Innovation through discrimination!?
A Formal Analysis of the Net Neutrality Debate

Jan Krämer, Lukas Wiewiorra*

University of Karlsruhe
Institute of Information Systems and Management
Karlsruhe, Germany

July 5, 2009

Abstract

We model the main arguments of the net neutrality debate in a two-sided market framework with network congestion sensitive content providers and Internet consumers on each side, respectively. The platform is controlled by a monopolistic Internet service provider, who may choose to sell content providers prioritized access to its customers. We explicitly consider the adverse effects of traffic prioritization to the remaining best-effort class and find that network discrimination has overall positive effects on welfare, because congestion is better allocated to those content providers with congestion inelastic advertisement revenues. In the long-run, network discrimination leads to infrastructure investments in transmission capacity and encourages innovation on the content provider side. In the short-run, however, discrimination has no effect on innovation because the ISP expropriates the content providers' increased surplus through the price for priority access. This is the downside of network discrimination: Albeit total welfare is increased, content providers will–at least in the short-run–be worse off than under network neutrality.

Keywords: Telecommunication, Network Neutrality, Two-Sided Market, Traffic Prioritization, Innovation, Broadband Investment

*Financial support by the Graduate School 895 "'Information Management and Market Engineering"' at the Universitt Karlsruhe (TH), funded by Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged by both authors.

1 Introduction

Without doubt, the Internet has become the central means of communications in almost all parts of the world and is virtually ubiquitously available, even in very remote areas. The Internet's popularity is in large due to the sheer endless number of available applications and services. The Internet not only accommodates all classical means of communications, such as telephony and video, but also an increasing number of new innovative services. The growing importance of the Internet is tightly intertwined with the innovations it has put forth. These innovations include new communications technology (e.g. instant messaging or video-on-demand), new business models (e.g. online retailing) and new forms of social interaction (e.g. social networking). It has been frequently argued[1] that these innovations could only have emerged because the Internet provides an open, standardized and above all non-discriminatory platform. More precisely, the so-called end-to-end principle has always been a major design principle of the Internet. It postulates that all "intelligence" (e.g. new protocols) should be located at the edges of the network (i.e. the sending and receiving end), while the backbone network shall be kept dedicated to simply relaying data on a non-discriminatory basis via the Internet protocol (IP). In particular, the end-to-end principle allows for innovations at the edge of the network without requiring prior "consent" of any party controlling the backbone network. In this spirit, the most radical proponents of network neutrality, oppose violations of the end-to-end principle which may result in any sort of discrimination of data packets based on their origin, destination or content type (cf. e.g. Crowcroft, 2007). Discrimination of whatever sort is seen as a threat to the innovational spirit of the Internet, because it would preclude a level playing field for providers of new innovative services. In particular, there is a deep-grounded fear that innovation and investment incentives of Internet start-ups would be negatively distorted or even completely destroyed because well-established content providers could buy themselves a preferred treatment of their data and thereby foreclose entry of new firms. Under this view, it is argued that some of the currently successful services, such as peer-to-peer filesharing or voice-over-IP would not have been possible if the respective data packets were put at a disadvantage to traffic from established content.

On the contrary, opponents of net neutrality comment that discrimination is a necessary means in order to be able to cope with the increasing infrastructure requirements of new innovative services, e.g. with respect to bandwidth and latency. In this sense, discrimination has a positive annotation-and might even be welfare enhancing, because it explicitly enables entry of those content

[1] In particular by consumer rights groups, such as `savetheinternet.com` or `openinternetcoalition.com`, and large content providers, such as Google, Yahoo and eBay (cf. e.g. Google, 2009), but also by academics (e.g. Crawford, 2007).

providers who crucially hinge on bandwidth and latency requirements which the traditional net neutral best-effort Internet cannot provide (in the future). Consequently, under this view net neutrality might even hinder innovation, because innovative business models would be required to be sustainable under the best-effort domain.

Of course the arguments are much more multi-layered than can be described here. Many recent publications, especially from the law domain, have attempted to cover the relevant facets of the net neutrality debate. More formal analyses of the issue are rare, however. We attempt to isolate the formal arguments in order to shed some light on the effect of network discrimination on innovation and welfare.

The majority of Internet users pays a monthly fee, either as a flatrate subscription or based on time or volume, for accessing content and services on the Internet. Certainly, for each user the value of his Internet connection increases with the availability of more content providers. However, generally users are fairly reluctant to pay additionally for specific content or services (Dou, 2004). The notion of a free-for-all Internet is still manifested deep in the minds of many and roots from the not-for-profit history of the early academically oriented Internet. Thus, for the most part content providers have to rely on online advertisement on their respective websites in order to generate revenues from their service offerings. Consequently, content providers value the presence of Internet users, because they will generate revenues through clicks on advertisements.

More formally, we capture this scenario by considering Internet users and content providers to be on one side of a two-sided market, respectively: The presence of users on one side of the market generates positive network externalities for content providers on the other side, and vice versa. Both sides are connected through an Internet service provider (ISP) who "controls" the market in the sense that he sets the terms of access for users and content providers strategically. Users are always charged a linear price for accessing the Internet. Under a neutral network regime, with a single best-effort data transmission service, the ISP does not charge the content provider in the market. Under a discriminatory regime, the ISP offers the content providers the additional option to buy priority access to the Internet for an extra charge. The content providers who exercise this option will be provided with a priority transmission service which enqueues request to their websites ahead of those request to websites of content providers in the best-effort class.

We investigate the effects of such discriminatory practice in the short-run and in the long-run. In the short-run, when the ISP's transmission capacity is fixed, network discrimination has at least two different effects.
On the one hand, speeding up some traffic will unmistakably lead to a reduction of average transmission speed in the best-effort class. Thus, those content providers who are not willing to pay for priority access are put at a

disadvantage twice: First, from the speeding up of other providers' content and second from the slowing down of their own content. This disadvantage lies at the heart of the concern of network neutrality proponents.[2]
On the other hand, the business model of some content providers will be more sensitive to transmission quality[3] than that of other providers. For example, a content provider who offers an on-demand video streaming service is certainly more sensitive to network congestion than a simple e-mail service provider. Consequently, when a content provider's service cannot be reliably offered such that the users' experience is unsatisfactory, online advertisement revenues will decline, possibly up to the point where the content provider is forced out of business. This argument is therefore in favor or network discrimination, because it would actually allow for entry of innovative, transmission quality demanding services.

Under some further assumptions, we formally investigate which one of the two arguments has more impact under a discriminatory network regime. Our main finding is that in the long-run *network discrimination will lead to more innovation*. Furthermore, we compare the overall welfare effects of discriminatory practice with respect to a network neutral regime and find that *network discrimination is generally welfare enhancing*. This is because congestion is better allocated to the congestion insensitive content providers, while providing some congestion relief to the content providers with congestion sensitive business models. However, *all content providers are worse off under network discrimination because this efficiency gain is fully appropriated by the ISP* through the welfare-neutral priority charge.

Finally, we explicitly compare the ISP's long-run incentive to invest into infrastructure (i.e. transmission capacity) under both network regimes. On the one hand, it can be argued that a network neutral regime provides the best incentives for network investment, because ISPs can only charge users higher prices for access if there is a wide variety of content available. Content

[2]In this paper we abstract from any form of direct service discrimination. Examples for such anti-competitive behavior would be the actual case of mobile phone operators blocking VoIP transmissions in their wireless networks (Hahn et al., 2007) or the degradation of traffic flows of certain protocols, such as P2P, in backbones of large Internet service providers. In our model every service provider can buy priority access on the same conditions and discrimination does not correspond to exclusion of services based on anticompetitive behavior. However, it shall be annotated that presently no generally accepted definition of net neutrality exists. For example, at the other extreme end, some net neutrality proponents regard only anti-competitive discrimination of specific services as a violation to net neutrality. In this line, Tim Berners-Lee, founder and director of the World Wide Web Consortium (W3C) argues: *"If I pay to connect to the Net with a certain quality of service, and you pay to connect with that or greater quality of service, then we can communicate at that level." (Berners-Lee, 2006)*. In our view, which is in line with the more formal literature, such practice would clearly violate net neutrality.

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

providers, on the contrary, will only be available if the best-effort transmission capacity is sufficient for their business models. Indeed, currently ISPs generally engage in costly *overprovisioning* of transmission capacity in order to accommodate for the most transmission-quality-demanding content providers in peak times. On the other hand, it seems logical that ISPs have an increased incentive to upgrade their infrastructure under a discriminatory regime, because then they can make the most demaning content providers pay for it.[4] Again, there are arguments in favor of either regime and the overall effect is unclear ex-ante. Our formal analysis reveals that *ISPs have a stronger incentive to invest into network infrastructure under a discriminatory regime.*

The remainder of this article is structured as follows. In Section 2 we compare our model assumptions and results to related work. In Section 3 the formal model is laid out and in Section 4 the short-term effects of network discrimination are considered. In Section 5 we consider the long-term effects of a discriminatory regime with respect to infrastructure investment incentives. We conclude in Section 6 by summarizing our results and providing some policy advice.

2 Related Work

With only few notable exceptions, the net neutrality debate has thus far been characterized by rather emotional and rhetorical publications, mostly coming from the law domain (e.g. Faulhaber and Rasmussen, 2006; Yoo, 2005; Wu, 2005). Within the more formal strand of the literature, Choi and Kim (2008) take an interesting approach, because they embody standard results from queuing theory in order to precisely model the relationship between priority and best-effort traffic. In our model, we will follow the same approach. However, the focus of Choi and Kim differs from ours. The authors investigate the effect of network discrimination among two *competing* content providers in a standard Hotelling model. It is assumed that customers dislike congestion and visit one of the two content providers exclusively. In reverse, the content providers can improve their competitive position by purchasing priority access from a monopolistic ISP. The hitch is, that the ISP will sell priority access only to one of the two content providers. The authors have to make this unconventional assumption, because otherwise the content providers engage in a prisoners' dilemma such that both end up purchasing the priority access: Clearly, individually each content provider prefers the priority access if the other remains in the best-effort class. But if both ex-

[4]Even if transmission capacity is not augmented, a discriminatory network regime might still be welfare enhancing because it avoids some of the the wasteful overprovisioning by alleviating congestion for the most demanding content providers when needed.

ercise this option neither gains an advantage and the price paid for priority access is forfeited. Furthermore, in order to resolve the arising coordination problem, i.e. which of the two content providers is awarded the priority access, Choi and Kim assume cost asymmetry between the two competitors, such that only the "low cost" provider can win the bargaining process with the ISP. The authors find that the content provider with the priority access obtains a larger market share and consumers have to pay a lower access fee under the discriminatory regime. In the short-run the overall effect of discrimination on welfare is ambiguous. On the one hand, consumers pay less for network access, but on the other hand, overall Hotelling transportation costs are higher. In the long-run, the ISP's investment incentives are held in check by two competing effects, which the authors call "rent extraction" and "network access fee" effect. The rent extraction effect provides the ISP with incentives to keep capacity scarce because this allows him to extract more of the content providers' rent. On the contrary, by the network access fee effect the ISP would like to extend capacity, such that he can charge consumers more for network access.

Economides and Tag (2008) analyze the effect of priority access in a network discriminatory regime by considering a two-sided market model, very much like ours. On one side of the market, there is a continuum of non-rivalry content providers 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. However, the authors do not consider a discriminatory regime in the sense that some content providers are prioritized over others. In their model, abandoning neutrality means that the platform owner charges a lump-sum fee for access to his network. They find that for a wide range of parameter values the incentives of the platform owner and a social planner are not aligned. Charging content providers for access leads to reduced market entry and consequently less externality to the consumer side. This, in turn, leads to a lower access charge, such that more consumers enter the market. The authors find that total welfare is higher under network neutrality, because content providers are the only party loosing from abandoning neutrality. This finding carries over to an extension with two ISPs, which makes them conclude that their results are robust under a certain level of competition.

Finally, Hermalin and Katz (2007) model network neutrality in a two-sided market as product line restrictions to the quality decision problem of the ISP. In contrast to our results they find that net neutrality regulation can drive out content providers from the market. Content providers who purchased a low-quality variant from the ISP will be excluded if there is only one single quality available. In their model, the overall welfare implications of neutrality regulation are very often negative. However, the authors do not check for investment incentives of the ISP and therefore the model is static

in its nature.

Our model complements previous work in this strand of the literature because we explicitly consider the adverse effects of traffic prioritization to the remaining best-effort class in a two-sided market model. Thereby, our focus is to study the effect of network discrimination on innovation and infrastructure investment.

3 The Model

We model the Internet as a two-sided market, with content providers (CP) on one side and Internet consumers (IC) on the other side, each of which value an increasing presence of the other side and dislike network congestion. In order to be able to isolate the arguments of the net neutrality debate we abstract from the full complexity of networks forming the Internet and consider a single monopolistic Internet service provider (ISP) providing access to consumers and content providers, respectively.

Internet service provider: The ISP controls the two-sided market through a number of variables which he sets strategically. First, he 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. As will be described below, both consumers and content providers dislike network congestion. The level of network congestion is captured by consumers' average waiting time for content, w, which is again controlled by the ISP through his choice of transmission capacity, μ. The ISP therefore faces the following basic tradeoff: On the one hand, he has an interest to reduce congestion, because this draws more content providers and consumers onto the platform, which in turn increases revenues from access. On the other hand, an increase of the participants on either side of the market will ceteris paribus increase congestion, which can only be counteracted by a costly increase of transmission capacity. Therefore, under a neutral network regime, ISP's long-run profit is[5]

$$\Pi_l^N = \alpha a - k\mu, \tag{1}$$

where α and k denote the share of consumers subscribing in equilibrium and the marginal costs of capacity expansion, respectively. Under a discriminatory network regime, the ISP has an additional source of revenue, which

[5]Throughout this paper, the network neutral regime is indicated by the superscript N, whereas superscript D denotes the network discriminatory regime. Moreover, subscript s and l denote short-run and long-run incentives, respectively.

stems from selling priority access (i.e. access with lower congestion) for a price of p to those content providers requiring it. More precisely,

$$\Pi_l^D = \alpha a + \beta \bar{\theta} p - k\mu, \tag{2}$$

where β is the share of content providers buying priority access and $\bar{\theta}$ denotes the total mass of active content providers.

In the short-run, the ISP's previous investment decisions into transmission capacity are sunk, so that μ can be considered an exogenous variable which is irrelevant for profit maximization. Thus, short-run ISP profits are

$$\Pi_s^N = \alpha a \tag{3}$$

$$\Pi_s^D = \alpha a + \beta \bar{\theta} p. \tag{4}$$

Content Providers: Whatever service content providers offer, they provide it for free and receive revenues only indirectly through online advertisements.[6] In the model, a content provider's advertisement revenue will depend on the gross advertisement revenue per unit of traffic, the average received traffic per user, and the individual sensitivity of his business model. More specifically, consider that a content provider receives a gross advertisement revenue of r per unit of traffic.[7] It is assumed that every content provider has an individual business model with an innate sensitivity to congestion. As argued before, the parameter reflects how crucial the quality of the transmission service is to the success of the individual online business. We differentiate this sensitivity through the parameter θ which is uniformly distributed on the unit interval. Depending on the individual congestion sensitivity, θ, and the current level of network congestion, w, a content provider's advertisement revenue is diminished by θw, such that net advertisement revenue per unit of traffic becomes $r - \theta w$.

Regarding the received traffic, we make two fundamental assumptions for simplification of the analysis:[8]

- Each content provider receives the same amount of traffic from each user, namely L. It is independent of a content provider's business model and consequently its innate sensitivity to network congestion.

- Each content provider's service is unique and therefore content providers are not in competition for traffic. Each content provider receives traffic of L per user, independent of the number of other content providers currently in the market.

[6] In particular, this means that we rule out the possibility of content providers charging consumers directly for access to their content.

[7] One might think of r as revenue-per-click, for example.

[8] We will comment on the model's sensitivity to these assumptions in Section 6.

Put together, under network neutrality, each content provider's profit is

$$U_{CP}^{N}(\theta) = \begin{cases} (r - \theta w^{N})\alpha L & \text{if active} \\ 0 & \text{if inactive.} \end{cases} \tag{5}$$

Notice, that a content provider's profit will also depend on the overall level of network congestion. In a discriminatory regime, differences in profit will therefore depend on the change of congestion level and, if applicable, on the additional price paid for priority treatment. Profits are

$$U_{CP}^{D}(\theta) = \begin{cases} (r - \theta w_1^{D})\alpha L - p & \text{if active with first priority service} \\ (r - \theta w_2^{D})\alpha L & \text{if active with best-effort service} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

As will be seen in detail later, it holds that $w_1^D < w^N < w_2^D$. Under a discriminatory regime, the content provider who is indifferent between choosing the priority and the best-effort transmission service is denoted by $\tilde{\theta}^D$. Furthermore, in both regimes, the content provider who is indifferent between becoming active and staying out of the market is characterized by a congestion sensitivity of $\bar{\theta}$. Recall that θ is normalized to the unit interval, such that $\bar{\theta}$ also reflects the mass of all active content providers. Thus, the share of content providers choosing priority service under a discriminatory regime is given by

$$\beta \equiv 1 - \frac{\tilde{\theta}^D}{\bar{\theta}^D}, \tag{7}$$

for which obviously $0 < \beta < 1$ must hold. Finally, see that the share of consumers connected in equilibrium, α, generates positive externality on the content providers' profit.

Network Congestion: Network congestion is measured through the consumers' average waiting time following a content request. In particular, like Choi and Kim (2008), we employ a $M/M/1$ queuing model to fix ideas on the relationship between average waiting time, network traffic and capacity.[9]

[9]The $M/M/1$ queuing model assumes that (*i*) service requests arrive according to a Poisson process (i.e. arrivals happen continuously and independently of one another), (*ii*) service time is exponentially distributed (i.e. request coming from a Poisson process are handled at a constant average rate) and (*iii*) that service requests are processed by a single server. 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.

Under a network neutral regime, i.e if all packets are treated equally independent of origin or destination, the $M/M/1$ model predicts that each consumer has an expected average waiting time of

$$w^N = \frac{1}{\mu - \lambda}. \tag{8}$$

Therein μ represents the average rate at which service requests are handled, which is interpreted as the overall *transmission capacity* here; whereas λ denotes the average rate at which consumers' aggregate content requests arrive at the ISP, which is interpreted as *network traffic*:

$$\lambda = \alpha\bar{\theta}L,$$

i.e. network traffic will depend on the share of connected consumers, α, the mass of available content providers, $\bar{\theta}$, and L, which denotes the average traffic of each user per content provider.[10] As explained before, we thereby make two important assumptions concerning the generation of traffic on the consumers side. First, each consumer generates more traffic as the number of content providers increases. Second, traffic generation grows linearly with the number of content providers.

Under a discriminatory regime, content providers are offered the choice between a priority and a best-effort transmission service. In the $M/M/1$ model this translates to introducing an additional queue which handles the request of the content providers in the priority class and which is processed ahead of the queue for the best-effort class. In this vein, the classical results of the $M/M/1$ queuing model represent the average waiting time in the priority class, w_1^D, and the best-effort class, w_2^D:[11]

$$w_1^D = \frac{1}{\mu - \beta\lambda} \tag{9}$$

$$w_2^D = \frac{\mu}{\mu - \lambda} w_1^D \tag{10}$$

It is easy to see, that relation

$$w_1^D < w^N < w_2^D \tag{11}$$

[10] More precisely, $\lambda = \int_{y=0}^{\alpha} \int_{x=0}^{\bar{\theta}} L \, dx \, dy = \alpha\bar{\theta}L$. As will be seen later, the consumer side will always be covered in equilibrium due to its homogeneity, i.e. $\alpha = 1$ and thus gross traffic will be $\lambda = \bar{\theta}L$.

[11] Remember that service providers are equally frequented by consumers and therefore the proportion of providers buying priority access is determined by the waiting time in the two service classes.

is always fulfilled, assuming a fixed capacity μ and $\beta < 1$.[12] This is an important feature of our model, because it shows formally that serving some content providers with priority will (in the short-run) unambiguously lead to a *degradation* of service quality for the remaining content providers in the best-effort class. These content providers are therefore put at a disadvantage twice. First, through the prioritization of foreign traffic and second, through the degradation of own traffic, compared to the network neutral regime.

Moreover, notice that the consumers' average waiting time is independent of the introduction of service classes,[13] because each consumer will visit every available content provider equally. Consumer requests to content providers in the first priority class will be processed within a time of w_1^D, whereas request to content providers in the best-effort class take w_2^D units of time. Consequently, average consumer waiting time is

$$w = \beta w_1^D + (1-\beta) w_2^D = w_1^D \left(\frac{\mu - \beta\lambda}{\mu - \lambda} \right) = w^N, \tag{12}$$

since $w_1^D = \frac{1}{\mu - \beta\lambda}$.

Consumers: Consumers are homogeneous and value basic connectedness to the Internet as well as the presence of content providers. In particular, we assume that connectedness adds a lump-sum utility of h whereas each additional content provider adds a marginal utility of v to the consumers' utility.[14] On the contrary, consumers dislike waiting for content because of network congestion. This is captured through consumers' average waiting time, w. To summarize, consumers' utility is given by

$$U_{IC} = \begin{cases} h + v\bar{\theta} - cw - a & \text{if connected} \\ 0 & \text{if not connected,} \end{cases} \tag{13}$$

where c denotes consumers marginal opportunity costs in time and a the access fee charged by the ISP.

Obviously, the homogeneity of consumers allows the ISP to extract all consumer surplus in equilibrium by setting

$$a^* = h + v\bar{\theta} - cw. \tag{14}$$

[12] For $\beta = 1$, when all content providers are in the first priority class, the model trivially collapses to $w_1^D = w^N$

[13] And independent of β for that matter.

[14] Recall that content providers are atomistic and consumers have no preference for specific content. Thereby we avoid to make any judgment about the value of a specific service innovation to consumers and attribute value to the whole mass of available content instead.

As a convention, we will assume that in this case all consumers connect to the ISP, such that

$$\alpha^* = 1. \tag{15}$$

Notice, that the ISP's revenue from consumer access will not directly depend on the network regime because of (12). However, w depends on the network capacity, μ, which is under the ISP's control, at least in the long-run. Furthermore, content variety, $\bar{\theta}$, has a direct influence, as well as an indirect influence via $w = \frac{1}{\mu - \theta L}$ on the optimal access charge. Thus, under a network neutral regime, the ISP's challenge is to set the optimal capacity μ, which will determine the congestion level and consequently content variety. Under a discriminatory regime, the ISP must additionally select a strategic price, p, at which he sells priority access to content providers.

4 Short-Run Effects of Network Discrimination

4.1 The Benchmark: Network Neutrality

Under network neutrality the platform offers only one transmission service class and no content provider's traffic is prioritized. As mentioned above, content providers are heterogeneous with respect to their sensitivity to network congestion, but all of them rely on advertising revenue. Service providers are arranged on the unit interval in order of ascending congestion sensitivity, θ. Those, providers with θ close to zero offer a service with waiting time insensitive advertisement revenues, whereas those with values of θ close to one are very sensitive to network congestion. Since it will always be optimal for the ISP to set the access price according to (14), all consumers will subscribe and $\alpha = 1$. Therefore, given the mass of active content providers and the ISP's transmission capacity,

$$w^N = \frac{1}{\mu - \bar{\theta}^N L}. \tag{16}$$

Given this congestion level, content providers will enter the market until their sensitivity to congestion is so severe that their business model is not sustainable anymore. Formally, the mass of active content providers in equilibrium is derived by equating (5) with zero (the value of the outside option) while

substituting (8) and (15) to yield:[15]

$$\bar{\theta}^{N^*} = \frac{r\mu}{Lr+1} \tag{18}$$

Recall, that under a network neutral regime, the ISP makes only profits by selling access to consumers. By substituting (8) and (18) into (14), one obtains the optimal access charge and thus, according to (3) short-run ISP profit:

$$a_s^{N^*} = \Pi_s^{N^*} = h + \frac{vr\mu}{rL+1} - \frac{c(rL+1)}{\mu} \tag{19}$$

Intuitively, it can be seen that the ISP's short-run profit increases with network capacity, μ, but decreases with individual traffic per content provider, L, which–of course–positively relates to network congestion and waiting time again.

4.2 Network Discrimination

In a discriminatory network the ISP can alleviate congestion for those content providers who have bought a priority access to users. According to (11), in the short-run, when transmission capacity is fixed, the remaining providers will be served at degraded best-effort access compared to the network neutral regime. We may now distinguish three types of content providers:

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 be foreclosed from market entry and remain inactive.

The content provider indifferent between the first two cases is denoted by $\tilde{\theta}^D$, whereas the content provider indifferent between the last two cases is denoted by $\bar{\theta}^D$. Obviously, it must hold that $0 < \tilde{\theta}^D < \bar{\theta}^D$.

[15]Since we normalized the mass of possible content provider to one, the condition

$$0 \leq r \leq \frac{1}{\mu - L} \tag{17}$$

has to be obeyed in an interior solution, which provides an upper bound on the feasible level of gross advertisement revenue per unit of traffic. In particular, notice that r must be smaller than the disutility of network congestion per unit of traffic of the last content provider located at $\theta = 1$, such that (at least) this provider will never enter.

Contrary to Choi and Kim (2008), in our model the "high cost" service providers are more likely to opt for the priority class. This has two reasons: First, we do not model the competitive aspect of obtaining a larger market share based on the prioritized connection. Second, in our model the incentive to buy first priority is based on the individual business model's innate need for a higher connection quality.

Formally, to determine $\tilde{\theta}^D$ and $\bar{\theta}^D$ from (6), one must compute the fulfilled expectations equilibrium by simultaneously considering the equations

$$\left(r - \tilde{\theta}^D w_2^D\right) L = \left(r - \tilde{\theta}^D w_1^D\right) L - p$$
$$\left(r - \bar{\theta}^D w_1^D\right) L - p = 0$$

as well as (7), (9) and (10) to obtain

$$\tilde{\theta}^{D^*} = \frac{p\mu}{L(Lr+1)(Lr-p)} = \frac{p}{(Lr-p)Lr}\bar{\theta}^{D^*} \tag{20}$$

$$\bar{\theta}^{D^*} = \frac{r\mu}{Lr+1}. \tag{21}$$

First, notice the intuitive result that $\tilde{\theta}^{D^*}$ decreases as the individual traffic volume, L increases, or likewise, as the price for priority transmission service, p, decreases, meaning that more content providers will chose priority access in both circumstances.

Furthermore, comparing equation (21) with (18) directly reveals that

$$\bar{\theta}^{D^*} = \bar{\theta}^{N^*}, \tag{22}$$

which immediately proofs the following result.

Proposition 1 *In the short-run, network discrimination has no effect on innovation. The number of active content providers does not change compared to network neutrality, independent of the price for priority access.*

With reference to the network neutrality debate this result has important implications, because under the present assumptions the claims of both parties are flawed in the short-run: *Network discrimination will neither lead to more, nor to less innovation.* What is even more surprising is the fact that the precise nature of the price for priority access is irrelevant.

Next, we investigate the ISP's incentives to engage in network discrimination. From equation (11) we know that consumers' average waiting time is not affected by network discrimination. Furthermore, equation (22) shows that also consumers' network utility has not changed compared to the network neutral regime. Therefore, in the short-run, the ISP cannot extract

extra rents from consumers and the optimal access charge is the same as under net neutrality:

$$a_s^{D^*} = a_s^{N^*} \tag{23}$$

However, in the discriminatory regime the ISP can additionally extract rents from content providers through the sales of priority access. In the short-run, he will do so by maximizing $\beta\bar{\theta}p^D$, which is achieved through

$$p^{D^*} = Lr\left(1 - \frac{1}{\sqrt{Lr+1}}\right). \tag{24}$$

It is easy to see that $0 < p^{D^*} < Lr$ for all positive values of L and r. Intuitively, this shows that through the sales of priority access, the ISP can extract a fraction of each content provider's gross advertisement revenue $L\,r$. Thus, the ISP will make an extra profit of

$$\Delta\Pi_s^D \equiv \beta\ \bar{\theta}^{D^*} p^{D^*} = \left(\frac{Lr+1-\sqrt{Lr+1}}{Lr}\right)\left(\frac{r\mu}{Lr+1}\right)p^{D^*} > 0 \tag{25}$$

compared to the network neutral regime.

Proposition 2 *The ISP has a short-run incentive to introduce a network discriminatory regime, because this allows him to collect extra profits from selling priority access to content providers. However, consumers pay the same price for network access than under the network neutral regime.*

Proposition 2 highlights that ISPs will indeed engage in network discrimination when capacity is fixed. This result is supported by numerous observations of recent ISP practices, which have in fact triggered the network neutrality debate in the first place. Among the most prominent examples is Comcast, one of the largest U.S. Internet service providers, who was reported to have engaged in discriminatory actions in late 2007 (Businessweek, 2008). However, it must be annotated, that Comcast did not receive any payments from content providers, at least officially, in order to accelerate or slow down certain types of traffic. Nevertheless, the incentives to do so are certainly very real and soon money will be changing hands if network discrimination is not prevented by regulatory bodies.

4.3 Welfare Analysis

In order to provide useful insights for the current policy debate, we conclude the short-run analysis of network discrimination by investigating the welfare effects. In the present context, total short-run welfare, W_s, is constituted by

three different elements: (i) Internet consumers' surplus, ICS_s, (ii) content providers' surplus, CPS_s, and (iii) the ISP's profit, Π_s:

$$W_s = ICS_s + CPS_s + \Pi_s \tag{26}$$

With respect to Internet consumers' surplus, it has been shown that $ICS_s = 0$ in both regimes, because consumers are homogeneous and thus the ISP can always extract their surplus fully. Regarding the ISP's profit, we established that $\Delta\Pi_s^D > 0$. Therefore, it remains to examine the effect of discrimination on content providers' surplus.

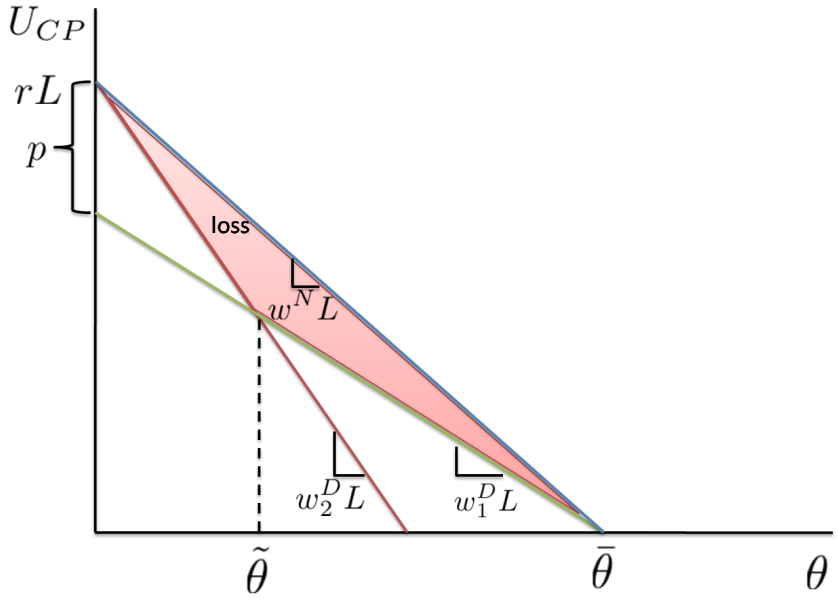

Figure 1: The Effect of Network Discrimination on Content Providers' Surplus

Consider Figure 1. First, notice that those content providers located at $\theta \in [0, \tilde{\theta}^{D^*})$ are evidently worse off under a discriminatory regime, because for them network congestion has increased from w^N to w_2^D. Second, the content providers' welfare loss increases with congestion sensitivity on the interval $\theta \in [0, \tilde{\theta}^{D^*})$. The business model of the provider located at $\theta = 0$ is not affected at all through congestion, while the provider at $\theta = \tilde{\theta}^{D^*}$ is already suffering so much, that he is indifferent between staying in the best-effort class and buying priority access. Third, by the converse argument, notice that the welfare loss is decreasing for those content providers in the priority class as $\theta \in [\tilde{\theta}^{D^*}, \bar{\theta}^{D^*})$ increases. To see this, recall from Proposition 1 that the last content provider to enter the market, $\bar{\theta}$, is identical under both regimes and receives a surplus of zero in both cases. For him, the benefit through reduced congestion (compared to the network neutral regime) is just offset by the price for priority access. Consequently, for all content providers

with less congestion sensitivity, i.e. $\theta \in [\tilde{\theta}^{D^*}, \bar{\theta}^{D^*})$, the price is higher than the benefit of being in the first priority class. Nevertheless, by definition of $\tilde{\theta}^{D^*}$, for these providers the welfare loss is yet less severe in the first priority class than in the best-effort class. In this line of argumentation, it is also obvious that content provider $\tilde{\theta}^{D^*}$ is worst off among all content providers, because he incurs the greatest welfare loss. In summary, we can conclude that in the short-run all active content providers are worse off under a discriminatory network regime.

Note that the charge for priority access is merely a welfare shift from the content providers to the ISP, so that the sign of the net welfare effect, ΔW_s, will only depend on the difference between the gross gain through less congestion of those content providers in the priority class and the gross loss through increased congestion of those providers remaining in the best-effort class.

$$\Delta W_s = \Delta \Pi_s^D + \int_{\theta=0}^{\bar{\theta}^{D^*}} U_{CP}^D(\theta)\ d\theta - \int_{\theta=0}^{\bar{\theta}^{N^*}} U_{CP}^N(\theta)\ d\theta \tag{27}$$

$$= \underbrace{\beta \bar{\theta} p^D}_{\text{priority charge}} - \underbrace{L(w_2^D - w^N) \int_{\theta=0}^{\tilde{\theta}} \theta\ d\theta}_{\text{congestion aggravation to best-effort class}} +$$

$$\underbrace{L(w^N - w_1^D) \int_{\theta=\tilde{\theta}}^{\bar{\theta}} \theta\ d\theta}_{\text{congestion alleviation to priority class}} - \underbrace{\int_{\theta=\tilde{\theta}}^{\bar{\theta}} p^D\ d\theta}_{\text{priority charge}} \tag{28}$$

$$= L(w^N - w_1^D) \int_{\theta=\tilde{\theta}}^{\bar{\theta}} \theta\ d\theta - L(w_2^D - w^N) \int_{\theta=0}^{\tilde{\theta}} \theta\ d\theta \tag{29}$$

$$= \frac{L}{2} \left[(w^N - w_1^D)(\bar{\theta}^2 - \tilde{\theta}^2) - (w_2^D - w^N)\tilde{\theta}^2 \right] \tag{30}$$

Equation (30) reveals nicely that the overall effect of network discrimination on welfare indeed depends on the relative size of the *congestion aggravation effect* to providers in the best-effort class versus the *congestion alleviation effect* to providers in the priority class. Obviously, the two effects relate directly to the main argument of proponents and opponents of net neutrality, respectively. This is exemplified in Figure 2.

Proposition 3 *In the short-run, network discrimination unambiguously increases welfare with respect to the network neutral regime, 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 discriminatory regime because the increased surplus is expropriated by the ISP.*

Proof:

$$\Delta W_s > 0 \;\Leftrightarrow\; \frac{w^N - w_1^D}{w_2^D - w_N} > \frac{\tilde{\theta}^2}{\bar{\theta}^2 - \tilde{\theta}^2} \;\Leftrightarrow$$
$$\frac{1-\beta}{\beta} > \frac{(1-\beta)^2}{1-(1-\beta)^2} \;\Leftrightarrow\; 0 < \beta < 1 \qquad \blacksquare$$

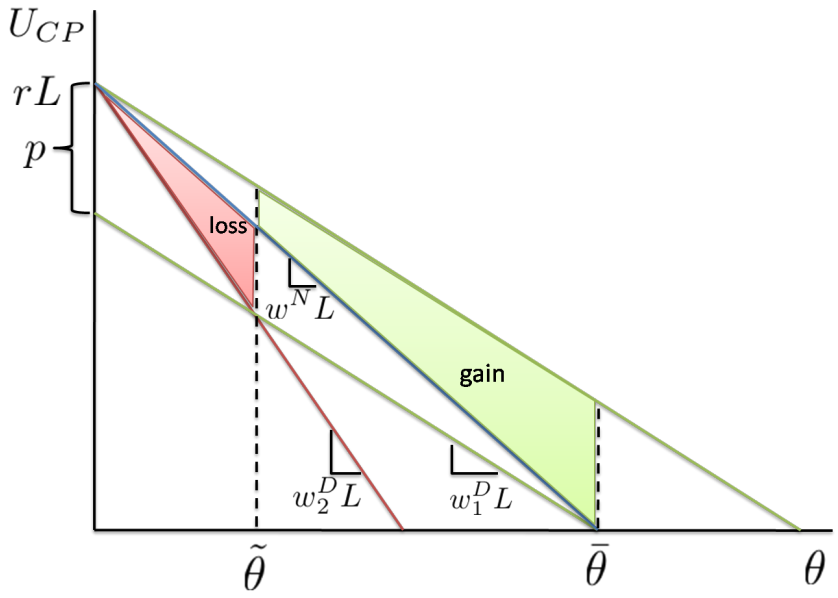

Figure 2: Congestion Alleviation vs. Congestion Aggravation Effect of Network Discrimination

Network discrimination will therefore generally be welfare improving in the short-run, because it "allocates" congestion more efficiently. Those content providers who are relatively inelastic to congestion with respect to their advertisement revenues are allocated more congestion than the providers with relatively congestion elastic revenues.

5 Long-Run Effects and Infrastructure Investments

In this section we extend our model to allow for long-run investments in network transmission capacity. Much of the neutrality debate is rooted in the ISPs' concerns about infrastructure investments. On the one hand, ISPs would like to accommodate for new innovative content because this is valued by customers. But, on the other hand, they do not want to let the content providers free-ride on their infrastructure investments. Therefore, network discrimination seems to be a plausible way out of this dilemma.

5.1 Transmission Capacity and Congestion

We now focus our analysis on the average service rate, μ, which is interpreted as a proxy for transmission capacity. An increase of μ allows the ISP to handle more service requests to content providers simultaneously. Since all available customers are already connected in equilibrium, capacity expansion cannot result in an increase of gross access revenue. Therefore, we can consider the effects of capacity expansion on the content provider side in isolation. In particular, we can distinguish three effects:

- The *expansion effect* denotes that an increase of capacity reduces the overall congestion level and thereby allows for the emergence of new innovative content providers.

- The *variety effect*, on the contrary, indicates the negative feedback loop of the expansion effect on congestion. The entry of new content providers generates higher network traffic, since customers visit all content providers equally, which in turn increases congestion.

- Likewise, an increase in network traffic can also be achieved via a *niveau effect* by which customers visit all content provider more frequently. In our model, this can be reflected by an exogenous increase of L. At first, it seems that content providers will unambiguously profit from an increase in customer visits, because this increases their respective gross advertisement revenues, rL. However, by equations (8) and (18) there is also a negative effect through

  $$\frac{\partial w^N}{\partial L} > 0 \qquad \text{and} \qquad \frac{\partial \bar{\theta}}{\partial L} < 0,$$

  which indicates that an increase in the traffic niveau will ceteris paribus increase the congestion level and drive the most congestion sensitive content providers out of the market.

In either regime the overall direction of these effects depends crucially on the ISP's optimal choice of μ. Formally, we model the investment decision as a discrete decision stage which precedes the previous analysis. Thus, the ISP chooses first its transmission capacity, μ, and then, in a second stage, the customer access charge, a, and if applicable, the priority price p.[16]

5.2 Network Neutrality

Recall from the short-run analysis that under a network neutral regime, the network operator makes profits from selling consumer access only. He captures the full consumer surplus, but does not tap content providers' surplus.

[16] Accordingly, the solution concept is that of subgame perfectness.

This does not change in the long-run. The ISP may therefore only increase his revenue through the expansion effect and reaping consumers' additional externality for content variety. Given the optimal access charge from (14), the ISP maximizes Π_l^N by choosing

$$\mu^{N^*}\left(a^{D^*}\right) = (rL+1)\sqrt{\frac{c}{k(rL+1)-vr}} \tag{31}$$

Intuitively, the ISP chooses a higher transmission capacity if consumer' opportunity costs in time, c, are high. Waiting costs directly reduce the possible access charge which the platform can extract from the users. The same intuition holds for the value of variety, v. In fact, if v is very large, the ISP has an excessive incentive to invest into capacity.[17] We therefore consider only the interesting case where the marginal costs of capacity expansion, k, are sufficiently high with respect to v, i.e.

$$k > \frac{vr}{(rL+1)}.$$

5.3 Network Discrimination

In a network discriminatory regime, the entry of additional content providers not only increases the ISP's revenues from access, but also revenues from priority sales. The necessary infrastructure upgrades caused by increased traffic through the entry of additional content providers can therefore be easier compensated. Given the equilibrium values of a and p from the short-run analysis, the ISP maximizes Π_l^N through

$$\mu^{D^*}\left(a^{D^*}, p^{D^*}\right) = (rL+1)\sqrt{\frac{c}{[k(rL+1)-vr] + r\left[2\sqrt{Lr+1} - (Lr+2)\right]}} \tag{32}$$

Notice that the second term in squared brackets is always negative.[18] Thus, compared to the network neutral regime, the ISP will invest more into infrastructure and provide more transmission capacity. Moreover, the capacity expansion will also increase the ISP's profit. To see this, recall from equation (25) that the ISP's short-run extra profit is $\Delta\Pi_s^D = \beta\,\bar{\theta}^D p^D$, where p^D and β are independent of μ. Only $\bar{\theta}$ is positively affected by μ, which means that,

[17] This is mainly due to the assumption that the ISP incurs linear costs of capacity expansion. Under a convex cost function, infinite capacity expansion can never be optimal.

[18] The stability condition in a discriminatory network therefore is

$$k > r\left(1 + \frac{v+1}{Lr+1} - \frac{2}{\sqrt{Lr+1}}\right)$$

in the long-run, the extra profits from network discrimination can be increased by encouraging new, congestion sensitive, content providers to enter the market. Of course, the increased revenue following a capacity expansion must at least compensate the ISP for the costs of additional infrastructure investments. More specifically, after reformulating (25), notice that the extra profits form selling priority access grow linearly in the transmission capacity:

$$\Delta\Pi_l^D \equiv \mu\, r^2\, \frac{\left(L+1-\sqrt{Lr+1}\right)\left(1-\sqrt{\frac{1}{Lr+1}}\right)}{Lr+1} \tag{33}$$

Proposition 4 *Under a discriminatory regime the ISP will invest more in network infrastructure and provide higher transmission capacity in the long-run. Thereby, ISP profits, the consumer access charge and content variety are higher than under a network neutral regime.*

5.4 Welfare and Price Regulation

From the short-run welfare analysis, it is obvious, that the general level of network congestion is crucial in driving the welfare effects.

Proposition 5 *Under the network discriminatory regime, the overall congestion level is reduced compared to network neutrality. The already positive short-run welfare effects of network discrimination are therefore even increased in the long-run.*

Proof: It remains to verify that $\frac{\partial w}{\partial \mu} = \frac{\partial \frac{1}{\mu-\lambda}}{\partial \mu} < 0$: First, see that $\lambda = \bar{\theta} L = \frac{Lr\mu}{Lr+1}$, so that $\frac{\partial \lambda}{\partial \mu} = \frac{Lr}{Lr+1} < 1$. Therefore, it holds that $\frac{\partial \mu - \lambda}{\partial \mu} > 0$ and consequently, $\frac{\partial \frac{1}{\mu-\lambda}}{\partial \mu} < 0$ ■

However, the downside of network discrimination is that most of the welfare gain is appropriated solely by the ISP. Once a discriminatory network regime is established the question will therefore arise if price regulation could enable content providers to retain more of their surplus. Assume a regulatory agency would lower the price for prioritized traffic under the level which the monopolistic platform owner has chosen according to (24). Certainly, this would result in more content providers subscribing to the priority service, which in turn increases the average waiting time for consumers using priority services. It is easy to verify that price regulation would indeed shift some of the surplus back from the ISP to the content providers. Unfortunately, when applying price regulation, the regulator also destroys welfare, because the ISP's revenues from priority access are in fact welfare neutral transfers. In the most extreme case, price regulation could impose a price of $p^D = 0$, which would in fact reestablish the network neutral regime. By Proposition 3 and 5 this cannot be welfare optimal.

Proposition 6 *In a discriminatory regime, price regulation can shift some surplus back to content providers but will inevitably lead to overall welfare reductions.*

6 Conclusion

We contribute to the net neutrality debate with a formal framework that incorporates Internet consumers, content providers and an Internet service provider. Our model complements previous work in this strand of the literature because we explicitly consider the adverse effects of traffic prioritization to the remaining best-effort class in a two-sided market model. Thereby, our focus is to study the effect of network discrimination on innovation of content providers and the ISP's incentives for transmission infrastructure investment.

Generally, we find that network discrimination has positive effects on welfare, because congestion is better allocated among content providers with different sensitivity to network congestion. Those content providers with congestion inelastic advertisement revenues remain with the free best-effort transmission service, in which they experience more congestion than under network neutrality. Conversely, those content providers with relatively congestion elastic ad revenues are alleviated from congestion and thereby generate overall welfare gains. In the long-run, network discrimination leads to infrastructure investments in transmission capacity and encourages innovation. In the short-run, however, innovation is not expected because the ISP expropriates the increased content provider surplus through the price for the priority transmission service. This is the downside of network discrimination. Although total welfare is increased, content providers will–at least in the short-run–be worse off than under network neutrality.

Although price regulation can shift some of the congestion alleviation gains back to content providers, it is inapt as a policy instrument, because welfare is proportionally destroyed in the process. Therefore, if regulatory correction is desired, it should address the expansion of transmission capacity directly, e.g. by subsidizing broadband roll-out.

To provide a benchmark for future research, we made the fundamental assumption that content providers are not in competition with each other and that they are equally visited by customers. We believe that our results are fairly robust to small variations of this assumption. For example, one could incorporate competition by assuming that content providers' gross advertisement revenue depends on the relative congestion level also. In this case one might presume that the relative effects of network discrimination are merely amplified.

References

Berners-Lee, T. (2006, June). Net neutrality: This is serious. `http://dig.csail.mit.edu/breadcrumbs/node/144`. Last visit 13.06.2009.

Businessweek (2008, Feb). The fcc, comcast, and net neutrality. `http://www.businessweek.com/technology/content/feb2008/tc20080225_498413.htm`. Last visit 13.06.2009.

Choi, J. P. and B.-C. Kim (2008). Net neutrality and investment incentives. mimeo No.2390, CESIFO.

Crawford, S. (2007). The internet and the project of communications law. mimeo, University of Michigan.

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

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

Economides, N. and J. Tag (2008). Net neutrality on the internet: A two-sided market analysis. mimeo, University of New York, Stern Business School.

Faulhaber, G. and B. Rasmussen (2006). Network neutrality and the economics of congestion. In *Conference Telecommunications Policy Research*, pp. 1847–1908.

Google (2009). Net neutrality. `http://www.google.com/help/netneutrality.html`. Last visit 11.06.2009.

Hahn, R., R. Litan, and H. Singer (2007, Juni). The economics of wireless net neutrality. *Journal of Competition Law and Economics 3*, 399–451.

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

Wu, T. (2005). Network neutrality, broadband discrimination. *Change 925*, 77–90.

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