Net neutrality and innovation at the core and at the edge*

Carlo Reggiani† Tommaso Valletti‡

March 2012

Abstract

We model an Internet broadband provider that can offer a different quality of service (priority) to content providers. Net neutrality regulation does not allow prioritization and all content is treated equally. Content providers derive their profits from advertising rates which differ with or without neutrality. We focus on the incentives to innovate in content by both large and small content providers, as well as on investment in core infrastructure to reduce congestion. Prioritization increases infrastructure investment as compared to regulation, except when the large content provider is considerably more inefficient than the small fringe providers. Prioritization is also desirable from a welfare perspective unless fringe content is particularly valuable to users. The results are reinforced if advertising rates for prioritized content are more sensitive to congestion than the rates for best-effort content.

JEL code: D4, L12, L43, L51, L52

Keywords: Internet, net neutrality, congestion, innovation.

*We thank Bruno Jullien, Fabio Manenti Joacim Tag, Joao Vareda, and seminar participants at Bern, Bologna, Manchester, Northwestern University, Rhodes (CRESSE), Stockholm (EARIE), Telecom ParisTech, and Valencia (JEI). Tommaso Valletti acknowledges funding from the Orange "Innovation and Regulation" Chair at Telecom ParisTech/Ecole Polytechnique.

†School of Social Sciences, University of Manchester, Manchester M13 9PL, UK. E-mail: carlo.reggiani@manchester.ac.uk.

‡Imperial College London, University of Rome and CEPR. Address: Imperial College Business School, South Kensington campus, London SW7 2AZ, UK. E-mail: t.valletti@imperial.ac.uk.

1 Introduction

The Internet has probably been the fastest developing industry of the last two decades. From the early development as an experimental network linking a limited number of computers, the Internet has now become one of the key priorities for policy makers around the world as it is seen as an engine to economic growth (Czernich et al., 2011; Mayo and Wallsten, 2011). The Internet is delivered by broadband providers who can use their infrastructure to set particular terms for access of Internet applications and content (e.g., websites, services, protocols). These access terms are discussed under the heading of "net neutrality" (henceforth, NN), generating one of the most hotly debated issues in communications policy in the U.S. and elsewhere.[1]

NN has often been linked to the "end to end" principle,[2] which is thought to have guaranteed openness and free access to the Internet; its operation, however, has been questioned by the establishment of broadband as the standard delivering technology. From an economic viewpoint, the issue is that broadband allows for web traffic management techniques that can potentially be used for quality discrimination of data packets, use of termination charges for network traffic, and several other practices that raise competitive concerns. From this angle, then, NN is mainly a data treatment (and its pricing) issue. On the one side stand proposers of a regulation that bans discrimination of data packets and guarantees open and equal access to the net (or "openists", according to Wu, 2004); on the other side it is believed that the Internet needs no regulation and will develop better by letting the market forces operate freely (or "deregulationists").

Valid arguments have been proposed by both sides. One of the main stances of "openists" is that NN is needed to protect the innovation of small start up content providers (CPs), where among those there may be tomorrow's giants like Google, Facebook or YouTube. Innovation at the "edge" of the network is one of the defining features of the Internet and discrimination constitutes a potential harm to it

[1] Recent developments include some mild forms of NN adopted by the U.S. FCC in November 2011, but already challenged in court by Verizon. The European Commission in 2011 issued a Communication that *de facto* declined to impose explicitly NN rules, adopting a wait-and-see approach. Some countries have begun to take more proactive positions: Chile (2010) and the Netherlands (2011) were the first two countries to adopt legislation establishing ex ante rules prohibiting NN violations.

[2] This was first expounded in Saltzer et al. (1984), and emerged as a design tool for use by network engineers. The initial principle was that the transmission and routing of Internet traffic should be "dumb", not interfering with information packets sent between sender and receiver.

(Lessig, 2001; Lee and Wu, 2009). On the other hand, the main counter argument of "deregulationists" is based on the need of Internet service providers (ISPs) to get an appropriate remuneration for the use of the infrastructure, which is seen as the best way to guarantee investment for maintenance and expansion of the capacity of the network (the "core" of the Internet), a prominent concern due to the increasing diffusion of bandwidth-intensive applications (Yoo, 2005; Van Schewick, 2006; Becker et al., 2010). Furthermore, NN can have a crowding out effect on CPs' innovation, hindering the development of new applications sensitive to delays and latency. The tension between these opposing views is fundamental to the debate and the model presented in this paper allows us to evaluate the arguments of both sides.

We develop a model where the funding to content providers comes from advertising revenues. These resources can be affected both by the priority regime and by network congestion. We show that, ultimately, the welfare properties of a discriminatory regime based on traffic prioritization, when contrasted to NN, depend on the ability to direct these resources to those content providers that can generate the highest number of applications.

The rest of the paper is structured as follows. Section 2 briefly reviews the relevant literature to locate the contribution of the paper. Section 3 introduces the basic model. Section 4 presents the results of the analysis. Section 5 extends the model to allow advertising rates to be affected by congestion. Section 6 concludes.

2 Related literature and contribution

NN has triggered a fierce debate and much has been written about it.[3] The vast majority, however, are policy and advocacy papers raising qualitative arguments. The economics literature is still relatively scarce but there are exceptions: early attempts at formalizing some aspects of the debate can be found in Hogendorn (2008), Kocsis and De Bijl (2008), and Musacchio et al. (2009).[4] Economides and Tag (2012) present a static model of charges imposed by the ISP to content providers for traffic termination to consumers. NN is captured by assuming that CPs are not charged for

[3] In March 2012, a casual SSRN search returned 267 papers with «net neutrality» in the title or abstracts. A similar Google search provided over 10m hits. See also Brennan (2011).

[4] Hermalin and Katz (2007) model NN as a restriction on the product line that an ISP can offer. Their results suggest that these restrictions are likely to reduce welfare. They do not explicitly consider Internet traffic.

termination and, in their results, this typically increases CPs' welfare. Congestion, or incentives for ISP's investment and CPs' innovation, however, are not addressed.

The structure of the industry naturally invokes a two-sided market approach: ISPs are the platforms that connect CPs to final users. Although the literature on two-sided networks has flourished in recent years (Armstrong, 2006; Rochet and Tirole, 2006), the issue of quality investment by platforms has been less explored.[5]

The contributions closest to ours are those that have modelled the key problem of traffic congestion and bandwidth allocation on the Internet.

Njoroge et al. (2010) consider vertically-differentiated duopolistic ISPs and assume that the quality of off-network exchanges is determined by the worst between the ISPs qualities.[6] NN is captured as a zero fee being imposed to CPs for off-net traffic. The investment strategy of the ISP is driven by the tension between the competitive effect, that can be reduced via quality differentiation, and the rent extraction from CPs. The first dominates under NN, while the second is more pronounced under priority pricing.

Economides and Hermalin (2012) assume that the "pipe" of a monopolist ISP has fixed capacity in the short run. Bandwidth is allocated in different proportions to CPs according to the pricing regime: equally under NN, with priority if discrimination is allowed. This feature and the elastic demand from final users are crucial for their results: uneven allocation of bandwidth under discrimination leads to a more than proportional increase of demand and increased traffic; discrimination can then lead to higher welfare only if it has an expansionary effect on CPs' supply.

Cheng et al. (2011), Choi and Kim (2010) and Kramer and Wiewiorra (2012) use, as we do, the M/M/1 approach: borrowed from queuing theory, it is considered a good proxy for actual congestion on the Internet.[7] Cheng et al. (2011) and Choi and Kim (2010) consider similar models in which users access exclusively only one of two content providers;[8] total supply of content is fixed so priority only affects the

[5] Fahri and Hagiu (2008) focus on investment decisions to deter or accommodate entry of a competing platform.

[6] Valletti and Cambini (2005) use a similar approach to model the quality of voice calls between competing telecommunications networks.

[7] McDysan (1999), cited by Kramer and Wiewiorra (2012).

[8] While this might be a characterization of particular situations where content providers are substitutes between each other (e.g., a subscriber will typically want to use only one search engine, and will decide, for instance, between either Google or Bing), it cannot capture the fact that most of the Internet content has a different nature, that is, subscribers want to see (and do see) both Google and

market shares. In Cheng et al. (2011) both CPs can get priority. This leads to a prisoners' dilemma: the individual incentive leads similarly efficient CPs to buy priority; the result is no effect on congestion and only more surplus extracted by the ISP. Choi and Kim (2010) consider the case in which CPs bargain with the ISP to obtain exclusive priority for their traffic; CPs are charged a fee only if they opt for priority. If the ISP has all the bargaining power, it is able to extract most of the surplus from both CPs. The impact of NN on investment, instead, crucially depends on the fact that CPs' content supply is inelastic: as more capacity means less value for priority, the ISP has less incentives to invest when NN is abandoned.

Kramer and Wiewiorra (2012) consider a continuum of CPs differently sensitive to congestion. Although not all CPs are served, NN has no effect on content supply in the short run: this implies that priority pricing is welfare enhancing as it leads to a better allocation of bandwidth. In the long run, the welfare superior regime is the one leading to higher investment; as NN reduces entry of new CPs, it prevails only if advertising revenues considerably increase with fewer CPs.

Our model shares with the last works cited the more accurate way to model congestion due to Internet traffic. However, we differ from each in several respects. First, unlike Cheng at al. (2011) and Choi and Kim (2010) but consistent with one of the defining features of the Internet, users are allowed to browse any content they wish once they connect to the net. Moreover, the market for content is not fully covered, thus we consider an elastic supply of CPs. In particular, one characteristic of the Internet is that CPs are very heterogeneous: a few CPs (e.g., Google, Facebook, YouTube) supply many applications and generate a lot of clicks and traffic, while there are many CPs that generate individually, but possibly not in aggregate, only a little traffic. Unlike the rest of the literature our model captures this feature by having a single large CP and a fringe made of many atomistic CPs: this assumption seems crucial to encapsulate the "innovation at the edge" argument that characterizes some policy debate. Furthermore, we assume that advertising revenues per click are the same under NN when all traffic is treated equally; instead advertising rates differ for prioritized and unprioritized traffic when NN is abandoned. Finally, we also consider that advertising revenues might be endogenous to the level of congestion; this is important as the (costly) intelligence that can be installed at the broadband network level can be used to make advertising itself more effective. In our model, we

YouTube, which cannot be modelled as mutually exclusive choices.

therefore study decisions both at the infrastructure "core" and at the Internet "edge", by looking at how the ISP invests in capacity and charges for it, in the anticipation of how many applications will be developed by CPs and funded by advertising revenues.

3 The model

Our model consists of a monopoly platform (ISP) that connects users with the content providers (CPs). This connection allows CPs to contact all available users, whose total mass in normalized to one, and derive advertising revenues from them. The advertising revenue per user contacted is denoted by a.

We introduce two sources of heterogeneity. First, there are two types of CPs: a continuum of "small" CPs that we call "fringe" and denote with the subscript F, and one "large" CP like Google, Facebook or YouTube, that we name firm G.[9] In the fringe, each CP supplies one unique application/content, while firm G can introduce several applications. Each CP has to pay a development cost for every application it introduces. These costs are also heterogeneous. In particular, firms in the fringe are distributed along a (unbounded) line, with the ISP located at zero. A CP located at x has to pay a linear transportation cost in order to supply its application, $t_F(x) = t_F x$. The profit of a firm in the fringe that gets advertising revenues a from a total unit mass of users is

$$\pi_F = a \cdot 1 - t_F x. \tag{1}$$

A free entry condition determines how many CPs enter into the fringe, namely a mass

$$x_F = a/t_F. \tag{2}$$

The total profits of the fringe are thus

$$\Pi_F = \int_0^{x_F} \pi_F \mathrm{d}x = \frac{a^2}{2t_F}.$$

Firm G also pays an entry cost t_G per application, but we assume that it can control many applications it eventually introduces along the line. That is, firm G maximizes w.r.t. x the total profit

$$\pi_G = a \cdot x \cdot 1 - t_G \int_0^x z \mathrm{d}z. \tag{3}$$

[9] Considering one large CP is meant to capture that particular applications generate a large part of the internet traffic (Sandvine, 2011).

Hence the mass of applications introduced by firm G will be

$$x_G = a/t_G, \tag{4}$$

with the corresponding profit

$$\pi_G = \frac{a^2}{2t_G}.^{10} \tag{5}$$

Note that we allow unit transportation costs t_i, $i = F, G$, to be different, in case firm G has application development costs different from the fringe. This distinction is introduced to discuss the extent to which a specific regime of neutrality can affect the incentives to develop content of more or less efficient providers.

3.1 Congestion

The unit mass of consumers connects to the entire content available over the Internet. Consumers pay a subscription fee p to the ISP. Consumers benefit from variety, which we model by assuming that each consumer enjoys a benefit v_F per available fringe application and v_G per firm G application. Consumers also care about congestion on the network.

Congestion depends on total traffic exchanged, on the capacity μ of the ISP, as well as on the traffic management techniques. We borrow from the extant literature the way congestion is affected by prioritization rules (Cheng et al., 2011; Choi and Kim, 2010; Kramer and Wiewiorra, 2012). Each user-CP exchange generates an amount of traffic λ. Under Net Neutrality (NN), congestion is

$$W(x_G, x_F) = \frac{1}{\mu - \lambda(x_G + x_F)}, \tag{6}$$

which is the waiting time W in a M/M/1 queuing system; the corresponding utility of the users is

$$U_{NN} = v_F x_F + v_G x_G - sW(x_G, x_F) - p, \tag{7}$$

where s is consumers' sensitivity to congestion.

With Priority Pricing (PP), the ISP can offer priority to traffic. If x_H and x_L are the masses of CPs that choose, respectively, to prioritize or not to prioritize their traffic, the users' utility is

$$U_{PP} = v_H x_H + v_L x_L - s\overline{W}(x_H, x_L) - p,$$

[10] An alternative interpretation, generating the same formalization, is that firm G has a single "large" application, whose size x_G is determined according to (3), leading to (4).

where v_H and v_L depend on the share of fringe providers opting for high and low priority. The congestion $\overline{W}(x_H, x_L)$ is given by the weighted average of waiting times. More specifically, waiting times of each type of traffic are

$$W_H = \frac{1}{\mu - \lambda x_H}, \qquad W_L = \frac{\mu}{\mu - \lambda x_H} \frac{1}{\mu - \lambda x_H - \lambda x_L} > W_H,$$

so that the average waiting time is

$$\overline{W}(x_H, x_L) = \frac{x_H}{x_H + x_L} W_H + \frac{x_L}{x_H + x_L} W_L. \tag{8}$$

There are two main properties of this way of modelling traffic. First, a M/M/1 system implies that the *average* congestion is the same in the two regimes, provided the capacity level and the total amount of traffic exchanged are also the same.[11] This is an important property that we must stress: PP, *per se*, does not lead to an efficiency improvement over NN, but just to a reallocation of capacity resources. However, the two regimes will give different incentives to invest in μ, and therefore will affect average congestion for an endogenous choice of μ. The second property of the queueing system is that, if some capacity is allocated to prioritized traffic, this must imply that, ceteris paribus, the non-prioritized traffic will experience a higher delay. Indeed this is a feature that is emphasized in the debate over net neutrality and that the model effectively captures.

3.2 Advertising

Differences in congestion and priority also affect the profitability of advertising rates. This is the mechanism that gives incentives to CPs to eventually opt for priority. With NN, the advertising rate is a for all the CPs, reflecting the fact that all applications are reachable with the same delay by end users. Without NN, there will be differences between the rates a_H and a_L for the prioritized traffic and for the best-effort content. For instance, targeted advertising is enhanced by prioritization techniques and deep packet inspections. For the initial analysis we do not need to put additional structure on these advertising functions, and simply posit that $a_L < a_H$, as traffic with priority suffers less from congestion problems.[12] The gap between the advertising rates created

[11] This can be checked immediately by comparing (6) and (8). When capacity is the same, it is $\overline{W} = W$ when $x_G + x_F = x_H + x_L$.

[12] Behavioral targeting is a technique used by advertisers to increase the effectiveness of their campaigns. It uses information collected on an individual's web-browsing behavior, such as the pages

by priority plays a crucial role in our model: as, contrary to Choi and Kim (2010) and Cheng et al. (2011), we do not impose single-homing of end users with respect to content, CPs would get no advantage from priority if that left unaffected their advertising revenue. In Choi and Kim (2010) priority is profitable as it allows to attract a higher share of (captive) users. Our case is complementary to theirs: as users can surf all the content available on the internet, CPs are not competing to attract them; but this implies that CPs opt for priority only if it allows to extract more advertising revenue per unit of content provided.[13]

At times we will also invoke the following property:

$$a = \gamma a_H + (1 - \gamma) a_L, \tag{9}$$

that is, the weighted average advertising rate does not change with and without NN. This property mirrors the previous result concerning the physical infrastructure whereby the average waiting time, when capacity and traffic are the same, does not change with the neutrality regime; similarly, we now imagine that the neutrality regime, as such, does not alter the average resources (from advertisers) that can be attracted by this economy, but it leads to a redistribution of these resources. Notice that a high (respectively, low) value of γ implies that a NN regime generates advertising resources that are closer to prioritized rates (respectively, best-effort rates): the lower the γ the greater the discrepancy between NN ad rates and priority ad rates. The weighting γ, in principle, can take any value between zero and one: in the following analysis, however, specific values of γ that may be reasonable in our context will be discussed. For instance, if both types of CPs are equally efficient and only one type opts for priority, then it seems natural to posit $\gamma = \frac{1}{2}$, so that (9) reduces to a simple average, $a = \frac{a_H + a_L}{2}$.

We will consider two regimes. With NN, all CPs access for free a best-effort Internet lane which treats everyone equally, and get a. With PP, CPs will have the choice of still paying nothing for best-effort and earning a_L; or paying a premium f_H for priority and getting advertising rates a_H from advertisers. In either regime, we

they have visited or the searches they have made, to select which ads to display to that individual. As properly targeted ads will fetch more consumer interest, the ad rates should command a premium over random advertising. Further discussion of the sensitivity of advertising rates to congestion is provided in Section 5.

[13] Kramer and Wiewiorra (2012) also share with us the assumption that end users can see all available content. In their paper, the mechanism that eventually gives incentives to prioritize content comes from the CPs' sensitivity to congestion.

consider a game where the monopolist chooses μ, and sets prices to CPs and to end users. We compare the long-run welfare properties of the two regimes in terms of impact on CPs, users, and ISP.

4 Analysis

The ISP can invest $I(\mu)$ to expand the capacity μ of the network and reduce the disutility linked to congestion and waiting times of data packets. For simplicity, we shall assume throughout the paper that $I(\mu) = \mu$; in other words, investment displays constant returns to scale with respect to capacity. Note, however, that we still have decreasing returns to scale of investment with respect to the average waiting time: this is due to the fact that, independently of the priority regime, the average waiting time decreases at a decreasing rate when capacity is expanded.

Under net neutrality, the profits of the ISP are obtained only from end-users:

$$\Pi_{ISP}^{NN} = \pi_{ISP}^{NN} - I(\mu) = p^{NN} - \mu,$$

while under no regulation a fee can be asked to those CPs who choose priority:

$$\Pi_{ISP}^{PP} = \pi_{ISP}^{PP} - I(\mu) = p^{PP} + f_H D_H - \mu,$$

where D_H denotes the demand for the high priority lane. If best effort is chosen in equilibrium only by the fringe while firm G opts for priority, it will be $D_H = 1$.

Since, under a priority regime, the individual profit of a generic CP in the fringe is either $a_L - t_F x$ or $a_H - t_F x - f_H$, all fringe providers opt for the best effort/low priority connection if:

$$f_H \geq a_H - a_L. \tag{10}$$

Provider G can opt for the high priority. The high priority is an option in case $\pi_G^H \geq \pi_G$, where the left hand side is the profit of G with priority, while the right hand side is its profit without priority. Notice that firm G is "pivotal", in that, if it does not choose priority, no one else will, and the best-effort regime will re-emerge, with advertising rates a. Instead, no one in the fringe is pivotal when firm G chooses priority, and this is why each fringe member compares a_H and a_L, as described by (10).[14]

[14] There cannot be an equilibrium where all the fringe firms opt for priority while firm G does not.

In order to induce G to choose the priority lane, the ISP will set the charge for priority such that it holds exactly $\pi_G^H = \pi_G$. Since the profit of firm G is given by (5), after substitution, the condition implies:

$$f_H = \frac{a_H^2 - a^2}{2t_G}. \tag{11}$$

In other words, the priority fee extracts all the extra rent from firm G. Condition (10) to ensure self-selection of the fringe to low priority then becomes:

$$t_G \leq \frac{a_H^2 - a^2}{2(a_H - a_L)}, \tag{12}$$

that we assume to hold, as otherwise a regime with prioritization will never emerge in equilibrium. This condition says that firm G should be "efficient" enough (low t_G), so that any redistribution of advertising resources towards prioritized traffic will induce the ISP to increase the corresponding premium fee more than proportionally, which ensures that the fringe will find it too costly to opt for priority.[15]

Under condition (12), firm G opts for priority while the fringe sticks to the unprioritized alternative. The ISP profit is then:

$$\Pi_{ISP}^{PP} = \pi_{ISP}^{PP} - I(\mu) = p^{PP} + f_H - \mu,$$

where (11) holds.

Finally, the ISP sets p to extract all surplus (7) from final users. Under network neutrality this implies:

$$p^{NN} = v_F x_F + v_G x_G - sW(x_F, x_G), \tag{13}$$

where $x_F = a/t_F$, while $x_G = a/t_G$ and $W(x_F, x_G)$ is given by (6). In case priority pricing is allowed, the charge to final users is:

$$p^{PP} = v_F x_L + v_G x_H - s\overline{W}(x_H, x_L), \tag{14}$$

where $x_L = a_L/t_F$ and $x_H = a_H/t_G$, while $\overline{W}(x_H, x_L)$ is given by (8).

Throughout the analysis, we concentrate only on the case where the ISP finds it optimal to supply both the fringe and firm G, instead of extracting all the surplus only

[15] When we invoke (9), condition (12) can be simplified. For example, if both types of CPs are equally efficient and $a = \frac{a_H + a_L}{2}$, then (12) becomes $4t_G \leq a + a_H$.

from firm G while neglecting the fringe. This is ensured by having the consumers' preference for variety which is strong enough. For this purpose, we assume

$$v_i > \lambda, \quad i = F, G. \tag{15}$$

As it will become apparent below, the condition tells that, for a given level of the waiting time, the marginal benefit from content provision exceeds the marginal cost to supply capacity.

Simple comparative statics lead to our first result on congestion, network capacity, and content supply.

Proposition 1 *In the long run: 1) The equilibrium average congestion is always the same under both regimes; 2) PP leads both to a higher capacity investment and to more total content than NN if and only if*

$$\frac{t_F}{t_G} > \frac{a - a_L}{a_H - a}. \tag{16}$$

If weighted advertising rates follow (9), this always holds as long as $\gamma < \widehat{\gamma} = \frac{t_F}{t_F + t_G}$.

Proof. The proof is very simple by doing a change of variable, as choosing μ also determines W. Under NN it is $W = \frac{1}{\mu - \lambda(x_F + x_G)}$, and hence

$$\mu = \frac{1}{W} + \lambda \left(x_F + x_G\right) = \frac{1}{W} + \lambda \left(\frac{a}{t_F} + \frac{a}{t_G}\right).$$

Notice that, for a given W, the capacity marginal cost when traffic x_i increases is λ, which clarifies the interpretation of assumption (15).

Similarly, under PP it is $\overline{W} = \frac{1}{\mu - \lambda(x_L + x_H)}$ and

$$\mu = \frac{1}{\overline{W}} + \lambda \left(\frac{a_L}{t_F} + \frac{a_H}{t_G}\right).$$

The first-order conditions in the two regimes are:

$$\begin{aligned} \frac{\partial \Pi_{ISP}^{NN}}{\partial W} &= -s - \frac{\partial \mu}{\partial W} = 0, \\ \frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} &= -s - \frac{\partial \mu}{\partial \overline{W}} = 0. \end{aligned} \tag{17}$$

These conditions are identical and thus determine the same average waiting time.[16]

[16] As $-\frac{\partial^2 \mu}{\partial W^2} < 0$, the second-order conditions are verified at the equilibrium under both regimes.

This means that $\mu - \lambda(x_F + x_G)$ under NN must equal $\mu - \lambda(x_L + x_H)$ under PP. The level of capacity therefore depends on the comparison of total traffic, which is generated by total content:

$$\mu^{PP} > \mu^{NN} \text{ iff } \frac{a_L}{t_F} + \frac{a_H}{t_G} > \frac{a}{t_F} + \frac{a}{t_G},$$

which gives (16) and is surely satisfied when t_G/t_F is low enough.

Under (9), (16) further simplifies to:

$$\frac{t_F}{t_G} > \frac{\gamma}{1-\gamma},$$

and therefore PP leads to higher investment and to higher total content iff $\gamma < \widehat{\gamma} = \frac{t_F}{t_F + t_G}$. **Q.E.D.**

The first part of the proposition is independent of any assumption on advertising rates. As the end users only care about average congestion, the neutrality regime has no bearing on the equilibrium average waiting time. The neutrality regime instead changes the amount of content provided and traffic generated. To keep the same waiting time, capacity has to adjust too.

The second part of the proposition focuses on the ISP's investment and on the CPs' supply of content. Condition (16) summarizes the general condition needed in order for PP to lead to higher investments compared to NN, still without making assumptions on advertising rates in the two regimes. The condition depends only on the relative efficiency of the CPs in producing content, and not on the value generated, since CPs derive their profits only from advertising and do not sell directly to end users. The condition is certainly satisfied when the ratio t_F/t_G is large. PP shifts advertising resources to firm G, and this produces an overall increase in total traffic when G is rather efficient in generating content. Therefore investment in capacity additionally increases compared to NN in order to keep the same average congestion. Conversely when the ratio t_F/t_G is small: it is only when resources, via PP, are directed to the "wrong" type of CP that the investment result can be reversed.

If the average amount of advertising resources is unaffected by the regime and the rates follow (9), PP results in higher investment if the weight γ does not exceed a threshold $\widehat{\gamma}$. Recall that γ captures the discrepancy between ad rates under PP and NN: if γ is low (high), then a_H is considerably higher (close to) than a. Condition (9) compares this to the relative degree of efficiency between firm G and the fringe in generating applications. For instance, if $\gamma = \frac{1}{2}$ in (9) and the advertising rate

under NN is a simple average, then the condition $\gamma < \widehat{\gamma}$ can be re-written as $t_G < t_F$. The result would then be particularly clear: total traffic depends on the average advertising funds available (that, under (9), do not differ in the two regimes), and on the relative efficiency if the CPs. When firm G is inefficient compared to the fringe ($t_G > t_F$), NN prevails over PP: advertising resources under PP are driven away from the smaller but more efficient CPs, so that the increase in the number of applications supplied by firm G does not compensate for the reduction of content supplied at the edge by the fringe.

Notice that the ISP's incentive to invest in capacity under PP do *not* depend on the premium fee: this is just used to extract firm G's rent, but does not affect traffic.

Having compared investment in capacity and total content provision under the two regimes, we now complete the characterization of the properties of the equilibrium.

Proposition 2 *The comparison between the equilibrium variables under the NN and PP regimes implies:*

$$
\begin{aligned}
x_G &< x_H,\ x_F > x_L, \\
W_H &< W(x_F, x_G) = \overline{W}(x_L, x_H) < W_L, \\
p^{NN} &< p^{PP} \text{ iff } v_G \geq \frac{t_G(a - a_L)}{t_F(a_H - a)} v_F, \text{ which under (9) simplifies to } v_G \geq \frac{t_G \gamma}{t_F(1-\gamma)} v_F, \\
\Pi_F^{NN} &> \Pi_F^{PP},\ \pi_G^{NN} = \pi_G^{PP}, \\
\Pi_{ISP}^{PP} &> \Pi_{ISP}^{NN} \text{ iff } v_G \geq v_{ISP},^{17} \text{ which under (9) is surely satisfied when } v_G \geq v_F \text{ and } \gamma \leq \widehat{\gamma}.
\end{aligned}
$$

Proof. See Appendix.

The results suggest that NN regulation is likely to have important redistributive effects on the sector that go beyond investment in infrastructure. The first part of the proposition is independent of any assumption made on the average advertising rates in the two regimes. The content decision of the fringe is determined only by advertising revenues: NN thus implies an increase in the participation at the edge. This also translates into higher aggregate profits for the fringe. Moving towards a regime of PP kills part of the innovation done at the edge by the fringe as small providers get reduced advertising rates. Conversely, firm G gets higher advertising revenues which leads it to invest in more applications; however, the ISP appropriates the extra rents by charging a premium fee, so that the net profits of firm G do not change. End users always have their consumer surplus completely extracted both

with and without NN, but the prices they pay differ. Provided the fringe content is not evaluated too highly relative to firm G's, the subscription price typically goes up with PP, reflecting the higher benefits they enjoy from more available applications. If instead the content of fringe firms is very important to users, NN leads to a higher price as its content supply best meets users' preferences.

If the average amount of resources from advertising is not affected by the pricing regime and (9) holds, further results can be established. The ISP typically benefits from supplying access with different priorities to CPs, but the result is not unequivocal. If the fringe content is particularly valuable to users, then their fees are higher under NN and the ISP may get higher profits overall. However, this requirement is rather stringent as a sufficient (but by no way necessary) condition for PP to generate higher profits for the ISP is that the content of firm G is at least as valuable as the fringe content, as long as the weight γ is not too high. To see this very clearly, imagine all CPs are equally efficient and generate the same value to end users. Also, set $\gamma = \frac{1}{2}$, so that average revenues from advertising do not change under either regime. Then total content is the same, and the price to end users is also the same. However, the ISP does strictly better under PP because it also earns the priority fee from G.

To summarize, the main effect of net neutrality regulation is therefore to direct advertising resources towards the fringe. The result is to induce more entry of new content providers in the fringe or, in other words, innovation at the edge, while it reduces content innovation done by large content providers.

We conclude this section with a simple exercise of comparative statics in the PP regime.

Corollary 1 *Imagine transportation costs are the same for all CPs and all content is equally valued by users. Then, for a given level of average advertising funds available, under PP, an increase in the dispersion in advertising rates leads to an increase in the premium price paid by firm G and in the profits of the ISP.*

Proof. Imagine that, when $t_F = t_G = t$, (9) reduces to $a_L + a_H = 2a$. Write $a_H = a + \Delta$ and $a_L = a - \Delta$, where Δ is a measure for dispersion. We now fix the level of a, and look at what happens when the gap Δ between the two rates under PP widens. From the proof of Proposition 2, when $v_F = v_G = v$ and $t_F = t_G = t$, it is $p = 2a\frac{v}{t} - \sqrt{s}$, which does not depend on Δ. The same applies to capacity investment and total content provision. However, the fee to firm G is $f_H = \frac{a_H^2 - a^2}{2t} = \frac{(2a+\Delta)\Delta}{2t}$, which increases with ads dispersion. By simple substitution, it is immediate

to find that the profits of the ISP also increase with the dispersion of advertising rates $\frac{\partial \Pi_{ISP}^{PP}}{\partial \Delta} = \frac{a+\Delta}{t} > 0$. **Q.E.D.**

A profit increase with the level of advertising funds is not surprising, as the ISP can appropriate more of these resources. More interestingly, under PP, for a given average level of these funds, the ISP benefits from an increase in their dispersion. It allows the ISP to make more money as it extracts higher premium profits from firm G. This is important for the ensuing analysis, in Section 5, where we assume that advertising rates change with the congestion level. The monopolist ISP will have an incentive to affect the level of advertising funds (under both regimes), as well as their dispersion, which is doable only under PP. Before turning to this case, we address the welfare implications of NN.

4.1 Welfare effects of NN regulation

As prices and fees are simple transfers, the expressions for social welfare are:

$$\begin{aligned} SW^{NN} &= v_F x_F + v_G x_G - sW(x_F, x_G) + \\ &\quad + a(x_F + x_G) - t_F \int_0^{x_F} x\mathrm{d}x - t_G \int_0^{x_G} x\mathrm{d}x - \mu, \\ SW^{PP} &= v_F x_L + v_G x_H - s\overline{W}(x_L, x_H) + \\ &\quad + a_H x_H + a_L x_L - t_F \int_0^{x_L} x\mathrm{d}x - t_G \int_0^{x_H} x\mathrm{d}x - \mu, \end{aligned}$$

under NN and PP respectively.

We shall start our analysis by comparing the private allocation with the first best under each regime. Under NN, the first best satisfies:

$$\frac{\partial SW^{NN}}{\partial x_F} = a - t_F x_F^* + v_F - \lambda = 0, \tag{18}$$

$$\frac{\partial SW^{NN}}{\partial x_G} = a - t_G x_G^* + v_G - \lambda = 0, \tag{19}$$

$$\frac{\partial SW^{NN}}{\partial W} = -s + \frac{1}{W^{*2}} = 0. \tag{20}$$

Using results from Proposition 2, the private equilibrium is characterized by:

$$a - t_F x_F = 0, \tag{21}$$

$$a - t_G x_G = 0, \tag{22}$$

$$-s + \frac{1}{W^2} = 0.$$

Comparisons are quite easy. The waiting time is socially optimal: the reason is that W impacts only on the cost of investment and on delay sensitivity. As all final users' surplus is extracted, the ISP internalizes the effect of delay, leading to the first best choice. The content supplied by all the CPs is instead suboptimally low. Each CP decides only on the basis of its advertising rate a and its development costs (t_i) hence, differently from a social planner, cannot internalize the impact on users (v_i) or on congestion costs (λ). In case $v_i > \lambda$, which is assumed under (15), the evaluation of CPs' content is higher than the corresponding marginal cost to keep waiting time constant: then each CP is underinvesting in content. Since both firm G and the fringe are underinvesting, while the waiting time is the same as in the first best, it immediately follows that investment in capacity is less than socially optimal: $\mu^{NN} < \mu^{*NN}$.

A similar analysis under PP implies that the first best allocations satisfy:

$$\begin{aligned} \frac{\partial SW^{PP}}{\partial x_F} &= a_L - t_F x_L^* + v_F - \lambda = 0, \qquad (23) \\ \frac{\partial SW^{PP}}{\partial x_G} &= a_H - t_G x_H^* + v_G - \lambda = 0, \qquad (24) \\ \frac{\partial SW^{PP}}{\partial \overline{W}} &= -s + \frac{1}{\overline{W}^{*2}} = 0, \end{aligned}$$

while the private equilibrium, from Proposition 2, is:

$$\begin{aligned} a_L - t_F x_L &= 0, \qquad (25) \\ a_H - t_G x_H &= 0, \qquad (26) \\ -s + \frac{1}{\overline{W}^2} &= 0. \end{aligned}$$

The comparisons to the first best under PP have exactly the same flavor as above: the average level of congestion $\overline{W}$ is optimal; the CPs' content is suboptimal and below x_i^* because $v_i > \lambda$. Hence capacity is also below its first best level, $\mu^{PP} < \mu^{*PP}$.

Next, we can state our main result on welfare: the following proposition establishes what regime is socially preferable, both when allocations are chosen by a social planner and by an unregulated ISP.

Proposition 3 *Imagine (9) holds. 1) PP is more efficient than NN iff $v_G \geq v^*$, with $v^* > v_{ISP}$. 2) When $\gamma \leq \widehat{\gamma}$, $v_G \geq v_F$ is a sufficient condition for PP to be the most efficient regime, both when allocations are determined by a social planner and by an unregulated ISP.*

Proof. See Appendix.

The proposition suggests that, from the point of view of a social planner, PP can do strictly better than NN unless the valuation for the fringe content is particularly high. PP welfare dominates NN when v_G is greater than some threshold value v^*.[18] To better grasp the intuition for this result, focus on the simpler case when firm G and the fringe are equally efficient both in their content development costs ($t_F = t_G$) and in the value generated ($v_F = v_G$), and also consider $\gamma = \frac{1}{2}$. Both regimes produce the same total amount of content. A PP regime would lead to an inefficient split of production of content: firm G would produce a higher proportion of the available content, therefore at higher marginal costs than in the symmetric split that would occur under NN. However, in that case proportionally more resources are generated via advertising, and this latter effect prevails overall and implies the superiority of PP.[19] The second part implies that in case γ is low enough, i.e. the advertising rate under priority is not too similar to the one under NN, a sufficient but not necessary condition for PP to welfare dominate is that firm G's content is more valuable to users than the fringe's.

These results clearly rely on the fact that all advertising is informative and increases the social value of the industry. Part of online advertising, however, may constitute a nuisance for final users, distracting them from the main reason they are surfing. In that case, if the nuisance increases sharply with advertising, it is easy to show that the welfare superiority of PP is less pronounced or may be overturned.

Finally, we can compare the regime preference of the social planner with the one of the monopolist ISP. Both the social planner and the ISP prefer PP if v is high enough. The preferences of both are reversed in favor of NN in case the value of the fringe's content v_F is relatively high. The monopolist's preferred regime is instead in contrast with welfare for intermediate values of the v_G/v_F ratio. In particular, the monopolist tends to choose PP "too" often when v_G is intermediate, $v_{ISP} < v_G < v^*$. As we show in the proof, the discrepancy between the threshold values v_{ISP} and v^* is larger the larger is the ratio of development costs, t_G/t_F. This is because the monopolist prefers PP to the extent that it can extract a premium fee from firm G,

[18] Notice that the result is independent of whether the allocation is determined by the monopolist ISP or by the social planner.

[19] With equally efficient CPs, it is $SW^{NN} = 2a\left(\frac{v-\lambda}{t}\right) + \frac{a^2}{t} - 2\sqrt{s}$ and $SW^{PP} = \frac{(a_L+a_H)(v-\lambda)}{t} + \frac{a_H^2+{a_L}^2}{2t} - 2\sqrt{s}$. When $2a = a_H + a_L$, it is immediate that $SW^{PP} > SW^{NN}$.

while it does not internalize the development costs, as a social planner would instead do.

5 Congestion-dependent advertising rates

We conducted the previous analysis under the assumption that advertising rates were given exogenously and that, when compared to NN, they would command a premium to those CPs that had chosen to prioritize their applications under PP. However, these premiums arise precisely because, when users suffer less from congestion problems, the applications they use work better, are more reliable, better preserve data integrity, and so forth. Lower congestion should be associated to better opportunities for those who place their ads over the Internet. For instance, smart banners and clips could be integrated with content. In this section, therefore, we make ad revenues dependent on congestion, both under NN and PP.[20] This extension is important since we want to go beyond the simple "rent extraction" mechanism of the premium fee that we discussed so far, which did not affect neither the supply of content nor the incentive to invest.

In particular, under NN, the (single) advertising rate takes the following general form

$$a = a(W),$$

with $a' < 0$. Similarly, under PP we have

$$a_L = a_L(\overline{W}), \quad a_H = a_H(\overline{W}),$$

with $a_L < a_H$, $a'_L < 0$, $a'_H < 0$.

Since our focus is now on the link which is being created between advertising funds and network congestion, from now onwards we assume that: a) all CPs have identical transportation costs, $t_F = t_G = t$, firm G and the fringe are therefore equally efficient in generating content; b) final users evaluate equally all the content provided on the Internet, no matter if it is supplied by firm G or by the fringe, $v_F = v_G = v$.

[20] Njoroge et al. (2010), similarly, suggest a positive relationship between advertising revenues and quality of the connection. Marketing research reveals that both advertising exposure and user involvement are crucial for recall and information processing (Danaher and Mullarkey, 2003); smooth and fast surfing should then increase users' involvement and, hence, the time spent on a website, leading to a better recall and processing of the information in the advert.

As before, we do not assume that departures from neutrality as such can increase the resources attracted to this economy. Hence, when average waiting time is the same, then also the average advertising revenues are the same

$$a_L(\overline{W}) + a_H(\overline{W}) = 2a(W) \qquad \text{when} \quad \overline{W} = W.$$

5.1 Investment in capacity

The analysis closely parallels the one with fixed advertising rates in Section 4: it is again very convenient to do a change of variable, so that the ISP sets prices and average waiting time. Consider first NN: for a given W, the free entry condition for the fringe and firm G's content maximization determine

$$x_F = x_G = \frac{a(W)}{t}.$$

The ISP solves

$$\max_W \Pi_{ISP}^{NN} = \pi_{ISP}^{NN} - \mu = p^{NN} - \mu,$$

where p^{NN} is given by (13), and from $W = \frac{1}{\mu - \lambda(x_F + x_G)}$ we obtain:

$$\mu = \frac{1}{W} + \frac{2\lambda}{t} a(W).$$

The equilibrium waiting time with variable advertising rates is then defined implicitly by the following first-order condition:

$$\frac{\partial \Pi_{ISP}^{NN}}{\partial W} = \frac{2(v-\lambda)}{t} a'(W) - s + \frac{1}{W^2} = 0. \tag{27}$$

Comparing (27) with (17), it is apparent that congestion now depends on the way it affects advertising rates. Only if advertising was not sensitive to congestion ($a' = 0$), would we obtain again $W = 1/\sqrt{s}$ as in (17). As $(v - \lambda) > 0$ the first term is negative, which implies that in equilibrium $W < 1/\sqrt{s}$.[21] The ISP has now further incentives to reduce congestion, because lower congestion increases advertising funds, which increase content provision, which finally increases the fee that users are willing to pay. For the same reason, the equilibrium value of W is now affected by all the parameters of the problem (t, v, λ, besides s).

[21] An interior equilibrium requires the second order condition on the waiting time: $\frac{v-\lambda}{t} a''(W) - \frac{1}{W^3} \leq 0$ to be satisfied.

Let us now turn to the analysis of PP. Following the same steps as above, for a given $\overline{W}$, we have:

$$x_L = \frac{a_L(\overline{W})}{t}, \quad x_H = \frac{a_H(\overline{W})}{t}.$$

The ISP solves

$$\begin{aligned} \max_{\overline{W}} \Pi_{ISP}^{PP} &= \pi_{ISP}^{NN} - \mu = p^{PP} + f_H - \mu \\ s.t. \quad f_H &= \frac{a_H^2(\overline{W}) - a^2(\overline{W})}{2t}, \end{aligned}$$

where p^{PP} is given by (14), and from $\overline{W} = \frac{1}{\mu - \lambda(x_L + x_H)}$ we obtain

$$\mu = \frac{1}{\overline{W}} + \frac{\lambda}{t}[a_L(\overline{W}) + a_H(\overline{W})].$$

The equilibrium waiting time with variable advertising rates is then defined implicitly by the following first order condition:

$$\frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} = \frac{v - \lambda}{t}[a'_H(\overline{W}) + a'_L(\overline{W})] + \frac{a_H(\overline{W})a'_H(\overline{W}) - a(\overline{W})a'(\overline{W})}{t} - s + \frac{1}{\overline{W}^2} = 0 \tag{28}$$

If advertising was not sensitive to congestion ($a'_i = 0$), again $\overline{W} = 1/\sqrt{s}$.

We can now state the following results.

Proposition 4 *With variable advertising rates, the following properties hold. 1) If advertising rates have the same sensitivity to congestion in both regimes, congestion is lower and investment in capacity and content is higher under PP compared to NN. 2) Under the assumption $a_H + a_L = 2a$, a sufficient condition for PP to lead to lower congestion and higher investment is $|a'_H| > |a'|$. A necessary, but not sufficient, condition for congestion to be lower and investment to be higher under NN compared to PP is $|a'| > |a'_H|$.*

Proof. See Appendix.

Proposition 4 is the main result of the paper, as it links the incentives to invest both in network expansion and total content to the sensitivity of advertising rates to congestion: our earlier intuition from Corollary 1 that the ISP has an incentive to adopt prioritized traffic and invest more, particularly when this allows to redirect advertising resources towards firm G, is confirmed. NN can favor investment only

when the sensitivity of advertisement rates is decreased by priority, arguably a rather unintuitive situation.

The results deserve some more detailed comments. Part 1) implies that, as advertising rates are equally affected by congestion, the ISP reduces congestion compared to the case with exogenous advertising rates. This happens particularly under PP: compared to NN, PP has a "level" effect that increases advertising funds overall; this induces the ISP to expand capacity and prioritize traffic. Clearly, the result is not a restatement of Proposition 2, which is obtained only as a limiting case when $a' = 0$.

Part 2) states the most general sufficient condition for PP to lead to lower congestion and higher investment: this simply requires that the sensitivity of advertising rates increases with prioritization. A decrease in congestion under PP increases the dispersion of advertising rates and leads to further funds attracted to firm G's applications: in other words, both the level and the dispersion effect go in the same direction. Indeed, a reversal in the sensitivity of ads to congestion is needed to overturn this finding: it is only if priority decreases, rather than increases, the sensitivity of ad rates to congestion that NN may lead to a lower congestion and a higher investment. It should be further noticed, however, that the condition devised ($|a'| > |a'_H|$) is a necessary but not sufficient result: this is due to the "level" effect, whereby advertising rates increase for firm G under PP, that is still operating.

The results on investment and content supply that we obtained for constant advertising rates are therefore likely to be sharpened when considering congestion-sensitive advertising rates: this is due to the joint operation of both the dispersion and level effects of advertising rates. The existence of these advertising sensitivity effects has also an impact on congestion: this contrasts with the case of fixed advertising rates, where waiting times remained constant in both regimes.

5.2 Welfare effects of NN regulation

To analyze the welfare implications of NN, we follow similar steps as in Section 4.1. For the sake of brevity, here we consider which priority regime generates the highest welfare when allocations are chosen privately, as this is also the policy question which ultimately matters.

Consider NN first. The expression for welfare is:

$$\begin{aligned} SW^{NN}(x_F, x_G, W) &= v(x_F + x_G) - sW + a(W)(x_F + x_G) \\ &\quad -\frac{t}{2}\left(x_F^2 + x_G^2\right) - \mu(x_F, x_G, W), \end{aligned}$$

where $\mu(x_F, x_G, W) = \frac{1}{W} + \lambda(x_F + x_G)$.

Under PP, social welfare can be written as:

$$\begin{aligned} SW^{PP}(x_L, x_H, \overline{W}) &= v(x_L + x_H) - s\overline{W} + a_L(\overline{W})x_L + a_H(\overline{W})x_H \\ &\quad -\frac{t}{2}\left(x_L^2 + x_H^2\right) - \mu(x_L, x_H, \overline{W}), \end{aligned}$$

where $\mu(x_L, x_H, \overline{W}) = \frac{1}{\overline{W}} + \lambda(x_L + x_H)$.

After substitution and rearranging terms, we obtain

$$\begin{aligned} \Delta SW &= SW^{NN} - SW^{PP} = \underbrace{\frac{v-\lambda}{t}(2a - a_H - a_L)}_{AD_1} + \\ &\quad \underbrace{\frac{1}{2t}\left(2a^2 - a_H^2 - a_L^2\right)}_{AD_2} - \underbrace{(W^{NN} - \overline{W}^{PP})(s - \frac{1}{W^{NN}\overline{W}^{PP}})}_{WT}. \end{aligned}$$

The welfare differential is decomposed into three parts: the first two terms capture the welfare effects of advertising rates (AD_1 and AD_2), and the last term captures the effect of the waiting times (WT). If ad rates were not sensitive to congestion, we know that waiting times would be identical, so $WT = 0$. Furthermore, under the assumption that the average ad revenues do not change in the two regimes, AD_1 would also be equal to zero and AD_2 negative. Hence we would find again the same result as in Proposition 3: a regime with priority has better welfare properties than a regime based on best-effort. More generally, when ad rates are sensitive to congestion, we have the following result on which regime is preferred for welfare.

Proposition 5 *The same conditions that lead to lower congestion under PP compared to NN (Proposition 4), are also sufficient for PP to be the most efficient regime.*

Proof. From (27) and (28), under either regime it is always $W < 1/\sqrt{s}$. Hence $s - \frac{1}{W^{NN}\overline{W}^{PP}} > 0$ and the sign of the term WT simply depends on the comparison of waiting times, as discussed in Proposition 4. Furthermore, if $W^{NN} > \overline{W}^{PP}$, it is also $a_H(\overline{W}^{PP}) + a_L(\overline{W}^{PP}) = 2a(\overline{W}^{PP}) > 2a(W^{NN})$, implying $AD_1 < 0$ and, a fortiori, $AD_2 < 0$. Thus $W^{NN} > \overline{W}^{PP}$ is a sufficient condition for $\Delta SW < 0$. **Q.E.D.**

An intuitive interpretation of the result goes as follows. For a given level of congestion, a more skewed distribution of advertising resources leads to a higher overall content supply under PP; hence, PP has a positive welfare effect that can only be outweighed by an increase in waiting time and the inconvenience that this has on final users. Since these effects are not present, then PP is clearly welfare superior to NN. The sufficiency result requires only PP to lead to lower congestion, as identified by our previous analysis on advertising rates' sensitivity to congestion.

The results obtained in the benchmark model are robust to the endogenization of the advertising rates: PP, through its redistribution of the advertising resources, is likely to lead to a welfare superior outcome with respect to NN regulation. The opposite result requires NN to reduce congestion, a scenario that requires a sharp reversal in the sensitivity of ads rates.

We have obtained a rather simple "rule of thumb" to assess the welfare properties of PP: if average congestion of the ISP is reduced with prioritization (something that could be monitored empirically), then PP is necessarily superior to NN in terms of total welfare.

6 Conclusions

The Internet industry is facing a crucial phase of its development. Since broadband has become the standard delivering technology, telephone and cable networks have become a gateway to content and applications. These ISPs can access a large amount of information about data packets and discriminate between them at a relatively low cost. The "net neutrality debate" has developed in several directions: from an economic standpoint, the debate focuses on the consequences that discrimination can have on pricing of ISPs to both content providers and final users. Both advocates of net neutrality regulation and opponents have put forward important arguments. One of the most controversial issues is whether regulation is needed to protect innovation at the "edge", i.e., from small and innovative CPs; on the other hand, investment incentives at the "core", i.e., ISPs' maintenance and upgrade of their networks, are also crucial in times of increased bandwidth demand.

Our paper contributes by developing a formal framework that, although stylized, seems well suited to capture the features of the Internet sector and analyze the arguments in favor of and against net neutrality. We proposed a model that formalizes prioritization as a tool that stands at the interface between Operations Management

and Marketing, especially in the context of clickstream tracking. Broadband network intelligence allows the ISP both to reduce waiting time of particular applications (which is directly enjoyed by end users) and to attract advertisers' interests via deep packet inspection (advertisers then fund CPs). The main idea that we have put forward is that a prioritization regime redirects resources towards particular players (the large CP and the ISP, in our model), and takes away resources from other stakeholders (the small CPs). This further affects their incentives to invest in either infrastructure or content, which has real effects.

In our framework, the main engine comes from differential advertising funds, but others can be thought of, e.g., paid-for content and applications in case CPs can directly charge end users. Our findings suggest that the ISP adjusts capacity to the level of traffic: net neutrality then is likely to slow investment at the core; however, regulation is likely to favor innovation at the edge while hindering the development of applications from large content providers. One of our results that should be of relevance to policy makers is that allowing prioritization implies that the large CP ("firm G") becomes even larger compared to the fringe of CPs, although not necessarily more profitable. Overall, we have identified conditions such that priority pricing leads to a better allocation of the resources available in the industry and, as such, it is welfare enhancing. The results are quite robust and are reinforced in case the advertising revenues of CPs are sensitive to congestion.

There are several ways in which our model can be improved. First, despite that the "last mile" of the Internet seems relatively uncompetitive, it would be desirable to extend our approach to the case of competing ISPs. Second, advertising rates could be further endogenized by modelling the demand and supply of advertising space: this extension constitutes a challenge for further research.

References

[1] Armstrong M. (2006), Competition in two-sided markets, RAND Journal of Economics, 37, 668-691.

[2] Becker G., Carlton D. and H. Sider (2010), Net Neutrality and Consumer Welfare, Journal of Competition Law & Economics, 6, 497-519.

[3] Brennan T.J. (2011), Net Neutrality or Minimum Quality Standards: Network Effects vs. Market Power Justifications, in Spicker I. and J. Kramer (Eds.), Network Neutrality and Open Access, Nomos Verlag, Baden Baden.

[4] Cheng H.K., Bandyopadyhay S. and H. Guo (2011), The Debate on Net Neutrality: A Policy Perspective, Information Systems Research, 22(1), 1-27.

[5] Choi J.P. and B. Kim (2010), Net neutrality and investment incentives, Rand Journal of Economics, 41(3), 446-471.

[6] Czernich N., Falck O., Kretschmer T. and L. Woessmann (2011), Broadband Infrastructure and Economic Growth, Economic Journal, 121(552), 505-532.

[7] Danaher P.J. and G.W. Mullarkey (2003), Factors affecting online advertising recall: a study of students, Journal of Advertising Research, 43(3), 252-267.

[8] Economides N. and J. Tag (2012), Net Neutrality on the Internet: A Two-sided Market Analysis, Information Economics & Policy, forthcoming.

[9] Economides N. and B.E. Hermalin (2012), The Economics of Net Neutrality on the Internet, Rand Journal of Economics, forthcoming

[10] Fahri E. and A. Hagiu (2008), Strategic Interactions in Two-Sided Market Oligopolies, Harvard University Strategy Unit Working Paper, 08-11.

[11] Hermalin B.E. and 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.

[12] Hogendorn C. (2008), Broadband Internet: net neutrality versus open access, International Economics and Economic Policy, 4(2), 185-208.

[13] Kocsis V. and P.W.J. De Bijl (2008), Network neutrality and the nature of competition between network operators, International Economics and Economic Policy, 4, 159-84.

[14] Kramer J. and L. Wiewiorra (2012), Network Neutrality and Congestion Sensitive Content Providers, Information Systems Research, forthcoming.

[15] Lee R. and T. Wu (2009), Subsidizing Creativity through Network Design: Zero-Pricing and Net Neutrality, Journal of Economics Perspectives, 23, 61-76.

[16] Lessig L. (2001), The future of ideas: the fate of commons in a connected world, Random House, New York, USA.

[17] Mayo J.W. and S. Wallsten (2011), From Network Externalities to Broadband Growth Externalities: a Bridge not yet Built, Review of Industrial Organization, 38(2), 173-190.

[18] McDysan D. (1999), QoS and Traffic Management in IP and ATM Networks, McGraw-Hill, New York, USA.

[19] Musacchio J., Schwartz G. and J. Warland (2009), A Two-Sided Market Analysis of Provider Investment Incentives with an Application to the Net-Neutrality Issue, Review of Network Economics, 8(1), 22-39.

[20] Njoroge P., Ozdaglar A., Stier-Moses N.E. and G.Y. Weintraub (2010), Investment in two sided markets and the net neutrality debate, Columbia Business School DRO Working Paper, 05.

[21] Rochet J.C. and J. Tirole (2006), Two sided markets: a progress report, RAND Journal of Economics, 37(3), 645-667.

[22] Saltzer J., Reed D. and D.D. Clark (1984), End-to-end arguments in system design, ACM Transactions of Computer Systems, 2(4), 277-288.

[23] Sandvine (2011), Global Internet Phenomena Spotlight, May 2011 (*accessible online at*: www.sandvine.com).

[24] Valletti T. and C. Cambini (2005), Investments and network competition, RAND Journal of Economics, 36, 446-467.

[25] Van Schewick B. (2006), Towards an Economic Framework for Network Neutrality Regulation, Journal of Telecommunications and High Technology Law, 5, 329-392.

[26] Wu T. (2004), The Broadband Debate: A User's Guide, Journal of Telecommunications and High Technology Law, 3, 69-96.

[27] Yoo C.S. (2005), Beyond network neutrality, Harvard Journal of Law and Technology, 19, 1-77.

A Appendix

Proof of Proposition 2. From the proof of Proposition 1 we can calculate $W = 1/\sqrt{s}$, and thus $\mu - \lambda(x_F + x_G) = \sqrt{s}$ under NN. This determines the capacity level and characterizes the equilibrium fully:

$$\begin{aligned} W &= \frac{1}{\sqrt{s}}, \quad \mu = \lambda(\frac{a}{t_F} + \frac{a}{t_G}) + \sqrt{s}, \\ p &= a\left(\frac{v_F}{t_F} + \frac{v_G}{t_G}\right) - \sqrt{s}, \\ x_G &= \frac{a}{t_G}, \quad x_F = \frac{a}{t_F}, \\ \Pi_F &= \frac{a^2}{2t_F}, \quad \pi_G = \frac{a^2}{2t_G}, \\ \Pi_{ISP} &= a\left(\frac{v_F - \lambda}{t_F} + \frac{v_G - \lambda}{t_G}\right) - 2\sqrt{s}. \end{aligned}$$

Under PP, the equilibrium is solved similarly:

$$\begin{aligned} \overline{W} &= \frac{1}{\sqrt{s}}, \quad \mu = \lambda(\frac{a_L}{t_F} + \frac{a_H}{t_G}) + \sqrt{s}, \\ p &= \frac{a_L v_F}{t_F} + \frac{a_H v_G}{t_G} - \sqrt{s}, \quad f_H = \frac{a_H^2 - a^2}{2t_G}, \\ x_H &= \frac{a_H}{t_G}, \quad x_L = \frac{a_L}{t_F}, \\ \Pi_F &= \frac{{a_L}^2}{2t_F}, \quad \pi_G = \frac{a^2}{2t_G}, \\ \Pi_{ISP} &= \frac{a_L(v_F - \lambda)}{t_F} + \frac{a_H(v_G - \lambda)}{t_G} + \frac{a_H^2 - a^2}{2t_G} - 2\sqrt{s}. \end{aligned}$$

The results follow from simple comparisons of the relevant expressions. In particular $a_H > a > a_L$ implies that $x_G < x_H$ and $x_F > x_L$.

The average waiting times are identical under both regimes while $\overline{W} < W_H$ and $\overline{W} > W_L$ follow immediately from the properties of the M/M/1 system.

As far as prices to end users are concerned, it is $p^{PP} > p^{NN}$ iff $v_G \geq \frac{t_G(a-a_L)}{t_F(a_H-a)} v_F$. Under (9), this further simplifies to $v_G \geq \frac{t_G \gamma}{t_F(1-\gamma)} v_F$.

Turning to the total profits of the fringe, again $a > a_L$ ensures that $\Pi_F^{NN} > \Pi_F^{PP}$. Firm G's profits instead do not change. Also the profits of the ISP in the two regimes can be compared:

$$\Pi_{ISP}^{PP} - \Pi_{ISP}^{NN} = \frac{(a_L - a)(v_F - \lambda)}{t_F} + \frac{(a_H - a)(v_G - \lambda)}{t_G} + \frac{a_H^2 - a^2}{2t_G}.$$

The first term is negative, the second is positive as well as the third. The second term always prevails over the first if v_G is high enough.

Using assumption (9), profits are higher under priority iff

$$v_G \geq v_{ISP} \equiv (v_F - \lambda)\frac{t_G\gamma}{t_F(1-\gamma)} + \lambda - \frac{a_H(1+\gamma) + a_L(1-\gamma)}{2}. \tag{29}$$

When $\gamma = \widehat{\gamma} = \frac{t_F}{t_F+t_G}$, the previous condition simpifies to

$$v_G \geq v_F - \frac{a_H(1+\widehat{\gamma}) + a_L(1-\widehat{\gamma})}{2},$$

so that $v_G \geq v_F$ is a sufficient condition for $\Pi^{PP}_{ISP} > \Pi^{NN}_{ISP}$ for all values of $\gamma \leq \widehat{\gamma}$. **Q.E.D.**

Proof of Proposition 3. We start with the priority regime choice by a social planner. Waiting time is the same under both regimes. By substituting the first best allocations (18), (19), (23), and (24) into the expressions for social welfare, and taking the difference, we obtain in general

$$\begin{aligned} SW^{NN} - SW^{PP} \quad = \quad & -\frac{(a_H - a_L)(a_H - a_L + 2v_G - 2v_F)}{2(t_F + t_G)} - \left(\underbrace{a_H\frac{t_F}{t_F+t_G} + a_L\frac{t_G}{t_F+t_G} - a}_{A} \right) \cdot \\ & \left(\underbrace{a_H\frac{t_F}{t_F+t_G} + a_L\frac{t_G}{t_F+t_G} - a}_{A} + \frac{2t_G(v_F-\lambda) + 2t_F(v_F-\lambda)}{t_F+t_G} \right) \frac{t_F+t_G}{2t_Ft_G}. \end{aligned}$$

Under assumption (9) the sign of second term in brackets depends simply on γ. If $\gamma \leq \widehat{\gamma} = \frac{t_F}{t_F+t_G}$ then for sure $A \geq 0$ and the whole term is not positive. Hence a sufficient condition for also the first term to be negative is $v_G \geq v_F$.

More in general, the welfare difference is still negative for any

$$v_G \geq v^* \equiv (v_F - \lambda)\frac{t_G\gamma}{t_F(1-\gamma)} + \lambda - \frac{a_H(1+\gamma) + a_L(1-\gamma)}{2} + \frac{\gamma t_G[a_L(2-\gamma) + a_H\gamma]}{2(1-\gamma)t_F}. \tag{30}$$

From inspection of (30) and (29), the first three terms are identical in both expressions. It then follows that $v_{ISP} - v^* = -\frac{\gamma t_G[a_L(2-\gamma)+a_H\gamma]}{2(1-\gamma)t_F} < 0$.

Consider next the expressions of the social welfare in the two different regimes, when allocations are determined by the ISP. As the fee charged to firm G does not affect any of the content decisions, the distortions to content implied by the comparison

of (18)-(19) with (21)-(22) and (23)-(24) with (25)-(26) are neutral between regimes and do not affect the social welfare properties of the first best as compared with the monopolist allocation. Thus, it is again $SW^{PP} \geq SW^{NN}$ iff $v_G \leq v^*$. **Q.E.D.**

Proof of Proposition 4. The results on total content follow from comparing $x_F + x_G = \frac{2a(W)}{t}$ under NN with $x_L + x_H = \frac{a_H(\overline{W}) + a_L(\overline{W})}{t}$ under PP. If one can prove that $\overline{W}^{PP} < W^{NN}$, then it immediately follows that total content supply increases under PP since $a_H(\overline{W}^{PP}) + a_L(\overline{W}^{PP}) = 2a(\overline{W}^{PP}) > 2a(W^{NN})$.

Consider next congestion, given by (27) and (28) in the two regimes. The only terms that matter are respectively

$$\begin{aligned} A^{NN} &= 2\frac{(v-\lambda)}{t}a', \\ A^{PP} &= \frac{v-\lambda}{t}(a'_L + a'_H) + \frac{a_H a'_H - a a'}{t}, \end{aligned}$$

where, to avoid clutter, we have dropped the dependence of ads on waiting time.

The results on capacity investment follow from noting that

$$\begin{aligned} \mu^{NN} &= \frac{1}{W} + 2\frac{\lambda}{t}a(W), \\ \mu^{PP} &= \frac{1}{\overline{W}} + \frac{\lambda}{t}\left[a_H(\overline{W}) + a_L(\overline{W})\right]. \end{aligned}$$

Since $a'_i < 0$, a *sufficient* general condition for PP to increase investment is that $\overline{W}^{PP} < W^{NN}$.

1) If the sensitivity of ads to congestion is the same ($a'_L = a'_H = a'$), then, as $a_H - a > 0$, we have that $A^{PP} < A^{NN}$, and hence $\overline{W}^{PP} < W^{NN}$.

2) Under $a_H + a_L = 2a$ one can write:

$$A^{PP} - A^{NN} = \frac{a'_H a_H - a' a}{t}.$$

As $a_H > a$, a *sufficient* condition for $A^{PP} < A^{NN}$ is $a'_H < a'$ or $|a'_H| > |a'|$. It is only when $|a'| > |a'_H|$ that the sign could be eventually reversed. **Q.E.D.**