Net neutrality and inflation of traffic*

Martin Peitz† Florian Schuett‡

First version: April 2013; this version: December 2013

PRELIMINARY AND INCOMPLETE

Abstract

Internet service providers may be required to carry data without any differentiation and at no cost to the content provider. However, the amount of traffic is endogenous implying that traffic volumes depend on the prevailing regime. We provide a simple framework in which under some conditions net neutrality which includes a single transport class and a zero price leads to socially inefficient inflation of traffic. By contrast, deviations from net neutrality may implement the socially optimal allocation.

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

*We thank Cédric Argenton, Jan Boone, Jan Krämer, Jens Prüfer, Bert Willems, Gijsbert Zwart, participants at a TILEC (Tilburg University) seminar and at the June 2013 "Economics of ICT"-conference in Mannheim for helpful comments. Martin Peitz gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (SFB TR 15).

†Department of Economics, University of Mannheim, D-68131 Mannheim, Germany. Email: martin.peitz@googlemail.com. Also affiliated with CEPR, CESifo, ENCORE, 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 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 indirect network externalities 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 inflate traffic in case traffic is underpriced. We show that strict net neutrality rules lead to traffic inflation and tend to lead to a loss in social welfare.

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. Second, some content is more sensitive to delay than other content. More 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* – 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.

From an economic point of view, congestion control techniques of this type create externalities in traffic generation: although they reduce individual delay, they increase aggregate congestion on the network. Under net neutrality (best effort for all traffic, no prioritization, zero prices on the content side), the network essentially constitutes a common property resource. Net neutrality therefore leads to excessive exploitation by time-sensitive CPs. By charging for time-sensitive traffic and handling it 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

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

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.

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, some congestion 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 net neutrality – 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 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 efficient level. Even 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.

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 available instruments of 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). 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.

Independently, Choi *et al.* (2013) consider congestion externalities on the internet. They consider 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.

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 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 homogenous with respect to content and derive a utility u from each content provider whose content is delivered in time.

The continuum of content providers is normalized to mass 1. Content providers come in two categories. Content providers of category 1 offer time-sensitive content, while content providers of category 2 only 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 as email and delayed on-demand movies may arrive a bit later without much loss 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 quality of 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 ISPs and how much content is sent by content providers. Network capacity acts as 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 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 and minimum quality of service) (fast lane and slow lane, use of slow lane free);
- regime 5: unregulated tiering without price restrictions or minimum quality of service (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. 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. CPs choose p_i.
2. ISP announces subscription price s and transmission fee t per unit of content (possibly conditional on priority classes).
3. CPs decide whether to be active and choose α_i. Consumers choose whether to buy internet access from ISP.
4. Consumers choose whether to connect to CPs and pay s.
5. Content of CP i is delivered to consumers on time with probability $\gamma(\alpha_i, \alpha)$. Consumers pay p_i for each unit delivered on time to CP i; CPs pay t to ISP for each unit of traffic carried (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

We postulate that viewers are homogeneous and have demand for a single unit of each content provider. The valuation for each such unit is denoted by u. Hence, each content provider i will set $p_i = u$, which is collected only if the content reaches the consumer (which happens with probability γ). Profit of a time-sensitive content provider is $\gamma(\alpha_i, \alpha)u - k\alpha_i$, where $\alpha \equiv (\int_0^\mu \alpha_i di)/\mu$ is the average number of packets sent by time-sensitive CPs.[2] 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 carried on the network. Sending a packet several times increases the probability that at least one packet arrives on time. To be more precise, let us distinguish two systems of content delivery: a one-tiered system, in which all traffic is routed according to a best-effort principle, and a two-tiered system, in which some traffic is prioritized in times of bandwidth shortage. Let B denote the ISP's available bandwidth (or network capacity). In a one-tiered system, the probability of reaching a consumer when sending a package α_i times is

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

where

$$\delta(\alpha) = \min\left\{\frac{B}{\mu\alpha + 1 - \mu}, 1\right\}.$$

In a two-tiered system in which time-sensitive content is prioritized, so that time-insensitive traffic does not occupy any bandwidth in times of shortage, the probability of on-time delivery is

$$\gamma(\alpha_i, \alpha) = 1 - (1 - \tilde{\delta}(\alpha))^{\alpha_i}, \tag{2}$$

where

$$\tilde{\delta}(\alpha) = \min\left\{\frac{B}{\mu\alpha}, 1\right\}.$$

Suppose that each content provider can send a package once, twice, or not at all, i.e., $\alpha_i \in \{0, 1, 2\}$.[3] Assume moreover that $B < 1$, which implies that in a one-tiered system, if each CP sends one packet (so $\alpha = 1$), not all time-sensitive content can be delivered on time.

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

[2] 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.

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

traffic is $\alpha = \mu(\lambda_1 + 2\lambda_2) + (1 - \mu)$. Then, consumers purchase internet access if and only if

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

Since, in an 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.[4]

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 on-time delivery is given by (2). We also consider a *second best* world in which all content is routed through a single tier according to a best-effort principle, and where the probability of delivery is thus given by (1).

Total surplus in the market for time-sensitive content is[5]

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

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 time-sensitive CPs randomize over $\alpha_i = 1$ and $\alpha_i = 2$ with probabilities $\lambda_1 > 0$ and $\lambda_2 = 1 - \lambda_1$, respectively, 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 $\tilde{\delta}(\alpha) = 1$ for $\alpha \leq \hat{\alpha}_{dp}$ and $\tilde{\delta}(\alpha) < 1$ 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 $\hat{\alpha}_{dp} = B/\mu$ and $\hat{\alpha}_{nn} = \max\{0, (B - (1 - \mu))/\mu\}$. If the 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 lemmata characterize first-best and second-best traffic volumes, respectively. They provide a natural benchmark to compare equilibrium outcomes with in the various regimes considered below.

[4]While we restrict our analysis to fixed capacity of the ISP an immediate consequence of this finding is that in this admittedly extreme setting the ISP has no incentive to increase capacity if expanding capacity is costly.

[5]The function W is based on 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, keeping α fixed. However, the increase in probability of delivery for those sending 2 packages is less than the decrease for those sending 0: $\gamma(2, \alpha) - \gamma(1, \alpha) < \gamma(1, \alpha) \Leftrightarrow \delta(\alpha)(1 - \delta(\alpha)) < \delta(\alpha)$.

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 the elasticity (in absolute value) of the delivery probability $\tilde{\delta}$ equals one:

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

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., $\alpha\tilde{\delta}(\alpha)$ is invariant with respect to α). The increase in available content is exactly offset by a decrease in delivery probability. While an increase in traffic has no effect on consumer surplus, it raises cost (αk) and is therefore undesirable from a total surplus perspective.

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*

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

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*

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

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-efforts 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:

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

That is, raising α beyond $\hat{\alpha}_{nn}$ leads to an increase in the amount of time-sensitive content delivered without delay because the reduction in the delivery probability is smaller than the increase in available content. The intuition for this result is that part of the congestion cost is borne by time-insensitive content. By definition, time-insensitive content can be delayed without reducing consumer surplus. Sending more time-sensitive traffic increases the probability that this content is delivered on time; although it 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.

3 Market equilibrium

3.1 Net neutrality

We now study equilibrium traffic in a regime of net neutrality, where all content is routed through a single tier. 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, $u\gamma(1,\alpha)-k$, to the profit from sending it twice, $u\gamma(2,\alpha) - 2k$, or not at all (yielding zero), taking as given the average traffic α. 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*

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

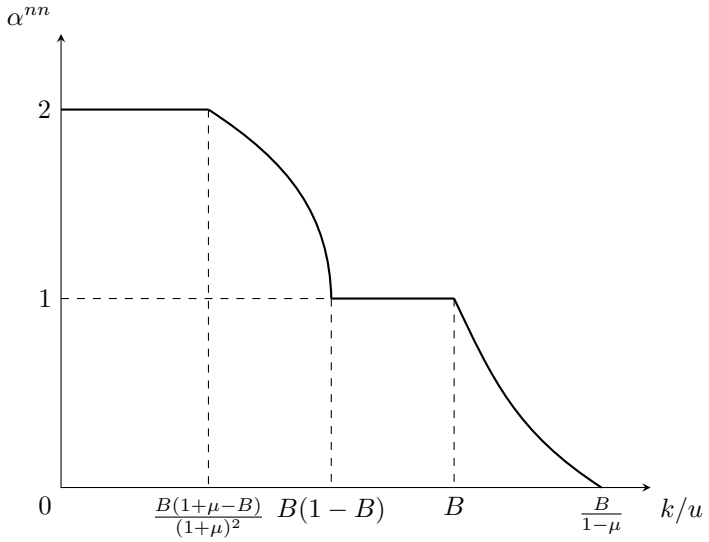

Figure 1: The equilibrium under net neutrality when $B < 1/2$

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 is at least one mixed-strategy equilibrium in which time-sensitive CPs randomize over* $\alpha_i = 1$ *(probability* $2-\alpha^{nn}$*) and* $\alpha_i = 1$ *(probability* $\alpha^{nn}-1$*), where* $\alpha^{nn} \in (1,2)$ *solves*

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

No other symmetric equilibrium exists.

According to Lemma 3, for a given value of k/u, it is possible that there is multiplicity of equilibria. There can be multiple pure-strategy equilibria: for some parameter values, both $\alpha^{nn} = 1$ and $\alpha^{nn} = 2$ can 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 $1/2 \notin (B/(1+\mu), B)$, $\delta(\alpha)(1-\delta(\alpha)) = k/u$ 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 $\alpha^{nn} \in (1,2)$.

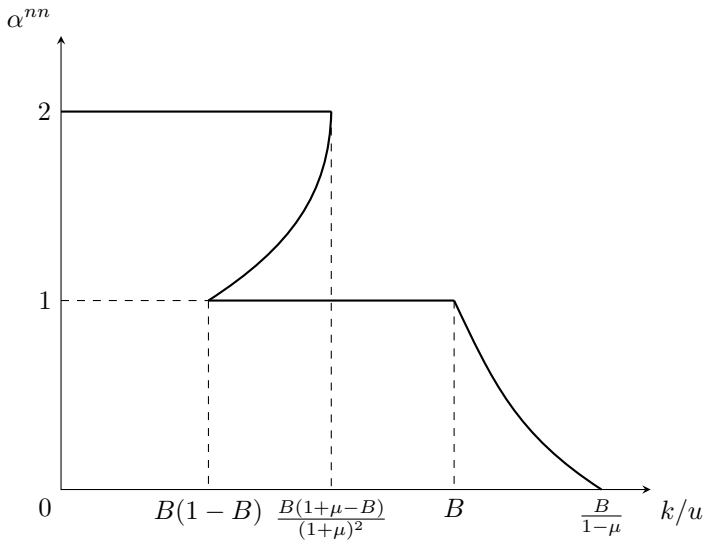

Figure 2: The equilibrium under net neutrality when $B/(1+\mu) > 1/2$

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

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 on 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 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. The probability that a given packet is delivered without delay is $\tilde{\delta}(\alpha) = \frac{B}{\mu\alpha}$, and the probability of on-time delivery for content sent α_i times is given by (2). Thus, time-sensitive content has a higher probability of being delivered on time for any given α. Letting $\bar{\tilde{\delta}} = \arg\max_{B/2\mu \leq \tilde{\delta} \leq B/\mu} \tilde{\delta}(1-\tilde{\delta})$, 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*

$$\tilde{\delta}(\alpha^{dp}) = \frac{k}{u} \qquad \textit{(partial availability)}, \tag{8}$$

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 $\min\{B(\mu-B)/\mu^2 B(\mu-B)/\mu^2, B(2\mu-B)/4\mu^2\} < k/u < \bar{\bar{\delta}}(1-\bar{\bar{\delta}})$, there is at least one mixed-strategy equilibrium in which time-sensitive CPs randomize over $\alpha_i = 1$ (probability $2 - \alpha^{dp}$) and $\alpha_i = 1$ (probability $\alpha^{dp} - 1$), where $\alpha^{dp} \in (1, 2)$ solves*

$$\tilde{\delta}(\alpha^{dp}) \left(1 - \tilde{\delta}(\alpha^{nn})\right) = \frac{k}{u} \qquad \textit{(partial duplication)}. \tag{9}$$

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 2 *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 2 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.[6] From (6) and (8), it must then be that $\delta(\alpha^{nn}) = \tilde{\delta}(\alpha^{dp})$. Using the definitions of δ and $\tilde{\delta}$, we find that

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

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

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 probability of delivery, 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, time-sensitive content has a higher overall probability of delivery under deep packet inspection, as there is a larger proportion of time-sensitive CPs sending their packets twice ($\alpha^{dp} > \alpha^{nn}$). Formally, $\delta(\alpha^{nn}) = \tilde{\delta}(\alpha^{dp})$ implies that

$$\gamma(\alpha_i, \alpha^{nn}) = 1 - (1 - \delta(\alpha^{nn}))^{\alpha_i} = 1 - (1 - \delta(\alpha^{dp}))^{\alpha_i} = \gamma(\alpha_i, \alpha^{dp}).$$

Because $\gamma(2, \alpha) > \gamma(1, \alpha)$, the overall delivery probability for time-sensitive content,

$$(\alpha - 1)\gamma(2, \alpha) + (2 - \alpha)\gamma(1, \alpha),$$

is higher under deep packet inspection than under net neutrality.

We find that deep packet inspection does not necessarily increase delivery probabilities, and may even decrease 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 $\gamma(2, 2) = 1-(1-B/(2\mu))^2$. 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 3 *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 constellation, deep packet inspection actually performs worse than (strict) net neutrality. Thus, deep packet inspection alone cannot reliably fix the problem of traffic inflation. We now turn to transmission fees.

3.3 A uniform transmission fee

Suppose that the ISP routes all traffic according to a best-efforts principle (no prioritization) and charges 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

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

Thus, the equilibrium is the same as under net neutrality (see Subsection 3.1) replacing k by $k+t$, and the demand from time-sensitive CPs $\alpha(t)$ facing the ISP for a given t is equal to the equilibrium traffic, i.e. $\alpha(t) = \alpha^{nn}$. The equilibrium is unique for all $k + t$ if and only $B < 1/2$, which we will assume is the case in what follows to ensure that demand is well behaved. The inverse demand is given by

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

The ISP's problem is

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

Using (10), we can compute

$$t'(\alpha) = \begin{cases} 0 & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{nn} \\ -\frac{\mu}{B}u(\delta(\alpha))^2 & \text{for } \hat{\alpha}_{nn} < \alpha \leq 1 \\ -\frac{\mu}{B}u(\delta(\alpha))^2(1 - 2\delta(\alpha)) & \text{for } 1 < \alpha \leq 2, \end{cases} \tag{11}$$

from which we deduce the ISP's marginal revenue, $MR(\alpha) = t'(\alpha)(1 + \mu(\alpha - 1)) + \mu t(\alpha)$, noting that $(1 + \mu(\alpha - 1))/B = 1/\delta(\alpha)$:

$$MR(\alpha) = \begin{cases} \mu(u - k) & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{nn} \\ -\mu k & \text{for } \hat{\alpha}_{nn} < \alpha \leq 1 \\ \mu(u(\delta(\alpha))^2 - k) & \text{for } 1 < \alpha \leq 2. \end{cases} \tag{12}$$

Thus, marginal revenue is constant and positive for $\alpha \in [0, \hat{\alpha}_{nn}]$, constant and negative for $\alpha \in (\hat{\alpha}_{nn}, 1]$, and decreasing for $\alpha \in (1, 2]$. There are thus three potential solutions to the ISP's problem: two corner solutions, $\alpha = \max\{0, \hat{\alpha}_{nn}\}$ and $\alpha = 2$, as well as one interior solution solving $\delta(\alpha) = \sqrt{k/u}$, which exists if and only if $B/(1+\mu) < \sqrt{k/u} < B$. As the following lemma shows, the ISP always chooses $\alpha = \max\{0, \hat{\alpha}_{nn}\}$.

Proposition 4 *The transmission fee that maximizes the ISP's profit prices out congestion, i.e., t is such that $\alpha = \hat{\alpha}_{nn}$.*

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 entirely. This result is reminiscent of Anderson and De Palma (2009), where a monopoly gatekeeper prices out information congestion. In contrast to that paper, however, here the fee is in general excessive, compared to what 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.

3.4 Bandwidth tiering

The next regime we examine is one where 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 this 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 $\alpha(t_f)$ is the demand for priority service when only time-sensitive content is transmitted via the fast lane, which is equivalent to the equilibrium traffic under deep packet inspection, α^{dp}, as derived in Lemma 4, replacing k by $k + t_f$. Because of multiplicity of equilibria, we need to specify which equilibrium is selected for each possible t_f 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 congestion.

If $B \geq \mu$, there is no congestion at $\alpha = 1$. Under the above selection rule, the inverse demand for traffic on the fast lane then is

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

owing to the fact that $\alpha = 2$ is an equilibrium for $(k + t_f)/u \leq \tilde{\delta}(2)(1 - \tilde{\delta}(2))$ by Lemma 4. If instead $B < \mu$, inverse demand for traffic on the fast lane is

$$t_f(\alpha) = \begin{cases} u - k & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{dp} \\ u\tilde{\delta}(\alpha) - k & \text{for } \hat{\alpha}_{dp} < \alpha \leq 1 \\ u/4 - k & \text{for } 1 < \alpha \leq \max\{1, B/(2\mu)\} \\ u\tilde{\delta}(\alpha)(1 - \tilde{\delta}(\alpha)) - k & \text{for } \max\{1, B/(2\mu)\} < \alpha \leq 2, \end{cases} \tag{14}$$

owing to the fact that the equilibrium selected is $\alpha^{dp} \in (1, 2)$ solving (9) for $\tilde{\delta}(2)(1-\tilde{\delta}(2)) < (k + t_f)/u < \max\{1/4, \tilde{\delta}(1)(1 - \tilde{\delta}(1))\}$ and $\alpha = 2$ for $(k + t_f)/u \leq \tilde{\delta}(2)(1 - \tilde{\delta}(2))$.

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 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 $MR(\alpha) = t_f'(\alpha)(1 + \mu(\alpha - 1)) + \mu t_f(\alpha)$.

If $B \geq \mu$, it follows from (13) that $t_f' = 0$ for all α, so marginal revenue is everywhere positive, and we can restrict attention to the corner solutions $\alpha = 1$ and $\alpha = 2$. If $B < \mu$, then using (14) and noting that $\tilde{\delta}' = -(\mu/B)\tilde{\delta}^2$, we can compute

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

Hence, the ISP's marginal revenue in this case, noting that $\mu\alpha/B = 1/\tilde{\delta}(\alpha)$, becomes:

$$MR(\alpha) = \begin{cases} \mu(u - k) & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{dp} \\ -\mu(k + (1 - \mu)\frac{u}{B}\tilde{\delta}(\alpha)) & \text{for } \hat{\alpha}_{dp} < \alpha \leq 1 \\ \mu(u/4 - k) & \text{for } 1 < \alpha \leq \max\{1, B/(2\mu)\} \\ \mu\left(u(\tilde{\delta}(\alpha))^2\left(1 - \frac{(1-\mu)(1-2\tilde{\delta}(\alpha))}{B}\right) - k\right) & \text{for } \max\{1, B/(2\mu)\} < \alpha \leq 2. \end{cases} \tag{16}$$

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, he 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 (14). If $B \geq \mu$, then $t_f' = 0$ (except at $\alpha = 1$, where it is undefined), so it suffices to look at the corner solutions $\alpha = 1$ and $\alpha = 2$. If $B < \mu$, then t_f' is as derived in (15), from which we

deduce marginal revenue $MR(\alpha) = \mu(\alpha t_f'(\alpha) + t_f(\alpha))$,

$$MR(\alpha) = \begin{cases} \mu(u-k) & \text{for } 0 \leq \alpha \leq \hat{\alpha}_{dp} \\ -\mu k & \text{for } \hat{\alpha}_{dp} < \alpha \leq 1 \\ \mu(u/4-k) & \text{for } 1 < \alpha \leq \max\{1, B/(2\mu)\} \\ \mu\left(u(\tilde{\delta}(\alpha))^2 - k\right) & \text{for } \max\{1, B/(2\mu)\} < \alpha \leq 2. \end{cases} \tag{17}$$

The following lemma derives 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 $t_s = t_f$.*

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 on (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.[7]

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 also 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 Conclusion

We present a model of congestion on the internet in which there is time-sensitive content, which needs to be delivered on time for consumers to derive utility from it, and time-insensitive content, for which 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

[7]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.

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 net neutrality may not be a good idea for network congestion. Net neutrality effectively turns the network into a common property resource and does not address the externalities in traffic generation. We show that departures from strict net neutrality can alleviate the problem. Deep packet inspection may eliminate congestion, and thus the incentive to inflate traffic. However, the result is ambiguous since deep packet inspection is possibly inferior to strict net neutrality. That is, deep packet inspection can backfire if CPs respond by increasing traffic. Alternatively, if the ISP can charge a transmission fees, it will price out congestion.

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. Even better outcomes can be achieved by allowing the ISP to engage in bandwidth tiering and price discrimination. This can implement the first-best allocation without any need for regulatory intervention.

Appendix: Proofs

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

$$W(\alpha) = \begin{cases} \alpha(u-k) & \text{for } \alpha \in [0, \min\{\hat{\alpha}_{dp}, 1\}) \\ \alpha(u\tilde{\delta}(\alpha) - k) & \text{for } \alpha \in [\min\{\hat{\alpha}_{dp}, 1\}, 1] \\ u - \alpha k & \text{for } \alpha \in [1, \max\{\hat{\alpha}_{dp}, 1\}] \\ u\tilde{\delta}(\alpha)\left[\alpha(1-\tilde{\delta}(\alpha)) + \tilde{\delta}(\alpha)\right] - \alpha k & \text{for } \alpha \in (\max\{\hat{\alpha}_{dp}, 1\}, 2]. \end{cases} \tag{18}$$

Differentiating (18) and using $\tilde{\delta}' = -(\mu/B)\tilde{\delta}^2$ yields

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

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

$$W(\alpha) = \begin{cases} \alpha[u\delta(\alpha) - k] & \text{for } \alpha \in [0,1] \\ u\delta(\alpha)\left[\alpha(1-\delta(\alpha)) + \delta(\alpha)\right] - \alpha k & \text{for } \alpha \in (1,2] \end{cases} \tag{19}$$

Since, for $\alpha \geq \hat{\alpha}_{nn}$, $\delta(\alpha)$ is strictly monotonic in α, it can be inverted. Let $\alpha(\delta) = (B/\delta - (1-\mu))/\mu$ denote its inverse 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 (19), we thus obtain

$$\hat{W}(\delta) = \begin{cases} \frac{1}{\mu}(B/\delta - (1-\mu))(u\delta - k) & \text{for } \delta \in \left[B, \min\left\{\frac{B}{1-\mu}, 1\right\}\right] \\ \\ \frac{1}{\mu}\left[u\left(\delta^2 - \delta(B+1-\mu) + B\right) - k\left(B/\delta - (1-\mu)\right)\right] & \text{for } \delta \in \left[\frac{B}{1+\mu}, B\right). \end{cases} \tag{20}$$

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 (20) and rearranging yields

$$\frac{B}{1+\mu}\left(1 - \mu - \frac{B}{1+\mu}\right) \leq \frac{k}{u}. \tag{21}$$

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

$$\frac{k}{u} \leq w(\delta_{\max}). \tag{22}$$

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

$$w(\delta_{\max}) < (1-\mu)\delta_{\max}/B \leq (1-\mu)B. \tag{23}$$

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 (21) that $\hat{W}(B) \geq \hat{W}(B/(1+\mu))$, and from (23) 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 (21) and rules out existence of a local maximum on $[B/(1+\mu), B)$ by (23) and (22), 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 (21) 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 (22) and (23). Hence, what remains to be shown is that there exists 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

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

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. [TO BE ADDED.] ■

Proof of Proposition 1. [TO BE ADDED.] ■

Proof of Lemma 4. [TO BE ADDED.] ■

Proof of Proposition 2. 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 4. The ISP's profit when $\alpha = \hat{\alpha}_{nn}$ is given by $\pi_0^{\text{ISP}} = \max\{B, 1-\mu\}(u-k)$. His profit when setting α such that $\delta(\alpha) = \sqrt{k/u}$ (which is greater or equal to his profit when setting $\alpha = 2$) is $\pi_1^{\text{ISP}} = B(u - 2\sqrt{uk})$. We have $\pi_0^{\text{ISP}} > \pi_1^{\text{ISP}}$ if and only if $k < 2\sqrt{uk}$, which is always satisfied. ■

Proof of Proposition 5. We start by considering the case where $B \geq \mu$. Under regulated tiering with $t_s = 0$, the ISP 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 $\min\ 2\mu^2 - B(2\mu - B) = B^2/2 > 0$. Hence, $\alpha = 1$ is optimal.

Without regulatory restrictions on t_s, we have seen that the ISP will set $t_s = t_f$. The ISP prefers $t_f(1)$ to $t_f(2)$ if and only if

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

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

a sufficient condition for which is $4\mu^2 - (1+\mu)B(2\mu - B) \geq 0$. The value of μ that minimizes this expression is $\mu = B/4$, yielding $\min\ 4\mu^2 - (1+\mu)B(2\mu - B) = B^2(3 + B/2)/4 > 0$. Again, $\alpha = 1$ is optimal.

We now turn to the case where $B < \mu$. Under regulated tiering with $t_s = 0$, we observe that marginal revenue is decreasing only on $\max\{1, B/(2\mu)\}, 2$. Thus an interior solution, if it exists, solves $\tilde{\delta}(\alpha) = \sqrt{k/u}$, yielding $\alpha = B/(\mu\sqrt{k/u})$. The ISP's profit when setting $\alpha = B/(\mu\sqrt{k/u})$ (which is greater or equal to his profit when setting $\alpha = \max\{1, B/(2\mu)\}$ or $\alpha = 2$) is $B(u - 2\sqrt{uk})$. His profit at $\alpha = \hat{\alpha}_{dp}$ is $\mu(u - k)$. Since $\mu > B$ and $k < 2\sqrt{uk}$, $\hat{\alpha}_{dp}$ is optimal.

Finally we establish that $\hat{\alpha}_{dp}$ is optimal also in the absence of regulation on t_s. When the ISP can set $t_s > 0$, 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 $t_f(\hat{\alpha}_{dp})$ is optimal when $t_s = 0$, it must be optimal *a fortiori* when $t_s = t_f$. ■

References

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.

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.

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.

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.

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