Net neutrality and innovation at the core and at the edge

Carlo Reggiani[*] Tommaso Valletti[†]

15th May 2011 - Very preliminary and incomplete

Abstract

Net neutrality is a hotly debated topic. A key point is the treatment of the data by internet providers: as users are affected by the waiting times of data delivery, internet providers have an incentive to offer content providers different priorities. Net neutrality regulation does not allow prioritization. We study the effect of regulation on the incentives to innovate of content and service providers. In the short run, regulation increases content provision at the edge by a fringe, while it decreases the number of applications of a large provider. In the long run, the internet provider adjusts capacity to maintain constant the average waiting time. Regulation leads to lower supply of capacity and overall content, although it fosters entry of new content providers.

JEL code: D4, L12, L4, L43, L51, L52.

Keywords: net neutrality, congestion, innovation.

[*] University of Manchester. Address: School of Social Sciences - Economics, University of Manchester, Sir Arthur Lewis Building, 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.

1 Introduction

Net neutrality (NN) is a hotly debated topic. President Obama expressed his views in favour of net neutrality and his government committed to approve a regulation. The UK recently supported a contrarian view, and also the EU decided not to take further measures to guarantee net neutrality. Why such huge differences? First of all because the concept of NN is a bit ambiguous and different people mean different things. Secondly, the debate has been mostly led by informal policy reports and advocacy pieces, therefore there is a need to conduct more rigorous analysis to clarify the various issues at stake – within the context of specific but hopefully relevant models.

What is network neutrality? Generally speaking, proponents identify it with a set of rules needed to guarantee the openness and freedom of access to the internet, the key elements of the early success of the internet, now possibly jeopardized in the broadband era. Although NN does not have a widely accepted definition, it usually refers to a restriction on how ISPs interact with the CPs and end-users. Kruse (2010) categorizes economic relevant meanings of net neutrality in five decreasing levels of neutrality. The stricter interpretation interprets neutrality as the identical treatment of every data packet at any router; a slightly weaker definition allows for different prices of different priorities but an equal price for the same priority; a third interpretation involves the same price for all services but a different treatment of data packets depending on the internet service provider (ISP) needs for traffic management; a fourth possibility is to let ISPs charge different prices for different services within the same data packet priority, allowing a full discrimination. Finally, the maximum breach of neutrality would consist in different treatment and service for data packets with the same priority. No matter the strictness in the definition adopted, the key point seems the treatment of data packets.

In this paper we do not consider vertical integration between the ISP and content providers (CPs). Thus we do not deal with foreclosure issues. Rather, we investigate NN defined as a restriction on particular pricing structures and network management techniques, and its implication on investment choices of both ISPs and innovation incentives of CPs.

While the economic literature is scarce, yet there is some relevant work available. In particular, Kim and Choi (2010) and Cheng et al. (2011) develop models where non-NN allows a ISP to charge differentially for access by CPs to its network, according to the priority chosen by the CPs individually. They find that non-NN induces higher investment in capacity on the side of the ISP.[1] In their approach, two advertising-funded CPs (CP1 and CP2) have to decide whether or not to pay for prioritized traffic when connecting to a monopolist ISP. A key feature of their models is that end users can see only one CP. This assumption holds both with and without NN. If – say – CP1 opts for priority and CP2 does not, this implies that a end user has to decide whether to see only the content of CP1 (with lower congestion) or only the content of CP2 (with higher congestion), but never both contents. This assumption of "exclusivity" of users to a single CP is needed to provide incentives to CPs to opt eventually for prioritized traffic (as end users will then react, since ceteris paribus they prefer a CP with lower congestion). While this might be a characterization of particular situations where content providers are a substitute between each other (e.g., a subscriber might 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 modeled as mutually exclusive

[1] Two of the main differences between these models lie in: (i) the contract space, Choi-Kim (2010) assume only one CP can get the priority; (ii) the characterization of the waiting times in presence of prioritization. Both Cheng et al. (2011) and Choi and Kim (2010) impose a zero price under NN, hence the ISP can make money from CPs only under non-NN. This assumption is also shared by Kramer and Wiewiorra (2010), who study a monopolist ISP and a continuum of vertically-differentiated CPs with different sensitivities to congestion.

choices. This is one of the very defining features of the Internet: end users can access all the content available. Even in the recent European debate, it was clearly stated that network management techniques, departing from NN, will be endorsed only to the extent that the Internet remains "open".[2]

Economides and Tag (2009) do allow for CPs to contact and be seen by all the users. In their analysis, NN or its absence essentially corresponds to a price that the ISP can charge to CPs (under NN, CPs do not pay to be connected to the ISP), but without any difference in the quality/priority of connections offered by the ISP in the different regimes. In this sense, their analysis is static and their model is not well suited to analyze investment choices. Overwhelmingly, the available literature deals with a monopoly ISP, as we also will. This is perhaps justifiable by the local monopoly that ISPs typically have over the last mile.[3]

A contribution that stands on its own is Hermalin and Katz (2007) which considers NN as a restriction on the product line that an ISP can offer. The results suggest that any quality restriction is likely to reduce welfare. The paper, however, does not consider congestion and this may explain the result. Canon (2009) consider a slightly different context: a two-sided market in which sellers trade with both the platform and with buyers. Net neutrality corresponds to the case where the platform cannot charge sellers. As regulation encourages content supply, it has a positive welfare effect.

In our model, we allow all end users to see all the content available, with or without NN. That is, the Internet is always "open" in our model. The

[2] For more on this, refer to the statements of the British culture minister Ed Vaizey (http://www.bbc.co.uk/news/uk-politics-11773574) and Neelie Kroes, slowing down on the need of the EU of new NN rules (http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/10/643).

[3] Economides and Tag (2009) also consider, in an extension, a duopoly model of competition between ISPs. In the duopoly model they reintroduce the "mutual exclusivity" between users and a particular CPs, in contrast with their monopoly model (and independently from having or not having NN).

incentive to prioritize or not arises from two sources. First, advertising revenues vary with the level of congestion.[4] Second, the difference in congestion induces the monopolist ISP to change its prices to users.

On top of studying the incentive of the ISP to invest in capacity or "core" infrastructure – which is one of the fundamental problems surrounding the debate on NN – we also look at the incentives on the side of CPs. In particular, one characteristic of the Internet is that CPs are very heterogeneous. A few CPs (e.g., google, youtube) generate a lot of traffic (thanks to the numerous applications they provide), while there are many CPs that generate, individually but possibly not in aggregate, little traffic. We capture this by having a large CP and a fringe made of many atomistic CPs. The debate over NN has mentioned the point that innovations are made at the "edge" by CPs.[5] Allegations have been made both that departures from NN may lead to less innovation at the "edge", meaning less entry by CPs, as well as opposite claim of "crowding out", i.e., some applications may actually not develop at all under NN (e.g., applications very sensitive to delays and latency). In our model, we therefore study decisions both at the "core" and at the "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. We find that, for a given level of capacity, net neutrality fosters the provision of content at the edge; the overall content available is however reduced in the short run and the only benefit for final users is less congestion. In the long run, the ISP adapts capacity to the congestion. Net neutrality reduces the

[4] According to Njoroge et al. (2010) advertising revenues and quality of content are related: consumers have a better experience with high-quality platforms and, therefore, they spend more time online which increases the advertisers' brand exposure; thus advertisers are willing to pay more. Our assumption goes along similar lines as the quality perceived by final users depends on congestion in our context. Athey et al. (2010) provide a thorough analysis of advertising in media markets.

[5] Sydell (2006) is the first to focus the attention on innovation and content supply at the edge; the issue is then more formally analyzed by Bandyopadyhay et al. (2010), and Kramer and Wiewiorra (2010).

profits and the investment of the ISP, and still increases content provision by the fringe at the edge but not overall. Both google and the fringe have higher profits with net neutrality.

The rest of the paper is structures as follows. Section 2 presents the model and Section 3 analyzes it. Section 3.1 considers the effects of regulation in the short run. Section 3.2 deals with regulation under endogenous capacity choice of the ISP. Section 4 conducts the analysis when congestion affects advertising revenues. Section 5 provides a few concluding comments.

2 The model

Our model consists of a monopoly platform (ISP) that connects users (the Internauts) with the content providers (CPs). The ISP first invests in capacity μ at a cost $I(\mu)$ which is increasing in μ. 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. 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 Hotelling 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, $t_F(x) = t_F x$.[6] If f denotes the connection fee paid to the ISP, then the profit

[6] The linearity assumption is adopted to simplify the following analysis. Many of the

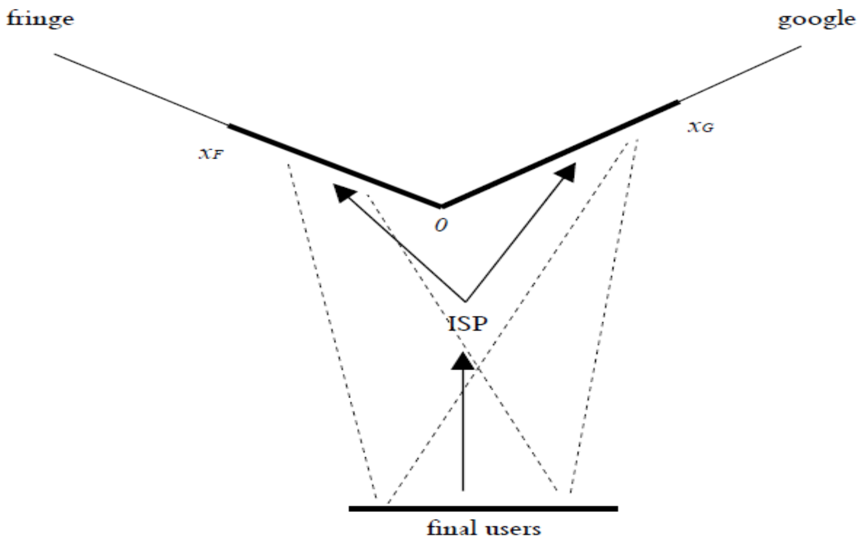

Figure 1: A stylized representation of the internet.

of a firm in the fringe that gets advertising revenues 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

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

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

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

results, however, can be generalized to a generic non-decreasing transport cost.

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. Notice how 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 is important as 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. Also notice that 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 ore less efficient providers.

2.1 Congestion

There is a unit mass of consumers that always connect to the entire content available over the Internet. This is true independently of possible prioritization of content; the latter, however, affects congestion. Consumers pay a connection fee p to the ISP. Consumers enjoy variety, which we model by assuming that each consumer enjoys a fixed benefit v per available application.[7] Consumers also care about congestion on the network.

Congestion depends on the total traffic exchanged, as well as 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,

[7] It is likely that consumers evaluate differently the content provided by google and by the fringe: the model can be generalized to allow for this.

2010; Kramer and Wiewiorra, 2010). Each user-CP exchange generates an amount of traffic λ. Under Net Neutrality (NN), congestion is

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

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

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

where d is consumers' sensitivity to congestion.

With Priority Pricing (PP), the ISP can offer priority to traffic. If a mass x_H (respectively, x_L) of CPs chooses to prioritize traffic (respectively, does not choose to prioritize), the utility of consumers is

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

where the congestion $\overline{W}(x_H, x_L)$ is given by the weighted average of waiting times (with and without priority). 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

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

We note here that this way of modelling traffic implies that the *average* congestion is the same in the two regimes, for the same capacity level and for the same total amount of traffic exchanged.[8] This is an important property that we must stress: PP, per se, does lead to an efficiency improvement over NN, but just to a reallocation of resources. However, the two regimes will give

[8] This is a property of M/M/1 queuing systems. It can be checked immediately by comparing (6) and (8). When capacity is the same, it is $\overline{W} = W$ when $x_G + x_F = x_H + x_L$.

different incentives to invest in μ, and therefore will affect average congestion for an endogenous choice of μ. Also notice another property of the queuing system whereby, 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 debated over net neutrality and that the model can capture.

2.2 Advertising

Differences in congestion and priority also affect the profitability of advertising rates. This is the mechanism that we introduce in order to give 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 equally valued by users, and they are arrive to the end users with the same congestion probability. Without NN, as motivated by our earlier discussion, there will be differences between the rates a_H and a_L for the prioritized traffic and for the best-effort content. For now, we do not put additional structure on these advertising functions and simply assume that $a_L < a_H$, as traffic with priority suffers less from congestion problems.

We also additionally assume that

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

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. This assumption mirrors the previous result concerning the physical infrastructure that 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 two-stage game where the monopolist first sets μ, then it sets prices to CPs and to end users.

3 Analysis

We first conduct a static analysis, considering capacity as fixed, and compare the welfare properties of the two regimes in terms of impact on CPs, users, and ISP. In Section 3.2, we then consider the long-run investment choice in capacity.

3.1 Fixed network capacity

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

$$p^{NN} = (x_F + x_G)v - dW(x_F, x_G). \tag{10}$$

In case priority pricing is allowed, the charge to final users is:

$$p^{PP} = (x_H + x_L)v - d\overline{W}(x_H, x_L). \tag{11}$$

The CPs subgame is more effectively analyzed by distinguishing between the cases of net neutrality and priority pricing.

3.1.1 Net Neutrality

The ISP profits are:

$$\pi_{ISP}^{NN} = p^{NN} + f + fx_F. \tag{12}$$

Substituting expressions (10), (2), and (4) into (12), and differentiating, one gets the following first order condition for f:

$$\frac{\partial \pi_{ISP}^{NN}}{\partial f} = 1 + x_F + \left[f + v - d\frac{\partial W}{\partial x_F} \right] \frac{\partial x_F}{\partial f} = 0, \tag{13}$$

where only the content of the fringe is affected by f, as $x_F = \frac{a-f}{t_F}$ while $x_G = \frac{a}{t_G}$. The first three terms corresponds to the marginal revenue from increasing f to (all) the CPs. The last two terms instead capture the adjustment to the consumers' price as a higher f both reduces the content available (which brings a value v to consumers), as well as improving congestion (which is weighted by d). 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.[9]

3.1.2 Priority Pricing

The ISP profit function is:

$$\pi_{ISP}^{PP} = p^{PP} + f_H D_H + f_L D_L \tag{14}$$

where D_L and D_H denote the demand for the high and the low priority respectively. Notice that, since a fee is charged per CP and not per application, if the low priority 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 low priority connection if:

$$f_H - f_L \geq a_H - a_L. \tag{15}$$

[9] At the endogenously chosen capacity level that we analyze in section 3.2, this condition is given by $(v - \lambda - t_F)^2 + 2at_G(v - \lambda + t_F) + a^2(t_G - 2t_F) > 0$, which is always satisfied when v is high enough.

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 LHS is the profit of google with priority, while the RHS is its profit without priority.[10] 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 takes the value (5), after substitution, this gives:

$$f_H = f_L + \frac{a_H^2 - a_L^2}{2t_G}. \tag{16}$$

In other words, the priority fee extracts all the extra rent from google, but it does not affect google's choice of content. Condition (15) to ensure self-selection of the fringe to low priority then becomes:

$$a_H + a_L \geq 2t_G, \tag{17}$$

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 $t_G = t_F$, we can further use the simplified version of (9), to obtain for (17) a very obvious condition

$$a \geq t_G.$$

To sum up, under condition (17), Google opts for priority while the fringe sticks to the unprioritized alternative. The ISP problem (14) is then:

$$\pi_{ISP}^{PP} = p^{PP} + f_H + f_L x_L,$$

where (11) and (16) hold. The low priority fee f_L is determined by:

$$\frac{\partial \pi_{ISP}^{PP}}{\partial f_L} = 1 + x_L + \left[f_L + v - d\frac{\partial \overline{W}}{\partial x_L} \right] \frac{\partial x_L}{\partial f_L} = 0, \tag{18}$$

where $x_L = \frac{a_L - f_L}{t_F}$ and $x_H = \frac{a_H}{t_G}$.

[10] Notice that, with this solution, we are assuming that google takes the dual advertising rates a_H and a_L as given. In other words, we do not consider that, if google deviates and adopts low priority, then this can induce the advertising market to realize that no CP has chosen priority, and hence advertising rates will fall back to the level under NN.

3.1.3 The short-run effects of NN regulation

To analyze the effects of net neutrality regulation, the equilibrium expressions in the two regimes are compared.

Proposition 1 *Imagine capacity is fixed, and average advertising rates with and without NN satisfy (9). Then the short-run relation between the equilibrium variables in the two regimes is given by:*

$$\begin{aligned}
f_L &< f < f_H \\
x_G &< x_H,\ x_F > x_L \\
x_F + x_G &< x_H + x_L \\
W(x_F, x_G) &< \overline{W}(x_L, x_H) \\
p^{NN} &\lesseqgtr p^{PP} \\
\pi_G^{NN} &> \pi_G^{PP},\ \overline{\pi}_F^{NN} > \overline{\pi}_F^{PP} \\
\pi_{ISP}^{NN} &< \pi_{ISP}^{PP}.
\end{aligned}$$

Proof. See the Appendix.

The results provided suggest that NN regulation has important effects already in the short run. First of all, as far as the connection fees are concerned, a best-effort regime reduces the charge of G while increasing the fee paid by firms in the fringe. NN thus implies an increase in the participation at the edge, also translating in higher overall profits for the fringe. G gets instead lower revenues from advertising under NN, which implies a decrease in its content supply. 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, which is lower than under PP. Since we conducted the analysis for a fixed level of capacity, this explains the result that there is less congestion and a decrease in the average waiting time under NN.

Moving towards a regime of PP, instead, kills part of the innovation done at the edge by the fringe. They do pay cheaper access fee, but they get a relatively higher penalty from reduced advertising rates. Conversely, google gets much higher advertising revenues which leads it to invest in more applications; however, the ISP appropriates most of these rents with the premium fee, which explains why net profits of google go down.

When the ISP is not able to discriminate, regulation decreases its profits, while overall profits increase under PP. End users always have their consumer surplus completely extracted both with and without NN. The price that they pay typically goes up with PP, reflecting the higher benefits they enjoy from more applications. Only if the disutility from higher congestion (due to higher content) is sufficiently high, then this result can be reversed.

From a static perspective, the main effect of net neutrality regulation is therefore to direct advertising resources toward the fringe. There is more entry of new content providers in the fringe or, in other words, innovation at the edge, while it kills innovation done by google. It remains to be seen what happens to investment in core infrastructure, but before turning the analysis to this we briefly recap our static results with a numerical example.

Example 1

Suppose that the rate of data arrival, the unit transport costs and the disutility of waiting are normalized to one, so that: $\lambda = t_F = t_G = d = 1$. Also, the final users evaluation is $v = 10$, which is sufficiently high for the ISP to serve all CPs, instead of extracting all the surplus from google alone. The advertising rates are $a = 10$ in the neutral case, $a_L = 8$ for the best effort and $a_H = 12$ if the CP opts for the priority. Also consider a capacity of $\mu = 20$.[11]

[11] As we show below, this value of μ is exactly the endogenous choice of capacity under NN when we solve this numerical example also in the long run, for a given specification of the investment cost.

The net neutral equilibrium is described by the following fee to CPs and price to users:

$$f = 1 \qquad p = 189.$$

This is reflected into content supply from the fringe and google:

$$x_F = 9 \qquad x_G = 10,$$

with a total supply of content equal to $x_F + x_G = 19$. With a capacity of 20, the average waiting time is 1.

The total profits of the fringe, of the ISP, and of google are respectively:

$$\overline{\pi}_F = 40.5 \qquad \pi_{ISP} = 199 \qquad \pi_G = 49.$$

The non-neutral equilibrium, instead, is characterized by:

$$f_L = 0.66 \qquad f_H = 40.66 \qquad p = 191.91$$

that is, the fringe pays less and google pays much more to connect to the Internet. End users also pay more.

Content supply from the fringe and google is now:

$$x_F = 7.34 \qquad x_G = 12,$$

showing that the reduced fee for the fringe does not compensate for the lower advertising revenues. Google, instead, produces more content as it gets more advertising revenues. The total supply of content increases to $x_F + x_G = 19.34$. This increase in traffic, under the same total capacity, pushes up the average waiting time which is now 1.52.

Google is incentivized to produce more content, but most of the profits go to the ISP via the premium fee. The total profits of the fringe, of the ISP, and of google are:

$$\overline{\pi}_F = 26.94 \qquad \pi_{ISP} = 237.39 \qquad \pi_G = 31.34.$$

3.2 Investment in capacity

We now consider the first stage of the game. Suppose that the ISP can invest an amount $I(\mu)$ to expand the capacity μ of the network and reduce the disutility linked to congestion and waiting times of data packets. Under net neutrality the net profits of the ISP become:

$$\Pi_{ISP}^{NN} = \pi_{ISP}^{NN} - I(\mu) = p^{NN} + f + f x_F - I(\mu), \tag{19}$$

while under no regulation they are:

$$\Pi_{ISP}^{PP} = \pi_{ISP}^{PP} - I(\mu) = p^{PP} + f_H + f_L x_L - I(\mu). \tag{20}$$

With simple comparative statics we obtain our next result.

Proposition 2 *In the long run: 1) The equilibrium average congestion is always the same under both regimes. 2) When average advertising rates also do not change, the abolition of NN always leads to higher investment in capacity.*

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

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

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

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

Applying the envelope theorem, the first order conditions are:

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

These FOCs 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 depends on the comparison of total traffic:

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

Using (9) this inequality can be rewritten as

$$\frac{f_L - f}{t_F} < 0.$$

From the Proposition 1, we know that $f_L < f$ and thus $\mu^{NN} < \mu^{PP}$ follows. **Q.E.D.**

The first part of the proposition is independent of advertising rates. As only the end users 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 is driven that the fact that, under (9), PP still shifts advertising resources to google, and this produces an overall increase in total traffic, despite the reduction in content supplied by the fringe. Therefore investment in capacity additionally increases compared to NN in order to keep the same average congestion.

Example 2

The results above can be illustrated through an example, where we can also calculate explicitly the equilibrium values of all the other variables. The investment in capacity is expressed by the function $I(\mu) = \mu$. Also suppose that, a part from capacity which is now endogenous, final users' evaluation, the rate of data arrival, the (identical) unit transport costs, the disutility of waiting, and advertising rates are the same as in Example 1.

In this case, the net neutral equilibrium is indeed characterized by a capacity $\mu = 20$, which we assumed exogenously in Example 1. Hence, all the same equilibrium values emerge as in Example 1 under NN, with the only exception of the net profits of the ISP that now also account for the capacity costs, and are equal to

$$\Pi_{ISP} = 179.$$

The non-neutral equilibrium is instead characterized by a higher endogenous choice of capacity, $\mu = 21$. The fees are now

$$f_L = 0 \qquad f_H = 40 \qquad p = 199.$$

Notice how the ISP now even offers the best-effort traffic for free (this value is optimal in this example, without constraining fees to be non-negative). Content supply from the fringe and google are respectively

$$x_F = 8 \qquad x_G = 12.$$

Since the content supply of google depends only on a_H, it obviously does not change in the short- or long-run. Instead the fee to the fringe decreases now, which explains why its total supply goes up. Also the total supply of content increases to $x_F + x_G = 20$. This increase in traffic is exactly matched by the increase in total capacity, and the average waiting time stays at 1. However the congestion of google's applications, with priority, is $W_H = 1/9$, while those without priority have a waiting time of $W_L = 7/3$.

The average profits of the fringe, of the ISP, and of google are:

$$\overline{\pi}_F = 32 \qquad \Pi_{ISP} = 199 \qquad \pi_G = 32.$$

Thus, in the long-run, only the ISP benefits from PP, while google and the fringe are worse off.

Using the investment $I(\mu) = \mu$ we can obtain closed-form solutions and characterize the long-run analysis completely.

Proposition 3 *Imagine capacity is endogenously determined with an investment cost $I(\mu) = \mu$. The long-run relation between the equilibrium variables* *in the two regimes is given by:*

$$\begin{aligned}
\mu^{NN} &< \mu^{PP} \\
f_L &< f < f_H \\
x_G &< x_H,\ x_F > x_L \\
x_F + x_G &< x_H + x_L \\
W_H &< W(x_F, x_G) = \overline{W}(x_L, x_H) < W_L \\
p^{NN} &< p^{PP} \\
\pi_G^{NN} &> \pi_G^{PP},\ \overline{\pi}_F^{NN} > \overline{\pi}_F^{PP} \\
\Pi_{ISP}^{NN} &< \Pi_{ISP}^{PP}.
\end{aligned}$$

PP leads to a higher investment than NN if and only if

$$2(a_H - a)\frac{t_F}{t_G} > a - a_L \tag{21}$$

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

Proof. See the Appendix.

This analysis has developed the idea that PP redirects advertising funds towards google, but diminishes those to the fringe. This typically creates more traffic overall, which leads to higher investment to keep the level of congestion constant. If all CPs are equally efficient in the way they create content, the result is clear: total traffic depends on the average advertising funds available (that 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. This result comes from our assumption (9): even when 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.

Without making any assumption on advertising rates, (21) summarizes the more general condition needed in order for PP to lead to higher investments compared to NN. The condition is certainly satisfied when the ratio t_F/t_G is large. It is only when resources, via PP, are directed to the "wrong" type of CP that the result can be reversed. For instance, if instead of (9), one alternatively assumed that $a = \frac{a_H+a_L}{2}$ (i.e., with no reference to transportation costs), then (21) would be re-written as

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

and therefore NN could lead to higher investment when $t_G > 2t_F$. In this case google is "very" inefficient compared to the fringe ($t_G > 2t_F$) 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 by the edge of the fringe.

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. 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 as well as investment in capacity does.*

Proof. Recall that when $t_F = t_G$ (9) reduces to $a_L + a_H = 2a$. We now fix the level of a, and look at what happens when the gap between the two rates under PP, i.e., $a_H - a_L$, widens. From the proof of Proposition 2, when $t_F = t_G = t$, it is $p = \frac{v(2a+a_H+v-t-\lambda)}{2t} - \sqrt{d}$, which increases with ads dispersion.

The same effect applies to investment, as $\mu = \frac{\lambda(2a+a_H+v-t-\lambda)}{2t} + \sqrt{d}$. The opposite is true for $f_L = \frac{a_L+t+\lambda-v}{2}$, which decreases. 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 (17), 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: call $a_H = a+\Delta$ and $a_L = a-\Delta$, where Δ is a measure for dispersion. 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 associate 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.**

That profits 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 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 overall from the fringe, but the first effects always prevail. This is important for the ensuing analysis where we imagine that advertising rates change with the congestion level. The monopolist ISP will have an incentive to affect the level of adverting funds (under both regimes), as well as their dispersion, which is doable only under PP.

4 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 then 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. Targeted advertising thanks to deep packet inspections allowed by prioritization techniques are another obvious case in point. In this section, therefore, we make ad revenues dependent on congestion, both under NN and PP. 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 all CPs have identical transportation costs, $t_F = t_G = t$. Google and the fringe are therefore equally efficient in generating content.

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

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

In order to obtain closed-form solutions, we will work at times with a specific example using the following functional form:

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

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

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

where δ represents the relative advantage from advertising revenues when priority is chosen over best effort. Notice that indeed the specification we have chosen 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 $a'_H < a' < a'_L < 0$.

To set grounds, we start working first with this specific example, and then turn to a more general analysis. Under NN, the free entry condition for the fringe CPs is still (1). However, with the feed-back effect from advertising rates, the mass of the fringe is now

$$\begin{aligned} 0 &= \alpha + \frac{\beta}{W(x_F, x_G)} - f - tx_F \\ &\Longrightarrow \quad x_F = \frac{\alpha + \beta\mu - f - \lambda\beta x_G}{t + \lambda\beta}. \end{aligned} \tag{22}$$

Notice how 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

$$\begin{aligned} 0 &= a + x_G \frac{\partial a}{\partial x_G} - tx_G \\ &\Longrightarrow \quad x_G = \frac{\alpha + \beta\mu - \lambda\beta x_F}{t + 2\lambda\beta}. \end{aligned} \tag{23}$$

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 (22) and (23), 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 neutral case.

With these preliminaries at hand, we conduct first a numerical analysis in the short run using the specific functional form. We then turn to a general long-run analysis.

Example 3

A simple example is solved to illustrate the static analysis. As in Example 1, let $\lambda = t = d = 1$, $v = 10$, and $\mu = 20$. Now the parameters for the advertising functions are $\alpha = 9$ and $\beta = 0.24$, that we have chosen for a reason that becomes apparent after solving for the equilibrium.

The net neutral equilibrium is described by the following fee to CPs and price to users:

$$f = 2.22 \qquad p = 158.19$$

This is reflected into content supply from the fringe and google:

$$x_F = 7.78 \qquad\qquad x_G = 8.06$$

with a total supply of content equal to $x_F + x_G = 15.84$. With a capacity of 20, the average waiting time is 0.24. The equilibrium advertising rate is $a = \alpha + \beta/W = 10$, which results in the same (exogenous) advertising rate as in Example 1, thus making the results comparable. Despite the advertising revenues for CPs are the same, in principle, the incentives to provide content are very different. As explained above, the CPs in the fringe are less elastic with respect to the fixed fee they pay. Google also supplies more content the higher is f. These effects imply that the ISP can charge more on the CPs' side, which reduces the amount of content supplied by the fringe in equilibrium. Google also invests less compared to Example 1, because of the congestion crowding out effect. This explains why the price to end users is lower compared to example 1. The total profits of the fringe, of the ISP, and of google are now respectively:

$$\overline{\pi}_F = 30.\,11 \qquad \pi_{ISP} = 177.66 \qquad \pi_G = 45.89.$$

The non-neutral equilibrium, instead, is characterized by:

$$f_L = 2.40 \qquad\qquad f_H = 5.77 \qquad\qquad p = 154.49.$$

Contrary to the case with exogenous advertising, now both fees to CPs go up. End users, instead, pay less compared to NN.

Content supply from the fringe and google is now:

$$x_F = 7.47 \qquad\qquad x_G = 8.00$$

showing that the fringe reduces supplied content. Google also produces less content, despite getting more advertising revenues. The total supply of content decreases to $x_L + x_H = 15.47$. This decrease in traffic, under the same total capacity, pushes down the average waiting time which is now 0.22. The advertising rates in equilibrium are $a_L = 9.87$ and $a_H = 10.30$. The reduction in waiting time increases now the average advertising rate in the industry.

The profits of the fringe, of the ISP, and of google are:

$$\overline{\pi}_F = 27.90 \qquad \pi_{ISP} = 178.19 \qquad \pi_G = 44.66.$$

4.1 Long-run analysis

As in Section 3.2, it is again very convenient to do a change of variable, where the ISP sets prices and average waiting time. Starting with NN, for a given W, we can use almost identically the same analysis as under exogenous adverting rates. 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

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

where p^{NN} is given by (10), 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 FOCs:

$$\begin{aligned} \frac{\partial \Pi_{ISP}^{NN}}{\partial f} &= 1 + x_F + (f + v)\frac{\partial x_F}{\partial f} + \frac{\lambda}{t}I' \\ &= 1 + \frac{a(W) - v - 2f + \lambda I'}{t} = 0, && (24) \\ \frac{\partial \Pi_{ISP}^{NN}}{\partial W} &= v\frac{\partial (x_F + x_G)}{\partial W} - d + f\frac{\partial x_F}{\partial W} + \frac{I'}{W^2} - 2\frac{\lambda}{t}I'a'(W) \\ &= \frac{2v - 2\lambda I' + f}{t}a'(W) - d + \frac{I'}{W^2} = 0, && (25) \end{aligned}$$

where it is apparent now that congestion depends on the way it affects advertising rates. From (24) we obtain

$$f = \frac{a(W) + t + \lambda I' - v}{2},$$

which has the same form as the equilibrium fee in Proposition 3 when $I(\mu) = \mu$. If advertising was not sensitive to congestion ($a' = 0$), from (25) we would also obtain again $W = 1/\sqrt{d}$. With variable advertising rates, the equilibrium waiting time is defined implicitly by the condition

$$\frac{3(v - \lambda I') + a(W) + t}{2t} a'(W) - d + \frac{I'}{W^2} = 0. \tag{26}$$

The first term is negative, which implies that, compared again to Proposition 3, in equilibrium $W < 1/\sqrt{d}$. 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. This is also the reason why the equilibrium value of W is affected by all the parameters of the problem (t, v, λ, besides d).

Let us now turn to the analysis of PP. 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

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

where p^{PP} is given by (11), and from $\overline{W} = \frac{1}{\mu - \lambda(x_L + x_H)}$ we obtain

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

The equilibrium is represented by the following two FOCs:

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

We thus obtain

$$f_L = \frac{a_L(\overline{W}) + t + \lambda I' - v}{2}.$$

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

$$\frac{v - \lambda I' - a_L(\overline{W}) + t}{2t} a'_L(\overline{W}) + \frac{v - \lambda I' + a_H(\overline{W})}{t} a'_H(\overline{W}) - d + \frac{I'}{\overline{W}^2} = 0. \quad (27)$$

We can now state the following results.

Proposition 4 *With variable advertising rates, in the long run: 1) If advertising rates have the same sensitivity to congestion in both regimes, congestion is lower and investment 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 in capacity 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, 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 the Appendix.

Proposition 4 makes formal 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. More in detail, notice that the part 1) of the proposition is not a restatement of Proposition 2 (which is obtained only as a limiting case when $a' = 0$). As advertising rates are now affected by congestion, the ISP reduces congestion compared to the case with exogenous advertising rates, and particularly so under PP compared to NN. PP has a "level" effect that increases advertising funds overall. Part 2) is also quite intuitive. When v is high enough, the effect that prevails must come from the price charged to end users which, in turn depends on total content created. Under NN this is $x_F + x_G = \frac{3a(\cdot)+v-t-\lambda}{2t}$. Under PP it is $x_L + x_H = \frac{2a(\cdot)+a_H(\cdot)+v-t-\lambda}{2t}$. This simple comparison tells that PP leads to higher prices and higher incentives to invest if and only if a reduction in congestion leads to higher ads revenues for prioritized traffic, $|a'_H| > |a'|$. Part 3) shows that, in our leading example where it holds that $|a'_H| > |a'| > |a'_L|$, that PP leads to higher investments and lower congestion in general, including lower values of v. A decrease in congestion under PP increases the dispersion of advertising rates, and leads to further funds attracted to google's applications. In this example, both the level and the dispersion effect go in the same direction. Finally, part 4) shows that the result is quite robust. Indeed, it needs quite some reversal in the sensitivity of ads to congestion to overturn the main finding. In fact, it is a necessary but not sufficient result as the level effect, whereby advertising rates increase for google under PP, is still present.

5 Conclusions

In light of the arguments for and against NN regulation, the debate seems stuck in the sense that, at this point, it is difficult to foresee which architecture will ultimately represent the best approach. The internet seems to be working well to encourage innovation and expansion. However, future demand growth, driven by more content-rich applications, will test the limits of existing networks. The main dispute concerns which of two policy options would generate greater welfare: the protection of innovation and competition in internet content, or the encouragement of greater investment in new capacity.

In an effort to advance the debate, which has mostly been of a qualitative nature, we provide a formal framework which incorporates the arguments of either side. We compare network management techniques with NN, with a focus on innovation. We study the effects on investment incentives of ISP and CPs, and its concomitant effects on consumer and firms' surplus, and CPs' market participation. Our results suggest that in the short run, regulation increases content provision at the edge by a fringe, while it decreases the number of applications of a large provider. In the long run, the internet provider adjusts capacity to maintain constant the average waiting time. Regulation leads to lower supply of capacity and overall content, although it fosters entry of new content providers. Average advertising revenues of content providers are a key driver of the results: when advertising is affected by the congestion on the network, the results are not as clear cut and conditions are identified under which net neutrality guarantees lower congestion and higher ISP investment. The value of economic models like ours for policy makers is that they help them to work out the balance of the argument between "content" and "capacity" under different assumptions.

References

[1] Athey S., Calvano E. and Gans J. (2010), **Can online advertising markets save the media?**, *mimeo.*

[2] Bandyopadyhay S., Guo H. and Cheng H.K. (2009), **Net Neutrality, Broadband Market Coverage and Innovation at the Edge**, *mimeo.*

[3] Cheng H.K., Bandyopadyhay S. and Guo H. (2011), **The Debate on Net Neutrality: A Policy Perspective**, *Information Systems Research*, forthcoming.

[4] Canon C. (2009), **Regulation Effects on Investment Decisions in Two-Sided Market Industries: The Net Neutrality Debate**, *Toulouse School of Economics Working Paper.*

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

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

[7] Hermalin B.E. and Katz M.L. (2007), **The Economics of Product Line Restrictions with an Application to the Network Neutrality Debate**, *Information Economics and Policy*, 19, 215-248.

[8] Kramer J. and Wiewiorra L. (2010), **Network Neutrality and Congestion Sensitive Content Providers: Implications for Service Innovation, Broadband Investment and Regulation**, *Karlsruhe Institute of Technology Working Paper.*

[9] Kruse K. (2010), **Net Neutrality, Priority Pricing, and Quality of Service**, *2nd International Symposium on Communications Regulation,* Karlsruhe Institute of Technology, *mimeo.*

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

[11] Sydell L. (2006), **Internet Debate - Preserving User Parity**, in: All Things Considered, NPR USA.

6 Appendix

Proof of Proposition 1. The first part of the proof is greatly simplified by operating a change of varibales so that the the ISP chooses the average waiting time W_i rather than f_i. For a given level of capacity, the fees to CPs can then be easily recovered. Expressing fees as a function of waiting times gives respectively under NN and PP:

$$\begin{aligned} f &= \frac{t_F}{\lambda W} - \frac{\mu t_F}{\lambda} + a + a\frac{t_F}{t_G} \\ f_L &= \frac{t_F}{\lambda \overline{W}} - \frac{\mu t_F}{\lambda} + a_L + a_H\frac{t_F}{t_G} \end{aligned}$$

For the same level of average W, it immediately follows from (9) that it would also be $f = f_L$. The first order conditions with respect to the average waiting times are:

$$\begin{aligned} \frac{\partial \Pi_{ISP}^{NN}}{\partial W} &= -d + \frac{\partial f}{\partial W}\left[1 + x_F + (v + f)\frac{\partial x_F}{\partial f}\right] = 0, \\ \frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} &= -d + \frac{\partial f_L}{\partial \overline{W}}\left[1 + x_L + (v + f_L)\frac{\partial x_L}{\partial f_L}\right] = 0. \end{aligned}$$

For a given W, as $\frac{\partial f_i}{\partial W_i} = -\frac{t_F}{\lambda W_i^2} < 0$, and $\frac{\partial x_F}{\partial f} = \frac{\partial x_F}{\partial f} = \frac{-1}{t_F}$, the two conditions are identical apart from $x_F = \frac{a-f}{t_F} > x_L = \frac{a_L - f}{t_F}$. Then, the (nonincreasing) function $FOC^{NN}(W)$ takes a lower value than $FOC^{PP}(W)$.This implies $W^{NN} < \overline{W}^{PP}$and consequently also $f > f_L$ follows.

Turning back to the equilibrium fees: recall from (16) that $f_H = f_L + \frac{a_H^2 - a_L^2}{2t_G}$. Calculate (13) at this level of f to get

$$\begin{aligned} \left.\frac{\partial \pi_{ISP}^{NN}}{\partial f}\right|_{f=f_H} &= 1 + \frac{a - f_L - (a_H^2 - a_L^2)/(2t_G)}{t_F} + \left[v + f_L + \frac{a_H^2 - a_L^2}{2t_G} - d\frac{\partial W}{\partial x_F} \right] \frac{\partial x_F}{\partial f} \\ &= 1 + x_F + \left[v + f - d\frac{\partial W}{\partial x_F} \right] \frac{\partial x_F}{\partial f} - \frac{a_H^2 - a_L^2}{t_G t_F} = -\frac{a_H^2 - a_L^2}{t_G t_F} < 0. \end{aligned}$$

Hence, $f < f_H$.

Since $W^{NN} < \overline{W}^{PP}$ and in the short-run capacity is fixed, it must be that $x_F + x_G < x_L + x_H$. As $a < a_H$, from (4) it also immediately follows that $x_G < x_H$. To show $x_F > x_L$ requires $a - a_L > f - f_L$. This is always satisfied since, from the comparison of the FOCs with respect to the fees (13)-(18), the difference $f - f_L$ is bounded above by $\frac{a - a_L}{2}$.

As $\Delta p = p^{NN} - p^{PP} = \left[v - d\frac{\lambda}{W(x_F, x_G)\overline{W}(x_H, x_L)} \right] (x_F + x_G - x_H - x_L)$, then $p^{NN} < p^{PP}$ iff $\frac{v}{d} > \frac{\lambda}{W(x_F, x_G)\overline{W}(x_H, x_L)}$.

Turning to profits, recall that for google $\pi_G^{PP} = \frac{a_H^2}{2t_G} - f_H = \frac{a_L^2}{2t_G} - f_L$. Under NN, it is instead $\pi_G^{NN} = \frac{a^2}{2t_G} - f$. Again, because f_L decreases at most by $\frac{a - a_L}{2}$ compared to f, it is $\pi_G^{PP} < \pi_G^{NN}$. The same reasoning applies to the fringe, both individually and globally, where the total fringe profits are $\overline{\pi}_F^{NN} = \int_0^{x_F} (a - f - t_F x)dx = \frac{(a-f)^2}{2t_F}$ and $\overline{\pi}_F^{PP} = \int_0^{x_L} (a_L - f_L - t_F x)dx = \frac{(a_L - f_L)^2}{2t_F}$.

As far as the ISP is concerned, it is $\pi_{ISP}^{NN} = p^{NN} + f + f x_F$ while $\pi_{ISP}^{PP} = p^{PP} + f_H + f_L x_L$. Under NN, the ISP makes more profits on the fringe. Compared to PP, the gain corresponds to $f x_F - f_L x_L = f\frac{a-f}{t_F} - f_L\frac{a_L - f_L}{t_F}$, which is at most equal to $\frac{a^2 - a_L^2}{4t_F}$. This is always dominated by the extra profits made on google, equal to $f_H - f = f_L + \frac{a_H^2 - a_L^2}{2t_G} - f$. **Q.E.D.**

Proof of Proposition 3. When $I' = 1$, from the proof of Proposition 2 we can immediately calculate $W = 1/\sqrt{d}$, and thus $\mu - \lambda(x_F + x_G) = \sqrt{d}$

under NN. Since $\frac{\partial W}{\partial x_F} = \frac{-\lambda}{[\mu-\lambda(x_F+x_G)]^2} = \frac{-\lambda}{d}\frac{\partial W}{\partial \mu}$, we can simplify (13), obtaining

$$\begin{aligned} 1 + x_F + (f+v-\lambda)\frac{\partial x_F}{\partial f} &= \\ 1 + \frac{a-f}{t_F} - \frac{f+v-\lambda}{t_F} &= 0, \end{aligned}$$

which is solved to get

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

This determines the capacity level and characterizes the equilibrium fully:

$$\begin{aligned} \mu &= \frac{\lambda[a(2\frac{t_F}{t_G}+1)+v-t_F-\lambda]}{2t_F} + \sqrt{d}, \\ p &= \frac{v[a(2\frac{t_F}{t_G}+1)+v-t_F-\lambda]}{2t_F} - \sqrt{d}, \\ x_G &= \frac{a}{t_G}, \quad x_F = \frac{a+v-t_F-\lambda}{2t_F}, \\ W &= 1/\sqrt{d}. \end{aligned}$$

Under PP, the equilibrium is solved similarly, to get

$$\begin{aligned} \mu &= \frac{\lambda(2a_H\frac{t_F}{t_G}+a_L+v-t_F-\lambda)}{2t_F} + \sqrt{d}, \\ f_L &= \frac{a_L+t_F+\lambda-v}{2}, \quad f_H = f_L + \frac{a_H^2-a_L^2}{2t_G}, \\ p &= \frac{v(2a_H\frac{t_F}{t_G}+a_L+v-t_F-\lambda)}{2t_F} - \sqrt{d}, \\ x_H &= \frac{a_H}{t_G}, \quad x_L = \frac{a_L+v-t_F-\lambda}{2t_F}, \\ \overline{W} &= 1/\sqrt{d},\ W_H = \overline{W}[1 - \frac{\lambda(a_F+v-t_F-\lambda)}{\lambda(a_F+v-t_F-\lambda)+2t_F\sqrt{d}/\lambda}], \\ W_L &= \overline{W}[1 + \frac{2\lambda a_G\frac{t_F}{t_G}}{\lambda(a_F+v-t_F-\lambda)+2t_F\sqrt{d}}]. \end{aligned}$$

The results follow from simple comparisons of the relevant expressions. In particular

$$\mu^{PP} > \mu^{NN} \quad \text{iff} \;\; 2a_H\frac{t_F}{t_G} + a_L > a(2\frac{t_F}{t_G}+1),$$

which, making use of (9), is rewritten as

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

and therefore PP leads to higher investment, especially when t_G/t_F is low. **Q.E.D.**

Proof of Proposition 4. The results on congestion follow from comparing (26) and (27) 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 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

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

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 (17), 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 (17), 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.**