, content providers, and the monopoly ISP. Consumers decide on subscription and the purchase and use of content; content providers sell their content to consumers and decide on the intensity of use of the Internet and possibly the type of contract offered by the ISP. Consumers are homogeneous with respect to content and derive a utility u from each content provider whose content is delivered on time.

There is a continuum of content providers whose mass is normalized to 1. Content providers come in two categories. Content providers of category 1 offer time-sensitive content, while content providers of category 2 offer time-insensitive content. Content of category 1 arrives "in good order" with probability γ, which depends on the capacity of the network, on the decision of the content provider in question about how to deliver the content, and on the total volume of traffic. Content providers of category 2 are not constrained by the limited capacity and their content is delivered with probability 1 since their delivery can be delayed to a moment in which there is no congestion in the network. A fraction μ of content providers is of category 1, while the remaining fraction $1-\mu$ is of category 2. This is arguably the simplest way to model heterogeneity between content providers. The heterogeneity reflects the fact that some types of content such as live digital television and video telephony are highly time-sensitive, while other types of content such as email and delayed on-demand movies and most streaming services are less time-sensitive. There is not much loss if email and delayed on-demand movies arrive a bit later, and most streaming services can be buffered and thus do not require immediate delivery from the point of view of consumers. Implicit in our model is that traffic volumes vary over time with the feature that there are always periods of spare capacity during

which time-insensitive content can be delivered without any loss of value.

The monopoly ISP offers subscriptions to consumers and, depending on the regime it is subject to, may offer contracts to content providers. In our setting the capacity of the ISP is given. Thus, an excessive use by content providers may lead to delays and a deterioration of the surplus consumers derive from time-sensitive content. More specifically, a content provider with time-sensitive content may increase its probability of being delivered in time, γ, by sending its content more than once.

The network may be congested, which depends on how content is treated by the ISP and how much content is sent by content providers. Network capacity constitutes a common property resource. The contribution of our base model to the net neutrality debate is to allow content providers to inflate traffic in order to increase their probability of successful delivery; the traffic volume of CP i is denoted by α_i, the total volume of traffic by A. The following subsection will specify the behavior of content providers and derive the delivery probability γ.

Motivated by the net neutrality discussion, we will consider the following regulatory regimes:

- regime 1: strict net neutrality (only fast lane);
- regime 2: deep packet inspection (fast lane and slow lane, with priority according to needs for speed);
- regime 3: uniform pricing on the content provider side (only fast lane, but at a price);
- regime 4: regulated tiering with zero pricing restriction for non-prioritized packages (fast lane and slow lane, use of slow lane free);
- regime 5: unregulated tiering without price restrictions (fast lane and slow lane, payments depending on lane).

Regime 1 is currently largely in place due to the historical development of the Internet if one abstracts from content delivery networks.[3] Regime 2 is partly practiced with respect to TV streaming services and VoIP. Regime 4 is foreseen in regulation e.g. in the European Union. Regimes 3 and 5 are currently not part of the policy debate, but appear natural possibilities in a two-sided market setting.

The timing of events is as follows:

1. ISP announces subscription price s and transmission fee t per unit of content, which may be conditioned on priority classes.
2. CPs decide whether to be active and choose p_i and α_i.
3. Consumers choose whether to buy Internet access from ISP at subscription price s and which content to request.

[3]This is true, in particular, under the FCC's new net neutrality rules, which prohibit prioritization agreements between ISPs and CPs.

4. Content of CP i is delivered to consumers on time with probability $\gamma(\alpha_i, A)$. For each requested unit delivered on time consumers pay p_i to CP i; for each unit of traffic carried CPs pay t to the ISP (possibly conditional on priority classes). Consumers realize net utilities, CPs and ISP obtain profits.

We solve for subgame perfect Nash equilibria (SPNE) of the associated game.

2.2 Congestion

Since we consider a market with homogeneous viewers who have unit demand for each content i and whose valuation for each such unit is u, each content provider i will set $p_i = u$, which is collected only if the content reaches the consumer (which happens with probability γ).[4] The profit of a time-sensitive content provider is $\gamma(\alpha_i, A)u - k\alpha_i$. To isolate the effect of redundancies and multiple routes, we consider the stylized situation in which a content provider has to deliver a single packet. We assume that the probability that a given packet is delivered on time is equal to the ratio between the ISP's bandwidth and the total traffic A carried on the network. Sending a packet several times increases the probability that at least one packet arrives on time. Here, packets are perfect substitutes in the sense that the consumers' utility is the same if the content is delivered once or twice on time. Let B denote the ISP's available bandwidth (or network capacity). The probability of reaching a consumer when sending a package α_i times is

$$\gamma(\alpha_i, A) = \delta(A) \sum_{\tau=1}^{\alpha_i} (1 - \delta(A))^{\tau-1} = 1 - (1 - \delta(A))^{\alpha_i}, \tag{1}$$

where

$$\delta(A) = \min \left\{ \frac{B}{A}, 1 \right\}. \tag{2}$$

We distinguish between two systems of content delivery: a one-tiered system, in which all traffic is routed according to the best-effort principle, and a two-tiered system, in which some traffic is prioritized in times of bandwidth shortage. In a one-tiered system $A = \mu\alpha + 1 - \mu$, where $\alpha \equiv (\int_0^\mu \alpha_i di)/\mu$ is the average number of packets sent by time-sensitive CPs. By contrast, in a two-tiered system time-sensitive traffic is prioritized and $A = \mu\alpha$.

Suppose that each content provider can send a package once, twice, or not at all, i.e., $\alpha_i \in \{0, 1, 2\}$.[5] Assume moreover that $B < 1$, which implies that in a one-tiered system, if each CP sends one packet (so $A = 1$), not all time-sensitive content can be delivered on time.

[4] To not further increase the number of parameters, the value u is assumed to be independent of the type of traffic. Clearly, introducing different values of u depending on the type would affect the allocation of capacity between the two types of content. This applies to the equilibrium capacity allocation as well as the capacity allocation in the first-best and second-best benchmark.

[5] Sending a package twice can be interpreted as including redundancies, even though in practice redundancies tend to increase the traffic volume by less than 100 %.

At stage 4, if consumers have purchased Internet access, they consume all content for which $u \geq p_i$ (presuming that the payment is only made if the delivery occurs on time). Suppose that a fraction λ_n of time-sensitive CPs, $n = 0, 1, 2$, chooses $\alpha_i = n$. Then, consumers purchase Internet access if and only if

$$\int_0^{\mu} [\lambda_2 \gamma(2, A) + \lambda_1 \gamma(1, A)](u - p_i)di + \int_{\mu}^{1} (u - p_i)di \geq s.$$

Since, in SPNE, $p_i = u$ for all i, this condition becomes $s \leq 0$. At stage 2, the ISP thus chooses $s = 0$. This implies that if ISPs can only charge on the consumer side, content providers absorb all the surplus generated from delivering content and the monopoly ISP will make zero revenues.[6]

2.3 Efficiency: first-best and second-best traffic volumes

We begin by considering two benchmarks. In the stylized environment we study, time-insensitive content does not need to be delivered on time for consumers to derive utility from it. This implies that the *first best* always involves prioritization of time-sensitive content, i.e., content delivery is two-tiered and the probability of delivery for a packet that is sent α_i times is $\gamma(\alpha_i, \mu\alpha)$. We also consider a *second best* world in which all content has to be routed through a single tier according to a best-effort principle; the probability of delivery for a packet that is sent α_i times is then given by $\gamma(\alpha_i, \mu\alpha + 1 - \mu)$.

Total surplus in the market for time-insensitive content is always equal to $(1-\mu)(u-k)$, independent of the number of tiers and the traffic volume. In the market for time-sensitive content, total surplus as a function of α is given by[7]

$$W(\alpha) = \begin{cases} u\alpha\gamma(1, A) - \alpha k & \text{for } \alpha \in [0, 1] \\ u\left[(\alpha - 1)\gamma(2, A) + (2 - \alpha)\gamma(1, A)\right] - \alpha k & \text{for } \alpha \in (1, 2]. \end{cases} \tag{3}$$

To understand the first line, note that when a share λ_1 of time sensitive CPs choose $\alpha_i = 1$ and a share λ_0 choose $\alpha_i = 0$, then $\alpha = \lambda_1$. To understand the second line, observe that when a share λ_1 of time-sensitive CPs choose $\alpha_i = 1$ and a share $\lambda_2 = 1 - \lambda_1$ choose $\alpha_i = 2$, then $\alpha = \lambda_1 + 2(1 - \lambda_1) = 2 - \lambda_1$. Thus, we can replace λ_1 by $2 - \alpha$ and λ_2 by $\alpha - 1$.

Let $\hat{\alpha}_{dp}$ denote the level of traffic in a two-tiered system above which the delivery probability falls below 1, i.e., $\hat{\alpha}_{dp}$ is such that $\delta(\mu\alpha) = 1$ for $\alpha \leq \hat{\alpha}_{dp}$ and $\delta(\mu\alpha) < 1$ for $\alpha > \hat{\alpha}_{dp}$. Similarly, let $\hat{\alpha}_{nn}$ denote the level of traffic in a one-tiered system above which the delivery probability drops below 1. (The reason for the use of the subscripts dp and nn will become clear below.) We have that $\hat{\alpha}_{dp} = B/\mu$ and $\hat{\alpha}_{nn} = \max\{0, (B - (1 - \mu))/\mu\}$. If the

[6] While we restrict our analysis to fixed capacity of the ISP, under net neutrality, an immediate consequence of this finding is that in this admittedly extreme setting the ISP has no strict incentive to increase capacity even if expanding capacity is costless.

[7] The function W reflects the fact that it can never be socially optimal to have CPs randomize between 0 and 2 packages. Consider for example a situation in which all CPs send 1 package. One may wonder whether it can be optimal to have some send 2 packages instead, and others zero, while leaving α unchanged. This is not the case because the increase in probability of delivery for those sending 2 packages is less than the decrease for those sending 0: $\gamma(2, A) - \gamma(1, A) < \gamma(1, A) \Leftrightarrow \delta(A)(1 - \delta(A)) < \delta(A)$.

traffic volume is less than $\hat{\alpha}$, then all content is delivered on time; otherwise some content is delayed. It is readily seen that $\hat{\alpha}_{dp} \geq \hat{\alpha}_{nn}$: when only time-sensitive content is carried, the volume needed to cause congestion is larger. The following lemmas characterize first-best and second-best traffic volumes, respectively. They provide natural benchmarks to compare equilibrium outcomes with in the various regimes considered below.

Throughout the paper we will refer to situations with $\alpha \in [0, 1)$ as *partial availability* and to situations with $\alpha = 1$ as *full availability*. This relates to whether or not all time-sensitive content is available to consumers. Similarly, we will refer to situations with $\alpha \in (1, 2)$ as *partial duplication* and to situations with $\alpha = 2$ as *full duplication*, which relates to whether some or all time-sensitive CPs send their content twice.[8]

Lemma 1 *The first-best traffic volume α^{FB} is such that there is no congestion and no duplication, i.e., each CP's content is sent at most once:*

$$\alpha^{FB} = \begin{cases} \hat{\alpha}_{dp} & \text{if } B < \mu \quad \textit{(partial availability)} \\ 1 & \text{if } B \geq \mu \quad \textit{(full availability).} \end{cases}$$

According to Lemma 1, the first-best level of traffic always avoids congestion. A social planner prefers a situation where all available content is delivered on time but some content is unavailable to a situation where more content is available but some of it delivered with delay. The intuition for this result is that, for $\alpha \geq \hat{\alpha}_{dp}$, the elasticity (in absolute value) of the delivery probability δ equals one:

$$-\frac{d\delta/d\alpha}{\delta/\alpha} = \frac{\mu\alpha}{B}\delta(\mu\alpha) = 1.$$

This implies that increasing α beyond $\hat{\alpha}_{dp}$ leaves the amount of time-sensitive content delivered on time – and thus gross consumer surplus – unchanged (i.e., $\alpha\delta(\mu\alpha)$ is invariant with respect to α). The increase in available content is exactly offset by a decrease in delivery probability. While it has no effect on consumer surplus, the increase raises cost (αk) and is therefore undesirable from a total surplus perspective.

Let us now determine the efficient allocation under the constraint that all traffic is routed according to the best-effort principle and that the traffic volume of time-insensitive content is given. To characterize the second-best level of traffic, let $w(\delta) \equiv \delta^2(B + 1 - \mu - 2\delta)/B$ and $\delta_{\max} \equiv \arg\max_{B/(1+\mu) \leq \delta \leq B} w(\delta)$.

Lemma 2 *The second-best traffic volume α^{SB} may involve congestion and duplication: there exists $\hat{k} \in [\min\{uw(B/(1+\mu)), uB(1-\mu^2-B)/(1+\mu)^2\}, uw(\delta_{max})]$ such that,*

1. *for $k/u \geq \min\{(1-\mu)/B, B/(1-\mu)\}$, $\alpha^{SB} = \hat{\alpha}_{nn}$ (partial availability),*
2. *for $(1-\mu)B \leq k/u < \min\{(1-\mu)/B, B/(1-\mu)\}$, $\alpha^{SB} \in (\hat{\alpha}_{nn}, 1)$ solves*

$$\frac{1-\mu}{B}\left(\delta(\mu\alpha^{SB} + 1 - \mu)\right)^2 = \frac{k}{u} \quad \textit{(partial availability),} \tag{4}$$

[8] In situations with (partial or full) duplication, we have full availability.

3. *for* $\hat{k}/u \leq k/u < (1-\mu)B$, $\alpha^{SB} = 1$ *(full availability),*

4. *for* $\min\{w(B/(1+\mu)), B(1-\mu^2-B)/(1+\mu)^2\} \leq k/u < \hat{k}/u$, $\alpha^{SB} \in (1,2)$ *solves*

$$w\left(\delta\left(\mu\alpha^{SB}+1-\mu\right)\right) = \frac{k}{u} \quad \textit{(partial duplication),} \tag{5}$$

5. *for* $k/u \leq \min\{w(B/(1+\mu)), B(1-\mu^2-B)/(1+\mu)^2\}$, $\alpha^{SB} = 2$ *(full duplication).*

Lemma 2 shows that when all traffic needs to be routed according to a best-effort principle, the surplus-maximizing traffic volume may be so high as to cause congestion on the network; moreover, the planner may want to send time-sensitive content more than once. This is in contrast with the result of Lemma 1, showing that when time-sensitive content can be prioritized, the planner avoids congestion and duplication. Here, as the cost k of sending packets decreases, the optimal volume of traffic tends to increase. This result can again be related to the elasticity of the delivery probability:

$$-\frac{d\delta/d\alpha}{\delta/\alpha} = \frac{\mu\alpha}{B}\delta(\mu\alpha+1-\mu) = \frac{\mu\alpha}{\mu\alpha+1-\mu} < 1.$$

That is, raising α beyond $\hat{\alpha}_{nn}$ leads to an increase in the amount of time-sensitive content delivered without delay.

The intuition for this result is that part of the congestion caused by increasing traffic above $\hat{\alpha}_{nn}$ is borne by time-insensitive content. By definition, time-insensitive content can be delayed without reducing consumer surplus. Although sending more time-sensitive traffic creates congestion, part of this comes at the expense of time-insensitive content, for which delay does not matter. This is worthwhile doing if k is sufficiently small.

To further illustrate, consider a hypothetical choice between two traffic volumes: $A = B$ (which corresponds to $\alpha = \hat{\alpha}_{nn}$) and $A = 1$ (which corresponds to $\alpha = 1$). With the first option, all content is delivered on time ($\delta(B) = 1$). With the second option, only a fraction B of time-sensitive content is delivered on time ($\delta(1) = B$). Thus aggregate surplus (gross of transmission costs) is uB with the first option and $u[B\mu + 1 - \mu] > uB$ with the second option, where the inequality is due to $B < 1$.

3 Market equilibrium

3.1 Net neutrality

In a regime of net neutrality all content is routed through a single tier and content providers do not make any payments to the consumers' ISPs. In this regime we characterize equilibrium traffic. We look for a symmetric equilibrium in which all time-sensitive CPs behave alike. This may involve randomizing between different $\alpha_i \in \{0, 1, 2\}$ (which is equivalent to fractions λ_n of time-sensitive CPs using pure strategies n). To begin, we make the following assumption:

Assumption 1 $k/u < B/(1-\mu)$.

This is a minimal assumption for the model to be interesting. Otherwise, it is not profitable for any time-sensitive CP to send a package even if all other time-sensitive CPs send zero packages.

Each time-sensitive CP compares its profit from sending the package once, $u\gamma(1,A)-k$, to the profit from sending it twice, $u\gamma(2,A)-2k$, or not at all (yielding zero), taking as given total traffic A. For the purposes of the following lemma, let $\bar{\delta} = \arg\max_{B/(1+\mu)\leq\delta\leq B} \delta(1-\delta)$.

Lemma 3 *Under net neutrality, depending on the parameters one or several symmetric and possibly degenerate mixed-strategy equilibria exist. The equilibrium traffic volume α^{nn} can be characterized as follows:*

1. *for $B < k/u < B/(1-\mu)$, there is a mixed-strategy equilibrium in which time-sensitive CPs randomize over $\alpha_i = 0$ (probability $1-\alpha^{nn}$) and $\alpha_i = 1$ (probability α^{nn}), where $\alpha^{nn} \in (\hat{\alpha}_{nn}, 1)$ solves*

$$\delta(\mu\alpha^{nn} + 1 - \mu) = \frac{k}{u} \qquad \textit{(partial availability)}, \tag{6}$$

2. *for $B(1-B) \leq k/u \leq B$, there is a pure-strategy equilibrium in which all CPs choose $\alpha_i = 1$ so that $\alpha^{nn} = 1$ (full availability),*

3. *for $k/u \leq B(1+\mu-B)/(1+\mu)^2$, there is a pure-strategy equilibrium in which all time-sensitive CPs choose $\alpha_i = 2$ so that $\alpha^{nn} = 2$ (full duplication),*

4. *for $\min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\} < k/u < \bar{\delta}(1-\bar{\delta})$, there are one or two mixed-strategy equilibria in which time-sensitive CPs randomize over $\alpha_i = 1$ (probability $2-\alpha^{nn}$) and $\alpha_i = 2$ (probability $\alpha^{nn}-1$), where $\alpha^{nn} \in (1,2)$ solves*

$$\delta(\mu\alpha^{nn} + 1 - \mu)\left(1 - \delta(\mu\alpha^{nn} + 1 - \mu)\right) = \frac{k}{u} \qquad \textit{(partial duplication)}. \tag{7}$$

No other symmetric equilibrium exists.

According to Lemma 3, for a given value of k/u, it is possible that there are multiple equilibria. There can be multiple pure-strategy equilibria: for some parameter values, both $\alpha^{nn} = 1$ and $\alpha^{nn} = 2$ form an equilibrium (namely, if $B(1+\mu-B)/(1+\mu^2) > B(1-B)$). There can also be multiple mixed-strategy equilibria: noting that, in general, $\delta(1-\delta)$ is inverse U-shaped, with a maximum at $\delta = 1/2$, we conclude that unless $1/2 \notin (B/(1+\mu), B)$, the equation $\delta(A)(1-\delta(A)) = k/u$ has two solutions, corresponding to two different mixed-strategy equilibria $\alpha^{nn} \in (1,2)$. Finally, there can be situations with (at least) one pure-strategy equilibrium and (at least) one mixed-strategy equilibrium. Figure 1 depicts the case where there is a unique equilibrium for all values of k/u, and α^{nn} decreases (weakly) with k/u over the whole range. Figure 2 depicts the case where for $k/u \in (B(1-B), B(1+\mu-B)/(1+\mu)^2)$, there are two pure-strategy equilibria ($\alpha^{nn} = 1$ and $\alpha^{nn} = 2$) as well as a mixed-strategy equilibrium with $\alpha^{nn} \in (1,2)$.

Drawing on Lemmas 2 and 3, the following proposition compares the equilibrium traffic under net neutrality with the traffic volume that is second-best efficient.

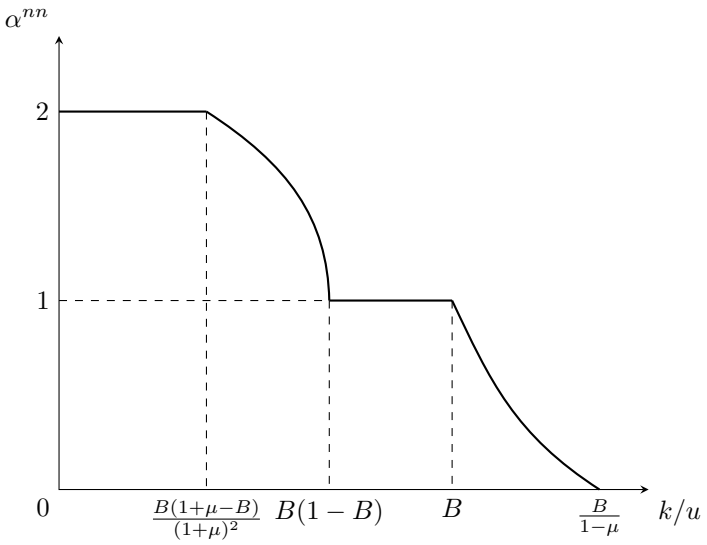

Figure 1: The equilibrium under net neutrality when $B < 1/2$

Proposition 1 *The equilibrium level of traffic under net neutrality always exceeds the second-best level: $\alpha^{nn} \geq \alpha^{SB}$, with strict inequality for at least part of the parameter space.*

According to Proposition 1, net neutrality generates inflation of traffic, leading to excessive congestion of the network. Time-sensitive CPs do not internalize the effect of the data they send on overall traffic, and therefore choose to send more than the socially optimal number of packets.

3.2 Uniform transmission fee

Suppose that the ISP routes all traffic according to a best-efforts principle (no prioritization), but charges content providers a uniform transmission fee t per unit of traffic it carries on its network. Type-1 (time-sensitive) CPs choose $\alpha_i \in \{0, 1, 2\}$ to maximize

$$\gamma(\alpha_i, A)u - \alpha_i(k + t),$$

where $A = \mu\alpha + 1 - \mu$. Thus, for given t, the equilibrium is the same as under net neutrality (see Subsection 3.1) replacing k by $k + t$, and the traffic from time-sensitive CPs $\alpha(t)$ facing the ISP for a given t is equal to the corresponding equilibrium traffic. Because of multiplicity of equilibria, we need to specify which equilibrium is selected for each possible t in order for the ISP's problem to be well defined. In what follows, we will assume that whenever there are multiple equilibria, the one with the highest traffic volume is selected. This is the most favorable selection rule for the ISP. We will then show that despite this favorable rule, the ISP will always choose a transmission fee that prevents

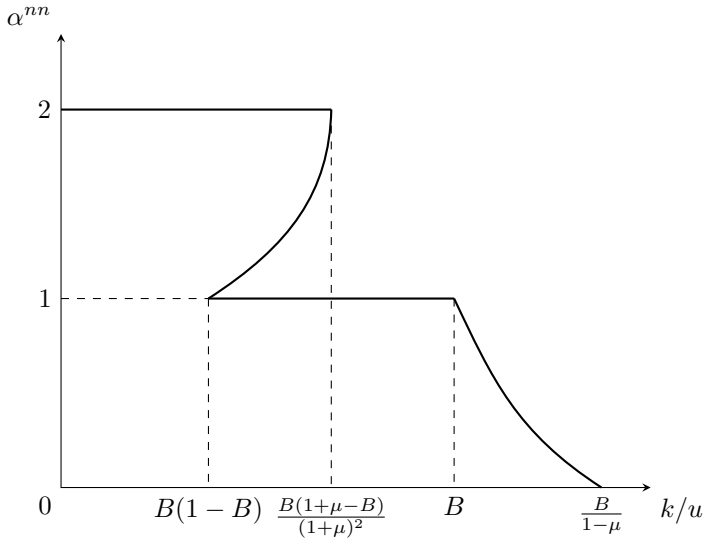

Figure 2: The equilibrium under net neutrality when $B/(1+\mu) > 1/2$

congestion. The inverse demand is given by

$$t(\alpha) = \begin{cases} u - k & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{nn} \\ u\delta(\mu\alpha + 1 - \mu) - k & \text{for } \hat{\alpha}_{nn} < \alpha \leq 1 \\ u\bar{\delta}(1 - \bar{\delta}) - k & \text{for } 1 < \alpha \leq \tilde{\alpha}_{nn} \\ u\delta(\mu\alpha + 1 - \mu)(1 - \delta(\mu\alpha + 1 - \mu)) - k & \text{for } \tilde{\alpha}_{nn} < \alpha \leq 2, \end{cases} \tag{8}$$

where $\tilde{\alpha}_{nn} = \max\{1, (2B - (1-\mu))/\mu\}$. The ISP's problem is

$$\max_{0 \leq \alpha \leq 2} t(\alpha)(\mu\alpha + 1 - \mu).$$

Using (8), we can compute

$$t'(\alpha) = \begin{cases} 0 & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{nn} \\ -\frac{\mu}{B} u\delta^2(\mu\alpha + 1 - \mu) & \text{for } \hat{\alpha}_{nn} < \alpha \leq 1 \\ 0 & \text{for } 1 < \alpha \leq \tilde{\alpha}_{nn} \\ -\frac{\mu}{B} u\delta^2(\mu\alpha + 1 - \mu)(1 - 2\delta(\mu\alpha + 1 - \mu)) & \text{for } \tilde{\alpha}_{nn} < \alpha \leq 2, \end{cases} \tag{9}$$

from which we deduce the ISP's marginal revenue, $MR(\alpha) = t'(\alpha)(\mu\alpha + 1 - \mu)) + \mu t(\alpha)$, noting that for $\mu\alpha + 1 - \mu \geq B$ (i.e., $\alpha \geq \hat{\alpha}_{nn}$), we have $(\mu\alpha + 1 - \mu)/B = 1/\delta(\mu\alpha + 1 - \mu)$:

$$MR(\alpha) = \begin{cases} \mu(u - k) & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{nn} \\ -\mu k & \text{for } \hat{\alpha}_{nn} < \alpha \leq 1 \\ \mu(u\bar{\delta}(1 - \bar{\delta}) - k) & \text{for } 1 < \alpha \leq \tilde{\alpha}_{nn} \\ \mu(u\delta^2(\mu\alpha + 1 - \mu) - k) & \text{for } \tilde{\alpha}_{nn} < \alpha \leq 2. \end{cases} \tag{10}$$

As the following proposition shows, the ISP always chooses $\alpha = \hat{\alpha}_{nn}$.

Proposition 2 *The transmission fee that maximizes the ISP's profit prices out congestion; i.e., t is such that $\alpha(t) = \hat{\alpha}_{nn}$.*

If the ISP is allowed to charge a uniform transmission fee, it responds to traffic inflation by charging a fee that eliminates congestion on its network. To see why, consider a hypothetical choice for the ISP between traffic volume $A = B$ (corresponding to $\alpha = \hat{\alpha}_{nn}$) and $A = 1$ (corresponding to $\alpha = 1$). With the first option, all content is delivered on time ($\delta(B) = 1$) but only a fraction of time-sensitive CPs are active. All active CPs are willing to pay $t(\hat{\alpha}_{nn}) = u - k$. The ISP's profit is $B(u - k)$. With the second option, all CPs are active and time-sensitive content is delivered with probability $\delta(1) = B$. Now, the *marginal* CP is time-sensitive and has willingness to pay $t(1) = Bu - k$. The ISP's profit is $Bu - k < B(u - k)$, where the inequality follows from $B < 1$. Although increasing traffic beyond $A = B$ increases the total amount of content delivered on time (see Section 2.3), unlike the planner the ISP only takes into account the effect on the marginal CP, who happens to be time-sensitive. The increase in the amount of time-sensitive content delivered on time does not compensate the decrease in surplus extracted from time-insensitive CPs.[9]

Recall that eliminating congestion entirely is generally not socially optimal in a single-tiered system. In other words, the fee chosen by the ISP tends to exceed the fee a social planner would choose. The profit-maximizing transmission fee implements the second-best level of traffic α^{SB} only if $k/u \geq \min\{(1-\mu)/B, B/(1-\mu)\}$. If instead $k/u < \min\{(1-\mu)/B, B/(1-\mu)\}$, then the profit-maximizing transmission fee leads to an inefficiently low level of traffic. Thus, it is not a priori clear whether allowing the ISP to charge a uniform transmission fee is better than net neutrality: while net neutrality leads to traffic inflation, freely set transmission fees lead to excessive contraction of traffic. The ISP may go as far as to price time-sensitive content out of the market (this happens if $B \leq 1 - \mu$).

The flip side of this argument is that a cap on the transmission fee can always implement the second-best efficient level of traffic. Thus, a departure from net neutrality that allows ISPs to set uniform transmission fees should be accompanied by a regulatory intervention in the form of a price cap.

3.3 Deep packet inspection

Deep packet inspection allows the ISP to identify whether a given packet contains time-sensitive or time-insensitive content. Therefore, under deep packet inspection, all available bandwidth in times of shortage can be allocated to time-sensitive content.[10] The probability that a given packet is delivered without delay is $\delta(\mu\alpha)$. Thus, time-sensitive content

[9]The amount of time-sensitive traffic delivered on time is $\mu\hat{\alpha}_{nn}$ when $A = B$ and μB when $A = 1$. The latter exceeds the former for $B < 1$ and $\mu \leq 1$. Consider the case $B > 1-\mu$ so that $\hat{\alpha}_{nn} > 0$. Then, as traffic goes from $A = B$ to $A = 1$, the ISP's profit from time-sensitive CPs increases by $(1-\mu)(1-B)u - k(1-B)/\mu$. However, the ISP's profit from time-insensitive CPs decreases by $(1-\mu)(1-B)u$. Hence, for any $k > 0$, the increase in profit from time-sensitive CPs is strictly less than the decrease in profit from time-insensitive CPs.

[10]We assume that the ISP carries out this prioritization although in the absence of transmission fees it does not benefit from this.

has a higher probability of being delivered on time for any given α. We assume that $B < 2\mu$, so that not all content is delivered on time if all time-sensitive CPs send their content twice. Letting $\bar{\bar{\delta}} = \arg\max_{B/2\mu \leq \delta \leq B/\mu} \delta(1-\delta)$, the following lemma characterizes the equilibrium under deep packet inspection.

Lemma 4 *Under deep packet inspection, depending on the parameters, one or several possibly degenerate symmetric mixed-strategy equilibria exist. The equilibrium traffic α^{dp} can be characterized as follows:*

1. *for $k/u > B/\mu$, there is a mixed-strategy equilibrium in which time-sensitive CPs randomize over $\alpha_i = 0$ (probability $1-\alpha^{dp}$) and $\alpha_i = 1$ (probability α^{dp}), where $\alpha^{dp} \in (0,1)$ solves*
$$\delta(\mu\alpha^{dp}) = \frac{k}{u} \qquad \textit{(partial availability)}, \tag{11}$$
2. *for $B(\mu-B)/\mu^2 \leq k/u \leq B/\mu$, there is a pure-strategy equilibrium in which all CPs choose $\alpha_i = 1$ so that $\alpha^{dp} = 1$ (full availability),*
3. *for $k/u \leq B(2\mu-B)/4\mu^2$, there is a pure-strategy equilibrium in which all time-sensitive CPs choose $\alpha_i = 2$ so that $\alpha^{dp} = 2$ (full duplication),*
4. *for $\min\{B(\mu-B)/\mu^2, B(2\mu-B)/4\mu^2\} < k/u < \bar{\bar{\delta}}(1-\bar{\bar{\delta}})$, there are one or two mixed-strategy equilibria in which time-sensitive CPs randomize over $\alpha_i = 1$ (probability $2-\alpha^{dp}$) and $\alpha_i = 2$ (probability $\alpha^{dp}-1$), where $\alpha^{dp} \in (1,2)$ solves*
$$\delta(\mu\alpha^{dp})\left(1-\delta(\mu\alpha^{dp})\right) = \frac{k}{u} \qquad \textit{(partial duplication)}. \tag{12}$$

No other symmetric equilibria exist.

Comparing the equilibrium level of traffic described in Lemma 4 with the first-best level identified in Lemma 1, the following proposition identifies a case in which deep packet inspection leads to efficiency.

Proposition 3 *If $B \geq \mu$, there exists an equilibrium under deep packet inspection in which the first-best level of traffic is transmitted irrespective of k/u; i.e., $\alpha^{dp} = \alpha^{FB} = 1$.*

Proposition 3 shows that deep packet inspection has the potential to alleviate traffic inflation. When $B \geq \mu$ and each CP sends one packet, then all content arrives on time. Thus, given the other CPs' behavior, no CP has an incentive to deviate and send more than one packet, regardless of k/u. Under net neutrality, even if $B \geq \mu$, the equilibrium may involve substantial inflation; in particular, full duplication ($\alpha^{nn} = 2$) may occur if k/u is low. In such a situation, introducing deep packet inspection can reduce traffic inflation and eliminate congestion, resulting in the efficient outcome (subject to multiplicity of equilibria and equilibrium selection).

A sufficient condition for deep packet inspection to improve efficiency is that $\alpha^{dp} \leq \alpha^{nn}$, but this is not necessarily the case. Deep packet inspection can actually lead time-sensitive CPs to increase the number of packets they send, at least partially dissipating

the efficiency gains from the prioritization of time-sensitive content. Suppose that CPs play a mixed-strategy equilibrium with $\alpha \in (0,1)$ under both net neutrality and deep packet inspection.[11] From (6) and (11), it must then be that $\delta(\mu\alpha^{nn} + 1 - \mu) = \delta(\mu\alpha^{dp})$ or, equivalently,

$$\mu\alpha^{nn} + 1 - \mu = \mu\alpha^{dp}.$$

Thus, the total traffic on the network (in times of shortage) is the same in both regimes. Intuitively, for CPs to be indifferent, the delivery probability for a given packet must be the same in both regimes, which requires higher volumes of time-sensitive traffic under deep packet inspection; i.e., $\alpha^{dp} > \alpha^{nn}$.

What we are ultimately interested in is whether deep packet inspection increases the overall amount of content delivered on time, which could be the case even if traffic increases. Consider again the situation where CPs play equivalent mixed-strategy equilibria. Even though total traffic (and thus the probability of delivery for a given packet) is the same under both regimes, there is a a larger proportion of time-sensitive CPs sending their packets ($\alpha^{dp} > \alpha^{nn}$). Thus the amount of delivered content is higher under deep packet inspection than under net neutrality in this case.

The above finding does not hold for all parameters; there are cases in which deep packet inspection does not increase delivery probabilities and even decreases them, as we show by example. Suppose that $\alpha^{nn} = 1$ and $\alpha^{dp} = 2$ are the respective equilibria under net neutrality and deep packet inspection; i.e., time-sensitive CPs generate twice as much traffic under deep packet inspection as under net neutrality. This situation can arise if $B < 2\mu$ and

$$B(1-B) \leq \frac{k}{u} \leq \frac{B}{2\mu}\left(1 - \frac{B}{2\mu}\right),$$

which to be possible, assuming that total traffic is greater under deep packet inspection (i.e., $2\mu > 1$), requires $2\mu/(1+2\mu) < B$. The probability of delivery under net neutrality is then $\gamma(1,1) = B$ while under deep packet inspection it is $\gamma(2,2\mu) = 1 - (1 - B/(2\mu))^2$. Thus, the probability of delivery is higher under net neutrality if $B > 1 - (1 - B/(2\mu))^2$ which is equivalent to $B < 4\mu(1-\mu)$. A value of B satisfying $2\mu/(1+2\mu) < B < 4\mu(1-\mu)$ exists if $\mu < (1+\sqrt{5})/4 \approx 0.81$. The following proposition summarizes the above finding.

Proposition 4 *There are parameter constellations such that the equilibrium probability of on-time delivery for time-sensitive content is lower under deep packet inspection than under net neutrality.*

While deep packet inspection may implement the efficient allocation, under some parameter constellations, deep packet inspection actually performs worse than (strict) net neutrality. Thus, deep packet inspection alone cannot reliably fix the problem of traffic inflation.

[11] This requires $B/\mu < k/u < B/(1-\mu)$.

3.4 Bandwidth tiering

Under bandwidth tiering the ISP can introduce two tiers of service (a fast, prioritized and a slower best-effort lane), and charge different transmission fees in each tier (regime 5). The ISP divides its bandwidth B into a slow lane B_s and a fast lane B_f such that $B_s + B_f = B$ and $B_f \geq B_s \geq 0$, where, as previously, B_f should be interpreted as the bandwidth allocated to priority service *in times of shortage* (and similarly for B_s and non-priority service). We start with the general case in which both t_s and t_f may be positive. Further below we look at regulated tiering, and, in particular, a zero-price rule for the slower lane (regime 4) before determining the solution under unregulated tiering.

Clearly, we must have $t_s \leq t_f$; otherwise, no one would ever choose the slow lane. Moreover, in the absence of minimum quality of service (QoS) requirements, the ISP has an incentive to make the slow lane as slow as possible: on the one hand, the willingness to pay of time-insensitive CPs is unaffected by B_s; on the other hand, the willingness to pay of time-sensitive CPs is increasing in B_f. Thus, the ISP will set $B_s = 0$ and $B_f = B$. (Note that this is efficient in our setup, as it does not mean that the slow lane will not deliver, but rather that the slow lane delivers with delay in times of high traffic.)

The ISP's problem is

$$\max_{t_s,t_f} (1-\mu)t_s + \mu\alpha(t_f)t_f \quad \text{subject to} \quad t_s \leq t_f,$$

where $\mu\alpha(t_f)$ is the demand for priority service when only time-sensitive content is transmitted via the fast lane. It is the same as the equilibrium traffic under deep packet inspection, $\mu\alpha^{dp}$, as derived in Lemma 4, after replacing k by $k + t_f$. Once again we assume that whenever there are multiple equilibria, the one with the highest traffic volume is selected. Under this selection rule, the inverse demand for traffic on the fast lane is

$$t_f(\alpha) = \begin{cases} u - k & \text{for } 0 \leq \alpha \leq \min\{1, \hat{\alpha}_{dp}\} \\ u\delta(\mu\alpha) - k & \text{for } \min\{1, \hat{\alpha}_{dp}\} < \alpha \leq 1 \\ u\bar{\bar{\delta}}(1-\bar{\bar{\delta}}) - k & \text{for } 1 < \alpha \leq \tilde{\alpha}_{dp} \\ u\delta(\mu\alpha)(1-\delta(\mu\alpha)) - k & \text{for } \tilde{\alpha}_{dp} < \alpha \leq 2, \end{cases} \tag{13}$$

where $\tilde{\alpha}_{dp} = \max\{1, 2B/\mu\}$.

The constraint $t_s \leq t_f$ must be binding at the ISP's profit maximum. Time-sensitive CPs will never switch to the slow lane since $B_s = 0$ means the probability of on-time delivery in times of high traffic is zero. Hence, $t_s = t_f$, allowing us to write the ISP's problem as

$$\max_{\alpha} (1-\mu)t_f(\alpha) + \mu\alpha t_f(\alpha),$$

from which we obtain marginal revenue

$$MR(\alpha) = t_f'(\alpha)(\mu\alpha + 1 - \mu) + \mu t_f(\alpha). \tag{14}$$

Using (13) and the fact that $\delta' = -(1/B)\delta^2$, we can compute

$$t'(\alpha) = \begin{cases} 0 & \text{for } 0 \leq \alpha \leq \min\{1, \hat{\alpha}_{dp}\} \\ -\frac{\mu}{B}u(\delta(\mu\alpha))^2 & \text{for } \min\{1, \hat{\alpha}_{dp}\} < \alpha \leq 1 \\ 0 & \text{for } 1 < \alpha \leq \tilde{\alpha}_{dp} \\ -\frac{\mu}{B}u(\delta(\mu\alpha))^2(1-2\delta(\mu\alpha)) & \text{for } \tilde{\alpha}_{dp} < \alpha \leq 2. \end{cases} \tag{15}$$

Noting that for $\mu\alpha \geq B$ (i.e., $\alpha \geq \hat{\alpha}_{dp}$), we have $\mu\alpha/B = 1/\delta(\mu\alpha)$, the ISP's marginal revenue is:

$$MR(\alpha) = \begin{cases} \mu(u-k) & \text{for } 0 \leq \alpha \leq \min\{1, \hat{\alpha}_{dp}\} \\ -\mu(k + (1-\mu)\frac{u}{B}\delta(\mu\alpha)) & \text{for } \min\{1, \hat{\alpha}_{dp}\} < \alpha \leq 1 \\ \mu\left(u\bar{\bar{\delta}}(1-\bar{\bar{\delta}}) - k\right) & \text{for } 1 < \alpha \leq \tilde{\alpha}_{dp} \\ \mu\left(u(\delta(\mu\alpha))^2\left(1 - \frac{(1-\mu)(1-2\delta(\mu\alpha))}{B}\right) - k\right) & \text{for } \tilde{\alpha}_{dp} < \alpha \leq 2. \end{cases} \tag{16}$$

Before deriving the optimal transmission fee on the fast lane under unregulated tiering we will first look at the case of regulated tiering (regime 4).

Regulated tiering. Consider a zero-price rule on the slow lane that restricts the ISP to charging $t_s = 0$. The ISP is free to choose t_f, as well as B_s and B_f. As previously, it will set $B_s = 0$ and $B_f = B$ to maximize the surplus that can be extracted from time-sensitive CPs. The ISP's profit is $\pi^{ISP} = \mu\alpha t_f(\alpha)$, where $t_f(\alpha)$ is defined in (13) and t'_f as derived in (15). From this we deduce marginal revenue $MR(\alpha) = \mu(\alpha t'_f(\alpha) + t_f(\alpha))$,

$$MR(\alpha) = \begin{cases} \mu(u-k) & \text{for } 0 \leq \alpha \leq \min\{1, \hat{\alpha}_{dp}\} \\ -\mu k & \text{for } \min\{1, \hat{\alpha}_{dp}\} < \alpha \leq 1 \\ \mu\left(u\bar{\bar{\delta}}\left(1-\bar{\bar{\delta}}\right) - k\right) & \text{for } 1 < \alpha \leq \tilde{\alpha}_{dp} \\ \mu\left(u(\delta(\mu\alpha))^2 - k\right) & \text{for } \tilde{\alpha}_{dp} < \alpha \leq 2. \end{cases} \tag{17}$$

The following lemma reports the profit-maximizing transmission fees under bandwidth tiering with and without regulatory restrictions on the price of the slow lane.

Proposition 5 *Irrespective of regulation, the profit-maximizing transmission fee on the fast lane prices out congestion, i.e., t_f is such that $\alpha = \min\{1, \hat{\alpha}_{dp}\}$. The profit-maximizing transmission fee on the slow lane, if unregulated, is $t_s = t_f$.*

This result is reminiscent of Anderson and De Palma (2009), where a monopoly gatekeeper prices out information congestion. For a better understanding, it is useful to look at the ISP's choice between implementing two levels of traffic on the fast lane, $A = B$ (i.e., $\alpha = B/\mu$) and $A = \mu$ (i.e., $\alpha = 1$), assuming $B < \mu$. With the first option, all content arrives on time but only a fraction of time-sensitive CPs are active. Active CPs are willing to pay $t_f(B/\mu) = u - k$, so the ISP's profit from the fast lane is $(B/\mu)(u-k)$. With the second option, all time-sensitive CPs are active but their content arrives on time only with probability B/μ. Thus their willingness to pay is $t_f(1) = (B/\mu)u - k$, and the ISP's profit from the fast lane is $(B/\mu)u - k < (B/\mu)(u-k)$ (since, by assumption, $B < \mu$). The intuition is that increasing α leaves the amount of content delivered on time unchanged, but in the second case more content is being sent so costs are higher.

Unregulated tiering. Proposition 5 shows that the ISP will prevent congestion on the network also under bandwidth tiering; this holds independently of regulatory restrictions on the price of the slow lane, t_s. If prices are unregulated the ISP will price the slow lane

exactly as (or just marginally below) the fast lane, so that time-insensitive CPs choose the slow lane and time-sensitive ones, for whom the slow lane is not an option, choose the fast lane.[12]

Comparing the equilibrium outcome when the ISP is allowed to charge for the fast lane to the first-best solution identified in Lemma 1, we see that the prices that maximize the ISP's profits implement the efficient solution: time-insensitive content is routed through the slow lane, time-sensitive content is routed through the fast lane, and the volume of traffic is at the efficient level: $\alpha = \min\{1, \hat{\alpha}_{dp}\}$. Unlike in the case of a uniform transmission fee, no regulatory intervention is required to ensure efficiency. Allowing the ISP to do bandwidth tiering and charge (at least) for the fast lane leads to the first-best allocation.

Note that in this simple model there is no efficiency rationale for implementing a minimum QoS requirement, i.e., imposing a lower bound $\underline{B}$ on the bandwidth allocated to the slow lane (so that $B_s \geq \underline{B}$).

4 Extensions

4.1 Congestion control techniques that reduce traffic

In our base model it is assumed that the congestion control technique available to content providers is such that individual delay can be reduced only at the expense of increasing the volume of traffic. In this subsection, we instead consider techniques that decrease the volume of traffic but have other drawbacks for the content provider: namely, compression and quality reduction. We modify the basic model by assuming that time-sensitive content providers have two packets of content to deliver, and that *both* must arrive on time for consumers to derive utility from the content. Time-insensitive CPs continue to send a fixed volume of traffic, which we set equal to one unit per CP. This reflects the idea that time-sensitive content is often more bandwidth-heavy as well (this is the case, e.g., for video telephony and online gaming). If a time-sensitive CP sends its content without compression and in high quality, the probability that both packets arrive on time is $\delta(A)^2$, and the CP's payoff is $u\delta(A)^2 - 2k$. In the following subsections, we consider compression and quality reduction as two alternative ways of trimming down the data volume of time-sensitive content and thereby enhancing the probability of on-time delivery.

4.1.1 Compression

Suppose that CPs can make use of a compression technology that reduces the number of packets required to transmit time-sensitive content from two to one. Time-sensitive CPs have to pay c per packet to use such a technology. With compression, the probability that the content arrives on time is thus $\delta(A)$, and the CP's payoff is $u\delta(A) - k - c$. Assume $c > k$ (otherwise sending one compressed packet would always be cheaper than sending

[12]The fact that both lanes are priced the same is an artefact of our somewhat extreme assumption that time-sensitive content is never delivered on time on the slow lane and that time-insensitive content does not benefit at all from faster delivery. In a more realistic setup, the result would be less extreme but similar in spirit.

two uncompressed packets) and $c < \min\{uB/(1-\mu), u\} - k$ (otherwise compression would never be profitable, even if no other time-sensitive CP were active).

Denote by λ_1 the fraction of time-sensitive CPs investing in compression and by λ_2 the fraction using uncompressed transmission. (Thus, $1 - \lambda_1 - \lambda_2$ gives the share of CPs remaining inactive.) The social planner solves

$$\max_{\lambda_1,\lambda_2} \lambda_1 \left(u\delta(A) - c - k\right) + \lambda_2 \left(u(\delta(A))^2 - 2k\right) \tag{18}$$

subject to $\lambda_1 + \lambda_2 \leq 1$, where $A = \mu(\lambda_1 + 2\lambda_2) + 1 - \mu$ in a one-tiered system and $A = \mu(\lambda_1 + 2\lambda_2)$ in a two-tiered system. A more insightful way of looking at this problem is the following: (a) fix δ and find the optimal combination of λ_1 and λ_2 for a given δ; (b) find the optimal δ. Formally, part (a) entails, in a one-tiered system,

$$\delta = \frac{B}{\mu(\lambda_1 + 2\lambda_2) + 1 - \mu},$$

or $\lambda_1 = (B/\delta - (1-\mu))/\mu - 2\lambda_2$. Substituting this into the objective, the optimal combination of λ_1 and λ_2 is obtained by solving

$$\max_{0\leq\lambda_2\leq(B/\delta-(1-\mu))/2\mu} \left(\frac{B/\delta - (1-\mu)}{\mu} - 2\lambda_2\right)(u\delta - c - k) + \lambda_2(u\delta^2 - 2k).$$

Differentiating with respect to λ_2 yields $2c - u\delta(2-\delta)$. Thus, for each $\delta \in [B/(1+\mu), 1]$, there exists a cutoff value $\hat{c} \equiv u\delta(1-\delta/2)$ such that $\left(\lambda_1 = 0, \lambda_2 = (B/\delta - (1-\mu))/2\mu\right)$ is optimal for $c > \hat{c}$ and $\left(\lambda_1 = \min\{(B/\delta - (1-\mu))/\mu, 1\}, \lambda_2 = 0\right)$ is optimal for $c < \hat{c}$. In words, for a given δ, the planner will use the same transmission technology for all CPs: if c is high, all content will be sent uncompressed, while if c is low, all content will be sent compressed. Note that the choice depends only on c and not on k. The intuition is that sending content uncompressed generates twice as much transmission costs ($2k$ versus k) but also twice as much traffic; to keep δ and thus traffic constant, half as much content can be sent as with compression.[13] Total transmission costs are thus the same and only the compression cost matters for the comparison.

Although fully characterizing the optimal policy is difficult, based on this insight we can derive a sufficient condition for the planner to use only compressed transmission at the second-best optimum. At $\delta = B/(1+\mu)$ (the lowest possible delivery probability), we have $\hat{c} = uB(1+\mu-B/2)/(1+\mu)^2$. Since $\delta(1-\delta/2)$ is increasing in δ, we conclude that if

$$c \leq uB(1+\mu-B/2)/(1+\mu)^2, \tag{19}$$

the optimal second-best policy necessarily involves all active time-sensitive CPs using compression technology ($\lambda_1 > \lambda_2 = 0$).

We now turn to equilibrium behavior. For simplicity we impose the following assumption:

[13] Strictly speaking, this is true only as long as $\alpha \geq \hat{\alpha}_{nn}$. For $\alpha < \hat{\alpha}_{nn}$, a marginal increase in traffic does not reduce the delivery probability. Here, however, this is irrelevant as $B < 1$. Thus, an outcome where all time-sensitive CPs use compression and $A < B$ cannot arise.

Assumption 2 *The transmission cost satisfies*

$$\frac{k}{u} \leq \frac{B^2}{2(1+\mu)^2}.$$

This implies that sending content uncompressed ($\alpha_i = 2$) is profitable even if all other time-sensitive CPs do the same. Thus, the only decision we have to consider is whether time-sensitive CPs use compression and not whether they are active. The superscript *ct* stands for compression technology.

Lemma 5 *Suppose that Assumption 2 holds. Then, under net neutrality, all time-sensitive CPs are active in equilibrium. The equilibrium use of compression technology is characterized as follows:*

1. *for $(c-k)/u \leq B(1-B)$, there is a pure-strategy equilibrium in which all time-sensitive CPs invest in compression ($\lambda_1^{ct} = 1, \lambda_2^{ct} = 0$),*
2. *for $(c-k)/u \geq B(1+\mu-B)/(1+\mu)^2$, there is a pure-strategy equilibrium in which all time-sensitive CPs choose uncompressed transmission ($\lambda_1^{ct} = 0, \lambda_2^{ct} = 1$),*
3. *for $\min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\} < (c-k)/u < \bar{\delta}(1-\bar{\delta})$, there are one or two mixed-strategy equilibria in which time-sensitive CPs randomize over compressed (probability λ_1^{ct}) and uncompressed (probability $1-\lambda_1^{ct}$) transmission, where $\lambda_1^{ct} \in (0,1)$ solves*

$$\delta(1+\mu(1-\lambda_1))\left(1-\delta(1+\mu(1-\lambda_1))\right) = \frac{c-k}{u}. \tag{20}$$

The following result extends Proposition 1 by showing that under net neutrality CPs tend to underinvest in compression technology.

Proposition 6 *Suppose that Assumption 2 holds. Then there exists a range of admissible values of $(c-k)/u$ such that, under net neutrality, all CPs use uncompressed transmission in equilibrium, while the second-best optimum calls for all active CPs to use compressed transmission.*

Proposition 6 is weaker than Proposition 1 in the sense that it only identifies a range of parameter values for which there is underinvestment in compression technology but does not show that there can never be overinvestment. The latter would require fully characterizing the optimal second-best policy, which is a complex task that we leave for future research.

Next we show that bandwidth tiering with a zero-price rule on the slow lane can solve the problem of underinvestment in compression. Assume that $B \leq \mu$ so that there is congestion even if all time-sensitive CPs use compression. We also impose the following assumption.

Assumption 3 *The compression cost satisfies*

$$c \leq u\frac{B}{2\mu}\left(1-\frac{B}{4\mu}\right). \tag{21}$$

This is the equivalent in a two-tiered system of condition (19), ensuring that the planner wants all time-sensitive CPs to use compression.

As in the baseline model, the ISP chooses to allocate bandwidth $B_f = B$ to the fast lane and $B_s = 0$ to the slow lane, charging a price $t_f \geq 0$ for the fast lane while the price for the slow lane t_s is exogenously set to zero. Letting $t_f(\alpha)$ denote the inverse demand for traffic on the fast lane, with $\alpha = \lambda_1 + 2\lambda_2$, the ISP solves

$$\max_{\alpha} \mu\alpha t_f(\alpha).$$

The demand for traffic is determined by the equilibrium of the compression-choice game we have just analyzed (see Lemma 5), replacing $\delta(1)$ by $\delta(\mu)$, $\delta(1+\mu)$ by $\delta(2\mu)$, and k by $k + t_f$. As before, in case of multiple equilibria we select the equilibrium with the largest demand for traffic.

For brevity we restrict attention to the case $B/\mu \geq 2/3$, implying that $\delta(2\mu)(1-\delta(2\mu)) \geq \delta(\mu)(1-\delta(\mu))$. CPs using compressed transmission earn $u\delta(\mu\alpha) - c - (k + t_f)$ while CPs using uncompressed transmission earn $u(\delta(\mu\alpha))^2 - 2(k + t_f)$. If the ISP charges $t_f = u\delta - k - c$, then all time-sensitive CPs prefer compressed over uncompressed transmission: compressed transmission yields zero, while uncompressed transmission yields $u\delta^2 - 2(k + t_f) = 2c - u\delta(2-\delta) < 0$, where the inequality follows from (21) and the fact that $\delta(1-\delta/2)$ is increasing in δ. Thus, inverse demand for the fast lane is[14]

$$t_f(\alpha) = \begin{cases} u-k-c & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ u\delta(\mu\alpha) - k - c & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ c - k - u\delta(\mu\alpha)\left(1-\delta(\mu\alpha)\right) & \text{for } \alpha \in (1, 2B/(2\mu - B)] \\ c - k - u\delta(2\mu)\left(1-\delta(2\mu)\right) & \text{for } \alpha \in (2B/(2\mu-B), 2], \end{cases}$$

owing to the fact that the equilibrium selected is $\alpha = 2$ for $(c - (k + t_f))/u \geq B(2\mu - B)/(4\mu^2)$, which happens for all α such that $\delta(\mu\alpha) \leq 1 - \delta(2\mu) \Leftrightarrow \alpha \geq 2B/(2\mu - B)$. From this we infer

$$t_f'(\alpha) = \begin{cases} 0 & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ -u\delta(\mu\alpha)/\alpha & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ -u\frac{\mu}{B}(\delta(\mu\alpha))^2\left(2\delta(\mu\alpha) - 1\right) & \text{for } \alpha \in (1, 2B/(2\mu - B)] \\ 0 & \text{for } \alpha \in (2B/(2\mu-B), 2]. \end{cases}$$

Putting both together, we obtain marginal revenue, $MR(\alpha) = \mu[\alpha t_f'(\alpha) + t_f(\alpha)]$, or

$$MR(\alpha) = \begin{cases} \mu[u-k-c] & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ -\mu[c+k] & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ \mu[c - k - u(\delta(\mu\alpha))^2] & \text{for } \alpha \in (1, 2B/(2\mu - B)] \\ \mu[c - k - u\delta(2\mu)(1-\delta(2\mu))] & \text{for } \alpha \in (2B/(2\mu-B), 2]. \end{cases}$$

[14]It can be verified that CPs' payoff is always nonnegative for these prices and traffic volumes. In particular, for $\alpha = 2$ so that $t_f = c - k - uB(2\mu - B)/(4\mu^2)$, a CP's payoff is $uB(4\mu - B)/(4\mu^2) - 2c \geq 0$, where the inequality follows from (21).

Proposition 7 *Suppose that Assumption 3 holds and* $2/3 \leq B/\mu \leq 1$. *Then, the profit-maximizing transmission fee on the fast lane prices out congestion and leads all time-sensitive CPs to use compression; i.e.,* t_f *is such that* $\alpha = \hat{\alpha}_{dp}$.

Proposition 7 shows that allowing the ISP to implement bandwidth tiering and charge for the fast lane can solve the problem of underinvestment in compression highlighted by Proposition 6.

4.1.2 Quality reduction

Suppose that instead of compressing their content, CPs have the option of reducing the quality of transmission (e.g., by using a lower resolution or a lower frame rate). More specifically, suppose that quality reduction cuts the volume of data that needs to be transmitted in half (one instead of two packets), but also decreases the surplus from consuming the content by a factor $\beta < 1$. Thus, a CP's profit from sending reduced-quality content is $\beta u\delta(A) - k$.

Let us again denote by λ_1 the fraction of time-sensitive CPs sending reduced-quality content and by λ_2 the fraction sending standard-quality content. Consider the social planner's problem expressed as a two-step procedure: (a) fix δ and find the optimal combination of λ_1 and λ_2 for a given δ; (b) find the optimal δ. In a one-tiered system, part (a) entails

$$\delta = \frac{B}{\mu(\lambda_1 + 2\lambda_2) + 1 - \mu},$$

or $\lambda_1 = (B/\delta - (1-\mu))/\mu - 2\lambda_2$. Substituting this into the objective, the optimal combination of λ_1 and λ_2 is obtained by solving

$$\max_{\lambda_2} \left(\frac{B/\delta - (1-\mu)}{\mu} - 2\lambda_2 \right) (\beta u\delta - k) + \lambda_2(u\delta^2 - 2k),$$

where $\lambda_2 \in [0, (B/\delta - (1-\mu))/2\mu]$. Differentiating with respect to λ_2 yields $-2(\beta u\delta - k) + u\delta^2 - 2k$. Thus, for each $\delta \in [B/(1+\mu), 1]$ there is a cutoff value $\hat{\beta} \equiv \delta/2$ above which it is optimal to send all content in reduced quality $\big(\lambda_1 = \min\{(B/\delta - (1-\mu))/\mu, 1\}, \lambda_2 = 0\big)$, and below which it is optimal to send all content in standard quality $\big(\lambda_1 = 0, \lambda_2 = (B/\delta - (1-\mu))/2\mu\big)$. In particular, for $\beta \geq 1/2$, it is second-best socially optimal that all active CPs send content in reduced quality.

The following lemma characterizes equilibrium behavior under net neutrality.

Lemma 6 *Suppose that* $\beta \geq 1/2$. *Under net neutrality, the equilibrium use of quality reduction is characterized as follows:*

1. *for* $k/u > \beta B$, *there is a mixed-strategy equilibrium in which time-sensitive CPs randomize over* $\alpha_i = 0$ *(probability* $1 - \lambda_1^{qr}$*) and* $\alpha_i = 1$ *(probability* λ_1^{qr}*), where* $\lambda_1^{qr} \in (0,1)$ *solves*

$$\beta\delta(\mu\lambda_1^{qr} + 1 - \mu) = \frac{k}{u}, \tag{22}$$

2. *for* $B(B-\beta) \leq k/u \leq \beta B$*, there is a pure-strategy equilibrium in which all time-sensitive CPs send content in reduced quality* ($\lambda_1^{qr} = 1, \lambda_2^{qr} = 0$)*,*

3. *for* $k/u \leq B(B-\beta(1+\mu))/(1+\mu)^2$*, there is a pure-strategy equilibrium in which all time-sensitive CPs send content in standard quality* ($\lambda_1^{qr} = 0, \lambda_2^{qr} = 1$)*,*

4. *for* $B(B-\beta(1+\mu))/(1+\mu)^2 < k/u < B(B-\beta)$*, there is a mixed-strategy equilibrium in which time-sensitive CPs randomize over reduced quality (probability* λ_1^{qr}*) and standard quality (probability* $\lambda_2^{qr} = 1 - \lambda_1^{qr}$*), where* $\lambda_1^{qr} \in (0,1)$ *solves*

$$\delta(1+\mu(1-\lambda_1))\left(\delta(1+\mu(1-\lambda_1)) - \beta\right) = \frac{k}{u}. \tag{23}$$

Putting together this result with the earlier observation on the optimal second-best policy, we can state the following.

Proposition 8 *Suppose that* $\beta \geq 1/2$*. Then, under net neutrality, for* $k/u < B(B-\beta)$*, at least some CPs use standard-quality transmission in equilibrium, while the second-best optimum calls for all active CPs to use reduced-quality transmission.*

The assumption that $\beta \geq 1/2$ means that the loss in utility from the quality reduction suffered by consumers is less than proportional to the reduction in traffic. Under this assumption, Proposition 8 shows that for sufficiently low costs of transmission the equilibrium is associated with insufficient quality reduction compared to what the social planner would choose.

Next we show that bandwidth tiering with a zero-price rule on the slow lane can solve the problem of insufficient quality reduction. Assume that $B \leq \mu$ so that there is congestion even if all time-sensitive CPs use quality reduction. Moreover, assume that $1/2 \leq \beta \leq B/\mu$. The first inequality ensures that the planner wants all time-sensitive CPs to use reduced-quality transmission, while the second inequality ensures that in the equilibrium under deep-packet inspection, some CPs use standard-quality transmission if k/u is low enough.

Again the ISP allocates bandwidth $B_f = B$ to the fast lane and $B_s = 0$ to the slow lane, charging a price $t_f \geq 0$ for the fast lane while on the slow lane the regulator imposes $t_s = 0$. Letting $t_f(\alpha)$ denote the inverse demand for traffic on the fast lane, with $\alpha = \lambda_1 + 2\lambda_2$, the ISP solves

$$\max_{\alpha} \mu\alpha t_f(\alpha).$$

The demand for traffic is determined by the equilibrium of the quality-choice game (see Lemma 6), replacing $\delta(1)$ by $\delta(\mu)$, $\delta(1+\mu)$ by $\delta(2\mu)$, and k by $k + t_f$.

The assumption that $\beta \geq 1/2$ implies that $\beta\delta \geq \delta(\delta - \beta)$ for all $\delta \leq 1$. Hence, for $(k+t_f)/u \geq (B/\mu)((B/\mu) - \beta)$, all active CPs use reduced-quality transmission. Inverse demand for the fast lane is

$$t_f(\alpha) = \begin{cases} \beta u - k & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ \beta u\delta(\mu\alpha) - k & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ \beta u\delta(\mu\alpha)\left(\delta(\mu\alpha) - \beta\right) - k & \text{for } \alpha \in (1, 2]. \end{cases}$$

From this we infer

$$t'_f(\alpha) = \begin{cases} 0 & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ -\beta u \frac{\mu}{B} \left(\delta(\mu\alpha)\right)^2 & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ -\beta u \frac{\mu}{B} \left(\delta(\mu\alpha)\right)^2 \left(2\delta(\mu\alpha) - \beta\right) & \text{for } \alpha \in (1, 2]. \end{cases}$$

Putting both together, we obtain marginal revenue, $MR(\alpha) = \mu[\alpha t'_f(\alpha) + t_f(\alpha)]$, or

$$MR(\alpha) = \begin{cases} \mu[\beta u - k] & \text{for } \alpha \in [0, \hat{\alpha}_{dp}] \\ -\mu k & \text{for } \alpha \in (\hat{\alpha}_{dp}, 1] \\ -\mu \left[\beta u \left(\delta(\mu\alpha)\right)^2 + k\right] & \text{for } \alpha \in (1, 2]. \end{cases}$$

Proposition 9 *Suppose that $1/2 \leq \beta \leq B/\mu \leq 1$. Then, the profit-maximizing transmission fee on the fast lane prices out congestion and leads all time-sensitive CPs to use quality reduction; i.e., t_f is such that $\alpha = \hat{\alpha}_{dp}$.*

Thus, allowing the ISP to implement bandwidth tiering and charge for the fast lane can also solve the problem of insufficient quality reduction established in Proposition 8.

4.1.3 Ad-financed content providers

So far, we have considered a revenue model of CPs whereby they charge users directly for their service. Many real-world CPs follow an alternative business model whereby revenues are generated not only from payments by users to CPs but also from payments by advertisers. In addition, content providers can adjust their data volume through their choice of advertising strategy: they can choose between unobtrusive text advertisements, which are less bandwidth heavy but also draw less attention, and flashy pop-up videos, which consume more bandwidth and draw more attention.

We now present a simple extension of our model that includes revenues from advertising and gives content providers a choice between different advertising strategies. Content providers offer content, which users value with v. For simplicity, we postulate that there is a fixed level of advertising. Advertising leads to a utility loss of a on the user side. Hence, a user's net utility of a bundle consisting of content and advertising is $v - a$. To avoid any interaction of CPs on the advertiser side, we postulate that users have unit demand for each advertised product (and all CPs advertise different products).[15] They obtain gross surplus z for any product that makes it into their memory. Advertisers make take-it-or-leave-it offers to users. In turn, CPs are able to extract all surplus generated on the user and on the advertiser side. Congestion is an issue if the ISP does not deliver all content including advertising; this happens for time-sensitive content that does not reach users on time.

We now consider the content provider's strategy with respect to advertising. Suppose that ads can be of high resolution such that users will remember for sure if it is delivered.

[15]Removing interactions among advertisers is in line with most of the literature on ad-financed media; for a seminal contribution on ad-financed media platforms, see Anderson and Coate (2005); for a survey on Internet media, see Peitz and Reisinger (2014).

Alternatively, the content provider can choose to provide only low-resolution ads such that users remember the ad only with probability κ. The advantage of the latter is that the data volume is lower. For simplicity, let us assume that the data volume for the bundle consisting of content and high-resolution ad is two packets, while the data volume for the bundle that includes the low-resolution ad is only one packet. Furthermore, to mirror our previous analysis, time-insensitive CPs can only offer the low-resolution version (this assumption is merely for convenience).

In this simple setting, advertisers extract all surplus in the advertiser-user interaction and the price they charge for their product or service is z. Content providers then charge z for high-resolution advertising and κz for low-resolution advertising. Thus, recalling that the delivery probability of time-sensitive traffic is $\delta(A)$ for each packet and the CP's transmission cost per packet is k, a CP with time-sensitive content obtains revenues $(v - a + z)\delta(A)^2 - 2k$ employing high-resolution advertising and $(v - a + \kappa z)\delta(A) - k$ employing low-resolution advertising. Let us now do the following change of variables. We define $u = v - a + z$ and β such that $\beta u = v - a + \kappa z$, which implies that $\beta = (v - a + \kappa z)/(v - a + z)$. Therefore, this setting is formally equivalent to the setting in which less traffic leads to a reduction of content quality and CPs obtain only revenues from users. In the special case $v = a$, $u = z$ and $\beta = \kappa$, CPs make revenues only from advertisers and do not charge users directly.

We can thus reinterpret the results from the previous subsection as follows. If the probability κ of a low-resolution ad being remembered is sufficiently high, a social planner would want to send only low-resolution advertising; yet under net neutrality the equilibrium will exhibit high-resolution advertising when transmission costs are low (k small). Bandwidth tiering combined with fast-lane pricing can lead CPs to switch to socially efficient low-resolution advertising.

4.2 Content providers facing privately informed users

Since, in the models considered so far, CPs can extract the full surplus in any provider-user interaction and the ISP does not offer any stand-alone utility, the ISP makes zero profit from providing access to content. This feature of our model may be criticized as being unrealistic. It also implies that an ISP does not have an incentive to make traffic more efficient, as long as it does not have price instruments to obtain revenues on the content provider side. In particular, it implies that, in the models considered so far, the ISP does not have any incentive to engage in deep packet inspection (even if it were costless to implement). Our model can be easily augmented to generate positive equilibrium subscription fees for Internet access.

Suppose consumers are heterogeneous with respect to the value they derive from consuming content. For any content i, each user draws her valuation from $\{u_L, u_H\}$, where high willingness to pay $u_H = u + \Delta u$ is drawn with probability λ and low willingness to pay $u_L = u$ with probability $1 - \lambda$. Draws are independent across consumers and across content. We consider the timing where consumers observe their tastes after making the decision whether to buy an Internet subscription. Thus, consumers are ex ante identical when buying Internet access. We assume that content providers prefer serving both con-

sumer types to serving only one; i.e., $u \geq \lambda(u + \Delta u)$. Hence, in each user-content match users obtain an expected net surplus of $\lambda \Delta u$. Aggregating over all content (taking into account the probability that a match is formed) gives the expected consumer net surplus gross of the subscription fee. The ISP can then use the subscription fee to extract this surplus.

We have shown in Section 3 that under deep packet inspection there is an equilibrium in which CPs send their traffic once if $B > \mu$, and this implements the first-best level of traffic. Since all time-sensitive traffic arrives on time in this case, whereas it does not under net neutrality, user expected net surplus (gross of the subscription fee) must be larger under deep packet inspection than under net neutrality. The ISP can extract this expected surplus via its subscription fee. Hence, the ISP optimally sets a higher subscription fee and makes a larger profit under deep packet inspection than under net neutrality.

5 Conclusion

We have presented a model of a congested Internet that delivers time-sensitive and time-insensitive content to users. The former needs to be delivered on time for consumers to derive utility from it. For the latter, timely delivery does not matter. The probability of on-time delivery for a given packet is equal to the ratio between bandwidth and total traffic. Content providers can increase the overall probability of timely delivery by sending a packet several times, thereby improving the chances that at least one of them arrives on time. However, this creates negative externalities for other CPs.

In such a framework, enforcing strict net neutrality rules may not be a good idea, as it worsens network congestion. Net neutrality effectively turns the network into an unmanaged common property resource. We show that departures from strict net neutrality can alleviate the overexploitation and misallocation problem, as the ISP is enabled to manage this resource. Deep packet inspection may eliminate congestion by solving the misallocation problem and at the same time removing the incentive to inflate traffic. However, the result is ambiguous since, under some circumstances, deep packet inspection may actually increase congestion. Thus, deep packet inspection can backfire if CPs respond by increasing traffic, possibly making it inferior to strict net neutrality. Alternatively, if the ISP can charge a transmission fee, it will price out congestion, while a misallocation problem remains. Moreover, we find that fully eliminating congestion is generally not socially optimal in a best-effort system. Regulating the transmission fee (by means of a price cap) therefore raises efficiency.

A better outcome than under the above regulated environments can be achieved by allowing the ISP to engage in bandwidth tiering and price discrimination, as this allows the ISP to address both the traffic inflation and the traffic misallocation problem. In our simple and stylized setting, such a regime can implement the first-best allocation. Therefore, regulatory intervention risks being welfare-reducing.[16]

[16] An exception is a regulatory intervention that fixes the price of the slow lane at zero, which we show to be welfare neutral.

We show that our main insights apply more broadly; this includes environments in which CPs can take costly actions to reduce the volume of traffic (through the use of compression techniques, a reduction of quality such as lower resolution or an adjustment of the traffic volume associated with advertising). Since part of the benefits from these costly actions accrue to others, CPs tend to underinvest in them.

We have considered only short-term effects of different regulatory regimes and did not address investment issues, neither by the ISP nor by CPs. In particular, bandwidth was exogenous in the analysis. One concern in the net neutrality debate is that under net neutrality an ISP does not have a strong incentive to invest since it finds it difficult to monetize its investment. However, the opposite fear is that absent net neutrality, an ISP does not have strong incentives to invest in alleviating congestion simply because congestion allows the ISP to extract rents from CPs with time-sensitive content whose success depends on prioritized access. In our model, under net neutrality, the ISP does not have an incentive to invest in bandwidth because the ISP cannot make any profit (see, however, the extension section with a specification that generates positive profits for the ISP).

Our analysis is provided within a simple and stylized model. We deliberately focused on a monopoly ISP with perfectly inelastic demand for subscription, atomless CPs and vertical separation between Internet connection and content. Future work may want to add investment incentives by ISPs in a setting that incorporates the feature of our model that CPs react to changes in bandwidth by adjusting their traffic volume.[17]

Appendix: Relegated proofs

Proof of Lemma 1. Using (1) and (2) to substitute for γ in (3) as well as the fact that $\delta(\mu\alpha) = 1$ for $\alpha \leq \hat{\alpha}_{dp}$, we can rewrite total surplus as

$$W(\alpha) = \begin{cases} \alpha(u-k) & \text{for } \alpha \in [0, \min\{\hat{\alpha}_{dp}, 1\}) \\ \alpha(u\delta(\mu\alpha) - k) & \text{for } \alpha \in [\min\{\hat{\alpha}_{dp}, 1\}, 1] \\ u - \alpha k & \text{for } \alpha \in [1, \max\{\hat{\alpha}_{dp}, 1\}] \\ u\delta(\mu\alpha)\left[\alpha(1-\delta(\mu\alpha)) + \delta(\mu\alpha)\right] - \alpha k & \text{for } \alpha \in (\max\{\hat{\alpha}_{dp}, 1\}, 2]. \end{cases} \tag{24}$$

Differentiating (24) and using $\delta' = -(1/B)\delta^2$ yields

$$W'(\alpha) = \begin{cases} u - k & \text{for } \alpha \in [0, \min\{\hat{\alpha}_{dp}, 1\}) \\ u\delta(\mu\alpha)\left(1 - \delta(\mu\alpha)\mu\alpha/B\right) - k & \text{for } \alpha \in [\min\{\hat{\alpha}_{dp}, 1\}, 1] \\ -k & \text{for } \alpha \in [1, \max\{\hat{\alpha}_{dp}, 1\}] \\ u\delta(\mu\alpha)\left[1 - \delta(\mu\alpha)\left(1 + \mu\alpha/B\right)\right. & \\ \qquad \left. +2\left(\delta(\mu\alpha)\right)^2(\alpha-1)\mu/B\right] - k & \text{for } \alpha \in (\max\{\hat{\alpha}_{dp}, 1\}, 2]. \end{cases}$$

Noting that for $\alpha \geq \hat{\alpha}_{dp}$ we have $\mu\alpha/B = 1/\delta(\mu\alpha)$, we observe that for $\alpha \in [\min\{\hat{\alpha}_{dp}, 1\}, 1]$, $W'(\alpha) = -k < 0$, and for $\alpha \in (\max\{\hat{\alpha}_{dp}, 1\}, 2]$, $W'(\alpha) = u\left(\delta(\mu\alpha)\right)^2\left(1 - 2\tilde{\delta}\mu/B\right) - k < 0$,

[17] Previous work has included changes of total traffic due to the participation decision of CPs but abstracted from the adjustment of traffic volumes by active CPs. The exception is Choi *et al.* (2013). By comparing environments with different bandwidth, they obtained some interesting insights on how incentives of CPs depend on available bandwidth; see our literature review in the introduction.

where the inequality follows from $B/(2\mu) \leq \delta(\mu\alpha)$ for $\alpha \leq 2$. Hence, $W'(\alpha) > 0$ for $\alpha < \min\{\hat{\alpha}_{dp}, 1\}$ and $W'(\alpha) < 0$ for $\alpha > \min\{\hat{\alpha}_{dp}, 1\}$. Together with continuity of W, this implies that welfare in a two-tiered system is maximized at $\alpha = \min\{\hat{\alpha}_{dp}, 1\}$. ■

Proof of Lemma 2. Since time-insensitive content yields the same utility as time-sensitive content and has a weakly greater probability of delivery, the second-best allocation is such that time-insensitive content is always sent while the traffic volume of time-sensitive content is adjusted. Using (1) to substitute for γ in (3), total surplus from time-sensitive content in a one-tiered system can be written as

$$W(\alpha) = \begin{cases} \alpha[u\delta(\mu\alpha + 1 - \mu) - k] & \text{for } \alpha \in [0, 1] \\ u\delta(\mu\alpha + 1 - \mu)\left[\alpha(1 - \delta(\mu\alpha + 1 - \mu)) + \delta(\mu\alpha + 1 - \mu)\right] - \alpha k & \text{for } \alpha \in (1, 2] \end{cases} \tag{25}$$

Since, for $\alpha \geq \hat{\alpha}_{nn}$, $\delta(\mu\alpha + 1 - \mu)$ is strictly monotonic in α, it can be inverted. Let $\alpha(\delta) = (B/\delta - (1 - \mu))/\mu$ denote α as a function of δ and define $\hat{W}(\delta) \equiv W(\alpha(\delta))$. Because $0 \leq \alpha \leq 2$, an upper bound on δ is $\min\{B/(1 - \mu), 1\}$ and a lower bound is $B/(1 + \mu)$. From (25), we thus obtain

$$\hat{W}(\delta) = \begin{cases} \frac{1}{\mu}(B/\delta - (1 - \mu))(u\delta - k) & \text{for } \delta \in \left[B, \min\left\{\frac{B}{1-\mu}, 1\right\}\right] \\ \\ \frac{1}{\mu}\left[u\left(\delta^2 - \delta(B + 1 - \mu) + B\right) - k\left(B/\delta - (1 - \mu)\right)\right] & \text{for } \delta \in \left[\frac{B}{1+\mu}, B\right). \end{cases} \tag{26}$$

Before establishing Claims (1) - (5), we make three preliminary observations. First, we show that $\hat{W}$ is strictly concave on $[B, \min\{B/(1 - \mu), 1\}]$. We have

$$\hat{W}'(\delta) = \frac{1}{\mu}\left(\frac{kB}{\delta^2} - u(1 - \mu)\right)$$
$$\hat{W}''(\delta) = -\frac{2kB}{\mu\delta^3} < 0.$$

Second, we derive the condition under which $\hat{W}(B/(1 + \mu)) \leq \hat{W}(B)$. Substituting into (26) and rearranging yields

$$\frac{B}{1 + \mu}\left(1 - \mu - \frac{B}{1 + \mu}\right) \leq \frac{k}{u}. \tag{27}$$

Third, we derive a necessary condition for the existence of a local maximum on $[B/(1 + \mu), B)$. We have

$$\hat{W}'(\delta) = \frac{u}{\mu}[2\delta - (B + 1 - \mu)] + \frac{k}{\mu}\frac{B}{\delta^2}$$
$$\hat{W}''(\delta) = \frac{2}{\mu}\left(u - \frac{kB}{\delta^3}\right).$$

The first-order condition for a local maximum is $\hat{W}'(\delta) = 0$, or $w(\delta) = k/u$. Hence, a necessary condition for the existence of a local maximum is $k/u \leq \max_{B/(1+\mu) \leq \delta \leq B} w(\delta)$.

The unconstrained maximizer of $w(\delta)$ is found by solving $w'(\delta) = [2\delta(B+1-\mu)-6\delta^2]/B = 0$, yielding a unique $\delta_w = (B+1-\mu)/3$ at which the second-order condition holds ($w''(\delta_w) = -2(B+1-\mu)/B < 0$). Taking into account the constraint $B/(1+\mu) \leq \delta \leq B$ and the fact that $w(\delta)$ has a local minimum at $\delta = 0$, we obtain

$$\delta_{\max} = \begin{cases} (B+1-\mu)/3 & \text{if } B/(1+\mu) \leq (B+1-\mu)/3 \leq B \\ B & \text{if } (B+1-\mu)/3 > B \\ B/(1+\mu) & \text{if } (B+1-\mu)/3 < B/(1+\mu), \end{cases}$$

and $\max_{B/(1+\mu)\leq\delta\leq B} w(\delta) = w(\delta_{\max})$. We conclude that existence of a local maximum on $[B/(1+\mu), B)$ requires

$$\frac{k}{u} \leq w(\delta_{\max}). \tag{28}$$

Note that, for all $\delta \in [B/(1+\mu), B]$,

$$w(\delta) = \frac{\delta^2(B+1-\mu-2\delta)}{B} < \frac{(1-\mu)\delta^2}{B} \quad \Leftrightarrow \quad B < 2\delta,$$

which is always satisfied since $\delta \geq B/(1+\mu) > B/2$. Because moreover $(1-\mu)\delta^2/B$ is increasing in δ for $\delta \geq 0$, it follows that

$$w(\delta_{\max}) < (1-\mu)\delta_{\max}/B \leq (1-\mu)B. \tag{29}$$

Claim (1): Concavity of $\hat{W}$ on $[B, \min\{B/(1-\mu), 1\}]$ implies that if $\hat{W}'(\min\{B/(1-\mu), 1\}) \geq 0$ or

$$\min\{B/(1-\mu), (1-\mu)/B\} \leq k/u,$$

then $\hat{W}' > 0$ for all $\delta \in (B, \min\{B/(1-\mu), 1\}]$ as well as $\hat{W}'_+(B) \equiv \lim_{\delta\searrow B} d\hat{W}/d\delta = (k/B - u(1-\mu))/\mu > 0$ and hence $\hat{W}(B) < \hat{W}(\min\{B/(1-\mu), 1\})$. Moreover, since $(1-\mu)B < \min\{B/(1-\mu), (1-\mu)/B\}$, it follows from (27) that $\hat{W}(B) \geq \hat{W}(B/(1+\mu))$, and from (29) that there is no local maximum on $[B/(1+\mu), B)$. Hence, $\hat{W}$ is maximum at $\delta = \min\{B/(1-\mu), 1\}$ which implies $\alpha^{SB} = \hat{\alpha}_{nn}$.

Claim (2): Concavity also implies that if $\hat{W}'(\min\{B/(1-\mu), 1\}) < 0 < \hat{W}'_+(B)$ or

$$(1-\mu)B < k/u < \min\{B/(1-\mu), (1-\mu)/B\},$$

then there exists a unique local maximum on $[B, \min\{B/(1-\mu), 1\}]$ solving $(1-\mu)\delta^2/B = k/u$, which corresponds to the value of α solving (4). Since $k/u > (1-\mu)B$ implies (27) and rules out existence of a local maximum on $[B/(1+\mu), B)$ by (29) and (28), α^{SB} solving (4) is a global maximum.

Claim (5): A necessary condition for $\alpha^{SB} = 2$ to be optimal is $\hat{W}'(B/(1+\mu)) \leq 0 \Leftrightarrow w(B/(1+\mu)) \geq k/u$. A condition that, in conjunction with the first, is both necessary and sufficient, is $\hat{W}(B/(1+\mu)) \geq \hat{W}(B)$. Using (27) thus establishes the claimed result.

From the above results, we infer that if $\min\{w(B/(1+\mu)), B(1-\mu^2-B)/(1+\mu)^2\} < k/u < (1-\mu)B$, the solution must be some $\delta \in (B/(1+\mu), B]$. We know that $\delta = B$ (and thus $\alpha = 1$) must be optimal for $w(\delta_{\max}) \leq k/u < (1-\mu)B$ by (28) and (29).

Hence, what remains to be shown is that there exists $\hat{k}$ with the claimed properties when $k/u < w(\delta_{\max})$.

Note first that the second-order condition for a local maximum at some $\delta_0 \in (B/(1+\mu), B)$ satisfying the first-order condition $w(\delta_0) = k/u$ is

$$\hat{W}''(\delta_0) = \frac{2}{\mu}\left(u - \frac{kB}{\delta_0^3}\right) \leq 0 \quad \Leftrightarrow \quad \delta_0 \leq \frac{B+1-\mu}{3}.$$

Thus, we can distinguish three cases:

- If $B \leq (B+1-\mu)/3$, any $\delta_0 \in (B/(1+\mu), B)$ satisfying $w(\delta_0) = k/u$ is both a local and global maximum. Hence, $\hat{k} = uw(\delta_{\max}) = uw(B)$.
- If $B/(1+\mu) \geq (B+1-\mu)/3$, no $\delta_0 \in (B/(1+\mu), B)$ satisfying $w(\delta_0) = k/u$ can be a local maximum. Hence, $\hat{k} = \min\{uw(B/(1+\mu)), uB(1-\mu^2-B)/(1+\mu)^2\}$.
- If $B/(1+\mu) < (B+1-\mu)/3 < B$, there exists a unique $\delta_0 \in (B/(1+\mu), B)$ satisfying both $w(\delta_0) = k/u$ and $\delta_0 < (B+1-\mu)/3$. This δ_0 is a local maximum but not necessarily a global maximum.

What remains to be shown is that, in the last case, there exists $\hat{k}$ such that $\hat{W}(\delta_0) \geq \hat{W}(B)$ for $k \leq \hat{k}$ and $\hat{W}(\delta_0) < \hat{W}(B)$ for $k > \hat{k}$. Because $\hat{W}(\delta_0) = \max_\delta \hat{W}(\delta)$, by the envelope theorem

$$\frac{d}{dk}\left[\hat{W}(\delta_0) - \hat{W}(B)\right] = 1 - \frac{B}{\mu\delta_0} < 0,$$

where the inequality follows from $\delta_0 < B$. This proves Claims (3) and (4). ■

Proof of Lemma 3. Let us first consider an equilibrium in which each CP sends one packet, so that $\alpha = 1$. Then, CP i's profit from $\alpha_i = 1$ is

$$u\gamma(1,1) - k = uB - k.$$

Hence, if $k/u > B$, $\alpha_i = 1$ for all i is not an equilibrium as CPs would make negative profit. The only equilibrium then involves mixing over $\alpha_i = 0$ and $\alpha_i = 1$. For each CP to be indifferent, it must be that

$$u\gamma(1, \mu\alpha + 1 - \mu) - k = 0,$$

yielding (6), which can be solved for a unique $\alpha^{nn} \in (0,1)$. Note that deviating to $\alpha_i = 2$ can never be profitable if $u\delta(A) - k \leq 0$ since, for any $\delta(A) \leq 1$, $u\gamma(2,A) - 2k = 2[u\delta(A)(1-\delta(A)/2) - k] < 2[u\delta(A) - k] \leq 0$.

If $k/u \leq B$, $\alpha_i = 1$ for all i (implying $\alpha^{nn} = 1$) is an equilibrium provided no one can gain from deviating to $\alpha_i = 2$, which requires

$$u\gamma(1,1) - k \geq u\gamma(2,1) - 2k \quad \Leftrightarrow \quad \gamma(2,1) - \gamma(1,1) \leq k/u.$$

Noting that

$$\gamma(2,A) - \gamma(1,A) = \delta(A)(1-\delta(A)), \tag{30}$$

this becomes $B(1-B) \leq k/u$. Thus, $\alpha_i = 1$ for all i is an equilibrium for $B(1-B) \leq k/u \leq B$.

Next, let us consider an equilibrium in which each time-sensitive CP sends two packets, i.e., $\alpha_i = 2$ for all i, implying $\alpha = 2$. No CP must have an incentive to deviate to either $\alpha_i = 0$ or $\alpha_i = 1$, which requires, respectively,

$$u\gamma(2, 1+\mu) - 2k \geq 0 \quad \Leftrightarrow \quad \delta(1+\mu)\left(1 - \frac{\delta(1+\mu)}{2}\right) \geq \frac{k}{u} \tag{31}$$

$$u\gamma(2, 1+\mu) - 2k \geq u\gamma(1, 1+\mu) - k \quad \Leftrightarrow \quad \delta(1+\mu)\left(1 - \delta(1+\mu)\right) \geq \frac{k}{u}, \tag{32}$$

where (32) follows from (30). Since (32) implies (31), it is both necessary and sufficient. Thus, $\alpha_i = 2$ for all i is an equilibrium for $k/u \leq \delta(1+\mu)(1-\delta(1+\mu)) = B(1+\mu-B)/(1+\mu)^2$.

Finally, consider a mixed-strategy equilibrium in which time-sensitive CPs randomize between sending one and two packets. Each CP i must be indifferent between $\alpha_i = 1$ and $\alpha_i = 2$, i.e., $u\gamma(1, \alpha) - k = u\gamma(2, \alpha) - 2k$ or

$$\gamma(2, A) - \gamma(1, A) = \frac{k}{u},$$

which, using (30), yields (7). Moreover, it must be that

$$u\gamma(1, \mu\alpha + 1 - \mu) - k \geq 0 \tag{33}$$

at α solving (7). We have $u\gamma(1, \mu\alpha + 1 - \mu) - k = u\delta(\mu\alpha + 1 - \mu) - k \geq u\delta(\mu\alpha + 1 - \mu)(1 - \delta(\mu\alpha + 1 - \mu)) - k = 0$ for any α since $\delta(\mu\alpha + 1 - \mu) \leq 1$, where the last equality follows from (7). Hence, (33) is satisfied.

Existence of a mixed-strategy equilibrium $\alpha^{nn} \in (1, 2)$ requires

$$\min_{\alpha \in (1,2)} \delta(\mu\alpha + 1 - \mu)(1 - \delta(\mu\alpha + 1 - \mu)) < \frac{k}{u} < \max_{\alpha \in (1,2)} \delta(\mu\alpha + 1 - \mu)(1 - \delta(\mu\alpha + 1 - \mu)).$$

Since $\delta(1-\delta)$ is strictly concave, the minimum is necessarily attained at one of the boundaries. The boundaries are $\delta(1) = B$ and $\delta(1+\mu) = B/(1+\mu)$. Hence, $\min_{\alpha \in (1,2)} \delta(\mu\alpha + 1 - \mu)(1 - \delta(\mu\alpha + 1 - \mu)) = \min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\}$. By definition, the maximum of $\delta(1-\delta)$ is attained at $\bar{\delta}$. Thus an equilibrium with $\alpha^{nn} \in (1, 2)$ in which each CP randomizes between $\alpha_i = 1$ (probability $2 - \alpha^{nn}$) and $\alpha_i = 2$ (probability $\alpha^{nn} - 1$) exists for $\min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\} < k/u < \bar{\delta}(1-\bar{\delta})$. ■

Proof of Proposition 1. Consider first the case $B < k/u < B/(1-\mu)$. By Lemma 3, the equilibrium traffic α^{nn} is then determined by $\delta(\alpha^{nn}) = k/u$, with $\alpha^{nn} \in (\hat{\alpha}_{nn}, 1)$. By Lemma 2, if $k/u \geq \min\{(1-\mu)/B, B/(1-\mu)\}$, the optimal traffic is $\alpha^{SB} = \hat{\alpha}_{nn} < \alpha^{nn}$. If instead $k/u < \min\{(1-\mu)/B, B/(1-\mu)\}$, then α^{SB} is determined by $(1-\mu)(\delta(\mu\alpha + 1 - \mu))^2/B = k/u$. Since δ is decreasing in α, what we need to show is that $\delta \geq (1-\mu)\delta^2/B$ or $\delta \leq B/(1-\mu)$. This inequality follows from the definition of $\delta = \min\{B/(\mu\alpha + 1 - \mu), 1\} \leq B/(1-\mu)$.

For $\hat{k}/u \leq k/u \leq B$, where $\hat{k}$ is defined in Lemma 2, we have $\alpha^{nn} \geq 1 \geq \alpha^{SB}$. Furthermore, $\min\{w(B/(1+\mu), B(1-\mu^2-B)/(1+\mu)^2\} \leq B(1-\mu^2-B)/(1+\mu)^2 \leq \min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\}$ because

$$1 - B - \mu^2 \leq (1-B)(1+\mu)^2.$$

Since $\alpha^{nn} = 2$ is the unique equilibrium for $k/u \leq \min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\}$, we thus have $\alpha^{nn} = \alpha^{SB} = 2$ for $k/u \leq \min\{w(B/(1+\mu), B(1-\mu^2-B)/(1+\mu)^2\}$ and $\alpha^{nn} = 2 > \alpha^{SB}$ for $\min\{w(B/(1+\mu), B(1-\mu^2-B)/(1+\mu)^2\} < k/u \leq \min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\}$.

The interval that remains is $\min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\} \leq k/u < \hat{k}/u$. We can distinguish two cases depending on whether $B \lessgtr (1-\mu^2)/(2-\mu)$. We first show that for $B \geq (1-\mu^2)/(2-\mu)$, the above interval is empty as $\hat{k}/u \leq \min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\}$. Then we show that for $B < (1-\mu^2)/(2-\mu)$, if there is a local maximum of $W(\alpha)$ on $(1,2)$, then it is such that $\alpha^{nn} > \alpha^{SB}$.

We have $B \geq (1-\mu^2)/(2-\mu)$ if and only if $B/(1+\mu) \geq (1-\mu+B)/3$. From the proof of Lemma 2, we know that a necessary condition for a local maximum at δ_0 is $\delta_0 \leq (1-\mu+B)/3$. Thus, if $B \geq (1-\mu^2)/(2-\mu)$, there exists no local maximum of $\hat{W}(\delta)$ (defined in the proof of Lemma 2) on $(B/(1+\mu), B)$. Hence, $\hat{k}/u = \min\{w(B/(1+\mu)), B(1-\mu^2-B)/(1+\mu)^2\} = B(1-\mu^2-B)/(1+\mu)^2$, where the last equality follows from

$$w\left(B/(1+\mu)\right) = \frac{B(1-\mu)(1+\mu-B)}{(1+\mu)^3} \geq \frac{B}{1+\mu}\left(1-\mu-\frac{B}{1+\mu}\right) \quad \Leftrightarrow \quad B \geq \frac{\mu(1-\mu^2)}{2},$$

which is implied by $B \geq (1-\mu^2)/(2-\mu)$. We conclude that $\alpha^{SB} = 1 \leq \alpha^{nn}$ for $B(1-\mu^2-B)/(1+\mu)^2 < k/u < (1-\mu)B$.

Now suppose that $B < (1-\mu^2)/(2-\mu)$. Then $\hat{W}(\delta)$ *may* have a local maximum on $(B/(1+\mu), B)$ (and thus $W(\alpha)$ a local maximum on $(1,2)$). Inspection of (26) reveals that $\hat{W}$ is first increasing, then decreasing, and then increasing again as δ varies from 0 to ∞. Hence, existence of a local maximum on $(B/(1+\mu), B)$ requires either $\hat{W}'_-(B) \equiv \lim_{\delta \nearrow B} d\hat{W}/d\delta < 0$ or $\delta_I < B$, where $\delta_I = \sqrt[3]{Bk/u}$ is the inflection point of $\hat{W}$. We have

$$\begin{aligned} \hat{W}'_-(B) < 0 \quad &\Leftrightarrow \quad \frac{k}{u} < B(1-B-\mu) \\ \delta_I < B \quad &\Leftrightarrow \quad \frac{k}{u} < B^2. \end{aligned}$$

Thus, if $k/u \geq \max\{B(1-B-\mu), B^2\}$, there cannot be a local maximum on $(B/(1+\mu), B)$, so $\hat{k}/u \leq \max\{B(1-B-\mu), B^2\}$. But this implies that if $B^2 \leq B(1-B) \Leftrightarrow B \leq 1/2$, there exists no value of k/u such that $\alpha^{nn} = 1$ and $\alpha^{SB} \in (1,2)$. For $B \leq 1/2$, we can thus restrict attention to $k/u \leq B(1-B)$ (if $k/u \geq B(1-B) \geq \max\{B(1-B-\mu), B^2\}$, then $\alpha^{SB} \leq 1 \leq \alpha^{nn}$). By Lemma 3, if $B(1-B) \leq B(1+\mu-B)/(1+\mu)^2$, the only equilibrium when $k/u \leq B(1-B)$ is $\alpha^{nn} = 2$. If instead $B(1-B) > B(1+\mu-B)/(1+\mu)^2$, then for $B(1+\mu-B)/(1+\mu)^2 < k/u < B(1-B)$, there exists a mixed-strategy equilibrium where α^{nn} is determined by

$$\delta(\mu\alpha^{nn} + 1 - \mu)(1 - \delta(\mu\alpha^{nn} + 1 - \mu)) = \frac{k}{u}.$$

Importantly, $B(1-B) > B(1+\mu-B)/(1+\mu)^2$ implies that α^{nn} must lie on the upward sloping part of $\delta(1-\delta)$ (i.e., $\delta(\mu\alpha^{nn} + 1-\mu) < 1/2$), where

$$\frac{\partial}{\partial\alpha}[\delta(\mu\alpha+1-\mu)(1-\delta(\mu\alpha+1-\mu))]|_{\alpha=\alpha^{nn}} < 0.$$

An interior $\alpha^{SB} \in (1,2)$ must satisfy the first-order condition

$$\underbrace{\gamma(2,\mu\alpha+1-\mu)-\gamma(1,\mu\alpha+1-\mu)}_{=\delta(\mu\alpha+1-\mu)(1-\delta(\mu\alpha+1-\mu))}+(\alpha-1)\gamma_2(2,\mu\alpha+1-\mu)+(2-\alpha)\gamma_2(1,\mu\alpha+1-\mu)=\frac{k}{u}$$

as well as the second-order condition $W'' \leq 0$. Since $\gamma_2(\alpha_i,\mu\alpha+1-\mu) = -\frac{\mu\alpha_i}{B}(\delta(\mu\alpha+1-\mu))^2(1-\delta(\mu\alpha+1-\mu))^{\alpha_i-1} < 0$ for $\alpha_i \in \{1,2\}$, we conclude that $\alpha^{SB} < \alpha^{nn}$.

Finally, consider the case $1/2 < B \leq (1-\mu^2)/(2-\mu)$ (which exists if $\mu < 1/2$). Then, $B^2 \geq B(1-B)$, so we need to show that there exists no local maximum of $\hat{W}(\delta)$ on $(B/(1+\mu), B)$ for $k/u \geq B(1-B)$ (where $\alpha^{nn}=1$ is an equilibrium). By Lemma 2, any such local maximum satisfies the first-order condition $w(\delta_0) = k/u$. Thus, it suffices to show that

$$\max_{B/(1+\mu)\leq\delta\leq B} w(\delta) < B(1-B). \tag{34}$$

From the proof of Lemma 2, we know

$$\max_{B/(1+\mu)\leq\delta\leq B} w(\delta) \leq w\left(\frac{B+1-\mu}{3}\right) = \frac{1}{B}\left(\frac{B+1-\mu}{3}\right)^3.$$

A sufficient condition for (34) is therefore

$$\Phi(B,\mu) \equiv \left(\frac{B+1-\mu}{3}\right)^3 - B^2(1-B) < 0.$$

Maximizing $\Phi(B,\mu)$ subject to $1/2 \leq B \leq (1-\mu^2)/(2-\mu)$ yields the maximizer $(B^*,\mu^*) = (1/2, 0)$, with $\Phi(1/2,0) = 0$. It follows that $\Phi(B,\mu) < 0$ for all (B,μ) satisfying $1/2 < B \leq (1-\mu^2)/(2-\mu)$. ■

Proof of Proposition 2. From inspection of (10), the ISP's profit is increasing for $\alpha \in [0, \hat{\alpha}_{nn})$ and decreasing for $\alpha \in (\hat{\alpha}_{nn}, 1]$, implying a local maximum at $\alpha = \hat{\alpha}_{nn}$. At $\alpha = 1$, there is a discontinuity in the ISP's profit. To see this, note that $t(1) = uB - k$ while $\lim_{\alpha\searrow 1} t(\alpha) \leq u\bar{\delta}(1-\bar{\delta}) - k$. Since

$$\bar{\delta}(1-\bar{\delta}) = \begin{cases} (B/(1+\mu))(1-B/(1+\mu)) & \text{for } 1/2 \leq B/(1+\mu) \\ 1/4 & \text{for } B/(1+\mu) < 1/2 \leq B \\ B(1-B) & \text{for } B > 1/2, \end{cases}$$

we conclude that there is a downward jump in the ISP's profit at $\alpha = 1$. This rules out $\alpha = 1$ as a solution.

The ISP's revenue is continuous on $\alpha \in (1,2]$. There is a local maximum either at one of the boundaries, $\alpha = \tilde{\alpha}_{nn}$ and $\alpha = 2$, or at the interior solution solving $\delta(\mu\alpha+1-\mu) =$

$\sqrt{k/u}$. The latter exists if and only if $B/(1+\mu) < \sqrt{k/u} < B$. We now show that the local maximum on $(1,2]$ gives the ISP a lower profit than $\hat{\alpha}_{nn}$. The ISP's profit when $\alpha = \hat{\alpha}_{nn}$ is given by $\pi_0^{\text{ISP}} = \max\{B, 1-\mu\}(u-k)$. Its profit when setting α such that $\delta(\mu\alpha + 1 - \mu) = \sqrt{k/u}$ is $\pi_1^{\text{ISP}} = B(u - 2\sqrt{uk})$. Note that π_1^{ISP} must be weakly greater than the ISP's profit when setting either $\alpha = \tilde{\alpha}_{nn}$ or $\alpha = 2$, since it is the unconstrained maximum of $[u\delta(\mu\alpha+1-\mu)(1-\delta(\mu\alpha+1-\mu)) - k](\mu\alpha+1-\mu)$. But we have $\pi_0^{\text{ISP}} > \pi_1^{\text{ISP}}$ if and only if $k < 2\sqrt{uk}$, which is always satisfied; hence, $\alpha = \hat{\alpha}_{nn}$ is the global maximizer of the ISP's profit. ■

Proof of Lemma 4. Let us first consider an equilibrium in which each CP sends one packet, so that $\alpha = 1$. Then, CP i's profit from $\alpha_i = 1$ is

$$u\gamma(1,\mu) - k = u \min\left\{\frac{B}{\mu}, 1\right\} - k.$$

Hence, if $k/u > B/\mu$, $\alpha_i = 1$ for all i is not an equilibrium as CPs would make negative profit. The only equilibrium then involves mixing over $\alpha_i = 0$ and $\alpha_i = 1$. For each CP to be indifferent, it must be that

$$u\gamma(1,\mu\alpha) - k = 0,$$

yielding (11), which can be solved for a unique $\alpha^{dp} \in (0,1)$. Note that deviating to $\alpha_i = 2$ can never be profitable if $u\delta(\mu\alpha) - k \leq 0$ since, for any $\delta(\mu\alpha) \leq 1$, $u\gamma(2,\mu\alpha) - 2k = 2[u\delta(\mu\alpha)(1-\delta(\mu\alpha)/2) - k] < 2[u\delta(\mu\alpha) - k] \leq 0$.

If $k/u \leq B/\mu$, $\alpha_i = 1$ for all i (implying $\alpha^{dp} = 1$) is an equilibrium provided no one can gain from deviating to $\alpha_i = 2$, which requires

$$u\gamma(1,\mu) - k \geq u\gamma(2,\mu) - 2k \quad \Leftrightarrow \quad \gamma(2,\mu) - \gamma(1,\mu) \leq k/u.$$

Using (30) this becomes $B(\mu - B)/\mu^2 \leq k/u$. Thus, $\alpha_i = 1$ for all i is an equilibrium for $B(\mu-B)/\mu^2 \leq k/u \leq B/\mu$.

Next, let us consider an equilibrium in which each time-sensitive CP sends two packets, i.e., $\alpha_i = 2$ for all i, implying $\alpha = 2$. No CP must have an incentive to deviate to either $\alpha_i = 0$ or $\alpha_i = 1$, which requires, respectively,

$$u\gamma(2,2\mu) - 2k \geq 0 \quad \Leftrightarrow \quad \delta(2\mu)\left(1 - \frac{\delta(2\mu)}{2}\right) \geq \frac{k}{u} \tag{35}$$

$$u\gamma(2,2\mu) - 2k \geq u\gamma(1,2\mu) - k \quad \Leftrightarrow \quad \delta(2\mu)\left(1 - \delta(2\mu)\right) \geq \frac{k}{u}, \tag{36}$$

where (36) follows from (30). Since (36) implies (35), it is both necessary and sufficient. Thus, $\alpha_i = 2$ for all i is an equilibrium for $k/u \leq \delta(2\mu)(1-\delta(2\mu)) = B(2\mu - B)/(2\mu)^2$.

Finally, consider a mixed-strategy equilibrium in which time-sensitive CPs randomize between sending one and two packets. Each CP i must be indifferent between $\alpha_i = 1$ and $\alpha_i = 2$, i.e., $u\gamma(1,\mu\alpha) - k = u\gamma(2,\alpha) - 2k$ or

$$\gamma(2,\mu\alpha) - \gamma(1,\mu\alpha) = \frac{k}{u},$$

which, using (30), yields (12). Moreover, it must be that

$$u\gamma(1,\mu\alpha) - k \geq 0 \tag{37}$$

at α solving (12). We have $u\gamma(1,\mu\alpha) - k = u\delta(\mu\alpha) - k \geq u\delta(\mu\alpha)(1-\delta(\mu\alpha)) - k = 0$ for any α since $\delta(\mu\alpha) \leq 1$, where the last equality follows from (12). Hence, (37) is satisfied.

Existence of a mixed-strategy equilibrium $\alpha^{dp} \in (1,2)$ requires

$$\min_{\alpha\in(1,2)} \delta(\mu\alpha)(1-\delta(\mu\alpha)) < \frac{k}{u} < \max_{\alpha\in(1,2)} \delta(\mu\alpha)(1-\delta(\mu\alpha)).$$

Since $\delta(1-\delta)$ is strictly concave, the minimum is necessarily attained at one of the boundaries. The boundaries are $\delta(\mu) = \min\{B/\mu, 1\}$ and $\delta(2\mu) = \min\{B/(2\mu), 1\}$. By definition, the maximum of $\delta(1-\delta)$ is attained at $\bar{\bar{\delta}}$. Thus an equilibrium with $\alpha^{dp} \in (1,2)$ in which each CP randomizes between $\alpha_i = 1$ (probability $2-\alpha^{dp}$) and $\alpha_i = 2$ (probability $\alpha^{dp}-1$) exists for $\min\{B(\mu-B)/\mu^2, B(2\mu-B)/(2\mu)^2\} < k/u < \bar{\bar{\delta}}(1-\bar{\bar{\delta}})$. ■

Proof of Proposition 3. By Lemma 1, the efficient level of traffic when $B \geq \mu$ is $\alpha^{FB} = 1$. By Lemma 4, $\alpha^{dp} = 1$ is an equilibrium for $B(\mu-B)/\mu^2 \leq k/u \leq B/\mu$. If $B \geq \mu$, then $\mu - B \leq 0$ and $B/\mu \geq 1$. Hence, $\alpha^{dp} = 1$ is an equilibrium for all $k/u \in [0,1]$. ■

Proof of Proposition 5. Note first that when the ISP can set $t_s > 0$ (unregulated tiering), compared to the case of regulated tiering the constraint $t_s \leq t_f$ creates an additional incentive not to decrease t_f: any price decrease on the fast lane must also be applied to the slow lane, and implies a reduction in revenue there. Thus, if $t_f(\min\{1, \hat{\alpha}_{dp}\})$ is optimal when $t_s = 0$, it must be optimal *a fortiori* when $t_s = t_f$. Hence it suffices to show that setting $\alpha = \min\{1, \hat{\alpha}_{dp}\}$ is optimal under regulated tiering.

From inspection of (17), the ISP's profit is increasing for $\alpha \in [0, \min\{1, \hat{\alpha}_{dp}\}]$ and decreasing for $\alpha \in (\min\{1, \hat{\alpha}_{dp}\}, 1]$, implying a local maximum at $\alpha = \min\{1, \hat{\alpha}_{dp}\}$. At $\alpha = 1$, there is a discontinuity in the ISP's profit. To see this, note that $t(1) = u\delta(\mu) - k$ while $\lim_{\alpha\searrow 1} t(\alpha) \leq u\bar{\bar{\delta}}(1-\bar{\bar{\delta}}) - k$. Since

$$\bar{\bar{\delta}}(1-\bar{\bar{\delta}}) = \begin{cases} (B/(2\mu))(1-B/(2\mu)) & \text{for } 1/2 \leq B/(2\mu) \\ 1/4 & \text{for } B/(2\mu) < 1/2 \leq B/\mu \\ (B/\mu)(1-B/\mu) & \text{for } B/\mu > 1/2, \end{cases}$$

we conclude that there is a downward jump in the ISP's profit at $\alpha = 1$.

The ISP's revenue is continuous on $\alpha \in (1,2]$. There is a local maximum either at one of the boundaries, $\alpha = \tilde{\alpha}_{dp}$ and $\alpha = 2$, or at the interior solution solving $\delta(\mu\alpha) = \sqrt{k/u}$. The latter exists if and only if $B/(2\mu) < \sqrt{k/u} < B/\mu$. We now show that the local maximum on $(1,2]$ gives the ISP a lower profit than $\min\{1, \hat{\alpha}_{dp}\}$. The ISP's profit when $\alpha = \min\{1, \hat{\alpha}_{dp}\}$ is given by $\pi_0^{\text{ISP}} = \mu \min\{1, \hat{\alpha}_{dp}\}(u-k)$. His profit when setting α such that $\delta(\mu\alpha) = \sqrt{k/u}$ is $\pi_1^{\text{ISP}} = B(u - 2\sqrt{uk})$. Note that π_1^{ISP} must be weakly greater than the ISP's profit when setting either $\alpha = \tilde{\alpha}_{dp}$ or $\alpha = 2$, since it is the unconstrained maximum of $[u\delta(\mu\alpha)(1-\delta(\mu\alpha)) - k]\mu\alpha$. There are two cases. For $B < \mu$ so that $\min\{1, \hat{\alpha}_{dp}\} = B/\mu$,

we have $\pi_0^{\text{ISP}} > \pi_1^{\text{ISP}}$ if and only if $B(u-2\sqrt{uk}) < B(u-k)$, or $k < 2\sqrt{uk}$, which is always satisfied. For $B \geq \mu$, we have $\tilde{\alpha}_{dp} \geq 2$ so there can be no interior solution; the ISP compares profits at $\alpha = \min\{1, \hat{\alpha}_{dp}\} = 1$ and $\alpha = 2$. It prefers charging $t_f(1)$ to $t_f(2)$ if and only if

$$\mu t_f(1) \geq 2\mu t_f(2) \quad \Leftrightarrow \quad \mu(u-k) \geq 2\mu\left(\frac{uB}{2\mu}\left(1-\frac{B}{2\mu}\right)-k\right)$$

$$\Leftrightarrow u\left(1-\frac{B}{\mu}\left(1-\frac{B}{2\mu}\right)\right)+k \geq 0,$$

a sufficient condition for which is $2\mu^2 - B(2\mu - B) \geq 0$. The value of μ that minimizes this expression is $\mu = B/2$, yielding $\min\ 2\mu^2 - B(2\mu - B) = B^2/2 > 0$. We conclude that $\alpha = \hat{\alpha}_{dp}$ is the global maximizer of the ISP's profit both under regulated and unregulated tiering. ■

Proof of Lemma 5. Assumption 2 means that if all CPs use uncompressed transmission ($\lambda_1 = 0, \lambda_2 = 1$) so that $A = 1 + \mu$, they all make positive profit: $u(\delta(1+\mu))^2 - 2k \geq 0$. Because $\delta' < 0$, this also holds for $A < 1 + \mu$, and we only have to consider two strategies: compressed and uncompressed transmission. For a given A, compressed transmission is more profitable than uncompressed transmission if and only if $u\delta(A) - c - k \geq u(\delta(A))^2 - 2k$, or

$$\delta(A)(1-\delta(A)) \geq \frac{c-k}{u}. \tag{38}$$

Claim (1): Evaluating (38) at $A = 1$, which corresponds to ($\lambda_1 = 1, \lambda_2 = 0$), we obtain $B(1-B) \geq (c-k)/u$. We further need to show that CPs make positive profit in equilibrium, which requires $uB - c - k \geq 0$, or $B - 2k/u \geq (c-k)/u$. It suffices that $B - 2k/u \geq B(1-B) \Leftrightarrow B^2/2 \geq k/u$, which is implied by Assumption 2.

Claim (2): Evaluating (38) at $A = 1 + \mu$, which corresponds to ($\lambda_1 = 0, \lambda_2 = 1$), we obtain $B(1+\mu-B)/(1+\mu)^2 \geq (c-k)/u$. If this inequality is not satisfied, it is an equilibrium for all CPs to use uncompressed transmission. Profitability is ensured by Assumption 2.

Claim (3): For CPs to be indifferent between compressed and uncompressed transmission, (38) must hold with equality. Existence of a mixed-strategy equilibrium $\lambda_1^{ct} \in (0,1)$ requires

$$\min_{\lambda\in(0,1)} \delta(1+\mu(1-\lambda_1))(1-\delta(1+\mu(1-\lambda_1)) < \frac{c-k}{u} < \max_{\lambda\in(0,1)} \delta(1+\mu(1-\lambda_1))(1-\delta(1+\mu(1-\lambda_1)),$$

or $\min\{B(1-B), B(1+\mu-B)/(1+\mu)^2\} < (c-k)/u < \bar{\delta}(1-\bar{\delta})$ (see the proof of Lemma 3). Profitability is again ensured by Assumption 2. ■

Proof of Proposition 6. By Lemma 5, the condition for an equilibrium where all CPs use uncompressed transmission ($\lambda_1 = 0, \lambda_2 = 1$) is

$$\frac{c-k}{u} \geq \frac{B}{1+\mu}\left(1-\frac{B}{1+\mu}\right). \tag{39}$$

A sufficient condition for the second-best optimum to call for compressed transmission is $c \leq uB(1+\mu-B/2)/(1+\mu)^2$, or

$$\frac{c-k}{u} \leq \frac{B}{1+\mu}\left(1-\frac{B}{2(1+\mu)}\right)-\frac{k}{u}. \tag{40}$$

Assumption 2 implies that

$$\frac{B}{1+\mu}\left(1-\frac{B}{2(1+\mu)}\right)-\frac{k}{u} \geq \frac{B}{1+\mu}\left(1-\frac{B}{1+\mu}\right).$$

Hence, there always exists a value of $(c-k)/u$ such that (39) and (40) are simultaneously satisfied. ■

Proof of Proposition 7. Marginal revenue is positive and constant on $[0, \hat{\alpha}_{dp}]$, negative and constant on $(\hat{\alpha}_{dp}, 1]$, increasing on $(1, 2B/(2\mu-B)]$, and constant on $(2B/(2\mu-B), 2]$. Thus, we only need to compare the three corner solutions $\alpha = \hat{\alpha}_{dp}$, $\alpha = 1$, and $\alpha = 2$. The ISP prefers charging $t_f(\hat{\alpha}_{dp})$ to $t_f(1)$ if and only if $\mu\hat{\alpha}_{dp}t_f(\hat{\alpha}_{dp}) \geq \mu t_f(1)$, or

$$B(u-k-c) \geq \mu[uB/\mu-k-c] \;\Leftrightarrow\; B \leq \mu,$$

which is satisfied by assumption. The ISP prefers charging $t_f(\hat{\alpha}_{dp})$ to $t_f(2)$ if and only if $\mu\hat{\alpha}_{dp}t_f(\hat{\alpha}_{dp}) \geq 2\mu t_f(2)$, or

$$B(u-k-c) \geq 2\mu\left[c-k-u\frac{B}{2\mu}\left(1-\frac{B}{2\mu}\right)\right]$$

$$\Leftrightarrow\; uB\left(2-\frac{B}{2\mu}\right) \geq c(2\mu+B)-k(2\mu-B).$$

Since $\mu \geq B$, a sufficient condition for this is

$$u\frac{B}{3\mu}\left(2-\frac{B}{2\mu}\right) \geq c.$$

Under condition (21), it suffices to show that

$$u\frac{B}{3\mu}\left(2-\frac{B}{2\mu}\right) \geq u\frac{B}{2\mu}\left(1-\frac{B}{4\mu}\right)$$

$$\Leftrightarrow\; 2\left(2-\frac{B}{2\mu}\right) \geq 3\left(1-\frac{B}{4\mu}\right) \;\Leftrightarrow\; 1 > \frac{B}{4\mu},$$

which is always satisfied since $B \leq \mu$. ■

Proof of Lemma 6. Reduced quality is more profitable than standard quality if and only if $u\delta(A)^2 - 2k \leq \beta u\delta(A) - k$, or

$$\delta(A)(\delta(A)-\beta) \leq k/u. \tag{41}$$

Reduced-quality transmission allows CPs to make nonnegative profit if and only if $\beta\delta \geq k/u$. The assumption that $\beta \geq 1/2$ implies that $\beta\delta \geq \delta(\delta - \beta)$ for all $\delta \leq 1$. Hence, if reduced-quality transmission is unprofitable, then standard-quality transmission is also unprofitable.

If all time-sensitive CPs are active and send reduced quality, we have $A = 1$ and $\delta = B$. Thus, for $k/u \geq \beta B$, it is not an equilibrium for all time-sensitive CPs to be active; instead, the equilibrium is such that they randomize in a way that makes them indifferent between being active and inactive, as captured by (22). For all time-sensitive CPs to make a profit with reduced quality transmission and prefer it to standard transmission, it must be that $\beta B \geq k/u \geq B(B - \beta)$.

If all time-sensitive CPs send standard quality, we have $A = 1 + \mu$ and $\delta = B/(1 + \mu)$. For all time-sensitive CPs to prefer standard-quality to reduced-quality transmission, it must be that $k/u \leq B(B-\beta(1+\mu))/(1+\mu)^2$. Moreover, CPs must make nonnegative profit, which requires $u\delta(A)^2 - 2k \geq 0$ or $k/u \leq B/(2(1 + \mu)^2)$. The assumption $\beta \geq 1/2$ implies that $\delta^2/2 \geq \delta(\delta - \beta)$ for all $\delta \leq 1$; hence the condition $k/u \leq B(B - \beta(1 + \mu))/(1 + \mu)^2$ is sufficient for an equilibrium with only standard quality transmission. Finally, randomizing between reduced and standard quality requires that CPs are indifferent between the two, i.e., $\delta(A)(\delta(A)-\beta) = k/u$, which is possible for $B(B-\beta(1+\mu))/(1+\mu)^2 < k/u < B(B-\beta)$. ■

Proof of Proposition 8. This follows immediately from Lemma 6, showing that $\lambda_2^{qr} > 0$ for $k/u < B(B - \beta)$, and the second-best optimal policy requiring $\lambda_2 = 0$ for $\beta \geq 1/2$. ■

Proof of Proposition 9. Marginal revenue is positive and constant on $[0, \hat{\alpha}_{dp}]$ and negative on $(\hat{\alpha}_{dp}, 2]$. Thus, the revenue-maximizing traffic volume for the ISP is $\hat{\alpha}_{dp}$. ■

References

Anderson, S.P., Coate, S. (2005): Market Provision of Broadcasting: A Welfare Analysis. *Review of Economic Studies* 72(4): 947–972.

Anderson, S.P., De Palma, A. (2009): Information congestion. *RAND Journal of Economics* 40(4): 688–709.

Cheng, H.K., Bandyopadhyay, S., Guo, H. (2011): The Debate on Net Neutrality: A Policy Perspective. *Information Systems Research* 22(1): 60–82.

Bourreau, M., Kourandi, F., Valletti, T. (2014): Net Neutrality with Competing Internet Platforms. *Journal of Industrial Economics*. Forthcoming.

Choi, J.P., Jeon, D.S., Kim, B.C. (2014): Internet Interconnection, Business Models, and Network Neutrality. *American Economic Journal: Microeconomics*. Forthcoming.

Choi, J.P., Jeon, D.S., Kim, B.C. (2013): Asymmetric Neutrality Regulation and Innovation at the Edges: Fixed vs. Mobile Networks. NET Institute Working Paper 13-24.

Choi, J.P., Kim, B.C. (2010): Net Neutrality and Investment Incentives. *RAND Journal of Economics* 41(3): 446–471.

De Cicco, L., Mascolo, S., Palmisano, V. (2011): Skype Video congestion control: An experimental investigation. *Computer Networks* 55(3): 558–571.

Economides, N., Hermalin, B.E. (2012): The economics of network neutrality. *RAND Journal of Economics* 43(4): 602–629.

Economides, N., Tåg, J. (2012): Net Neutrality on the Internet: A Two-Sided Market Analysis. *Information Economics and Policy* 24(2): 91–104.

Economist (2014): Video in Demand, The Economist Technology Quarterly, December 6, 2014, page 5.

Hermalin, B.E., Katz, M.L. (2007): The Economics of Product-Line Restrictions with an Application to the Network Neutrality Debate. *Information Economics and Policy* 19: 215–248.

Jullien, B., Sand-Zantman, W. (2013): Pricing Internet Traffic: Exclusion, Signalling, and Screening. Working Paper, Toulouse School of Economics.

Kourandi, F., Krämer, J., Valletti, T. (2014): Net Neutrality, Exclusivity Contracts and Internet Fragmentation. *Information Systems Research*. Forthcoming.

Krämer, J., Wiewiorra, L. (2012): Network neutrality and congestion sensitive content providers: Implications for content variety, broadband investment, and regulation. *Information Systems Research* 23(4): 1303–1321.

Peitz, M., Reisinger, M. (2014): The Economics of Internet Media. Working Paper 14-23, University of Mannheim.

Van Zandt, T. (2004): Information overload in a network of targeted communication. *RAND Journal of Economics* pp. 542–560.