Network neutrality and congestion sensitive content providers: Implications for content variety, broadband investment and regulation

Jan Krämer and Lukas Wiewiorra

Karlsruhe Institute of Technology

5. September 2009

Forthcoming at *Information Systems Research*
manuscript ISR-2011-069

Network Neutrality and Congestion Sensitive Content Providers: Implications for Content Variety, Broadband Investment and Regulation

Jan Krämer, Lukas Wiewiorra
Karlsruhe Institute of Technology, kraemer@kit.edu, wiewiorra@kit.edu

We study departures from network neutrality through implementing a Quality of Service tiering regime in which an ISP charges for prioritization on a non-discriminatory basis. We find that Quality of Service tiering may be more efficient in the short run, because it better allocates the existing network capacity, and in the long run, because it provides higher investment incentives due to the increased demand for priority services by the entry of new congestion sensitive content providers. Which network regime is the most efficient depends on the distribution of congestion sensitivity among content providers, but a guideline is that the regime that provides higher incentives for infrastructure investments is more efficient in the long run.

Key words: telecommunications; net neutrality; quality of service; content variety; investment; regulation

1. Introduction

The most controversial part of the net neutrality debate is the question of the future relationship between Internet service providers (ISPs) and content providers (CPs). We seek to investigate in particular whether ISPs should be allowed to offer CPs differentiated service classes for the transmission of their data packets to end customers—known as Quality of Service (QoS) tiering (Lessig 2001, p.46; Hahn and Wallsten 2006). Under a network neutrality regime the prioritization of paid-for traffic would be prohibited, even if QoS tiering was offered on a non-discriminatory basis.[1]

Proponents of network neutrality argue that only this regime can ensure a level playing field for competition among CPs and will thus lead to more content variety (Lessig 2001, p.168–175; Wu 2003, Van Schewick 2006, Sydell 2006). At a given transmission capacity, the acceleration of priority traffic will lead unmistakably to a deceleration of the remaining best-effort traffic. Thus, those CPs who are not willing to pay for priority access are put at a disadvantage because other

[1] Here, 'non-discrimination' means that CPs can self-select whether they want to buy priority treatment for their data packets or not. Within each traffic class, all data packets are handled equally. However, in the past network providers have often discriminated data packets based on their content type. Examples for such anti-competitive behavior are the blocking of voice-over-IP transmissions by mobile network operators (Hahn et al. 2007) and the degradation of peer-to-peer traffic (O'Connell 2005).

providers' content is accelerated in lieu of their own content. This disadvantage lies at the heart of the concern of network neutrality proponents. Moreover, advocates claim that in the long run, broadband infrastructure investments are likely to be higher under network neutrality, because ISPs are forced to provide sufficient bandwidth in order to bring new content types on line and keep consumers satisfied (Lessig 2001, p.47). They even express the fear that QoS tiering may in fact hinder the roll-out of additional transmission capacity, because ISPs seek to charge CPs for exactly this resource, which is only possible if it is scarce (Wu and Yoo 2007, Choi and Kim 2010).

Opponents of a network neutrality regime argue to the contrary that QoS tiering will stimulate more content variety and broadband investment. A CP who offers an Internet telephony service, for example, is certainly more sensitive to network congestion than a simple e-mail service provider.[2] Consequently, opponents argue that the best-effort one-size-fits-all transmission regime of a neutral network is not appropriate anymore (Yoo 2005). If customers' experience of use is unsatisfactory because a CP's service cannot be reliably offered, this CP's advertisement revenues will decline, possibly up to the point where it is forced out of business (Crowcroft 2007). Hence, QoS tiering may in fact be welfare-enhancing, because it explicitly enables entry by those innovative CPs who crucially hinge on transmission quality requirements which the traditional neutral best-effort Internet may soon be unable to provide. By contrast, net neutrality could in fact hinder entry of innovative CPs, because congestion sensitive services can only be offered if the they are sustainable under the best-effort domain. Furthermore, supporters of QoS tiering argue that investments in broadband infrastructure would be higher under this regime because CPs can be billed for the transmission quality they are using (Van Schewick 2006, Yoo 2005). Even if transmission capacity is not extended,[3] QoS tiering would still handle the existing capacity more efficiently.

In light of the arguments for and against network neutrality regulation, some observers have noted that the debate seems stuck in the sense that "at this point, it is impossible to foresee which architecture will ultimately represent the best approach"(Wu and Yoo 2007).[4] Nevertheless, the Federal Communications Commission (FCC) has issued a regulatory framework which prohibits QoS tiering (FCC 2010) and is challenged in courts. In an effort to advance the debate, we provide a formal economic framework which incorporates the arguments of either side. This allows us to compare QoS tiering with network neutrality in terms of their impacts on content variety, broadband investment and overall welfare. More specifically, we model the Internet as a two-sided market (Armstrong 2006, Rochet and Tirole 2006) that connects congestion sensitive CPs with consumers, and which is controlled by a monopolistic ISP.[5] Under network neutrality regulation

[2] For expositional simplicity, in the following we will often use 'congestion' or 'speed' as a proxy for different transmission quality measures, such as bandwidth, latency or jitter.

[3] Crowcroft (2007), for example, suggests that QoS tiering is a "zero sum game at any instant".

[4] For a similar argument see Owen and Rosston (2006)

[5] In the context of network neutrality, the two-sided market framework was first suggested by Sidak (2006a,b).

all CPs experience the same transmission quality, whereas under the QoS tiering regime, every CP can choose to buy priority access to consumers on the same non-discriminatory conditions. That is, discrimination occurs only between the best-effort and the priority class, but not within each class. In addition to the network externalities that are generated by either side, we explicitly consider the adverse effect that traffic prioritization has on the transmission quality of the remaining best-effort class as well as the positive effect that congestion is allocated away from the most congestion sensitive CPs. In this framework, we investigate the effects of QoS tiering both in the short run, when network capacity is fixed, as well as in the long run, when the ISP can strategically invest in broadband infrastructure.

Our main results are that in the short run QoS tiering will lead to the same level of content variety as network neutrality if CPs' congestion sensitivity is uniformly distributed in the Internet economy. However, because QoS tiering allocates congestion better to the congestion insensitive CPs, overall short run welfare is generally higher under this regime. Nevertheless, it should also be clear that QoS tiering enables ISPs to expropriate some of the CPs' revenues and thus, in the short run, all CPs are worse off under QoS tiering than under network neutrality. Indeed, this fact has driven much of the emotionality in the debate.[6] Although the shift of revenues from CPs to the ISP is welfare neutral per se, it will generally still need to be scrutinized by policy makers in order to evaluate the consequences.

Furthermore, our analysis reveals that QoS tiering is likely to result into higher investments in network infrastructure in the long run. The reason is that higher network capacity encourages more entry by CPs, whose additional demand keeps the value of the priority service high. This result contrasts the findings of Cheng et al. (2011) and Choi and Kim (2010) who do not consider entry of new (congestion sensitive) CPs and therefore find that the ISP has an incentive to keep the value of the priority service high by maintaining network capacity scarce.

Finally, we also investigate the threat of strategic quality degradation and the effectiveness of minimum quality standards (MQS) in this context. Proponents of net neutrality are concerned that the ISP may have an incentive to degrade the transmission quality of the best-effort class even below its technical ability in order to drag CPs into a pay for priority agreement. We find that strategic quality degradation is only a profitable strategy for the ISP if consumers' marginal valuation for content variety is sufficiently small. In this case, an MQS policy can safeguard the positive welfare effects of QoS tiering. However, if strategic quality degradation is not an issue (e.g.,

[6] The debate was particularly stimulated after a blunt statement by Ed Whitacre, the Chief Executive Officer of ATT, who said: "Now what [content providers] would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it" (O'Connell 2005). Similar statements have been released by major European network operators (Lambert 2010, Schneibel and Farivar 2010).

in the presence of effective transparency obligations), we find that an MQS policy that requires the ISP to guarantee a congestion level in the best-effort class under QoS tiering which is at least as good as the best-effort congestion level under network neutrality is not sufficient to guarantee efficient infrastructure investments.

The remainder of this article is structured as follows. In Section 2 we discuss our framework in the context of related work before we formally introduce the model in Section 3. Next, we investigate the differences between the QoS tiering and network neutrality regimes in the short run with respect to content variety (Section 4), and in the long run with respect to broadband investments (Section 5). In Section 6 we consider the scope for regulatory intervention and particularly discuss whether an MQS is an appropriate policy instrument in this context. Finally, in Section 7 we comment on the possibility of strategic quality degradation under QoS tiering before we conclude in Section 8 by summarizing our results.

2. Related Work

Compared to the total number of academic papers that have been published in the context of the net neutrality debate, the number of formal economic papers within this domain is rather small. Schuett (2010) provides a comprehensive overview of this literature. The first formal approach to investigating net neutrality regulations is due to Economides and Tåg (2008), who consider a simple two-sided market model. On one side of the market, there is a continuum of non-competing CPs and, on the other side of the market, there is a continuum of consumers. Each side experiences positive network externalities through the presence of the other side. This is similar to our set-up, however, the authors do not consider a QoS tiering regime and instead see a violation of network neutrality in the ISP's practice of charging CPs a termination fee for access to its customers.

Cheng et al. (2011) and Choi and Kim (2010) investigate the head-to-head competition of CPs and the ISP's incentive to invest in network infrastructure under QoS tiering. Like us, they employ standard results from queuing theory to formalize the relationship between priority and best-effort traffic. However, their model set-up differs substantially from ours. The authors investigate the effect of QoS tiering on the competition of CPs that offer similar services. In their models, exactly two competing CPs are located at the end of a standard Hotelling line and it is assumed that customers dislike congestion and visit one of the two CPs exclusively (e.g., consumers either use *Google* or *Bing*, but never both). In contrast, our model studies the impact of QoS tiering on the variety of the available content on the Internet. The CPs in our model are not in direct competition to each other, but offer heterogeneous services that differ in their sensitivity towards congestion. In other words, while Cheng et al. (2011) and Choi and Kim (2010) intend to study the impact of QoS tiering on a particular content submarket, we seek to study the effect of QoS tiering on the content market as a whole.

One of the important features of our model is that content variety (i.e., how many CPs choose to join the network in equilibrium) is determined endogenously. This allows us to study the effect of QoS tiering on content variety, which is not possible in Cheng et al. (2011) and Choi and Kim (2010). In this respect, our model is similar to that of Jamison and Hauge (2008) and Hermalin and Katz (2007). However, in contrast to our model, Jamison and Hauge (2008) focus on the question whether transmission quality can substitute for content quality. Moreover, they assume that the current network capacity will inevitably increase with the introduction of QoS, such that the transmission quality of the best-effort class is not affected (non-degradation condition). We study the ISP incentives to invest absent this condition, but also consider a similar case where a minimum quality standard is enforced. In the model of Hermalin and Katz (2007) CPs differ in their value to consumers, but do not differ in their sensitivity towards congestion. Also, their model neither explicitly studies investment incentives, nor considers the inter-class externality that the high priority class exerts on the remaining best-effort class under fixed network capacity. In the subsequent paper of Economides and Hermalin (2010), which rests on the same principal modeling assumptions as Hermalin and Katz (2007), inter-class externalities are explicitly considered, however, much of the analysis is now based on the implicit assumption that content variety is exogenous and the same under net neutrality and QoS tiering. The clear focus of Economides and Hermalin (2010) is to show the negative impact of the so-called re-congestion effect on overall congestion under QoS tiering. The re-congestion effect describes that those CPs that are prioritized under a QoS tiering regime will receive even more consumer requests and thus generate more traffic than under net neutrality, which in turn re-congests the network. Our paper instead focuses on the *re-allocation effect*, by which QoS tiering enables to allocate congestion away from the congestion sensitive and to the congestion insensitive CPs.

Although all of the models discussed here consider important facets of the net neutrality debate, none has addressed the issues of congestion sensitivity, inter-class externality, endogenous entry by CPs (content variety) and investment incentives by the ISP together. Accordingly, previous results with respect to content variety, network investment and welfare are mixed: Hermalin and Katz (2007), who neglect inter-class externality, find that network neutrality leads to less content variety in the short-run and has a tendency to be welfare reducing. Jamison and Hauge (2008), who assume that the ISP invests more under QoS tiering, find that QoS tiering increases content variety. Cheng et al. (2011) and Choi and Kim (2010), who neglect endogenous entry and exit of CPs, show for a large range of parameters that the ISP's incentive to invest in infrastructure is higher under network neutrality, whereas QoS tiering is generally welfare-enhancing in the short run.

Our paper complements these previous approaches and finds that due to inter-class externalities and better allocation of congestion, QoS tiering may be the more efficient regime in the short run. Also in the long run, it provides higher incentives for broadband investments, because the entry by new, congestion sensitive CPs creates additional demand for the priority service that is absent under network neutrality. However, if the mass of congestion sensitive CPs is very large, then the priority service might get so overcrowded that the overall situation under QoS tiering is worse than under net neutrality. In effect this is similar to a re-congestion of the priority lane and thus the welfare conclusions of Economides and Hermalin (2010) are similar to ours: If the re-congestion effect is not too strong, QoS tiering provides higher investment incentives, leads to more content variety and is thus likely to be the more efficient regime in the long run.

3. The Model

We model the Internet as a two-sided market, with CPs and Internet customers on either side, each of which value an increasing presence of the other side and dislike network congestion. We assume that the ISP has a terminating monopoly over its customers (e.g., due to the customers' lack of alternative ISPs or high switching costs), which is reasonable for many regions in the US and Europe. Therefore, the only way for the CPs to reach these customers is through the ISP's network. Although the CPs' customer base is probably comprised from customers of many different ISPs, each of which might have a terminating monopoly, it is still insightful to investigate the relationship between CPs and a single ISP, particularly if that ISP is thought to be large. For example, it would certainly have a substantial impact on CPs' business model if they would not have access to customers' on AT&T's network. Note that we only consider charges to the CPs that are over and beyond those for access to the Internet. Thus, we consider net neutrality as a zero price rule which implies that the ISP cannot charge CPs additionally for terminating traffic in its network. Furthermore, we consider the politically relevant case where the best-effort lane under QoS tiering remains to be offered for zero additional costs.

Content Providers We consider a continuum of CPs. Whatever service the CPs offer, they provide it for free and receive revenues only indirectly through online advertisements.[7] In the model, a CP's advertisement revenue will depend on the average received traffic, the per-click advertisement revenue, and its individual click-through-rate, which is determined by the CP's innate sensitivity towards network congestion. Before these measures are formally introduced below, we make one fundamental assumption:

[7] In particular, this means that we rule out the possibility that CPs charge consumers directly for access to their content. However, this seems to be the more relevant case as empirical evidence suggest that customers are generally fairly reluctant to pay extra for specific content or services (Dou 2004, Sydell 2007).

ASSUMPTION 1. *Each CP receives the same average traffic from each customer, denoted by λ. This is independent of a content provider's business model and consequently its innate sensitivity to network congestion.*

For the remainder of this article, it will often be convenient to think of λ as the number of 'clicks' that a customer generates on each CP's website. This assumption provides a neutral reference case with respect to the traffic that is generated by the specific CPs and with respect to the value of the individual content of the CPs. The relationship between congestion sensitivity of a CP and the amount of traffic that this CP generates is far from obvious. For example, VoIP services are highly congestion sensitive (in terms of jitter, delay, packet loss), but generate comparably very little traffic. Likewise, file hosting services are highly traffic intensive but not congestion sensitive. On the one hand, Assumption 1 avoids to establish such a relationship between the traffic that a CP generates and its congestion sensitivity, which would otherwise inevitably bias the analysis. This allows us to assess QoS tiering based on its core ability, i.e., increasing transmission quality (not bandwidth). On the other hand, Assumption 1 also avoids to make any judgment about the value of specific content or services to consumers. In our model, CPs offer heterogeneous services which are all equally cherished by customers. Therefore it is reasonable to assume that customers distribute their clicks evenly among the available CPs.

In this context, it is important to highlight that we do not intend to study the effect of QoS tiering on the direct competition between otherwise similar CPs. This is done by Cheng et al. (2011) and Choi and Kim (2010). Instead, we seek to complement their analysis and study the effect of QoS tiering on content variety. Therefore, Assumption 1 implies that we abstract from any business stealing effects. More specifically, in what follows, we assume that λ is constant and thus, as the number of active CPs increases, consumers also increase their total number of clicks accordingly. Alternatively we could have assumed that the consumers' total number of clicks is fixed and thus λ diminishes as the variety of content increases. This would introduce a general notion of competitive pressure among the CPs, i.e., as more CPs enter the market, the revenue of each CP is reduced, everything else being constant. However, we believe that a constant λ is more intuitive in the context of our analysis, and note that the results of either assumption are qualitatively the same.[8] Instead, we have modeled such competitive pressure through diminishing ad revenues as described below.

Eventually, on the CP's end, only a fraction of these clicks can be turned into advertisement revenue. This measure is known as the click-through rate. We assume that each CP's click-through rate diminishes as network congestion increases. Moreover, each CP's business model has an innate

[8] A formal proof is provided in Section A.3 of the appendix.

sensitivity as to what extent network congestion affects the click-through rate. For example, a web-based e-mail provider is likely to be relatively insensitive to network congestion. Consumers that arrive on the website are satisfied with the service even under high network congestion, and more likely to click on advertisements. In contrast, consumers of a highly congestion sensitive web service (e.g., TV streaming) may still arrive at the CP's website, but are in the presence of network congestion less satisfied with the service and therefore less likely to view advertisements. This individual congestion sensitivity is denoted by θ and the corresponding click-through rate of a CP is assumed to be $(1-\theta w)$, where w denotes the CP's perceived average level of network congestion.[9] There exists a continuum of CPs with unit mass and distribution function $F(\theta):[0,1]\rightarrow[0,1]$. Let r be the average revenue-per-click on advertisements depending on the mass of active CPs in the market, then each CP's profit under net neutrality is[10]

$$\Gamma_N(\theta)=\begin{cases}(1-\theta\ w_N)\,\lambda\bar{\eta}r & \text{if active}\\ 0 & \text{otherwise,}\end{cases} \tag{1}$$

where $\bar{\eta}$ denotes the share of Internet customers in equilibrium. Under network neutrality all CPs perceive the same level of congestion, w_N. In the QoS tiering regime, however, CPs can opt for the priority transmission class with $w_{Q1}<w_N$ at a price of p per click. The CPs that remain in the best-effort class, on the other hand, experience a higher congestion level $w_{Q2}>w_N$.

$$\Gamma_Q(\theta)=\begin{cases}(1-\theta\ w_{Q2})\,\lambda\bar{\eta}r & \text{if active in best-effort class}\\ (1-\theta\ w_{Q1})\,\lambda\bar{\eta}r-\lambda\bar{\eta}p & \text{if active in priority class}\\ 0 & \text{otherwise.}\end{cases} \tag{2}$$

The CP that is indifferent between choosing the priority and the best-effort transmission class under a QoS tiering regime, is denoted by $\tilde{\theta}$. Furthermore, in both regimes, the CP that is indifferent between becoming active and staying out of the market is characterized by a congestion sensitivity of $\bar{\theta}$. Thus $F(\bar{\theta})$ reflects the mass of all active CPs (content variety) and the share of CPs choosing the priority class under a QoS tiering regime is given by $\beta\equiv 1-F(\tilde{\theta})/F(\bar{\theta})$.

We assume a competitive advertisement market, which introduces an indirect element of competition between the CPs. More specifically, it is assumed that the level of CPs' gross advertisement revenues depends on the mass of active CPs, i.e. $r(F(\bar{\theta}))$, and that $\partial r(\cdot)/\partial F(\bar{\theta})\ \leq 0$.

[9] Note that the click-through rate follows a Poisson thinning process. The thinning probability depends on the average waiting time (w) as a proxy for congestion in a transmission class and the sensitivity of the service (θ) itself. Therefore a CP with a high innate sensitivity has a lower probability of making money than a CP with a low innate sensitivity at any given congestion level.

[10] Throughout this paper, we distinguish between the network neutrality regime and the QoS tiering regime by subscript N and Q, respectively. However, in order to reduce the notational burden, we will omit the subscripts wherever the referenced network regime is unambiguous.

Customers Internet customers value basic connectedness to the Internet as well as the presence of many CPs. In particular, we assume that connectedness adds a base utility of $b > 0$ whereas each additional CP adds a marginal utility of $v > 0$ to a customer's utility. In reverse, congestion diminishes a customer's utility of using the CPs' services. To keep the analysis as clear as possible, we assume that consumers' utility is determined only by the average congestion level $w_Q = \beta w_{Q1} + (1-\beta) w_{Q2}$, or w_N, respectively. This implies that $w_Q = w_N$ in the short run (when $\mu_Q = \mu_N$) whenever $\bar{\theta}_N = \bar{\theta}_Q$. Alternatively, we could have assumed that customers are congestion sensitive as well and instead evaluate the level of congestion as $\hat{w}_N = \int_{\theta=0}^{\bar{\theta}} w_N \theta f(\theta) d\theta$ and $\hat{w}_Q = \int_{\theta=0}^{\tilde{\theta}} w_{Q2} \theta f(\theta) d\theta + \int_{\theta=\tilde{\theta}}^{\bar{\theta}} w_{Q1} \theta f(\theta) d\theta$, respectively. Although reasonable, this assumption would qualitatively not change our analysis, but merely emphasize the advantageousness of the QoS tiering regime whenever this is the case.[11] The reason is, as will be seen later, that the QoS tiering regime allocates congestion more efficiently such that $\hat{w}_Q \leq \hat{w}_N$. Our assumption is therefore more conservative and tipped in favor of the network neutrality regime. Formally,

$$U = \begin{cases} b + v\bar{\theta} - \iota w - a & \text{if connected} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\iota > 0$ denotes a consumer's marginal disutility due to congestion, and a the Internet access fee charged by the ISP.

As outlined before, in our analysis we intend to focus on the effect of QoS tiering on the relationship between CPs and the ISP. Thus, for expositional clarity we assume that customers are homogeneous and therefore the ISP is able to set an access fee such that all consumers connect to the ISP in equilibrium. This does not violate the two-sided market property (cp. Rochet and Tirole 2006) and is not a crucial limitation of the model per se . First, departing from this assumption would foremost allow for a more fine grained analysis of the rent distribution between customers and the ISP. With homogeneous customers, the ISP is able to fully extract the consumer surplus and thus no dead weight loss occurs.[12] By contrast, with heterogeneous customers, consumer rent will be positive, but also possibly generates a dead weight loss. Recall that the consumers' access fee is the only source of revenue for the ISP under net neutrality. Thus, this fee is likely to be higher under net neutrality than under QoS tiering where the ISP can collect additional rents from CPs through the priority fee. In fact, under QoS tiering the ISP may find it profitable to subsidize the consumer side by lowering the access fee, possibly down to zero, in order to stimulate customer subscriptions which in turn allows the ISP to make higher profits on the CPs' side. Consequently, the customer access fee and associated dead weight loss is lower under QoS tiering. Indeed, the

[11] This extension and formal proof is provided in Section A.4 of the appendix.

[12] We thank an anonymous referee for this comment and his helpful insights for the following discussion.

dead weight loss may even be zero if the access fee is zero. Under net neutrality, the consumers' access fee (and thus the dead weight loss) can never be zero, because otherwise the ISP would not make any profit. Consequently, it is likely that more consumers will subscribe under QoS tiering, and that consumers' welfare is higher than under net neutrality.[13]

Second, despite of customers' homogeneity, demand-side effects can be modeled through further assumptions on the distribution function F. For example, if it is assumed that priority providers will receive relatively more demand (clicks) by consumers than non-prioritized CPs, which in turn leads to a re-congestion of the priority lane, then this effect is for the purpose of our analysis qualitatively similar to a situation where a decrease in congestion to the priority class evokes such re-congestion by the overproportional entry of new, congestion sensitive CPs. The latter can be achieved by assuming that F' is sufficiently increasing in θ.[14]

Network Congestion Network congestion is measured through Internet consumers' average waiting time following a content request. We employ the well-known $M/M/1$ queuing model (Kleinrock 1976) to fix ideas on the relationship between average waiting time, network traffic and capacity.[15] Under a network neutral regime the $M/M/1$ model predicts that each consumer has an expected average waiting time of

$$w_N = \frac{1}{\mu - \Lambda}. \tag{4}$$

Here μ represents the average rate at which service requests are handled, which is interpreted as the overall *transmission capacity*; whereas $\Lambda = \lambda\ \bar{\eta}\ F(\bar{\theta})$ denotes the average rate at which customers' aggregate content requests arrive at the ISP's network, which is interpreted as *network traffic*. For the queuing system to be stable, we must assume that $\mu > \Lambda$.

Under a QoS regime, CPs are offered the choice between a priority and a best-effort transmission class. In the $M/M/1$ model this translates to introducing an additional queue which handles the request of the CPs in the priority class and which is processed ahead of the queue for the best-effort

[13] Further proof and a more detailed analysis of the model for heterogeneous consumers is provided in Section A.1 of the appendix.

[14] Of course, in general the re-congestion effect is not identical to the effect that arises from the overproportional entry of new, congestion sensitive CPs. Although total traffic of the priority lane increases in both cases, the second case (new CP entries) has one extra content variety effect on consumers' utility. We thank an anonymous referee for pointing this out. However, with respect to the comparison of the impact of network neutrality and QoS tiering on the relationship between CPs and ISP, which is in the focus of this paper, both cases qualitatively yield the same results. Further proof for this claim is provided in Section A.2 of the appendix.

[15] The $M/M/1$ queuing model assumes that (1) service requests arrive according to a Poisson process (i.e., arrivals happen continuously and independently of one another), (2) service time is exponentially distributed (i.e., request coming from a Poisson process are handled at a constant average rate) and (3) that service requests are processed by a single server. This last assumption is equivalent to assuming that network performance is dominated by a bottleneck component. Furthermore it is assumed that the length of the queue as well as the number of users is potentially infinite. This model is standard and considered to be a good proxy for actual Internet congestion (McDysan 1999).

class. However, in each class the queue is cleared on a first-come first-served basis. In this vein, the classical results of the $M/M/1$ queuing model represent the average waiting time in the priority class, w_{Q1}, and the best-effort class, w_{Q2}:

$$w_{Q1} = \frac{1}{\mu - \beta\Lambda}, \qquad w_{Q2} = \frac{\mu}{\mu - \Lambda} w_{Q1} \tag{5}$$

It is easy to see that relation $w_{Q1} < w_N < w_{Q2}$ is always fulfilled, assuming a fixed transmission capacity $\mu = \mu_Q = \mu_N$ and $\beta < 1$.[16] This is an important feature of our model, because it shows formally that serving some CPs with priority will (in the short run) unambiguously lead to a *degradation* of service quality for the remaining CPs in the best-effort class.[17]

Internet Service Provider The ISP controls the (two-sided) Internet market, over which it has a terminating monopoly, through a number of strategic variables. First, it charges an access fee, a, from connected consumers. Under a network neutral regime, the consumer access fee is the only source of revenue for the ISP. Second, in the long run the ISP also sets the level of network capacity, μ. As outlined before, customers and CPs dislike network congestion. The level of network congestion is captured by customers' average waiting time for content, w, which is again controlled by the ISP through its choice of network capacity.

Hence, under a network neutrality regime, the ISP's profit is

$$\Pi_N = \bar{\eta} a - c(\mu), \tag{6}$$

where $c(\mu)$ denotes the costs of capacity expansion.[18] Under a QoS tiering regime, the ISP has an additional strategic variable, p, the price which it charges CPs to transmit data packets with priority. The ISP will choose p in order to maximize its additional revenues from selling priority access. More precisely, under QoS tiering the ISP's profit function is

$$\Pi_Q = \bar{\eta} a + \beta\Lambda p - c(\mu). \tag{7}$$

We consider the ISP's previous investment decisions in transmission capacity as sunk in all regimes. Therefore, in the short run μ can be considered an exogenous variable which is irrelevant for profit maximization.

4. Short Run Effects on Content Variety and Welfare

First, we compare the two network regimes in the short run, i.e., when network capacity, μ, is exogenous and equal in both regimes.

[16] For $\beta = 1$, when all CPs are in the first priority class, the model trivially collapses to $w_{Q1} = w_N$

[17] Degradation of the best-effort class is an unavoidable consequence of traffic prioritization here. In Section 7 we consider the effect of an additional (strategic) degradation to the best-effort class under QoS tiering.

[18] To ensure the existence of an interior solution to the ISP's investment decision, we assume a non-concave cost function, i.e. $\partial c/\partial \mu \geq 0$ and $\partial^2 c/\partial \mu^2 \geq 0$.

4.1. Short run Equilibrium and Content Variety

Network Neutrality Regime First, it is obvious that the ISP will set an optimal customer access charge of $a = b + v\bar{\theta} - \iota w_N$, such that all customers will connect to the network ($\bar{\eta} = 1$) and total consumer surplus is appropriated by the ISP. Under network neutrality all CPs expect the same congestion level of w_N and enter the network only if they have non-negative utility at this level. Consequently, the last CP to enter the network is located at:[19]

$$\bar{\theta}_N = \frac{1}{w_N} = \mu - \lambda F(\bar{\theta}) \tag{8}$$

Hence, an increase in network traffic per CP, λ, has an adverse effect on network congestion ($\partial w_N/\partial\lambda > 0$) and content variety ($\partial\bar{\theta}_N/\partial\lambda < 0$). This is central to the debate on network neutrality, because it exemplifies the network operators' concerns with respect to the expected increase in traffic.

Quality of Service Tiering Regime In the QoS tiering regime, the ISP can alleviate congestion for the most congestion sensitive CPs through the provision of differentiated transmission classes. We may now distinguish three types of CPs: (1) CPs whose business model is relatively insensitive to network congestion. They will remain in the free-of-charge best-effort class. (2) CPs whose business model is sufficiently sensitive to network congestion. They will opt for priority access at a price of p. (3) CPs whose business model is extremely sensitive to network congestion. They will remain inactive as entry is not profitable. Remember that the CP indifferent between the first two cases is denoted by $\tilde{\theta}$, whereas the CP indifferent between the last two cases is denoted by $\bar{\theta}_Q$. Obviously, it must hold that $0 \leq \tilde{\theta} \leq \bar{\theta}_Q$.[20] In a fulfilled expectations equilibrium, the last CP to enter is located at

$$\bar{\theta}_Q = \frac{1 - p/r}{w_{Q1}} = \frac{r-p}{r}\left(\mu - \lambda\left(F(\bar{\theta}_Q) - F(\tilde{\theta})\right)\right). \tag{9}$$

From

$$\frac{\partial\bar{\theta}_Q}{\partial p} = \underbrace{-\frac{1}{r}\left(\mu - \lambda\left(F(\bar{\theta}_Q) - F(\tilde{\theta})\right)\right)}_{\text{First Order Effect}} + \underbrace{\lambda\frac{r-p}{r}\left(\frac{\partial F(\tilde{\theta})}{\partial p}\frac{\partial\tilde{\theta}}{\partial p} - \frac{\partial F(\bar{\theta}_Q)}{\partial p}\frac{\partial\bar{\theta}}{\partial p}\right)}_{\text{Second Order Effect}} \tag{10}$$

it is easy to see that an increase in the price for priority transmission, p, has an unambiguously negative first order effect on content variety. This is the central concern of net neutrality proponents, who argue that starting from a zero price under net neutrality, the introduction of a positive

[19] We restrict our analysis to the interesting case where (at least) the most congestion sensitive CP, located at $\theta = 1$, remains inactive in equilibrium. This is ensured iff the average congestion level satisfies $w_N < 1$.

[20] In Section B.1 of the appendix we provide the general condition under which this relation holds, as well as proof that it is always satisfied under our assumptions.

price under QoS tiering has negative first order effects on content variety. However, this argument neglects that there is a second order effect as well: An increase in p will induce more CPs to choose the free best-effort class and therefore alleviate congestion in the priority class. This in turn may encourage new, congestion sensitive CPs to enter, which drives congestion in the priority class up again. The size and direction of the second order effect hinges on the mass of CPs that is located locally at $\tilde{\theta}$ and $\bar{\theta}$ (i.e., $\frac{\partial F(\tilde{\theta})}{\partial p}\frac{\partial \tilde{\theta}}{\partial p} - \frac{\partial F(\bar{\theta}_Q)}{\partial p}\frac{\partial \bar{\theta}}{\partial p}$) and cannot be determined more specifically for a general distribution function. For the purpose of our analysis, let us therefore assume a particular density function of θ that exemplifies the effect of having a non-uniform distribution of θ. To this end, we consider the density function $f : [0,1] \to [0,1], f(\theta) := \alpha + 2\theta(1-\alpha)$, with $\alpha \in [0,2]$. Let F be the distribution function to f and notice that for $\alpha = 1$ we obtain a uniform distribution with $F(\theta) = \theta$. Otherwise, if $\alpha > 1$, there exists a relatively larger mass of congestion *insensitive* CPs ($F(\theta) > \theta$) and if $\alpha < 1$ there is a relatively larger mass of congestion *sensitive* CPs ($F(\theta) < \theta$). A variation of α is therefore equivalent to a gradual shift of mass from the congestion sensitive portion ($\theta > 0.5$) to the congestion insensitive portion ($\theta < 0.5$) of the CPs and vice versa.

In the appendix we show that under a uniform distribution ($\alpha = 1$) the first order and second order effect are exactly offset, at any price level, such that the price for priority transmission has no effect on content variety under QoS tiering. Thus, under a uniform distribution, net neutrality and QoS tiering will exactly yield the same level of content variety, i.e., $\bar{\theta}_Q = \bar{\theta}_N = \mu/(\lambda+1) = 1/w_N$. However, if the mass of congestion sensitive CPs is relatively large ($\alpha < 1$), an increase in price for priority will not lead to an equally large congestion alleviation for the priority class such that the first order effect prevails. Consequently, under QoS tiering less CPs will enter in equilibrium than under net neutrality. Conversely, if the mass of congestion sensitive CPs is comparably small ($\alpha > 1$), then the second order effect dominates and QoS tiering leads to more content variety than net neutrality. The following proposition, whose proof can be found in the appendix, summarizes these results.

PROPOSITION 1 **(Content Variety)**. *If content providers congestion sensitivity is uniformly distributed, QoS tiering has no effect on content variety in the short run: The number of active content providers is the same as under network neutrality. In both regimes the number of active content providers is inversely proportional to the average level of congestion in the network. However, if the mass of congestion sensitive content providers is comparably small (large), then QoS tiering is likely to lead to more (less) content variety.*

Therefore, it is useful to assume a uniform distribution of θ as the reference case for the subsequent analysis, from which it is then easy to draw more general conclusions.

ASSUMPTION 2. *Content providers' congestion sensitivity, θ, is uniformly distributed such that $F(\theta)=\theta$.*

Under the present assumptions QoS tiering will lead to neither more nor less content variety. However, under a QoS tiering regime the ISP can additionally extract rents from CPs through sales of priority access. In the short run, it will do so by maximizing revenues from priority sales ($\Lambda\beta p$) which is achieved by

$$p=\left(1-\sqrt{\frac{\bar{\theta}_Q}{\mu}}\right)r=\left(1-\frac{w_{Q1}}{w_Q}\right)r. \tag{11}$$

Intuitively, this shows that the ISP can extract a fraction of the CPs' gross advertisement revenue r, depending on the congestion alleviation to the priority class compared to the average congestion level in the network.

PROPOSITION 2 **(ISP Preferred Regime)**. *The ISP always prefers the QoS tiering regime because it can make extra profits by selling a priority transmission service to content providers.*

The proof is in the appendix.

4.2. Short Run Welfare Implications

Now we investigate the short run effect of QoS tiering on welfare. Total welfare, W, is the sum of consumers' surplus, CPs' surplus, and the ISP's profit. Thus, the difference in social surplus between QoS tiering and network neutrality is given by

$$\Delta W=(U_Q-U_N)+(\Gamma_Q-\Gamma_N)+(\Pi_Q-\Pi_N) \tag{12}$$

Recall that $U_Q=U_N=0$, because consumers' surplus is always fully appropriated by the ISP. However, notice that changes in consumers' gross surplus are reflected in changes of the ISP's profit. Furthermore, $\Pi_Q-\Pi_N>0$ according to Proposition 2. What remains to be examined is the short run effect of QoS tiering on CPs' surplus.

[ENTER FIGURE 1 ABOUT HERE]

To this extent, consider Figure 1 and notice that those CPs located at $\theta\in[0,\tilde{\theta})$ are evidently worse off under a QoS tiering regime, because for them network congestion has increased from w_N to w_{Q2}. Second, the CPs' welfare loss increases with congestion sensitivity on the interval $\theta\in[0,\tilde{\theta})$. The business model of the provider located at $\theta=0$ is not affected at all through congestion, while the provider at $\theta=\tilde{\theta}$ is already suffering so much that it is indifferent between staying in the best-effort class and buying priority access. Third, by the converse argument, notice that the welfare loss decreases for the CPs in the priority class as $\theta\in[\tilde{\theta},\bar{\theta})$ increases. To see this, recall from

Proposition 1 that the last CP to enter the market, $\bar{\theta}$, is identical under both regimes and receives a surplus of zero. For this CP, the benefit through reduced congestion (compared to the network neutrality regime) is just offset by the price that it pays for priority access. Consequently, for all CPs with less congestion sensitivity ($\theta \in [\tilde{\theta}, \bar{\theta})$) the price that is paid for priority is higher than the benefit of being in the priority class. Nevertheless, by definition of $\tilde{\theta}$, for these providers the welfare loss is still less severe in the first priority class than in the best-effort class. In this line of argumentation, it is also obvious that CP $\tilde{\theta}$ incurs the greatest welfare loss. In summary, we can conclude that in the short run all active CPs are (weakly) worse off under a QoS tiering regime.

However, the price that CPs pay for priority access is merely a welfare shift to the ISP (hatched area in Figure 1). The sign of the overall welfare effect will therefore only depend on the difference between the gross surplus gain through less congestion of those CPs in the priority class and the gross surplus loss through increased congestion of those providers remaining in the best-effort class. In the appendix we show that this difference is always positive.

PROPOSITION 3 **(Short run Welfare)**. *If content providers congestion sensitivity is uniformly distributed, QoS tiering unambiguously increases welfare with respect to the network neutrality regime in the short run, because congestion is alleviated for the most congestion sensitive content providers in lieu of the less congestion sensitive content providers. However, all content providers are worse off under a QoS tiering regime because the increased surplus is expropriated by the ISP.*

Furthermore, it is easy to see that this welfare conclusion is not as clear-cut under a non-uniform distribution. If QoS tiering leads to more content variety, then those CPs who are newly active in the market will enjoy a higher surplus than under network neutrality and Proposition 3 is even strengthened. However, if there is a relatively large mass of congestion sensitive CPs in the economy such that QoS tiering leads to less content variety, the associated welfare loss must be counterweighted with the welfare gain from better congestion allocation. In this case it is likely that Proposition 3 does not hold anymore.

5. Long Run Effects on Broadband Investments, Innovation and Welfare

Much of the neutrality debate is rooted in the ISPs' concerns about infrastructure investments. On the one hand, ISPs would like to accommodate new (congestion sensitive) content because this is valued by customers. However, on the other hand ISPs disapprove of CPs who free-ride on their infrastructure investments. QoS tiering seems to be a plausible way out of this dilemma, but it is unclear whether in the long run this regime will lead to more or less incentives for infrastructure investments than will network neutrality regulation. Thus, in this section we extend our analysis to long run investments in network transmission capacity. In our model, transmission capacity is

represented by the average service rate, μ, at which customer requests can be handled. An increase of μ allows the ISP to handle more service requests to CPs at a time.

5.1. Investment Incentives

Formally, the ISP's investment decision is a discrete decision stage which precedes the previous analysis. The ISP chooses the network capacity level, μ, first, and subsequently sets the customer access charge, a, and the priority price, p, if applicable. In the subgame perfect equilibrium, the ISP will set the optimal capacity level at the point where the marginal revenues of capacity expansion, $MR \equiv \partial\Pi/\partial\mu$, equal marginal costs, $MC \equiv \partial c(\mu)/\partial\mu$. Consequently, the ISP's optimal capacity level will be higher if marginal revenues from capacity expansion are higher.[21] In both network regimes the following two marginal effects of capacity expansion on ISP revenue can be distinguished:

- The *variety incentive:* $\left(v \cdot \partial F(\bar{\theta})/\partial\mu\right)$ denotes the ISP's marginal revenue effect on the customer access fee that comes from the entry of new, congestion sensitive CPs.
- The *congestion incentive* $(-\iota \cdot \partial w/\partial\mu)$ denotes the ISP's marginal revenue effect on the customer access fee that comes from a change of the overall congestion level.

Furthermore, notice that under the assumption of a uniform distribution of CPs' congestion sensitivity, these investment incentives are always positive and identical under both network regimes. Hence, potential differences in investments between the two regimes may only be a result of an additional investment incentive that an ISP has only under QoS tiering:

- The *priority revenue incentive* $(\partial(\beta\Lambda p)/\partial\mu)$ denotes the ISP's marginal revenue effect from selling priority access.

Consequently, the sign of the priority revenue incentive is definitive for the comparison between investment incentives under QoS tiering and network neutrality. The result is summarized by the following proposition, whose proof can be found in the appendix.

Proposition 4 **(Investment Incentives)**. *If the congestion sensitivity of content providers is uniformly distributed, the ISP's optimal capacity level is higher under QoS tiering.*

This finding contrast the results of Cheng et al. (2011) and Choi and Kim (2010). The reason is that we explicitly account for the fact that more network capacity encourages the entry of new CPs, whose additional demand keeps the value of the priority service high. By contrast, in Cheng et al. (2011) and Choi and Kim (2010) entry of new CPs is not possible and therefore it is more profitable to exploit the current CP base and to keep network capacity scarce.

[21] Thereby we assume that the ISP's marginal revenues with respect to μ are decreasing, while marginal costs are increasing. The conditions for the former assumption are shown in the appendix, whereas the latter is warranted by the assumption of a convex cost function.

Note that for non-uniform distribution functions, the mass of active CPs may differ between the two network regimes and thus the variety and congestion incentive will generally not coincide. In particular, if the mass of congestion sensitive CPs is very small, then the priority incentive can even be negative. This is because the ISP has only few congestion sensitive CPs to which it can sell priority and hence it seeks to make the priority class attractive to less congestion sensitive CPs by keeping network capacity scarce and the congestion level high. To see this, consider Figure 2 which presents a numerical example of the marginal investment incentives for varying distributions of CPs' congestion sensitivity. In line with Proposition 4, the variety and congestion incentive under either regime coincide under the uniform distribution of CPs' congestion sensitivity ($\alpha = 1$), such that the positive priority revenue incentive is decisive for the higher investment incentives under QoS tiering. However, the more congestion sensitive CPs are in the Internet economy ($\alpha < 1, \alpha \to 0$), the stronger is the variety incentive under net neutrality compared to QoS tiering. Notwithstanding, also the priority revenue incentive, which is only present under QoS tiering, increases. As the variety incentive grows linearly in v and the priority revenue incentive grows linearly in r, net neutrality may only lead to more infrastructure investments for $\alpha < 1$ if v is sufficiently larger than r. On the other hand, when there are relatively less congestion sensitive CPs in the economy ($\alpha > 1$), the variety and congestion incentive are slightly larger under QoS tiering while the priority incentive remains positive. In this case, QoS tiering provides unambiguously higher incentives for infrastructure investments. However, when the mass of congestion sensitive CPs becomes very small ($\alpha \gg 1$), the priority revenue incentive can indeed become negative, and eventually also the variety incentive under QoS tiering drops below the level under net neutrality. In this case, it is likely that net neutrality promotes investments in network infrastructure more. In summary, we can conjecture that Proposition 4 holds locally around $\alpha = 1$ (Assumption 2), i.e., if the proportion of congestion sensitive CPs to congestion insensitive CPs is balanced.

[ENTER FIGURE 2 ABOUT HERE]

5.2. Innovation at the Edge and Long Run Welfare

The ISP's investments in network infrastructure have direct ramifications for welfare. At higher capacity levels customers enjoy lower network congestion (congestion incentive) and higher network benefits (variety incentive). Figure 3 illustrates the effect of capacity expansion for CPs under QoS tiering. The reduction of network congestion increases CPs' click-through rate and thus the slope of their surplus curve in both transmission classes. The CPs in the best-effort class and also some CPs in the first priority class may still be worse off than under network neutrality. However, as a consequence of the overall decreased congestion level, both marginal CPs, $\tilde{\theta}$ and $\bar{\theta}$, are shifted to the right. This means that new, highly congestion sensitive CPs are able to enter the network. This

has been referred to as 'innovation at the edge' in the present context (Jamison and Hauge 2008). Obviously the surplus of the new CPs (crosswise hatched area), but also the surplus of some of the previously most congestion sensitive CPs (vertically hatched area), are thus increased compared to a network neutrality regime.

[ENTER FIGURE 3 ABOUT HERE]

Accordingly, higher capacity levels will ceteris paribus lead to higher gross utility for consumers and CPs and are thus beneficial for welfare. This is also shown formally in the appendix.

PROPOSITION 5 **(Long Run Welfare)**. *The regime that provides more incentives for infrastructure investments is more efficient in the long run. If the congestion sensitivity of content providers is uniformly distributed, QoS tiering is more efficient and provides more content variety than net neutrality.*

QoS tiering is consequently the more efficient regime in the long run and, by Proposition 3, also in the short run if CPs' congestion sensitivity is uniformly distributed. However, the fact remains that a non-negligible share of this surplus is immediately expropriated by the ISP.

6. Minimum Quality Standards

Price controls are not a suitable policy instrument in this context, because in the short run social and private incentives are in line: To see this, note that the social planner seeks to set the regulated priority price such that the socially optimal share of CPs selects the priority transmission class. In this vein CPs' gross surplus is maximized. The ISP, however, pursues the same goal, because it can subsequently extract a fraction of the CPs' surplus.

With regard to the ISP's investments in infrastructure, there generally is reason for regulation, however: Opponents of net neutrality regulation have often objected that net neutrality forces ISPs to invest above the efficient level, which is known as overprovisioning. On the contrary, opponents of QoS tiering argue that QoS tiering induces ISPs to keep transmission capacity scarce, and thus broadband investments are likely to be below the efficient level (underprovisioning). In the appendix we show that the difference between the efficient and private level of infrastructure investments is in fact independent of the network regime.

PROPOSITION 6 **(Efficient Investments)**. *The social planner has a higher incentive to invest in network capacity than the ISP. This result holds for both network regimes, QoS tiering and network neutrality.*

Minimum Quality Standards It has therefore been argued that a minimum quality standard (MQS) could be an appropriate policy instrument in this context (Brennan 2010); and a MQS policy is also already feasible under the new European legislative framework. After all, MQSs have found to be generally welfare-enhancing in competitive settings (Ronnen 1991). For example, it has been argued that the MQS could be set such that the ISP is required to offer CPs under QoS tiering a congestion level in the best-effort class that is as least as low as the equilibrium best-effort congestion level under network neutrality. Consequently, under the QoS tiering regime no CP would be set at a disadvantage anymore. Moreover, in order to meet this MQS, the ISP is required to increase the network's capacity, potentially to the extent that the gap between the level of private and efficient investments is closed. More precisely, by requiring the MQS $w_N(\mu_N^*) \equiv w_{Q2}(\mu_{MQS})$ the regulator implicitly defines the new capacity level $\mu_{MQS} > \mu_N^*$.[22]

By Propositions 4 and 6 the order of relevant capacity levels is $\mu_Q^{**} > \mu_Q^* > \mu_N^*$. Remember that $\mu_{MQS} > \mu_N^*$, and thus we can differentiate between three different cases. First, if $\mu_Q^* \geq \mu_{MQS}$ the MQS is not a binding condition for the ISP's capacity choice an hence is simply ineffective. Second, if $\mu_Q^{**} \geq \mu_{MQS} > \mu_Q^*$ the MQS is effective in raising the ISP's network capacity level, potentially up to the efficient level. Third, if $\mu_{MQS} > \mu_Q^{**}$ the MQS policy may lead to an excessive investment in network infrastructure. In summary, MQS are only effective in one out of three cases and thus for now their use is questionable.[23]

PROPOSITION 7 **(Minimum Quality Standard Regulation)**. *An MQS policy, which requires the ISP to guarantee a best-effort congestion level under QoS tiering which is equal to the equilibrium congestion level under network neutrality, may increase welfare, but may also lead to excessive investments or be ineffective.*

7. Strategic Quality Degradation

In the preceding analysis we have neglected the possibility that the monopolistic ISP may also engage in non-price discrimination, for example by degrading the quality of the best-effort class under QoS tiering. The concern for strategic quality degradation under a QoS tiering regime has been expressed by network neutrality proponents, but also previous empirical and theoretical research has identified several circumstances under which such practice is indeed profitable (Economides 1998, Foros et al. 2002, Crawford and Shum 2007). Absent the possibility to degrade the quality of the best-effort class, the ISP's only control over the share of CPs that buy priority

[22] One asterisk denotes the equilibrium capacity level, whereas two asterisks denote the socially optimal capacity level.

[23] In the next section we consider strategic quality degradation which provides a much stronger case for an MQS policy.

transmission in equilibrium (β) is through the price p. Quality degradation, however, provides the ISP with an additional means through which it can manipulate the relative attractiveness of the priority class over the best-effort class and thus the mass of CPs that buy priority transmission in equilibrium. It is inevitable that such practice will destroy some CPs' surplus and therefore questions the previously positive welfare results of QoS tiering. However, ex-ante it is not clear whether there exist scenarios under which quality degradation may actually be profitable to the ISP in the first place.

To this end, consider the extreme scenario where the ISP degrades the best-effort class under a QoS tiering regime maximally ($w_{Q2} \to \infty$), such that in equilibrium no CP wants to remain in the best-effort class ($\tilde{\theta} \to 0, \beta \to 1$). Furthermore, let $r(\bar{\theta}) = r$ be constant in this example.[24] We will show that there exist circumstances under which even this extreme form of quality degradation is profitable. More precisely, by rendering the best-effort class useless, the ISP effectively forces all CPs into the priority transmission class. It is easy to see that this is equivalent to a scenario in which the ISP demands a termination fee from each CP for transmitting content to its connected consumers.[25] At a fixed transmission capacity, this has detrimental effects on content variety and welfare.

To see this, recall that without quality degradation, the last CP to enter the network was located at $\bar{\theta}_{N,Q} = 1/w$, independent of the network regime and independent of the price for priority transmission. In contrast, the last CP to enter under quality degradation is located at $\bar{\theta}_D = (1 - p_D/r)\, w$.[26] Since all CPs are forced into the priority transmission class, congestion is the same for all CPs and at a similar level than under network neutrality. Thus, maximum quality degradation not only destroys the source of the positive welfare effects of the QoS tiering regime, but also forces the most congestion sensitive CPs out of the network: CPs experience a similar congestion level as under network neutrality, but have to pay a price $p > 0$ as if they were under QoS tiering. To be precise, it must be mentioned that the smaller mass of active CPs will also slightly reduce the average congestion level compared to network neutrality or QoS tiering. However, this type of congestion alleviation cannot outweigh the detrimental effect to content variety and welfare. Proofs are relegated to the appendix.

[24] This does not affect the generality of the existence of settings in which the ISP prefers to degrade the best-effort class under QoS tiering. In fact, as will be readily seen later, assuming r to be constant, is the most conservative assumption one can make in this context.

[25] Consequently the analysis in this section bridges the gap between the formal strand of the literature that considers net neutrality as a zero-price rule (i.e., no termination fees) and the literature that associates net neutrality with a non-discrimination rule (see Schuett 2010).

[26] Subscript 'D' denotes the QoS tiering regime with maximum quality degradation.

PROPOSITION 8 **(Content Variety and Welfare under Quality Degradation)**. *When the ISP degrades the quality of the best-effort class under QoS tiering such that all CPs choose to buy priority transmission, then, compared to network neutrality, less content providers enter the network in equilibrium and overall welfare is lower.*

Consequently, quality degradation is undesirable from a policy perspective and tarnishes the short run welfare results of QoS tiering. The question remains, however, whether quality degradation is in fact a profitable option to the ISP under QoS tiering and thus constitutes an actual source of concern to policy makers. The effect of quality degradation on the ISP's revenue depends on the trade-off of two opposing effects. By Proposition 8 quality degradation results in less content variety and consequently the ISP can charge consumers less for access. On the other hand, quality degradation forces *all* CPs to pay for their traffic and thus revenues from priority sales are potentially larger than before. Obviously, this trade-off is driven by the relative size of the marginal valuations of consumers and CPs, respectively. This can be exemplified by the equilibrium price formula:

$$p_D = \frac{r(1+\lambda) - \sqrt{r\left(v + r(1+\lambda)\right)}}{\lambda}. \tag{13}$$

Prices are positive as long as $v < r\lambda(1+\lambda)$, i.e., as long as the consumers' marginal utility for variety is not too large with respect to the CPs' marginal valuation for gross traffic (generated by consumers). On the contrary, if $v > r\lambda(1+\lambda)$, the ISP would theoretically like to subsidize the CPs and thus promote their entry in order to extract consumers' high utility from variety.

PROPOSITION 9 **(Profitability of Quality Degradation)**. *For all $v < \underline{v}$, where $\underline{v} < r\lambda(1+\lambda), \forall\lambda > 0$, the ISP makes larger profits under a QoS tiering regime in which the transmission quality of the best-effort class is degraded, such that all content providers choose to buy priority transmission in comparison to a QoS tiering regime without quality degradation.*

Proposition 9 establishes first that the ISP never subsidizes CPs by imposing negative prices under maximum quality degradation, but rather prefers to refrain from quality degradation and revert to the unhampered QoS tiering regime instead. This also implies that the ISP will not privately establish a network neutrality regime, which could be the result of QoS tiering with maximum quality degradation and a price of zero. Secondly, and more importantly, the proposition highlights that strategic quality degradation is in fact a profitable strategy for the ISP as long as consumers' marginal valuation for variety is sufficiently small.[27]

[27] If instead we would have assumed again that $\partial r(\bar{\theta})/\partial\bar{\theta} < 0$, then the profitability of quality degradation would even be increased. This is because the aggregate loss in CPs' gross advertisement revenues that is caused by the reduction in content variety according to Proposition 8, would be less pronounced.

Given the detrimental welfare consequences of quality degradation under a QoS tiering regime, policy makers should be aware of this strategic option. In particular, if policy makers suspect the ISP to engage in quality degradation, some of the previously reviewed policy instruments may now regain attention. Price regulation (i.e., $p_D = 0$) can at least ensure the current status quo of the network neutrality regime. However, such regulation also excludes the potentially positive welfare effects of an unhampered QoS tiering regime. In this context, minimum quality standards and transparency obligations seem to provide a more appropriate policy tool. If applied effectively, such obligations can preclude the ISP's negative strategic incentives under QoS tiering, while maintaining the generally positive welfare effects of this regime; after all, the ISP is still left better off than under network neutrality.

8. Conclusions and Policy Implications

Network neutrality has become a prime topic for many regulatory authorities, but the effect of such regulation is still unclear. Scholarly papers often find contradictory results with respect to the consequences of network neutrality on content variety, broadband investments and welfare. We contribute to the debate on network neutrality by providing a formal framework that incorporates the relevant arguments of net neutrality proponents and opponents in a two-sided market framework with Internet customers, CPs and an ISP. Our analysis focuses on the relationship between CPs and a monopolistic ISP, and compares network neutrality to a QoS tiering regime in which CPs may pay for the prioritized transmission of their data packets on a non-discriminatory basis. We explicitly consider the negative externality that prioritization has on the remaining best-effort class, but acknowledge that CPs' services differ in their sensitivity toward network congestion, and may offer their services only if they are sustainable under the given congestion level.

We find that the comparison between the two network regimes depends on the distribution of CPs' congestion sensitivity in the Internet economy. In particular, we have thoroughly investigated the neutral reference case where CPs' congestion sensitivity is uniformly distributed and find that QoS tiering increases welfare in the short run because the installed level of network capacity is used more efficiently: Network congestion is re-allocated, such that it is alleviated for the most congestion sensitive CPs. This offsets the congestion aggravation for the CPs in the remaining best-effort class. However, QoS tiering does not immediately promote the entry of new content provides with innovative services that are even more congestion sensitive. In fact, in the short run, all CPs likely to be worse off under a QoS tiering regime because the ISP is able to expropriate some of the CPs' surplus through priority pricing. Consequently, the ISP always prefers the QoS tiering regime. It is subject to the authority of policy makers to evaluate the shift of surplus from CPs to ISPs, which is welfare neutral per se, but lies at the heart of the net neutrality debate. On

the other hand, ISPs argue that they will use the additional revenues to invest more in broadband infrastructure. We show that this is true for the reference case of uniformly distributed congestion sensitivities, but may not hold for more skewed distribution functions. In sum, QoS tiering is likely to be the more efficient regime if the proportion of congestion sensitive to congestion insensitive CPs is balanced. In this case, the ISP invests more in broadband infrastructure, and thereby allows for entry of new, congestion sensitive CPs in the short run. Therefore, particularly the very congestion sensitive CPs will be better off under QoS tiering, and hence it is not surprising that Google and Verizon have privately agreed on a tiered system (Wyatt 2010).[28]

Furthermore, our analysis reveals that the level of private investments is generally not efficient. We show that an MQS policy that requires the ISP to guarantee a congestion level in the best-effort class under QoS tiering which is at least as good as the best-effort congestion level under network neutrality is not sufficient to guarantee efficient infrastructure investments. However, if the ISP has an incentive to strategically degrade the quality of the best-effort class, a MQS may be an appropriate policy instrument to mitigate the detrimental effects that this practice can have on content variety and welfare. Strategic quality degradation can also possibly be counteracted by Internet transparency obligations. Such obligations are already explicitly incorporated in the new US and European regulatory framework (FCC 2010, European Commission 2009, art. 21).

In conclusion, while our results show that some of the objections to QoS tiering are justified, we also find a strong case for a tiered network. The potential dangers of a QoS tiering regime, such as strategic quality degradation, can be overcome by transparency obligations or minimum quality standards. Furthermore, because strategic quality degradation reduces content variety, it is even less profitable when ISPs are in competition for customers. To the contrary, under competition ISPs will try to attract customers by offering them more content variety and a lower average congestion level than their competitor. This will boost their investment incentives. Likewise, the ISPs will also lower the customers' access charge and therefore some of the ISPs' rent is shifted towards the consumers. However, competition between ISPs does not change the main insights of our analysis under monopoly if we make the reasonable assumption that CPs multihome (i.e., are connected with best-effort to every ISP) whereas consumers singlehome (i.e., are connected to one ISP exclusively). In this case, every CP would again face a terminating monopoly over the connected consumers at each ISP, leaving the previously described relationship between the ISP and the CPs intact. Consequently, our result that QoS tiering is the ISPs' preferred regime and that it will lead to more investment and content variety due to the additional priority revenue incentive, remains unchanged. Hence, there is no reason to believe that competition between ISPs

[28] Interestingly, Google CEO Eric Schmidt argues that such an agreement would be in line with net neutrality, because it does not discriminate against specific CPs (Fehrenbacher 2010).

will warrant network neutrality. In reverse, the prohibition of QoS tiering (pay-for priority), which has been proposed by the FCC for fixed line networks, can eventually be harmful to content variety, broadband investment and welfare.

Acknowledgments

We would like to thank Ingo Vogelsang, Michal Grajek, Marc Bourreau and two anonymous referees as well as the editors for helpful comments. Financial aid by the German Research Foundation (DFG) and the NET Institute is gratefully acknowledged.

Appendix A: Alternative Model Variants

A.1. Model with Heterogeneous Internet Customers

Internet customers are now considered to be heterogeneous with respect to their willingness to pay for Internet connectivity. To this end, customers' utility (3) is modified as follows

$$U = \begin{cases} b - t\eta + v\bar{\theta} - \iota w - a & \text{if connected} \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where t is the degree of heterogeneity and η is assumed to be uniformly distributed on the unit interval. The results for this model cannot be presented in closed form solutions. Therefore we provide numerical results in Figure 4. Therein the outcomes under net neutrality (black) and QoS tiering (gray) are compared for a range of values of r (the CP's marginal valuation for customers) and v (the customers marginal valuation for CPs), the parameters that are relevant for the cross-side network effects in our two-sided market model. Compared to net neutrality, we find that QoS tiering indeed results in lower access fees to consumers, and thus more consumer subscriptions, as well as higher consumer welfare. At the same time, under QoS tiering less CPs are active in the network, and CPs' surplus is likely to be lower. Interestingly, if r is much larger than v, CPs' surplus may even be higher under QoS tiering. This is because CPs are charged relatively less for priority and customers relatively more for access in this case. Finally, it is evident that the ISP still prefers QoS tiering over net neutrality and that total short run welfare remains higher under QoS tiering.

[ENTER FIGURE 4 ABOUT HERE]

A.2. Model with Re-congestion Effect

The re-congestion effect describes that CPs in the priority class receive more traffic (clicks) than CPs in the best-effort class. We model this through the following modification of (2).

$$\Gamma_Q(\theta) = \begin{cases} (1 - \theta\ w_{Q2})\, \lambda_{BE}\, \bar{\eta}\, r & \text{if active in best-effort class} \\ (1 - \theta\ w_{Q1})\, \lambda_{QOS}\, \bar{\eta}\, r - \lambda_{QOS}\, \bar{\eta} p & \text{if active in priority class} \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where λ_{BE} and λ_{QOS} denote the clicks received by each CP in the best-effort and priority class, respectively. By definition of the re-congestion effect $\lambda_{BE} < \lambda_{QOS}$. ISP's profits under QoS tiering are adapted accordingly. In the following we intend to show that this model yields qualitatively the same results as the standard model

with a relatively large mass of congestion sensitive CPs ($\alpha < 1$). To this end, we conduct a numerical analysis in which we compare the results of both models side-by-side (Figure 5). The figure shows that the two models yield qualitatively the same results with respect to the comparison of QoS tiering and net neutrality. Thus, it can be concluded that an increase in the re-congestion effect (i.e., an increase of λ_{QOS}) acts very similar to an increase in the mass of congestion sensitive CPs (i.e., a decrease of α) within the scope of our analysis.

[ENTER FIGURE 5 ABOUT HERE]

A.3. Model with Competitive Clicks

We now assume each customer spends an exogenous amount of clicks on the Internet, say Λ, which he evenly distributes among the available CPs. That is, $\lambda = \Lambda / F(\bar{\theta})$, and consequently, λ decreases as the number of active CPs increases. Notice that this model is in line with Assumption 1, because each active CP receives the same number of clicks from each customer. In conjunction with Assumption 2 this model variant yields the following results:

$$F(\bar{\theta}_N) = \bar{\theta}_N = \mu - \Lambda \tag{16}$$

$$F(\bar{\theta}_Q) = \bar{\theta}_Q = \mu - \Lambda \tag{17}$$

$$F(\tilde{\theta}) = \tilde{\theta} = \frac{p}{r-p}\frac{\mu-\Lambda}{\Lambda}\bar{\theta}_Q, \tag{18}$$

Obviously, it holds that $\bar{\theta}_N = \bar{\theta}_Q = 1/w_N$, which is exactly the result that is denoted by Proposition 1. Furthermore, the optimal priority price, which maximizes $\Lambda\beta p$, is given by

$$p = \frac{\mu - \sqrt{\mu(\mu-1)}}{\mu} r = \left(1 - \sqrt{\frac{\bar{\theta}_Q}{\mu}}\right) r = \left(1 - \frac{w_{Q1}}{\hat{w}_Q}\right) r, \tag{19}$$

which is exactly the price structure that is described by (11). Consequently this model variant must yield the same qualitative results.

A.4. Model with Congestion Sensitive Consumers

Consider the Internet customers' utility function from (3), but now assume that consumers are congestion sensitive and, instead of the average congestion level, evaluate congestion by $\hat{w}_N = \int_{\theta=0}^{\bar{\theta}} w_N \theta f(\theta) d\theta$ and $\hat{w}_Q = \int_{\theta=0}^{\tilde{\theta}} w_{Q2} \theta f(\theta) d\theta + \int_{\theta=\tilde{\theta}}^{\bar{\theta}} w_{Q1} \theta f(\theta) d\theta$, respectively. It follows that $\hat{w}_N \geq \hat{w}_Q$ for $\bar{\theta}_Q = \bar{\theta}_N = \bar{\theta}$ and $\mu_N = \mu_Q = \mu$:

$$\hat{w}_N \geq \hat{w}_Q \Leftrightarrow \frac{w_N - w_{Q1}}{w_{Q2} - w_N} \geq \frac{\int_{\theta=0}^{\tilde{\theta}} \theta f(\theta) d\theta}{\int_{\theta=\tilde{\theta}}^{\bar{\theta}} \theta f(\theta) d\theta}$$

Substituting $w_N = \beta w_{Q1} + (1-\beta) w_{Q2}$ and $\beta = 1 - F(\tilde{\theta})/F(\bar{\theta})$ and integrating the right hand side yields:

$$\frac{F(\tilde{\theta})}{F(\bar{\theta}) - F(\tilde{\theta})} \geq \frac{\tilde{\theta}F(\tilde{\theta}) - \int_{\theta=0}^{\tilde{\theta}} F(\theta) d\theta}{\bar{\theta}F(\bar{\theta}) - \tilde{\theta}F(\tilde{\theta}) - \int_{\theta=\tilde{\theta}}^{\bar{\theta}} F(\theta) d\theta}$$

$$\Leftrightarrow F(\bar{\theta})(\bar{\theta} - \tilde{\theta}) + \frac{F(\bar{\theta}) - F(\tilde{\theta})}{F(\tilde{\theta})} \int_{\theta=0}^{\tilde{\theta}} F(\theta) d\theta \geq \int_{\theta=\tilde{\theta}}^{\bar{\theta}} F(\theta) d\theta$$

From $F(\bar{\theta})(\bar{\theta} - \tilde{\theta}) \geq \int_{\theta=\tilde{\theta}}^{\bar{\theta}} F(\theta) d\theta$ and from $\bar{\theta} \geq \tilde{\theta}$ as well as the monotonicity of the distribution function F, it follows that the inequality is always satisfied.

Consequently, if consumers are congestion sensitive, QoS tiering will warrant consumers a higher gross utility than net neutrality. Therefore, the specification with congestion sensitive consumers introduces an additional short run welfare gain in favor of the QoS tiering regime which is not present in the base model.

Appendix B: Proofs

B.1. Relation between Indifferent Content Providers

Recall $\bar{\theta}_Q$ from (9) and notice that

$$\tilde{\theta} = \frac{p\left(\mu - \lambda(F(\bar{\theta}_Q)\right)}{r\lambda F(\bar{\theta}_Q)} \left(\mu - \lambda\left(F(\bar{\theta}_Q) - F(\tilde{\theta})\right)\right). \tag{20}$$

It follows that

$$\tilde{\theta} \leq \bar{\theta}_Q \quad \Leftrightarrow \quad p \leq r\frac{\lambda F(\bar{\theta}_Q)}{\mu}. \tag{21}$$

Under Assumption 2, where p is determined by (25) and $F(\bar{\theta}_Q)$ by (23) this condition becomes $\sqrt{\lambda+1} \leq \lambda+1$, which is always true.

B.2. Proposition 1

It is easy to verify that for $\alpha = 1$

$$F(\bar{\theta}_N) = \bar{\theta}_N = \frac{\mu}{\lambda+1} \tag{22}$$

$$F(\bar{\theta}_Q) = \bar{\theta}_Q = \frac{\mu}{\lambda+1} \tag{23}$$

$$F(\tilde{\theta}) = \tilde{\theta} = \frac{p}{\lambda(r-p)}\bar{\theta}_Q, \tag{24}$$

which proves the first part of the proposition. Furthermore from (11) it follows that

$$p = \left(1 - \sqrt{\frac{1}{\lambda+1}}\right) r = \left(1 - \sqrt{\frac{\bar{\theta}_Q}{\mu}}\right) r \tag{25}$$

For $\alpha \neq 1$, $\bar{\theta}_Q$ will generally depend on p. First see that p cannot exceed p_{max} which solves $\Gamma_Q(\bar{\theta}_Q) = \Gamma_Q(\tilde{\theta})$. This price is

$$p_{max} = r\left(1 + \frac{\lambda\alpha + 1 - \sqrt{(\lambda\alpha+1)^2 + 4\mu\lambda(1-\alpha)}}{2\mu\lambda(1-\alpha)}\right)$$

The feasible values of $F(\bar{\theta}_Q)$ and $F(\bar{\theta}_N)$ in the interval $p \in [0, p_{max}]$ are plotted in Figure 6. It can be readily seen that $F(\bar{\theta}_N) = F(\bar{\theta}_Q)$ for $\alpha = 1$, irrespective of the value of p, and for $\alpha \neq 1$ whenever $p = 0$ or $p = p_{max}$. In all other cases $F(\bar{\theta}_N) \neq F(\bar{\theta}_Q)$ according to the proposition.

[ENTER FIGURE 6 ABOUT HERE]

B.3. Proposition 2

Given the fact that $F(\bar{\theta}_N) = F(\bar{\theta}_Q)$ under our assumptions, it follows that

$$\Pi_Q - \Pi_N = \Lambda\beta p = \mu r\left(1 + \frac{1}{1+\lambda} - \frac{2}{\sqrt{1+\lambda}}\right),$$

which is always greater than zero for $\mu, r, \lambda > 0$.

B.4. Proposition 3

$$\begin{aligned}\Delta W &= (\Pi_Q - \Pi_N) + (\Gamma_Q - \Gamma_N) \\ &= \lambda r \left(\underbrace{(w_N - w_{Q1}) \int_{\tilde{\theta}}^{\bar{\theta}} \theta \; d\theta}_{\substack{\text{congestion alleviation} \\ \text{to priority class}}} - \underbrace{(w_{Q2} - w_N) \int_{0}^{\tilde{\theta}} \theta \; d\theta}_{\substack{\text{congestion aggravation} \\ \text{to best-effort class}}} \right) \\ &= \frac{\lambda r}{2} \left(\left(\bar{\theta}^2 - \tilde{\theta}^2 \right) (w_N - w_{Q1}) - \tilde{\theta}^2 \left(w_{Q2} - w_N \right) \right).\end{aligned} \tag{26}$$

Thus,

$$\Delta W > 0 \Leftrightarrow \frac{w_N - w_{Q1}}{w_{Q2} - w_N} > \frac{\tilde{\theta}^2}{\bar{\theta}^2 - \tilde{\theta}^2} \Leftrightarrow \frac{1-\beta}{\beta} > \frac{(1-\beta)^2}{1-(1-\beta)^2} \Leftrightarrow 0 < \beta < 1$$

Equation (26) reveals that the overall effect of QoS tiering on welfare depends on the relative size of the congestion alleviation effect to providers in the priority class (vertically hatched area in Figure 7) versus the congestion aggravation effect to providers in the best-effort class (horizontally hatched area in Figure 7). These effects relate directly to the main argument of proponents and opponents of net neutrality, respectively.

[ENTER FIGURE 7 ABOUT HERE]

B.5. Proposition 4

Incentives to invest into network capacity are higher under QoS tiering iff marginal revenues from priority sales are greater than zero, provided that the ISP revenues are concave, and the costs of capacity expansion convex in μ. The latter is warranted by assumption. To ensure that the ISP's revenues are concave the property $\partial^2 \Pi_Q / \partial \mu^2 \leq 0$ has to be fulfilled. The second-order condition is thus given by

$$\frac{\partial^2 \Pi_Q}{\partial \mu^2} = -\frac{\iota (1+\lambda)}{\mu^3} + \underbrace{\left[\frac{\partial^2 r(\bar{\theta})}{\partial \bar{\theta}^2} \frac{\bar{\theta}}{2} + \frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \right]}_{A} \frac{\partial \bar{\theta}}{\partial \mu} \underbrace{\left(1 + \frac{1}{(1+\lambda)} - \frac{2}{\sqrt{1+\lambda}} \right)}_{B} < 0.$$

Since $B \geq 0$ always holds, the ISP's revenues are concave if

$$\frac{\partial^2 \Pi_Q}{\partial \mu^2} \leq 0 \begin{cases} A \leq 0 & \text{always} \\ A > 0 & \text{if} \quad \frac{\iota (1+\lambda)^2}{\mu^3 B} \geq A. \end{cases}$$

It is easy to see that $A \leq 0$ is warranted if ad revenues are decreasing (which is given by assumption) and concave (or not too convex). Otherwise we must assume, that the condition in the second case holds. However, alternatively it can also be assumed that the second-order condition holds locally around μ^*. Now consider $\Pi_Q - \Pi_N = \Lambda \beta p$. Differentiating with respect to μ yields

$$\frac{\partial (\Pi_Q - \Pi_N)}{\partial \mu} = \frac{\sqrt{\lambda+1} \left((\lambda+1) - \sqrt{\lambda+1} \right)}{(\lambda+1)\sqrt{\lambda+1}} \left[\frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \frac{\partial \bar{\theta}}{\partial \mu} \mu + r(\bar{\theta}) \right]$$

The sign of the derivative is determined by the part in square brackets. Notice from (22) and (23) that $\partial \bar{\theta} / \partial \mu = 1/(\lambda+1)$ and $\mu = \bar{\theta}(\lambda+1)$. Consequently,

$$\frac{\partial (\Pi_Q - \Pi_N)}{\partial \mu} > 0 \Leftrightarrow \varepsilon^r = \frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \frac{\bar{\theta}}{r(\bar{\theta})} > -1$$

Note that the gross industry advertisement revenue under net neutrality is given by $R(\bar{\theta}) = \lambda r(F(\bar{\theta}))F(\bar{\theta})$. It is sensible to assume that $R(\cdot)$ does not decrease as more content becomes available. Thus, under the uniform distribution,

$$\frac{\partial R(\cdot)}{\partial \bar{\theta}} = \lambda \frac{\partial r(\cdot)}{\partial \bar{\theta}} \bar{\theta} + \lambda r(\cdot) > 0,$$

which holds iff $\varepsilon^r > -1$, in which case QoS tiering leads to more investments.

B.6. Proposition 5

To see that the overall congestion level, w decreases with capacity expansion, we show that $\partial w/\partial \mu = \partial(1/(\mu-\Lambda))/\partial \mu < 0$: Notice that $\Lambda = \bar{\theta}\lambda = \lambda\mu/\lambda+1$, so that $\partial\Lambda/\partial\mu = \lambda/\lambda+1 < 1$. Therefore, it holds that $\partial(\mu-\Lambda)/\partial\mu > 0$ and consequently, $\partial(1/(\mu-\Lambda))/\partial\mu < 0$. By equation (2) and (3) it is immediately obvious that the gross utility of customers and CPs increases as the congestion level decreases. The homogeneity of customers allows the ISP to fully expropriate the additional customer utility. Capacity expansion also increases the amount of active CPs, since $\partial\bar{\theta}/\partial\mu > 0$ by equation (23). Before the capacity expansion occurred, these CPs had a surplus of zero and are therefore unambiguously better off.

If QoS tiering provides a higher capacity level ($\mu_Q^* > \mu_N^*$), the critical CP that is just equally well off as under network neutrality is determined by the equation $(1 - \breve{\theta} w_N)\lambda r_N = (1 - \breve{\theta} w_{Q1})\lambda r_Q - \lambda p$. Inserting (11) and reformulating yields:

$$\breve{\theta} = \frac{(r_N - r_Q)\, w_Q + (w_Q - w_{Q1})}{w_Q\,(w_N r_N - w_Q r_Q)}. \tag{27}$$

Because $\mu_Q^* > \mu_N^*$, it immediately follows that that $w_N > w_Q$, $\bar{\theta}_N < \bar{\theta}_Q$ and thus $r_N \geq r_Q$. It is easy to see that $0 < \breve{\theta} < \bar{\theta}_Q = 1/w_Q$. Therefore, all CPs in the interval $(\breve{\theta}, \bar{\theta}_Q]$ are better off than under network neutrality.

B.7. Proposition 6

We consider each regime separately and show that the conditions with respect to efficient investments coincide. First, we derive the conditions for which $\partial(W_N - \Pi_N)/\partial\mu$ is larger than zero:

$$\begin{aligned} W_N - \Pi_N &= \frac{\lambda}{2\,(\lambda+1)} \mu r(\bar{\theta}) \\ \frac{\partial\,(W_N - \Pi_N)}{\partial \mu} > 0 \quad &\Leftrightarrow \quad \frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \frac{\bar{\theta}}{r(\bar{\theta})} > -1 \quad \Leftrightarrow \quad \varepsilon^r > -1 \end{aligned}$$

The difference of private and efficient investment incentives under the QoS tiering regime is:

$$\begin{aligned} W_Q - \Pi_Q &= \frac{\sqrt{\lambda+1}-1}{\lambda+1} \mu r(\bar{\theta}) \\ \frac{\partial\,(W_Q - \Pi_Q)}{\partial \mu} &= \frac{\sqrt{\lambda+1}-1}{\lambda+1} \left(\frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \mu + r(\bar{\theta}) \right) > 0 \quad \Leftrightarrow \\ \frac{\partial r(\bar{\theta})}{\partial \bar{\theta}} \frac{\bar{\theta}}{r(\bar{\theta})} &> -1 \quad \Leftrightarrow \quad \varepsilon^r > -1 \end{aligned}$$

By the same argument as in the proof of Proposition 4, only $\varepsilon^r > -1$ is feasible and thus the proposition obtains.

B.8. Proposition 7

To show that the ISP under a minimum quality standard enforcement of $w_{Q2} = w_N$ has a higher incentive to invest in capacity than under network neutrality we have to show that $\mu_{MQS} > \mu_N^*$.

$$\begin{aligned} w_{Q2} = w_N \quad &\Leftrightarrow \\ \frac{\mu_{MQS}}{\mu_{MQS} - \lambda\bar{\theta}} \frac{1}{\mu_{MQS} - \lambda\beta\bar{\theta}} = \frac{1}{\mu_N^* - \lambda\bar{\theta}_N} \quad &\Leftrightarrow \\ \mu_{MQS} = \frac{1+\lambda}{1+\lambda(1-\beta)}\mu_N^* & \end{aligned}$$

Since $\beta < 1$ it is easy to see, that $\mu_{MQS} > \mu_N^*$ always holds true.

B.9. Proposition 8

In the QoS tiering regime with maximum quality degradation the last CP to enter the market is located at $\bar{\theta}_D = \frac{\mu(1-p_D/r)}{1+\lambda(1-p_D/r)}$. In contrast, the last CP to enter under network neutrality, or equivalently under the unhampered QoS tiering regime, is located at $\bar{\theta}_{N,Q} = \frac{\mu}{1+\lambda}$. Obviously, for $p_D = 0$, which corresponds to a network neutrality regime, the indifferent CPs coincide. However, $\forall p_D > 0$ it easy to see that $\bar{\theta}_D < \bar{\theta}_{N,Q}$ for $\lambda > 0 > r/(p_D - r)$. This proves the first part of the proposition.

The ISP's profit under maximum quality degradation is $\Pi_D = a(p_D) + \lambda\bar{\theta}_D(p_D)\, p_D$, which is maximized by a price of $p_D = [r(1+\lambda) - \sqrt{r(v + r(1+\lambda))}]/\lambda$. At this price level, $W_D < W_N$ iff $v < r\lambda(1+\lambda)$, which is the same condition as for a positive equilibrium price. Thus, as long as $p_D > 0$ (which is shown in Proposition 9), welfare is lower under QoS tiering with maximum quality differentiation compared to network neutrality (which again has lower welfare than the unhampered QoS tiering regime in the short-run).

B.10. Proposition 9

Inserting optimal prices and solving $\Pi_D > \Pi_Q$ for v yields

$$v < r\left(\lambda\left(3 + 2\lambda - 2\sqrt{\lambda+1}\right) - 2\sqrt{\lambda\,(\lambda+1)^2\left(\lambda + 2 - 2\sqrt{\lambda+1}\right)}\right) \equiv \underline{v} \tag{28}$$

Furthermore, $\forall\lambda > 0$ it holds that $\underline{v} < r\lambda(1+\lambda)$ at which $p = 0$. Thus, the ISP never engages in maximum quality degradation at $p_D \leq 0$, but prefers the unhampered QoS tiering regime instead.

References

Armstrong, M. 2006. Competition in two-sided markets. *The RAND Journal of Economics* **37**(3) 668–691.

Brennan, Timothy. 2010. Net neutrality or minimum quality standards: Network effects vs. market power justifications. URL `http://ssrn.com/abstract=1622226`. Mimeo, Social Science Research Network.

Cheng, Hsing Kenneth, Subhajyoti Bandyopadhyay, Hong Guo. 2011. The Debate on Net Neutrality: A Policy Perspective. *Information Systems Research* **22** 60–82.

Choi, Jay Pil, Byung-Cheol Kim. 2010. Net neutrality and investment incentives. *RAND Journal of Economics* **41**(3) 446–471.

Crawford, G.S., M. Shum. 2007. Monopoly quality degradation and regulation in cable television. *The Journal of Law and Economics* **50** 181–219.

Crowcroft, Jon. 2007. Net neutrality: The technical side of the debate - a white paper. *International Journal of Communication* **1** 567–579.

Dou, Wenyu. 2004. Will internet users pay for online content? *Journal of Advertising Research* **44**(4) 349–359.

Economides, N. 1998. The incentive for non-price discrimination by an input monopolist. *International Journal of Industrial Organization* **16**(3) 271–284.

Economides, N., B. Hermalin. 2010. The economics of network neutrality. Mimeo, NET Institute Working Paper 10-25.

Economides, N., J. Tåg. 2008. Net neutrality on the internet: A two-sided market analysis. Mimeo, University of New York, School of Law.

European Commission. 2009. Directive 2009/136/ec of the european parliament and of the council of 25 november 2009. *Official Journal of the European Union* **L337** 11–36. URL `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:337:0011:0036:En:PDF`.

FCC. 2010. Report and order: In the matter of preserving the open internet broadband industry practices. FCC 10-201.

Fehrenbacher, Katie. 2010. Caught on video: Google ceo dishes on google wave, verizon & social strategy. Gigaom.com. URL `http://gigaom.com/2010/08/05/google-ceo-dishes-on-google-wave-verizon-social-strategy/`.

Foros, Ø., H.J. Kind, L. Sørgard. 2002. Access pricing, quality degradation, and foreclosure in the Internet. *Journal of Regulatory Economics* **22**(1) 59–83.

Hahn, R., S. Wallsten. 2006. The economics of net neutrality. *The Berkeley Economic Press - Economists' Voice* **3(6)** 1–7.

Hahn, Robert, Robert Litan, Hal Singer. 2007. The economics of wireless net neutrality. *Journal of Competition Law and Economics* **3** 399–451.

Hermalin, B.E., M.L. Katz. 2007. The economics of product-line restrictions with an application to the network neutrality debate. *Information Economics and Policy* **19** 215–248.

Jamison, M., J. Hauge. 2008. Getting what you pay for: Analyzing the net neutrality debate. URL `http://ssrn.com/abstract=1081690`. Mimeo, Social Science Research Network.

Kleinrock, L. 1976. *Queuing Systems, Volume 1*. John Wiley & Sons, Inc., New York.

Lambert, Paul. 2010. Vodafone and telefonica are overplaying their hand with google. Telecoms.com. URL `http://www.telecoms.com/18389/vodafone-and-telefonica-are-overplaying-their-hand-with-google/`. Last accessed on 12/08/2010.

Lessig, Lawrence. 2001. *The future of ideas*. Random House New York.

McDysan, D. 1999. *QoS and traffic management in IP and ATM networks*. McGraw-Hill, Inc. New York, NY, USA.

O'Connell, Patricia. 2005. At sbc, it's all about "scale and scope". Businessweek. URL `http://www.businessweek.com/magazine/content/05_45/b3958092.htm`.

Owen, Bruce, Gregory Rosston. 2006. Local broadband access: Primum non nocere or primum processi? a property rights approach. Thomas Lenard, Randolph May, eds., *Net Neutrality or Net Neutering: Should Broadband Internet Services be Regulated*. Springer US, 163–194. URL `http://dx.doi.org/10.1007/0-387-33928-0_5`.

Rochet, J.C., J. Tirole. 2006. Two-sided markets: A progress report. *The RAND Journal of Economics* **37**(3) 645–667.

Ronnen, U. 1991. Minimum quality standards, fixed costs, and competition. *The RAND Journal of Economics* **22**(4) 490–504.

Schneibel, Gerhard, Cyrus Farivar. 2010. Deutsche telekom moves against apple, google and net neutrality. Deutsche Welle. URL `http://www.dw-world.de/dw/article/0,,5439525,00.html`.

Schuett, Florian. 2010. Network neutrality: A survey of the economic literature. *Review of Network Economics* **9**(2) Article 1.

Sidak, Gregory J. 2006a. Hearing on Network Neutrality. Testimony before the United States Senate, Committee on Commerce, Science and Transportation.

Sidak, J.G. 2006b. A Consumer-Welfare Approach to Network Neutrality Regulation of the Internet. *Journal of Competition Law and Economics* **2**(3) 349.

Sydell, Laura. 2006. Internet debate - preserving user parity. "All Things Considered" - National Public Radio, USA. URL `http://www.npr.org/templates/story/story.php?storyId=5362403`.

Sydell, Laura. 2007. Firms abandon online subscription plans. "All Things Considered" - National Public Radio, USA. URL `http://www.npr.org/templates/story/story.php?storyId=14537587`.

Van Schewick, B. 2006. Towards an economic framework for network neutrality regulation. *Journal on Telecommunications & High Technology Law* **5** 329.

Wu, T., C.S. Yoo. 2007. Keeping the Internet Neutral?: Tim Wu and Christopher Yoo Debate. *Federal Communications Law Journal* **59**(3) 575–592.

Wu, Tim. 2003. Network neutrality, broadband discrimination. *Journal on Telecommunications & High Technology Law* **2** 141.

Wyatt, Edward. 2010. Google and verizon near deal on web pay tiers. New York Times.

Yoo, Christopher. 2005. Beyond network neutrality. *Harvard Journal of Law & Technology* **19** 1–77.

Figures

Figure 1 The short run effect of QoS tiering on CPs' surplus.

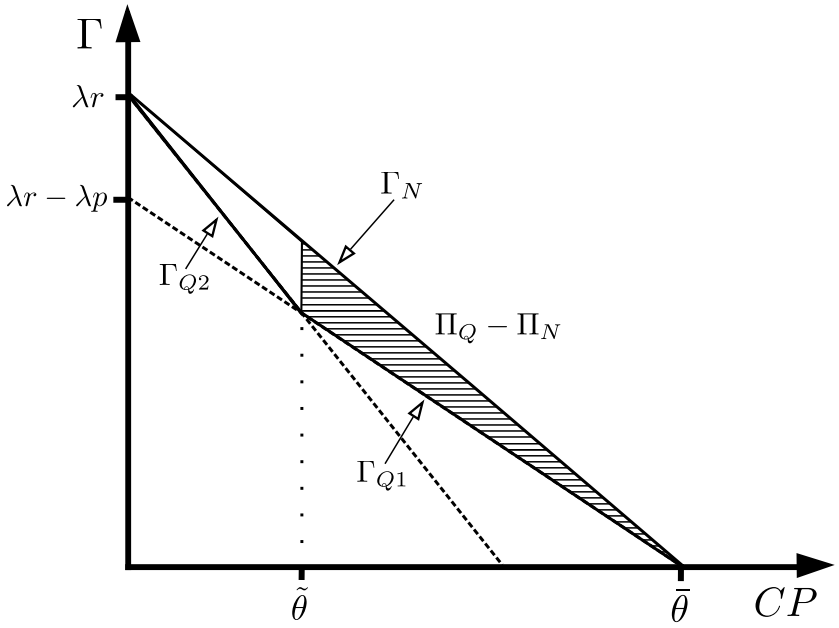
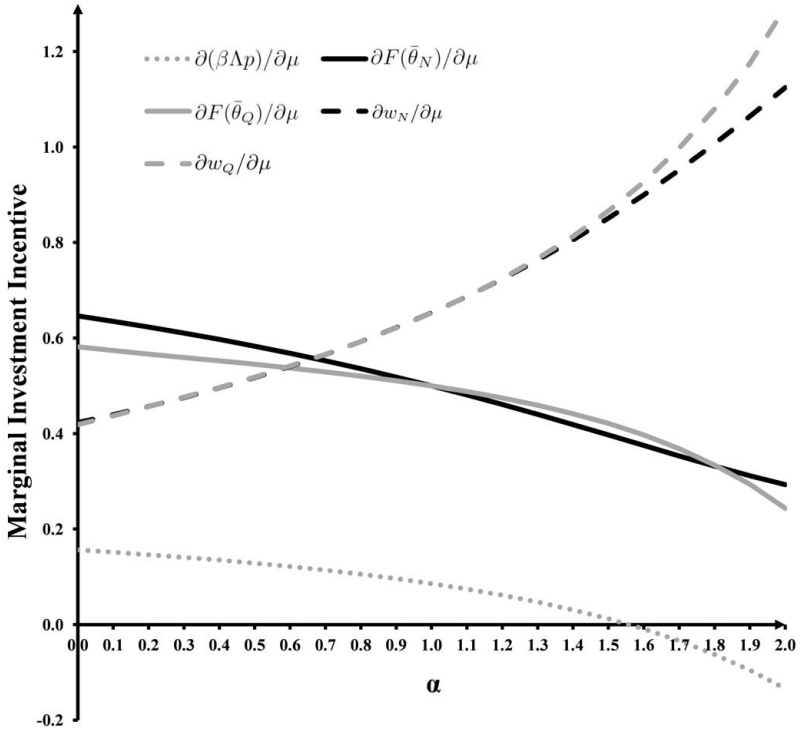


Figure 2 **ISP's marginal investment incentives under QoS tiering (gray) and net neutrality (black) for different distributions of CPs' congestion sensitivity (α).**



Note. The figure is derived for $\mu = 7/4$, $\lambda = 1$, $r = 1$, $v = 1$, but qualitatively identical results are obtained for other parameter values.

Figure 3 **The long run effect of QoS tiering on innovation and welfare.**

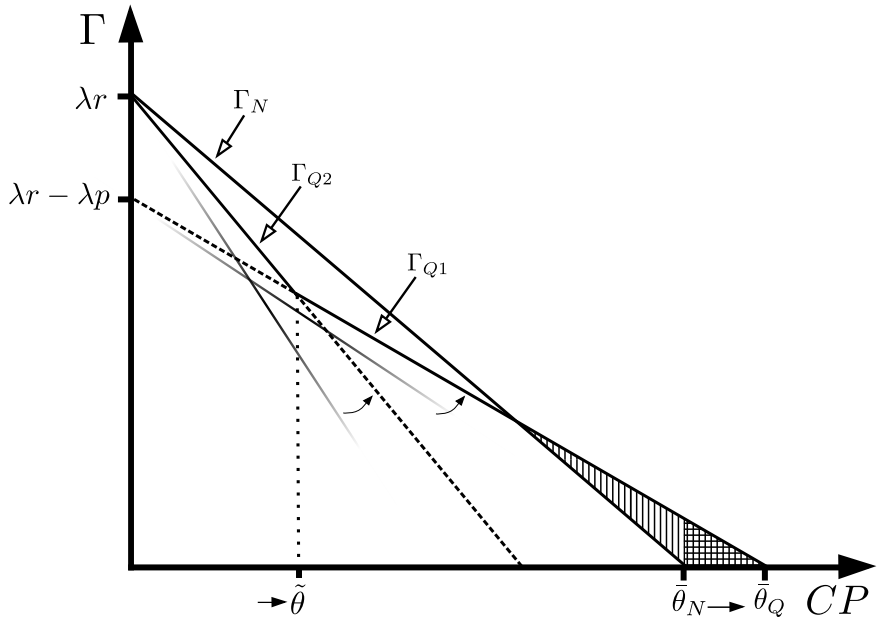
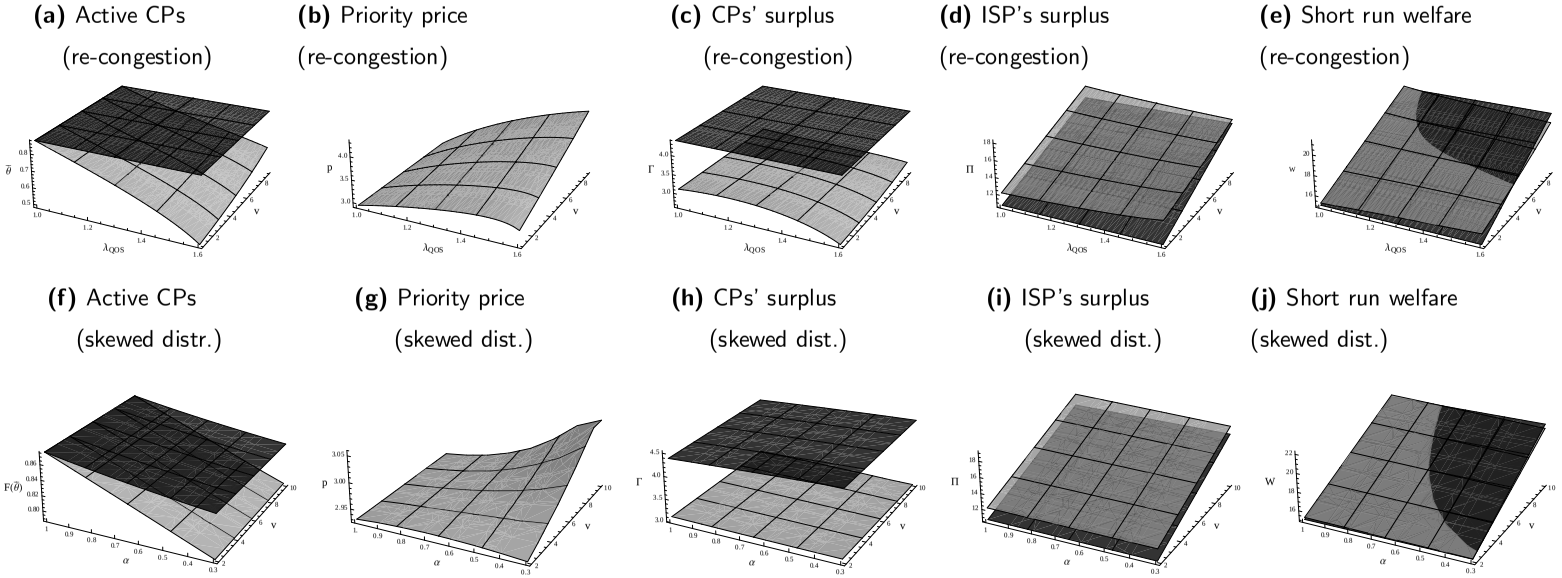


Figure 4 Comparison between QoS tiering (gray) and net neutrality (black) with heterogeneous customers.

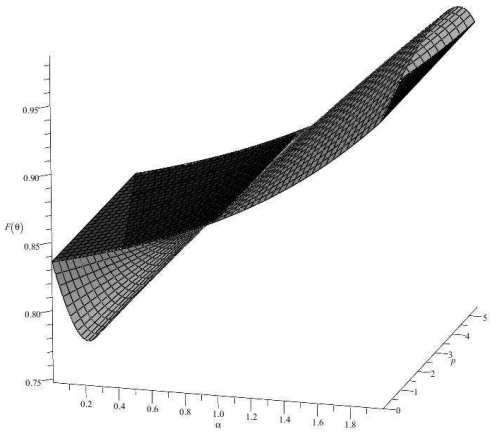
Note. The figures are derived for $\mu = 1$, $\lambda = 1$, $t = 10$, $b = 1$, but qualitatively identical results are obtained for other parameter values.

Figure 5 **Comparison of the re-congestion effect and the effect of a skewed distribution of CPs congestion sensitivity ($\alpha < 1$) with respect to the outcome of the comparison between QoS tiering (gray) and net neutrality (black).**



Note. The figures are derived for $\mu = 7/4$, $\lambda = 1$, $b = 10$, $r = 10$, but qualitatively identical results are obtained for other parameter values.

Figure 6 **Active CPs under QoS tiering (gray) and net neutrality (black) for different distributions of CPs congestion sensitivity (α).**



Note. The figure is derived for $\mu = 7/4$, $\lambda = 1$, $r = 10$, but qualitatively identical results are obtained for other parameter values.

Figure 7 **Congestion alleviation vs. congestion aggravation effect of QoS tiering.**