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

Carlo Reggiani† Tommaso Valletti‡

September 2011

Abstract

We model an Internet broadband provider that can offer 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, unless 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.

1 Introduction

The Internet has probably been the fastest developing industry of the last two decades. In the early times an experimental network linking a limited number of computers,

*We thank Bruno Jullien, Joacim Tag, Joao Vareda, and seminar participants at Bern, Manchester, the Conference on Internet Search and Innovation at Northwestern University, CRESSE (Rhodes), EARIE (Stockholm) and Telecom ParisTech.

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

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

the Internet has become one of the key priorities of policy makers around the world as, among other things, 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 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), 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 (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

[1] Recent developments include a Report and Order issued by the U.S. FCC in 2010, promulgating some mild forms of NN, which have been challenged in court by Verizon. The European Commission issued in 2011 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) are 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.

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 (2009) 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 termination and, in their results, this typically increases CPs' welfare. Congestion, or incentives for both investment of the ISP and innovation of CPs, however, are not addressed.

[3] In June 2011, a casual SSRN search returned 149 papers with «net neutrality» in the title or abstracts. A similar Google search provided 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.

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. They can be classified according to three approaches followed so far. First, the speed of data exchange is determined by the slowest of the connections between the start and the end point; second, there is a limited amount of bandwidth to be allocated between differently clicked CPs; third, further structure on congestion is imposed by using a M/M/1 queuing system.

Njoroge et al. (2010) consider vertically-differentiated duopolistic ISPs and employ the first approach: the quality of on-net exchanges corresponds to the quality of the ISP, while off-network exchanges' quality is determined by the worse quality between the ISPs.[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 (2010) adopt the second approach in modelling bandwidth allocation: the "pipe" of a monopolist ISP has fixed capacity in the short run. Bandwidth, then, is allocated in different proportions to CPs according to the pricing regime: equally under NN, with priority if discrimination is allowed. This feature, together with the elastic demand from final users, is key for the results: uneven allocation of bandwidth under discrimination leads to a more than proportional increase of demand and increased traffic; the only case in which discrimination can lead to higher welfare is when it has an expansionary effect on CPs' supply.

Cheng et al. (2011), Choi and Kim (2010), Kramer and Wiewiorra (2010) 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

[5] Fahri and Hagiu (2007) 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 in telecommunications networks.

[7] McDysan (1999), cited by Kramer and Wiewiorra (2010).

Kim (2010) consider similar models in which users access exclusively only one of two content providers;[8] total supply of content is fixed and only the market shares are affected by priority. 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 fixed 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 (2010) 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 in case 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 in several respects. First, and 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. Second, the market for content is not fully covered, thus pricing has effects on CPs' entry. 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 little traffic. We capture this 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. Since, contrary to Choi and Kim (2010) and Cheng et al. (2011), we do not impose single-homing of end users with respect to CPs, the CPs' incentive to prioritize or not arises from

[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 YouTube, which cannot be modelled as mutually exclusive choices.

a different mechanism, which represents the third element of our model. 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. We also consider that advertising revenues might be endogenous to the level of congestion; this reflects the fact that the (costly) intelligence that can be installed at the broadband network level can be used to make advertising itself more effective. Lastly, since CPs must be physically connected to an ISP platform under any regime, we allow for CPs to be charged by the ISP also in case NN regulation is imposed, which is a better description of a "local" model of the Internet. In our model, we 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). The ISP first invests in capacity μ at a cost $I(\mu)$; then it charges the two sides of the market (more on this below; see also Figure 1).

CPs pay a connection fee to the ISP since they need a physical connection supplied by the ISP to be on the Internet. 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 that we call "Google" and denote with the subscript G.[9] In the fringe, each CP supplies one unique application/content, while Google 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, $t_F(x) = t_F x$. If f denotes the connection fee paid to the ISP, then the profit of a firm in the fringe that gets

[9] Naming "Google" the large CP is simply for expositional convenience. This is meant to capture that particular applications generate a large part of the internet traffic (Sandvine, 2011).

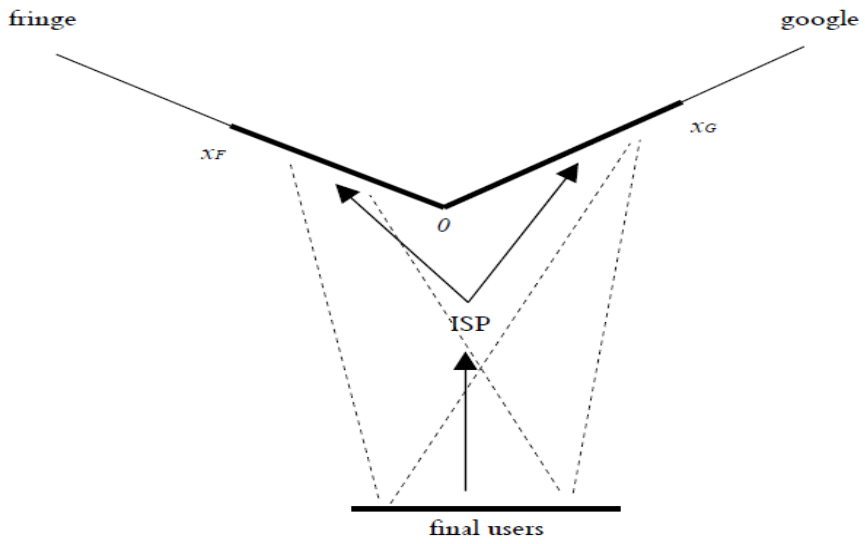

Figure 1: A stylized representation of the internet.

advertising revenues a from a total unit mass of users is

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

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

$$x_F = \frac{a - f}{t_F}. \tag{2}$$

The total profits of the fringe are thus

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

Google 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, Google maximizes w.r.t. x the total profit

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

Hence the mass of applications introduced by Google will be

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

to the extent that the corresponding profit

$$\pi_G = \frac{a^2}{2t_G} - f \tag{5}$$

is non-negative.[10] Three remarks are in order. First, the connection fee f enters only the determination of the mass of applications from the fringe (2), but not the mass of application (4) from Google. This implies that the marginal CP, and hence the elasticity of content with respect to f, is dictated by the very "edge" in the fringe, not by Google. Second, each CP pays the connection fee only once, even under NN. As there is a single ISP, and CPs must be connected to it, we do not find very realistic the assumption of zero connection fees employed by some of the literature, although we could easily also adopt it without affecting the main results. In our model, while a connection fee is always set, the ISP does not charge based on traffic volumes, for instance because the basic connection available is enough to support the traffic generated by any individual CP, including the larger Google. A flat "hook up" fee like ours is also found, e.g., in Economides and Hermalin (2010) and Njoroge et al. (2010). Third, we allow unit transportation costs t_i, $i = F, G$, to be different, in case Google 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 connection 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 Google application. Consumers also care about congestion on the network.

Congestion depends on the total traffic exchanged, 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, 2010). Each user-CP exchange generates an amount of traffic λ. Under Net Neutrality (NN), congestion is

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

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

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

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

where d 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 traffic, the utility of users is

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

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

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

so that the average waiting time is

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

Two are the 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

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

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 deep packet inspections, allowed by prioritization techniques. 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] At times, however, we will further assume:

$$a = a_H \frac{t_F}{t_F + t_G} + a_L \frac{t_G}{t_F + t_G}, \tag{9}$$

that is, the weighted average advertising rate does not change with and without NN, where the weights are given by the relative degree of efficiency between Google and the fringe to generate applications. If both types of CPs are equally efficient, then it reduces to a simple average, $a = \frac{a_H + a_L}{2}$. This assumption 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 total resources (from advertisers) that can be attracted by this economy, but it leads to a redistribution of these resources.

To sum up, we will consider two regimes. With NN, all CPs pay the same fee f, and get a. With PP, CPs will have the choice of paying differential fees: f_L 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 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

[12] Behavioural 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 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 on congestion is provided in Section 5.

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 net profits of the ISP are:

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

while under no regulation they are:

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

where D_L and D_H denote the demand for the high and the low priority respectively. If best effort is chosen in equilibrium only by the fringe while Google opts for priority, it will be $D_L = x_L$ and $D_H = 1$. Indeed, the fringe providers opt for the best effort/low priority connection if:

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

Provider G can opt for the high priority. The high priority is an option in case $\pi_G^H \geq \pi_G^L$, where the left hand side is the profit of Google with priority, while the right hand side is its profit without priority. In order to induce G to choose it, the ISP will set the charge for priority such that it holds exactly $\pi_G^H = \pi_G^L$. Since the profit of Google is given by (5), after substitution, the condition implies:

$$f_H = f_L + \frac{a_H^2 - a_L^2}{2t_G}. \tag{11}$$

In other words, the priority fee extracts all the extra rent from Google, but it does not affect Google's choice of content. Condition (10) to ensure self-selection of the fringe to low priority then becomes:

$$a_H + a_L \geq 2t_G, \tag{12}$$

that we assume to hold. This condition says that Google 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. Notice that, when $t_F = t_G$, we can use a simplified version of (9), to obtain for (12) the very simple condition

$$a \geq t_G.$$

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

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

where (11) holds.

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

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

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

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

where $x_L = \frac{a_L - f_L}{t_F}$ and $x_H = \frac{a_H}{t_G}$, 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 Google, instead of extracting all the surplus only from Google 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.

With simple comparative statics we obtain 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 investment and to more total content than NN if and only if*

$$2(a_H - a)\frac{t_F}{t_G} > a - a_L \tag{16}$$

which is always satisfied when average advertising rates follow (9).

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

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

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

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

The first order conditions in the two regimes are:

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

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

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:

$$\mu^{PP} > \mu^{NN} \text{ iff } \frac{a_L - f_L}{t_F} + \frac{a_H}{t_G} > \frac{a - f}{t_F} + \frac{a}{t_G}.$$

From the first order conditions identifying f and f_L one finds, respectively: $f = \frac{a + t_F + \lambda - v_F}{2}$ and $f_L = \frac{a_L + t_F + \lambda - v_F}{2}$ so that the latter can be rewritten as:

$$2a_H \frac{t_F}{t_G} + a_L > a(2\frac{t_F}{t_G} + 1),$$

which gives (16) and is surely satisfied when t_G/t_F is low enough. Under (9), (16) further simplifies to:

$$\frac{a_H - a_L}{1 + t_G/t_F} > 0,$$

and therefore PP leads to higher investment and to higher total content. **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 is certainly satisfied when the ratio t_F/t_G is large. PP shifts advertising resources to Google, and this produces an overall increase in total traffic when Google 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.

The condition for PP to result in higher investment is certainly satisfied under (9). For instance, if all CPs are equally efficient in the way they create content, the result is particularly clear: total traffic depends on the average advertising funds available (that, under (9), we assumed not to differ in the two regimes) and on the fee paid by the fringe: this fee goes down under PP, causing total traffic to increase. The result is even stronger in case Google is more efficient than the fringe in producing content, while it is diluted when Google is inefficient compared to the fringe ($t_G \gg t_F$), yet investment still increases under PP. The result comes from our assumption (9): if Google is very inefficient, then under NN the advertising rate would be very close to a_L and the fringe would not change much its behavior in the two regimes. Instead, Google would strictly bring more applications under PP.

Alternative assumptions on adverting rates in the two regimes would still keep the flavor of our findings. As an example, if instead of (9), one alternatively assumed that $a = \frac{a_H+a_L}{2}$ (i.e., with no reference to transportation costs), then (16) would be re-written as

$$(a_H - a_L)\left(\frac{t_F}{t_G} - \frac{1}{2}\right) > 0,$$

and therefore NN could lead to higher investment iff $t_G > 2t_F$. In this case google is "very" inefficient compared to the fringe so that advertising resources under PP are considerably driven away from the smaller but more efficient CPs: the increase in the number of applications supplied by google does not compensate for the reduction of content supplied at the edge by the fringe.

Having compared investment in capacity and total content provision under the

two regimes, we now complete the characterization of the properties of the long-run equilibrium:

Proposition 2 *Imagine capacity is endogenously determined. The long-run relation between the equilibrium variables in the two regimes is as follows:*

$$\begin{aligned}
f_L &< f < f_H, \\
x_G &< x_H,\ x_F > x_L, \\
W_H &< W(x_F, x_G) = \overline{W}(x_L, x_H) < W_L, \\
\Pi_F^{NN} &> \Pi_F^{PP}, \\
\pi_G^{NN} &> \pi_G^{PP} \textit{ iff } a + a_L > t_G, \textit{ which is always satisfied under } (9), \\
p^{NN} &< p^{PP} \textit{ iff } v_G \geq \frac{t_G(a - a_L)}{2t_F(a_G - a)} v_F, \textit{which simplifies to } v_G \geq v_F/2 \textit{ under } (9), \\
\Pi_{ISP}^{PP} &> \Pi_{ISP}^{NN} \textit{ iff } v_G \geq v_{ISP}, \textit{ where under } (9) \textit{ a sufficient condition is } v_G \geq v_F.
\end{aligned}$$

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. As far as the connection fees are concerned, a regulated regime increases the fee paid by firms in the fringe. However, what matters for the fringe is ultimately the higher advertising revenues. NN thus implies an increase in the participation at the edge. This also translates in higher aggregate profits for the fringe. Overall, a regime of NN benefits the CPs but does not increase overall content provision, given by the sum of x_F and x_G (see Proposition 1), which is lower than under PP because Google supplies less content. Moving towards a regime of PP, instead, kills part of the innovation done at the edge by the fringe. Small providers do pay a cheaper access fee, but they get a relatively higher penalty from reduced advertising rates. Conversely, Google gets higher advertising revenues which leads it to invest in more applications; however, the ISP appropriates some of Google's advertising rents by charging a premium fee that is higher than under NN which explains why the net profits of Google are likely to decrease. 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 Google's, the 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 amount of resources from advertising is not affected by the pricing regime and (9) holds, further results can be established. Google, despite developing more applications, now loses for sure under PP. As far the ISP is concerned, it typically benefits from supplying access with different priorities and price discriminate 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 Google is at least as good as the fringe's content.

To summarize, the main effect of net neutrality regulation is therefore to direct advertising resources toward 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 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 price paid by final users and by Google, and a decrease in the price paid by the fringe. Ultimately, the profits of the ISP increase and so do investment in capacity and total content supply.*

Proof. Recall that, when $t_F = t_G$, (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 $p = \frac{v(3a+\Delta+v-t-\lambda)}{2t} - \sqrt{d}$, which increases with ads dispersion. The same effect applies to capacity investment, as $\mu = \frac{\lambda(3a+\Delta+v-t-\lambda)}{2t} + \sqrt{d}$, and therefore also to total content provision. The opposite is true for $f_L = \frac{a-\Delta+t+\lambda-v}{2}$, which decreases in Δ. The fee to Google is $f_H = f_L + \frac{a_H^2 - a_L^2}{2t} = f_L + \frac{a(a_H - a_L)}{t}$. The first term decreases and the second increases in ads dispersion. However, from (12), it is $t < a$ and hence the second term always prevails. By simple substitution, it is immediate to find that the

profits of the ISP also increase both in the level and in the dispersion of advertising rates. It is in fact instructive to decompose the effects on total profits. Profits are made from consumers, with $\frac{\partial p}{\partial \Delta} = \frac{v}{2t} > 0$; from Google with $\frac{\partial f_H}{\partial \Delta} = \frac{4a-t}{2t} > 0$; from the fringe, with $\frac{\partial x_L f_L}{\partial \Delta} = -\frac{a-\Delta}{2t} < 0$. Investment also changes, with associated cost $\frac{\partial I}{\partial \Delta} = \frac{\lambda}{2t} > 0$. Overall, $\frac{\partial \Pi_{ISP}^{PP}}{\partial \Delta} = \frac{3a+\Delta+v-\lambda-t}{2t} > 0$. **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. This leads to both more content and more investment. Thus, it allows the ISP to make more money from the charges to end users, as well as extracting higher premium profits from Google. There is also a decrease in the amount that can be obtained from the fringe, but the first effects always prevail. 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:

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

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:

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

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

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

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

$$\begin{aligned} a + v_F - t_F(2x_F + 1) - \lambda &= 0, \quad (21) \\ a - t_G x_G &= 0, \quad (22) \\ -d + \frac{1}{W^2} &= 0. \end{aligned}$$

Comparisons are quite easy. The waiting time is socially optimal: the reason is that W impacts only on cost and 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 the fringe x_F is suboptimal and below x_F^*: on the one hand, the privately optimal fee f does take into account that the fringe's traffic has an impact on the waiting cost; however, the ISP does not extract all the surplus from the fringe, so the f it charges is "too high" from a social welfare perspective. The content x_G supplied by Google is also suboptimal: Google decides only on the basis of its advertising rate a and the f does not affect its choice; hence the ISP, differently from a social planner, cannot internalize the impact on users (v_G) or on congestion costs (λ). In case $v_G > \lambda$, which is assumed under (15), the evaluation of Google's content is higher than the corresponding marginal cost to keep waiting time constant: then also Google is underinvesting in content. Since both Google and the fringe are underinvesting, while the waiting time is the same as in the first best, it immediately follows that also investment in capacity is less than socially optimal: $\mu^{NN} < \mu^{*NN}$.

A similar analysis under PP implies that the first best allocation satisfies:

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

while the private equilibrium from Proposition 2 is:

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

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 fringe content is suboptimal and

below x_L^*, while Google content is also suboptimal and below x_H^* when $v_G > \lambda$. 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 *Under assumption (9): 1) When allocations are chosen by a social planner, PP is more efficient than NN iff $v_G \geq v^*$. 2) When allocations are chosen by an unregulated ISP, PP is more efficient than NN iff $v_G \geq \widetilde{v}_{ISP}$, with $\widetilde{v}_{ISP} < v^* < v_F$. Hence $v_G \geq v_F$ is a sufficient condition for PP to be the most efficient regime, both when allocations are chosen 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 evaluation for the fringe content is particularly high. PP welfare dominates NN, in particular, when Google and the fringe are equally efficient both in their content development costs ($t_F = t_G$) and in the value generated ($v_F = v_G$); in this case, while a PP regime would lead to an inefficient split of production of content (since Google would produce a higher proportion of the available content, therefore at higher marginal costs than in the symmetric split that would occur under NN), proportionally more resources are generated via advertising, and this effect prevails overall.

The second part of the proposition analyzes private allocations. In comparison to the socially optimal ones, the underinvestment of Google is *less* severe under PP: the extra terms $v_G - \lambda$ in expressions (24)-(26) are relatively smaller compared to a_H, with respect to (19)-(22) where $v_G - \lambda$ has to be compared to $a < a_H$. For similar reasons, the underinvestment of the fringe becomes instead exacerbated, as the terms $v_F - \lambda$ are relatively larger compared to a_L in (23)-(25) than to a in (18)-(21). As long as the content of the fringe is not too highly evaluated by users, Google ends up being larger: this effect makes PP more desirable. Notice finally that, according to Proposition 2, the monopolist prefers PP which is also the preferred regime of a social planner when v_G is high enough. The preference of both the monopolist and the social planner are reversed in favour of NN in case the value of 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.

5 Variable 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, and a decrease in advertising revenues otherwise. However, these premiums and penalties 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.[13] 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, $t_F = t_G = t$, Google 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 Google or by the fringe, $v_F = v_G = v$.

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

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

To set grounds and better illustrate the effects of endogenous advertising rates, we start working first with a specific example, and then turn to a more general analysis.

[13] Njoroge et al. (2010), similarly, suggest a positive relationship between advertising revenues and quality of the connection. Marketing research suggests 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.

We adopt the following functional form:

$$a(W) = \alpha + \frac{\beta}{W(x_F, x_G)} \tag{27}$$

under NN, where β is the sensitivity of advertising rates to (the inverse of) congestion. Under PP the dual rates are

$$a_L(\overline{W}) = \alpha + \frac{\beta(1-\delta)}{\overline{W}(x_H, x_L)} \qquad a_H(\overline{W}) = \alpha + \frac{\beta(1+\delta)}{\overline{W}(x_H, x_L)}, \tag{28}$$

where δ represents the relative advantage from advertising revenues when priority is chosen over best effort. Notice that indeed this specification preserves two properties, namely that $a_H > a_L$ and that, if average waiting time is the same, $a_H + a_L = 2a$. In addition, this specification has the property that $a'_H < a' < a'_L < 0$.

Under NN, the free entry condition for the fringe CPs is still obtained from (1). However, with the feed-back effect from advertising rates, the mass of the fringe is

$$\begin{aligned} 0 &= \alpha + \frac{\beta}{W(x_F, x_G)} - f - t x_F \\ \Longrightarrow \quad x_F &= \frac{\alpha + \beta\mu - f - \lambda\beta x_G}{t + \lambda\beta}. \end{aligned} \tag{29}$$

The demand from CPs in the fringe is less elastic compared to the case with exogenous advertising rates (it is still $\partial x_F/\partial f < 0$, but with $\partial^2 x_F/\partial f \partial\beta > 0$). The reason is simple: when f increases, there are less CPs entering the fringe, but this reduces congestion which, in turn, increases advertising rates. Hence the marginal provider in the fringe is made more price inelastic the higher is β.

Google maximizes (3), resulting in

$$\begin{aligned} 0 &= a + x_G \frac{\partial a}{\partial x_G} - t x_G \\ \Longrightarrow \quad x_G &= \frac{\alpha + \beta\mu - \lambda\beta x_F}{t + 2\lambda\beta}. \end{aligned} \tag{30}$$

In other words, also Google considers the impact that its content choice has on advertising revenues. While the fee f does not enter directly the mass of applications chosen by Google, it still has an indirect effect since the applications of Google and of the fringe are strategic substitutes. Though they do not compete directly against each other, applications cause congestion and reduce advertising revenues to all other CPs, creating an interdependency. Demand from CPs is thus obtained from solving (29) and (30), simultaneously. Due to the reduced congestion from the fringe when f

is increased, Google supplies more content the higher is f ($\partial x_G/\partial f > 0$). Recall instead that, if advertising was insensitive to congestion, then f would have no bearing on Google's choice.

Under PP, the demand for content is obtained in a conceptually similar way. We omit the details but it is immediate to prove that, as δ becomes positive, the effects are magnified: when f_L is increased, Google further increases its content. The reduced congestion from the fringe is particularly valuable to Google : under a priority scheme, in fact, the increase in advertising revenues due to the lower traffic is larger than in the neutral case and this acts to further increase the supply of Google. On the contrary, an increase in the fee for the best effort service hits fringe providers not only directly but also through the indirect effect on the advertising rate, more negative than in the NN case.

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 Google's content maximization determine:

$$\begin{aligned} x_F &= \frac{a(W)-f}{t}, \\ x_G &= \frac{a(W)}{t}. \end{aligned}$$

The ISP solves

$$\max_{f,W} \Pi_{ISP}^{NN} = \pi_{ISP}^{NN} - \mu = p^{NN} + f + f x_F - \mu$$

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

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

The equilibrium is represented by the following two first order conditions:

$$\begin{aligned} \frac{\partial \Pi_{ISP}^{NN}}{\partial f} &= 1 + x_F + (f+v)\frac{\partial x_F}{\partial f} + \frac{\lambda}{t} \\ &= 1 + \frac{a(W) - v - 2f + \lambda}{t} = 0, \qquad (31) \\ \frac{\partial \Pi_{ISP}^{NN}}{\partial W} &= v\frac{\partial (x_F + x_G)}{\partial W} - d + f\frac{\partial x_F}{\partial W} + \frac{1}{W^2} - 2\frac{\lambda}{t}a'(W) \\ &= \frac{2v - 2\lambda + f}{t}a'(W) - d + \frac{1}{W^2} = 0, \qquad (32) \end{aligned}$$

From (31) we obtain:

$$f = \frac{a(W) + t + \lambda - v}{2},$$

which has the same form as the equilibrium fee in Proposition 2. Comparing (32) with (17), it is apparent that congestion now depends on the way it affects advertising rates. Only if advertising was not sensitive to congestion ($a' = 0$), we would obtain again $W = 1/\sqrt{d}$ as in (17). With variable advertising rates, however, the equilibrium waiting time is defined implicitly by the condition

$$\frac{3(v-\lambda) + a(W) + t}{2t} a'(W) - d + \frac{1}{W^2} = 0. \tag{33}$$

The first term is negative, which implies that, compared again to Proposition 2, in equilibrium $W < 1/\sqrt{d}$.[14] The ISP has now further incentives to reduce congestion, on top of the benefits it can appropriate from the consumer's side, because lower congestion increases advertising funds that are appropriated by extracting rents from CPs. For the same reason, the equilibrium value of W is now affected by all the parameters of the problem (t, v, λ, besides d).

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

$$\begin{aligned} x_L &= \frac{a_L(\overline{W}) - f_L}{t}, \\ x_H &= \frac{a_H(\overline{W})}{t}. \end{aligned}$$

The ISP solves

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

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

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

[14] Notice that an interior equilibrium requires the second order condition on the waiting time: $\frac{3(v-\lambda)+a+t}{2t} a'' + \frac{a'^2}{2t} - \frac{2}{W^3} \leq 0$ to be satisfied.

The equilibrium is represented by the following two FOCs:

$$\begin{aligned}
\frac{\partial \Pi_{ISP}^{PP}}{\partial f_L} &= 1 + x_L + (f_L + v)\frac{\partial x_L}{\partial f} + \frac{\lambda}{t} \\
&= 1 + \frac{a_L(\overline{W}) - v - 2f_L + \lambda}{t} = 0, \\
\frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} &= v\frac{\partial (x_L + x_H)}{\partial \overline{W}} - d + f_L\frac{\partial x_L}{\partial \overline{W}} + \frac{a_H(\overline{W})a'_H(\overline{W}) - a_L(\overline{W})a'_L(\overline{W})}{t} + \\
&\quad + \frac{1}{\overline{W}^2} - \frac{\lambda}{t}[a'_H(\overline{W}) + a'_L(\overline{W})] \\
&= \frac{v - \lambda + f_L - a_L(\overline{W})}{t}a'_L(\overline{W}) + \frac{v - \lambda + a_H(\overline{W})}{t}a'_H(\overline{W}) - d + \frac{1}{\overline{W}^2} = 0.
\end{aligned}$$

We thus obtain:

$$f_L = \frac{a_L(\overline{W}) + t + \lambda - v}{2}.$$

If advertising was not sensitive to congestion ($a'_i = 0$), again $\overline{W} = 1/\sqrt{d}$. With variable advertising rates, the equilibrium waiting time is defined implicitly by

$$\frac{v - \lambda - a_L(\overline{W}) + t}{2t}a'_L(\overline{W}) + \frac{v - \lambda + a_H(\overline{W})}{t}a'_H(\overline{W}) - d + \frac{1}{\overline{W}^2} = 0. \qquad (34)$$

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) If v is large enough, a necessary and sufficient condition for the abolition of NN to lead to lower congestion and higher investment is that advertising rates for prioritized traffic are more sensitive to congestion than advertising rates under NN, i.e., $|a'_H| > |a'|$. 3) In the example (27)-(28), congestion is always lower and investment higher under PP compared to NN. 4) A necessary, but not sufficient, condition for congestion to be lower and investment to be higher under NN compared to PP is $|a'_L| > |a'| > |a'_H|$.*

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 that the ISP has an incentive to adopt prioritized

traffic and invest more, particularly when this allows to redirect advertising resources towards Google, is confirmed. NN favours 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 extend 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) has also an intuitive explanation. If v is high enough, the effect that prevails must come from the price charged to end users which, in turn, depends on the sensitivity of total content to congestion. PP leads to higher incentives to invest if and only if $|a'_H| > |a'|$, i.e., when advertising rates for prioritized traffic are more sensitive to congestion than advertising rates under NN.

Part 3) shows that, if it holds that $|a'_H| > |a'| > |a'_L|$, PP leads to higher investments and lower congestion more in general. In fact, the result holds no matter the value of v. In this case, a decrease in congestion under PP increases the dispersion of advertising rates, and leads to further funds attracted to Google's applications: in other words, both the level and the dispersion effect go in the same direction.

Finally, part 4) can be seen as a robustness check. Indeed, a reversal in the sensitivity of ads to congestion is needed to overturn the main finding: if priority decreases, rather than increases, the sensitivity of ad rates net neutrality leads to lower congestion and higher investment. Notice further that the condition devised is a necessary but not sufficient result as the "level" effect, whereby advertising rates increase for Google under PP, is still present.

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 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 welfare function is:

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

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

Under PP, social welfare can be written as:

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

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

After substitution and rearranging terms, we obtain

$$\begin{aligned} \Delta SW &= SW^{NN} - SW^{PP} = \underbrace{-(a_H - a_L)\frac{5(a_H - a_L) + 4(a_H + t + v - \lambda)}{32t}}_{AD_1 < 0} \\ &\quad \underbrace{-(a_H + a_L - 2a)\frac{14a + 7a_H + 7a_L - 4t + 28(v - \lambda)}{32t}}_{AD_2} \\ &\quad \underbrace{-(W^{NN} - \overline{W}^{PP})(d - \frac{1}{W^{NN}\overline{W}^{PP}})}_{WT}. \end{aligned}$$

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). AD_1 is always negative. 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_2 would also be equal to zero. 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 in general, when ad rates are sensitive to congestion, we have the following result on which regime is welfare dominant.

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 (33) and (34), under either regime it is always $W < 1/\sqrt{d}$. Hence $d - \frac{1}{W^{NN}\overline{W}^{PP}} > 0$ 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 $a_H(\overline{W}^{PP}) + a_L(\overline{W}^{PP}) = 2a(\overline{W}^{PP}) > 2a(W^{NN})$ and $AD_2 < 0$. Thus $W^{NN} > \overline{W}^{PP}$ is a sufficient condition for $\Delta SW < 0$. **Q.E.D.**

An intuitive interpretation for 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 increased waiting time 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.

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 favour and against net neutrality. We proposed a model that formalizes prioritization at the intersection between Operations Management and Marketing. 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 economic effects.

In our model, 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 favour 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 ("Google") 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 the "last mile" of the Internet seems relatively uncompetitive, it would be desirable to extend our model to the case of competing ISPs. Second, advertising rates could be endogenized by modelling the demand and supply of advertising space: this extension constitutes a challenge for further research. Finally, as our focus is on innovation and investment incentives, all (identical) users access the Internet and are extracted all surplus. A discussion of users' welfare would require to capture both their heterogeneity and more sophisticated pricing structure faced by users.

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] 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.

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

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

[8] Economides N. and J. Tag (2009), Net Neutrality on the Internet : A Two-sided Market Analysis, NET Institute Working Paper, 07-45.

[9] Fahri E. and A. Hagiu (2007), Strategic Interactions in Two-Sided Market Oligopolies, Harvard University Strategy Unit Working Paper, 08-11.

[10] 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.

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

[12] 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.

[13] Kramer J. and L. Wiewiorra (2010), Network Neutrality and Congestion Sensitive Content Providers, mimeo.

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

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

[16] 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.

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

[18] 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.

[19] Njoroge P., Ozdaglar A., Stier-Moses N.E. and G.Y. Weintraub (2010), Investment in two sided markets and the net neutrality debate, mimeo.

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

[21] 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.

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

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

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

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

[26] 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 $W = 1/\sqrt{d}$, and thus $\mu - \lambda(x_F + x_G) = \sqrt{d}$ under NN. The FOC w.r.t. f is:

$$1 + x_F + (f + v_F - \lambda)\frac{\partial x_F}{\partial f} = 0,$$

which, after substitution, is solved to get:

$$f = \frac{a + t_F + \lambda - v_F}{2}.$$

This determines the capacity level and characterizes the equilibrium fully:

$$\begin{aligned}
W &= \frac{1}{\sqrt{d}}, \quad \mu = \frac{\lambda\left[a(2\frac{t_F}{t_G}+1) + v_F - t_F - \lambda\right]}{2t_F} + \sqrt{d}, \\
p &= a\left(\frac{v_F}{2t_F} + \frac{v_G}{t_G}\right) + v_F\left(\frac{v_F - t_F - \lambda}{2t_F}\right) - \sqrt{d}, \\
x_G &= \frac{a}{t_G}, \quad x_F = \frac{a + v_F - t_F - \lambda}{2t_F}, \\
\Pi_F &= \frac{(a-f)^2}{2t_F} = \frac{(a - \lambda - t_F + v_F)^2}{8t_F}, \ \pi_G = \frac{a^2}{2t_G} - \frac{a + \lambda + t_F - v_F}{2}, \\
\Pi_{ISP} &= \frac{(a + v_F - \lambda - t_F)^2}{4t_F} + \frac{a\,(v_G + t_G - \lambda)}{t_G} - 2\sqrt{d}.
\end{aligned}$$

Under PP, the equilibrium is solved similarly, to get

$$\begin{aligned}
\overline{W} &= \frac{1}{\sqrt{d}}, \ \ \mu = \frac{\lambda(2a_H\frac{t_F}{t_G} + a_L + v_F - t_F - \lambda)}{2t_F} + \sqrt{d}, \\
f_L &= \frac{a_L + t_F + \lambda - v_F}{2}, \quad f_H = \frac{a_L + t_F + \lambda - v_F}{2t_G} + \frac{a_H^2 - a_L^2}{2t_G}, \\
p &= a_L\frac{v_F}{2t_F} + a_H\frac{v_G}{t_G} + v_F\left(\frac{v_F - t_F - \lambda}{2t_F}\right) - \sqrt{d}, \\
x_H &= \frac{a_H}{t_G}, \quad x_L = \frac{a_L + v_F - t_F - \lambda}{2t_F}, \\
W_H &= 2\frac{t_F}{\lambda a_L - \lambda^2 - \lambda t_F + \lambda v_F + 2\sqrt{d}t_F}, \\
W_L &= \frac{\lambda^2 t_G - 2\sqrt{d}t_F t_G - 2\lambda a_H t_F - \lambda a_L t_G + \lambda t_F t_G - \lambda t_G v_F}{t_G\sqrt{d}\left(\lambda^2 - \lambda a_L + \lambda t_F - \lambda v_F - 2\sqrt{d}t_F\right)}, \\
\Pi_F &= \frac{(a_L - f_L)^2}{2t_F} = \frac{(a_L - \lambda - t_F + v_F)^2}{8t_F}, \quad \pi_G = \frac{a_L^2}{2t_G} - \frac{a_L + \lambda + t_F - v_F}{2}, \\
\Pi_{ISP} &= \frac{(a_L + v_F - \lambda - t_F)^2}{4t_F} + \frac{a_H\,(v_G + t_G - \lambda)}{t_G} + \frac{(a_H - a_L)(a_H + a_L - 2t_G)}{2t_G} - 2\sqrt{d}.
\end{aligned}$$

The results for the first part follow from simple comparisons of the relevant expressions. In particular $f > f_L$ since $a > a_L$. Also $f < f_H$ if $(a_H + a_L)(a_H - a_L) > t_G(a - a_L)$ which is surely satisfied under (12). $a_H > a > a_L$ 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 $v_G \geq \frac{t_G(a-a_L)}{2t_F(a_G-a)} v_F$. Under (9), this further simplifies to $v_G \geq \frac{v_F}{2}$.

Turning to the total profits of the fringe, again $a > a_L$ ensures that $\Pi_F^{NN} > \Pi_F^{PP}$. Google's profits compare as follows:

$$\pi_G^{NN} - \pi_G^{PP} = \frac{1}{2}(a - a_L)\frac{a + a_L - t_G}{2t_G} > 0 \text{ iff } a + a_L > t_G.$$

Under (12), the last inequality is rewritten as $t_F(a_L + a_H) + t_G(2a_L - t_F - t_G) > 0$, which is satisfied under (9).

Also the profits of the ISP in the two regimes can be compared:

$$\Pi_{ISP}^{PP} - \Pi_{ISP}^{NN} = \frac{(a_L - a)(a + a_L + 2v_F - 2\lambda - 2t_F)}{4t_F} + \frac{(a_H - a)(v_G + t_G - \lambda)}{t_G} + \frac{(a_H - a_L)(a_H + a_L - 2t_G)}{2t_G}.$$

The first term is negative, the second is positive as well as the third (because of (12)). Using assumption (9), profits are higher under priority iff

$$v_G \geq v_{ISP} = \frac{v_F + \lambda + t_F}{2} - \frac{t_F(a_L + a_H) + t_G a_H}{2t_G} - \frac{(a_H - a_L)t_F}{4(t_F + t_G)}.$$

A sufficient condition for this inequality to be always satisfied is $v_G \geq v_F$. **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

$$\begin{aligned} SW^{NN} - SW^{PP} &= -\frac{(a_H - a_L)(a_H - a_L + 2v_G - 2v_F)}{2(t_F + t_G)} \\ &\quad - \left(\underbrace{a_H \frac{t_F}{t_F + t_G} + a_L \frac{t_G}{t_F + t_G} - a}_{=0} \right) \cdot \\ &\quad \left(a_H \frac{t_F}{t_F + t_G} + a_L \frac{t_G}{t_F + t_G} - a - 2\lambda + \frac{2t_G v_F + 2t_F v_G}{t_F + t_G} \right) \frac{t_F + t_G}{2t_F t_G} \end{aligned}$$

Under assumption (9) the second term in brackets is equal to zero. Hence a sufficient condition for also the first bracket to be negative is $v_G \geq v_F$. For lower values, the welfare difference is still negative for any

$$v_G \geq v^* = v_F - \frac{a_H - a_L}{2}.$$

Consider next the expressions of the social welfare in the two different regimes, when allocations are determined by the ISP. Under assumption (9), the expression for the difference is:

$$\begin{aligned} & SW^{NN} - SW^{PP} \\ = \; & -\frac{1}{8}\frac{a_H - a_L}{(t_F + t_G)^2}\left[2(t_F + t_G)(3(v_G - v_F) + t_F + v_G - \lambda) + (5a_H - 3a_L)\, t_F + (4a_H - 2a_L)t_G\right] \end{aligned}$$

which implies that $SW^{PP} \geq SW^{NN}$ holds for sure when $v_G \geq v_F$. For lower values, the welfare difference is still negative for any

$$v_G \geq \widetilde{v}_{ISP} = v_F - \frac{a_H - a_L}{2} - \frac{(a_H - a_L)t_F}{8(t_F + t_G)} - \frac{a_L + v_F + t_F - \lambda}{4},$$

with $\widetilde{v}_{ISP} < v^*$. **Q.E.D.**

Proof of Proposition 4. The results on total content follow from comparing $x_F + x_G = \frac{3a(W)+v-t-\lambda}{2t}$ under NN with $x_L + x_H = \frac{2a_H(\overline{W})+a_L(\overline{W})+v-t-\lambda}{2t} = \frac{2a(\overline{W})+a_H(\overline{W})+v-t-\lambda}{2t}$ under PP. If one can prove that $\overline{W}^{PP} < W^{NN}$, then it immediately follows that total content supply increases under PP since then $a_H(\overline{W}) > a(\overline{W}) > a(W)$.

Consider next congestion, given by (33) and (34) in the two regimes. The only terms that matter are respectively

$$\begin{aligned} A^{NN} &= \frac{3(v-\lambda) + t + a}{2t} a', \\ A^{PP} &= \frac{v-\lambda}{2t}(a'_L + 2a'_H) + \frac{1}{2}a'_L + \frac{1}{2t}(2a_H a'_H - a_L a'_L), \end{aligned}$$

where, to avoid clutter, we have dropped the dependence of ads on waiting time. The results on capacity investment follow from noting that

$$\begin{aligned} \mu^{NN} &= \frac{1}{W} + \frac{\lambda}{t}\frac{3a(\cdot) - t + v - \lambda}{2t}, \\ \mu^{PP} &= \frac{1}{\overline{W}} + \frac{\lambda}{t}\frac{2a(\cdot) + a_H(\cdot) - t + v - \lambda}{2t}. \end{aligned}$$

Since $a' < 0$ and $a'_H < 0$, 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 $2a_H - a_L > a$, we have that $A^{PP} < A^{NN}$, and hence $\overline{W}^{PP} < W^{NN}$.

2) If v is high enough, only the first terms in A^{NN} and A^{PP} above matter for the comparison. Since from $a_L(\cdot) + a_H(\cdot) = 2a(\cdot)$ it is also $a'_L + a'_H = 2a'$. Thus $a'_L + 2a'_H = 2a' + a'_H < 3a'$ iff $a'_H < a'$.

3) In the specific example we obtain

$$\begin{aligned} A^{NN} &= \frac{3(v-\lambda)+t+a}{2t}a', \\ A^{PP} &= \frac{(v-\lambda)(3+\delta)+t(1-\delta)+a(1+3\delta)+\frac{\delta\beta}{\overline{W}}(3+\delta)}{2t}a'. \end{aligned}$$

Every single term is bigger in absolute value in A^{PP} compared to the corresponding one in A^{NN}. There is only once exception ($-\delta t$), but because of assumption (12), which can now be re-written as $a > t$, this term is more than compensated by δa. Hence $A^{PP} < A^{NN}$, and $\overline{W}^{PP} < W^{NN}$ and $\mu^{PP} > \mu^{NN}$. In particular, in this case, after simple substitution, it is $\mu^{PP} - \mu^{NN} = \frac{W^{NN} - \overline{W}^{PP}}{W^{NN}\overline{W}^{PP}} + \frac{\beta\lambda[3(W^{NN} - \overline{W}^{PP}) + \delta W^{NN}]}{2tW^{NN}\overline{W}^{PP}} > 0$.

4) For the last part, imagine $a_H = a(1+x)$ and $a_L = a(1-x)$ where $x > 0$. Also, write $a'_H = a'(1+X)$ and $a'_H = a'(1-X)$, where a priori the sign of X is not determined. Then

$$A^{PP} = \frac{(v-\lambda)(3+X)+t(1-X)+a[1+3(X+x)+xX]}{2t}a'_L.$$

Using again (12), then a sufficient condition for $A^{PP} < A^{NN}$ is $X > 0$. It is only when $X < 0$ that the sign could be eventually reversed. **Q.E.D.**