Net neutrality and innovation at the core and at the edge

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

Carlo Reggiani[†] Tommaso Valletti[‡]

December 2015

Abstract

How would abandoning Internet net neutrality affect content providers that have different size? We model an Internet broadband provider that can offer a different quality of service (priority) to heterogeneous content providers. Internet users can potentially access all content, although they browse and click ads with different probabilities. Net neutrality regulation effectively protects innovation done at the edge by small content providers. Prioritization, instead, increases both infrastructure core investment and welfare only if it sufficiently stimulates innovation from the large content provider.

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. From the early development as an experimental network linking a limited number of computers, it has now become one of the key priorities for policy makers around the world, as it is seen as an engine to economic growth (Czernich *et al.*, 2011; Mayo and Wallsten, 2011). The Internet is delivered by broadband providers who can use their infrastructure to set particular terms for access to 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 US, the EU and elsewhere.

NN is commonly defined as the principle for which all the traffic on the Internet should be treated equally and it has often been linked to the "end to end" principle.[1] These principles are thought to have guaranteed openness and free access to the Internet; their operation, however, has been questioned by the establishment of broadband as the standard delivering technology.[2] The US Federal Communications Commission (FCC) published in April 2015 the final rule on its new regulations on "Promoting and Protecting the Open Internet".[3] The NN debate received massive attention in the US: a record 4 million comments were submitted to the FCC. Even President Obama made very strong comments in favor of NN in November 2014.[4]

From an economic viewpoint, the issue related to NN is that broadband allows for web traffic management techniques. These techniques can be used, e.g., for quality discrimination of data packets or use of termination charges for data traffic. From this angle, NN is mainly a data treatment (and pricing) issue with possible redistributive consequences. While the debate is complex, the following schematization is useful. On the one side stand proposers of a regulation that bans discrimination of data packets and guarantees open and equal access to the net (or "openists", according to Wu, 2004); on the other side it is believed that the Internet needs no regulation and will develop better by letting the market forces operate freely (or "deregulationists").

Valid arguments have been proposed by both sides. One of the main stances of

---

[1]The principle was that the transmission and routing of Internet traffic should be "dumb", not interfering with information packets sent between sender and receiver (Saltzer *et al.*, 1984).

[2]Broadband adopts the TCP/IP protocols that allow discrimination between data packages.

[3]https://www.fcc.gov/document/protecting-and-promoting-open-internet-nprm.

[4]http://www.nytimes.com/2014/11/11/technology/obama-net-neutrality-fcc.html?_r=0.   In Europe, the European Commission introduced highly debated rules in October 2015.

"openists" is that NN is needed to protect the innovation of small start up content providers (henceforth, CPs), where among those there may be tomorrow's giants like Facebook or Google. Innovation at the "edge" of the network is one of the key attributes 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 (henceforth, ISPs) to get an appropriate remuneration for the use of the infrastructure, which is seen as the best way to guarantee investment for maintenance and expansion of the capacity of the network (the "core" of the Internet). This concern is becoming more prominent due to the increasing diffusion of bandwidth-intensive applications (Yoo, 2005; Van Schewick, 2006; Becker *et al.*, 2010).

In this paper we develop a two-sided model of the Internet to analyze the possible tensions at the "core" versus the "edge". We focus on the two polar cases of NN regulation and priority pricing (henceforth, PP). In the model, a monopolist ISP allows CPs to reach final users of the Internet. Importantly, the model captures one of the defining features of the Internet, that is, the heterogeneous size of CPs. In particular, we assume there is one large, established CP and also a number of small CPs that constitute a fringe. The main contribution of our paper is to investigate how a NN regulation would affect the division of resources between different players and their incentives to innovate. To the best of our knowledge, the heterogeneous size of the CPs has not been explicitly addressed in the formal literature on NN. In our model, CPs fund themselves through advertising revenues that are related to the clicks received on the content supplied. The likelihood of being clicked, and consequently the resources available through advertising, can be affected by the priority regime. Final users, on their side, desire to potentially access as much content and applications as possible. The ISP owns the infrastructure to connect the CPs to end users. Its incentives to invest in maintaining and extending the network, which affects congestion, is also ultimately related to the regime adopted.

The analysis of the model provides important results, including some "rules of thumb" for policy makers. First, NN or PP do not *per se* affect congestion: for a given level of traffic, average congestion is unaffected by the priority regime. The priority regime, however, alters the incentives of the ISP to build capacity. In particular, PP leads to more investment at the core only if it sufficiently boosts innovation from the established large CP: we show that this is a necessary but not sufficient condition.

2

Second, NN is the appropriate regime to protect innovation at the "edge", as it guarantees more supply of content and apps from small CPs. On the other hand, there are circumstances where the overall content supplied is higher under PP as, in that case, the large provider expands its supply to more than compensate for loss of content from the fringe of small CPs. This result is related to the crucial role played in our model by the advertising revenues per click. When the revenues per click are not too high, the larger CP reacts positively to the decrease in the content supplied by the fringe, implied by PP: this may lead to higher overall content. However, as the ad revenues increase, the strategic response of the large provider to reduced content from the fringe is to reduce its content supply as well: this avoids a "cannibalization" effect and a reduction of the clicks received overall. The size of advertising revenues is thus critical in our results and is related to the effectiveness of ads in online platforms: the more effective adverts, the less likely that PP can be welfare enhancing. Thus we also study the extent to which private incentives of the ISP to adopt a particular regime are not aligned with those of a social planner. There is only one limiting case, when content is "king" (that is, content from any CP is highly valued), in which the private and social choice of the regime always coincide.

While the model is developed having in mind the specificities of the NN debate that centres around CPs and ISPs, we note that it can also be reinterpreted to deal with other platform environments. For instance, think of Facebook that invests in a platform that hosts different user-generated content as well as advertising banners. Facebook can use data analytics to allow for more targeted advertising. The regime equivalent to prioritization would be one where Facebook is allowed to price discriminate among advertisers based on data analytics. Instead, a neutral regime is one where all advertisers must be treated equally. Of course there are differences, as Facebook users do not pay (contrary to the ISP's subscribers), and even uniform advertising rates would be positive (while, under NN, termination fees are set as zero). Still, the analogy lies in an economic environment with a multi-sided platform and where some policy regulation affects prices and transactions on one side, having deep consequences over the entire ecosystem.

The rest of the paper is structured as follows. Section 2 briefly reviews the most closely related work to better locate the contribution of the paper. Section 3 introduces the basic model. Section 4 presents the results of the analysis. Section 5 discusses the robustness of the findings. Section 6 concludes with policy implications.

2 Related work and contribution

Two facts have been long recognized about work on NN: first, as the debate is fierce, much has been written about it;[5] second, there is a disproportion between the law, policy and advocacy papers and the formal economic analysis. In recent years, however, the economics literature has developed at a fast pace and in different directions.[6]

The structure of the Internet industry naturally invokes a two-sided market approach: ISPs are the platforms that connect CPs to final users. Economides and Tag (2012) present a static model of charges imposed by the ISP to CPs for "traffic termination" to consumers. Their paper shares with ours the approach but key problems of the Internet such as traffic congestion and bandwidth allocation and ISP's investment are not addressed.

Cheng *et al.* (2011), Choi and Kim (2010) and Kramer and Wiewiorra (2012) use, as we do, the M/M/1 approach: borrowed from queuing theory, it is considered a good proxy for actual congestion on the Internet. Cheng *et al.* (2011) and Choi and Kim (2010) consider similar models in which users access exclusively only one of two content providers;[7] total supply of content is fixed so priority only affects the market shares. In Cheng *et al.* (2011) both CPs can get priority. This leads to a prisoners' dilemma and similar CPs both buy priority: the result is no effect on congestion and more surplus extracted by the ISP. Choi and Kim (2010) consider the case in which CPs bargain with the ISP to obtain exclusive priority for their traffic; CPs are charged a fee only if they opt for priority. The impact of NN on investment crucially depends on the inelastic content supply of the CPs: as more capacity means less value for priority, the ISP has less incentive to invest when NN is abandoned.

---

[5] In November 2015, a casual SSRN search returned 450 papers with "network neutrality" in the title or abstracts. A similar Google search provided over 11m hits.

[6] For example, recently, Njoroge *et al.* (2013), Bourreau *et al.* (2015) and Choi *et al.* (2015) have looked at the role of competition between ISPs in the context of NN, while Kourandi *et al.* (2015) and D'Annunzio and Russo (2015) consider competition among CPs. Peitz and Schuett (2015) and Jullien and Sand-Zantman (2015) study traffic inflation in presence of congestion sensitive data. Gans (2015) considers direct payments from consumers to CPs, which are ignored in other models. Schuett (2010), Kramer *et al.* (2013) and Greenstein *et al.* (2016) provide overarching surveys on NN.

[7] 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 Facebook, which cannot be modelled as mutually exclusive choices.

Kramer and Wiewiorra (2012) consider a continuum of CPs differently sensitive to congestion. In the long run, the welfare superior regime is the one leading to higher investment; as NN reduces entry of new CPs, it prevails only if advertising revenues considerably increase with fewer CPs.

Our model shares with the works cited the way to model congestion due to Internet traffic.[8] However, we differ from each in several respects. Unlike Cheng *et al.* (2011) and Choi and Kim (2010) but consistent with one of the defining aspects of the Internet, users are allowed to potentially browse any content they wish once they connect to the net. Moreover, the market for content is not fully covered, thus we consider an elastic supply by CPs: the market expansion or contraction of content associated to the neutrality regime is central to our results. Furthermore, one characteristic of the Internet is that CPs are very heterogeneous: a few CPs (e.g., Google or Facebook) supply or host on their platform many applications and generate many clicks and, possibly, traffic; on the other hand, there are many CPs that generate individually, but possibly not in aggregate, only a limited amount of traffic. Unlike the rest of the literature, we capture this feature that is crucial to address the "innovation at the edge" argument. Besides, priority affects CPs' revenues endogenously by influencing the probability that ads are clicked. This probability is modelled as a Tullock contest. Contrary to Kramer and Wiewiorra (2012), in our model the large CP faces a trade-off and may or may not increase its supply and, as a consequence, the overall content available. In contrast with them, we find that PP is always detrimental for "innovation at the edge".

Our model is therefore set up to 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 the content and applications that will be developed by heterogeneous CPs, funded by advertising revenues.

3 The model

We present a stylized two-sided model of the Internet that captures the heterogeneity among CPs and the effect this has on a proposed NN regulation. The Internet is constituted of three types of agents: (1) one ISP, (2) several CPs, and (3) a large

---

[8]Economides and Hermalin (2012) follow a different approach and assume that the "pipe" of a monopolist ISP has a fixed capacity in the short run. In their analysis, unlike in ours, content is mostly exogenous and identical between the two regimes.

number of Internet users.

Internet Service Provider. The monopolist ISP owns the infrastructure that allows reaching final users. The assumption of monopoly reflects realistically the market environment in both the US and the EU, where usually one company serves the "last mile", creating a bottleneck in the supply of Internet to final users.[9] The ISP has a capacity denoted by μ and it is subject to congestion, depending on the traffic. The content available on the Internet affects the ISP profits in three ways: through the price charged to users, through the cost of investment and, if allowed by regulation, through the fee charged for priority.

Content Providers. Content and applications are supplied by CPs. Their business model is exclusively based on advertising. In particular, CPs derive advertising revenues proportionally to the clicks they receive.[10] Content is heterogeneous in two respects. First, there are two types of CPs: a continuum of "small" CPs that we call fringe and denote by F, and one large, established CP like Google or Facebook, that we name firm G. The distinction is aimed at capturing the difference in the size of CPs. Second, each unit of content has a differential cost of development. More in detail, CP x in the fringe produces only one unit of content, for which it has to pay a production cost $t_F(x) = t_F x$, where t_F is the fringe CPs' unit cost of development, and $x \in [0, +\infty)$ measures the heterogeneity in development costs for fringe CPs. The higher is x, the less likely that content will be developed as, *ceteris paribus*, it is costly to do so.[11] Firm G, the large and established CP, can control and supply many applications and units of content, x_G, and pays a unit cost t_G per each application/unit developed.[12] We allow the unit development costs t_i, $i = F, G$, to be different: it is realistic to assume that firm G has development costs that may differ from fringe CPs. Still, among the fringe CPs, some are actually very efficient in developing content (those characterized by a low value of x).[13]

---

[9] See Musacchio *et al.* (2009) for an alternative approach where global CPs connect with several local ISPs.

[10] In reality, Internet content is partly funded via advertising and partly via fees charged to users. Our assumption is also adopted in large part of the literature and it makes the model more tractable without affecting the main message of the paper. See Gans (2015) for further discussion.

[11] The right side of the distribution of x is irrelevant for the ensuing analysis. As it becomes clear below, only those CPs with a cost below a cut-off point are going to play an active role.

[12] We thus describe the business of large CPs like Google or Facebook which offer different applications and services (e.g., mail, video, office apps, etc.). Our model would not apply to 'single-purpose' large CPs that increase the quality instead of the variety of its applications.

[13] t_i is a measure of productive efficiency, and the distinction between F and G is introduced to

On the revenue side, we denote the (exogenous) advertising revenue by a and we assume that it is the same per each unit of content. We denote by p_F the probability that the ad associated to a particular fringe CP is clicked by end users, and by p_G the probability that one of the ads of firm G is clicked. These probabilities are further discussed below. The expected profit of firm x in the fringe that gets advertising revenues from a total unit mass of users is thus

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

A free entry condition determines the last CP that develops an application, implicitly

$$x_F = a p_F / t_F, \tag{2}$$

where x_F is therefore also the total content/applications developed in aggregate by the fringe. Under a uniform distribution of content development costs, the total profits of the fringe are

$$\Pi_F = \int_0^{x_F} \pi_F \mathrm{d}x.$$

Firm G sets the number of applications it develops by maximizing total profit

$$\pi_G = a \cdot p_G \cdot x - t_G x, \tag{3}$$

which shall determine a total content, denoted by x_G. In what follows, we interpret total supply of content by CPs (given by the sum of x_F and x_G) as innovation and the content supplied by fringe CPs is considered a proxy for "innovation at the edge".

Final users. A large number of users, whose mass is normalized to one for analytical convenience, wish to connect to the Internet. To this end, they pay a subscription fee P to the ISP. Consumers benefit from the variety of applications available, so we assume that they get a utility v_F for each application of the fringe they can access and v_G for each application of firm G. When looking for an Internet subscription, users know they can browse, say, the site of the Rio 2016 Olympics or check the news on a Russian or Australian portal, or whatever web content is available.

In order to be clear from the outset, we need first a model for consumer behavior and, second, a model of the effects on CPs of prioritization. As said above, applications and contents are non-rival, so consumers visit *every* website. All applications

---

find the extent to which a specific regime of neutrality can affect the incentives to develop content of more or less efficient providers.

and contents are seen with probability one.[14] Turning to prioritization: this is a technique that does not affect consumer behavior *per se*, as consumers still see all available applications. It does however affect the CPs, as navigability and speed of broadband access are important determinants of the success of online advertising. Imagine a consumer is on some application, say Facebook (with probability one). On Facebook there are also video clips, ad banners, etc., embedded with content. Importantly, users do not always click on the ad banners available on the web pages they browse. We assume that each ad banner associated to a unit of content is clicked with a certain probability p_i, $i = F, G$. These probabilities depend on traffic *and* on the neutrality regime which is discussed next.

Net neutrality vs priority pricing. We consider two regimes. Under NN, all CPs access for free a best-effort Internet lane which treats everyone equally. Under PP, CPs have the choice of still paying nothing for best-effort or paying a premium fee f_H for priority. The pricing and the traffic regime affect, in our model, both congestion and the probability of browsing and clicking on the ads. Congestion depends on the total traffic exchanged, on the capacity μ of the ISP, as well as on the traffic management techniques. We borrow from the extant literature the way congestion is affected by prioritization rules. Each user-CP exchange generates an amount of traffic λ. Under NN, congestion is

$$W(x_G, x_F) = \frac{1}{\mu - \lambda (x_G + x_F)}, \tag{4}$$

which is the average waiting time W in a M/M/1 queuing system;[15] the corresponding utility of the users is

$$U_{NN} = v_G x_G + v_F x_F - sW(x_G, x_F) - P, \tag{5}$$

where s is users' sensitivity to congestion. Notice that consumers' utility depends only on the average waiting time. This formulation should be viewed as an approximation: as consumers care only about total available content, utility depends only on the

---

[14]This follows extant literature, as Kramer and Wieworra (2012), Bourreau *et al.* (2015), Choi *et al.* (2015), and Jullien and Sand-Zantman (2015). In Section 5.2 we consider an extension where ads associated to more popular content may be clicked more often.

[15]The M/M/1 formula is the solution to a well-defined queuing system which, in particular, makes the assumption that the arrival of packets follows a Poisson process. In reality, packets move through a complex network of routers (packets from the same origin might take different paths), but tractable economic models have not yet been developed to account for this.

average speed of the connection and is independent of the congestion-sensitivity of each specific CP.

Under PP, the ISP can offer priority to traffic. If x_H and x_L are the masses of applications and contents developed by CPs that choose, respectively, to prioritize or not to prioritize their traffic, the users' utility is

$$U_{PP} = v_H x_H + v_L x_L - s\overline{W}(x_H, x_L) - P, \tag{6}$$

where v_H and v_L depend on the share of providers opting for high and low priority.[16] 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

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

so that the average waiting time is

$$\overline{W}(x_H, x_L) = \frac{x_H}{x_H + x_L} W_H + \frac{x_L}{x_H + x_L} W_L. \tag{7}$$

There are two main properties of this way of modelling traffic. First, a M/M/1 system implies that the *average* congestion is the same in the two regimes, provided the capacity level and the total expected exchanged traffic are also the same.[17] This is an interesting property also highlighted, *inter alia*, by Choi and Kim (2010) and Cheng *et al.* (2011): 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 the ISP for investment in capacity, μ: therefore, the regime will affect average congestion for an endogenous choice of μ. The second property of the queuing 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. This is a feature that is emphasized in the NN debate and that the model captures.

Finally but crucially, the neutrality regime affects the probability that ad banners are clicked. For expositional convenience, we shall use a uniform distribution for the click-through probability, according to which under NN all content is treated equally and thus banners have the same probability of being clicked

$$p_F = p_G = \frac{1}{x_F + x_G}. \tag{8}$$

---

[16] In general, v_H can be thought of as a weighted average of v_G and v_F, weighted by the amount of content of the fringe and of firm G that is prioritized. As it becomes apparent below, in equilibrium only firm G opts for priority, thus $v_H = v_G$ and $v_L = v_F$.

[17] This can be checked immediately by comparing (4) and (7).

A feature of this specification is that the probability depends *negatively* on the total amount of content supplied by each type of CP. For an atomistic fringe CP, this feature bears little consequences. For firm G, however, the decision of introducing more content implies two effects on profits: the first is positive as, according to (6), it increases final users' utility and therefore the amount of content they want to access and the related ad revenues; the second is a negative externality on its own existing content which is less likely to be clicked, i.e., a "cannibalization" effect.

Under PP, CPs opting for priority increase the probability that their webpage's advertising is clicked. For CPs whose business model is heavily relying on ad revenues, the speed of access is one of the crucial determinants of their success.[18] In other words, speed has an impact on the effectiveness of advertising. In case of the generalized uniform distribution of clicks, the probability distributions in presence of prioritization change from (8) to

$$p_L = \frac{1-\delta}{(1-\delta)\,x_L + (1+\delta)\,x_H}, \ p_H = \frac{1+\delta}{(1-\delta)\,x_L + (1+\delta)\,x_H} \tag{9}$$

where $\delta \in (0,1)$ captures the effect of priority on the probability of a click. While we fall short of providing a full microfoundation of click probabilities, this is a parsimonious way of capturing a central aspect of prioritization which is amenable to analytical solutions. Also, an important feature of this specification is that, clearly, NN is obtained as a special case when $\delta = 0$.[19]

Summary of the multi-sided linkages in the model. To sum up, we set up a model with the following features.

- For users, content is non-rival. They enjoy whatever content is available. Total content enters directly their utility function (possibly with different weights if they value content of the fringe and of firm G differently). Users also suffer from (average) delays because of congestion. PP or a NN regime does not affect their utility directly, but only indirectly via the type of content supplied. The ISP's investment also affects utility directly (since it is a determinant of congestion), and this investment varies with the neutrality regime.

- For CPs, instead, content has a degree of rivalry. The more content available,

---

[18] This issue was covered by the New York Times: "For impatient web users, even an eye blink may be too long", 29/02/12.

[19] In a more general formulation, the click probabilities could be made more directly dependent on the capacity investment, although this extension would also complicate the analysis considerably.

the less likely that a certain CP collects ad revenues. There could be several mechanisms that generate this. For instance, when content is watched by users and it is bundled with several ads, not all ads will be actually clicked. Equivalently, ad recall may be less effective in the presence of multiple contents. In other words, the CPs participate in a contest for users' attention and ad revenues.

- *Additionally*, we assume that a PP regime tilts the contest in favor of the CP that opts for priority, *vis-à-vis* the rivals that do not. This is because, for instance, prioritized capacity allows banners to be more visible, easier to download, etc., and these are more likely to be clicked.[20] Hence, for the CPs and contrary to the users, the priority regime directly affects their profits, via the parameter δ. The ISP's investment instead does not affect CPs' profits: this is a limitation of our model, but makes the analysis much more tractable.

We compare the long-run equilibrium properties of the NN and PP regimes in terms of impact on CPs, users, ISP and welfare in a game with the following timing:

- The monopolist chooses μ, and sets price P to end users and, under PP, also a fee f_H for priority to CPs.
- Each CP decides whether to purchase priority (if available); small CPs also decide whether to enter and the large CP chooses its total content x_G.
- Consumers decide whether or not to join the ISP.

4 Analysis

The ISP can invest $I(\mu)$ to expand the capacity μ of the network and reduce the disutility linked to congestion and waiting times of data packets. For simplicity, we shall assume that $I(\mu) = \mu$; in other words, investment displays constant returns to scale with respect to capacity. Note, however, that the specification displays decreasing returns to scale of investment with respect to the average waiting time.[21]

Under NN, the profits of the ISP are obtained only from end-users that are charged

---

[20] As another example, videos on Youtube come bundled with advertisements that viewers can skip after a few seconds, or view until the end. Without prioritization, it is more likely that these ads will be skipped as the viewer would have to waste quite some time to download and see them in full.

[21] In Section 5.1 we discuss the robustness of our result to a general investment function.

an access fee P:

$$\Pi_{ISP}^{NN} = P^{NN} - \mu^{NN},$$

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

$$\Pi_{ISP}^{PP} = P^{PP} + f_H D_H - \mu^{PP},$$

where D_H denotes the demand for the high priority lane. Under PP we shall focus on the case in which best effort is chosen in equilibrium only by the fringe while firm G opts for priority: this implies that $D_H = 1$, as this is the only firm that pays for it. The conditions under which this happens are derived and discussed in what follows; however, we shall underline that this is the most interesting scenario as it captures the concern that a deviation from NN would lead only the established and financially sound CPs to access the "Internet superhighways".[22]

Most of the analysis in this section exploits the fact that NN is obtained as a special case of PP: in other words, the analysis is performed by capturing PP as a small deviation from NN, i.e., starting with $\delta = 0$, and then looking at changes when there is a small effect of priority on the probability of clicking an advert, i.e., $\delta > 0$ with δ being small.[23]

We now identify the conditions such that, under PP, different types of CPs opt in equilibrium for the contract designed for them. Under a priority regime, the individual profit of a generic CP in the fringe is either $\pi_F = ap_L - t_F x$ if the best-effort lane is chosen, or $\pi_F^H = ap_H - t_F x - f_H$ if the priority lane is chosen. Since x does not affect this choice, and recalling (9), all fringe providers indeed opt for the best-effort/low priority connection if

$$f_H \geq a(p_H - p_L) = \frac{2a\delta}{(1 - \delta) x_L + (1 + \delta) x_H}. \tag{10}$$

The fee for a high priority connection has to be sufficiently high, in a way that fringe CPs do not find convenient to pay for it. Provider G, instead, opts for the high priority

---

[22] Note that priority is sold through a fixed fee. This is important to ensure self-selection (see Lemma 1 below): only the large firm G pays for priority, while none of the small firms in the fringe would do so. An alternative would be to set a linear priority fee on each unit of content. This alternative pricing is however more problematic since a system with priority and self-selection may not be sustainable in equilibrium. Since it is ultimately the ISP that, in our model, sets the fees, it is credible that it will opt for a fixed fee, as this indeed generates a situation which is incentive-compatible and results in a dual system with a best-effort alongside a priority lane.

[23] In Section 5.3 we discuss how the results are affected when δ takes a wider range of values.

in case $\pi_G^H \geq \pi_G^D$, where the left-hand side is the profit of G with priority, while the right-hand side is its profit of a unilateral deviation of firm G to no priority, taking as given the choice of the fringe. Notice that firm G is "pivotal", in that, if it does not choose priority, no one else will, and the best-effort regime will re-emerge.[24] Firm G's expected revenue in the presence of a unilateral deviation is $ap_G^D x_G^D$, where p_G^D is given by $\frac{1}{x_L + x_G^D}$ since we go back to a NN regime, and x_G^D is the unique maximizer of π_G^D. As the profit of firm G is given by (3) and given the choices of fringe providers, after simplification, the no deviation condition implies

$$f_H \leq \frac{ax_L}{\left(x_L + x_G^D\right)} \frac{(1+\delta)x_H - (1-\delta)x_G^D}{(1-\delta)\,x_L + (1+\delta)\,x_H} - t_G\left(x_H - x_G^D\right). \tag{11}$$

Since the ISP has all the bargaining power and is a profit maximizer, then it will set the priority fee f_H that exactly extracts all the extra rent from firm G and (11) holds with the equality sign. Bringing together (10) and (11) and focusing on small deviations from NN, i.e., small δ, self-selection of both types of firms translates into the following requirement.

Lemma 1 *For small values of δ, firm G opts for the priority lane and the fringe for the best-effort lane if the unit advertising revenue is such that*

$$a > \widetilde{a} = \frac{(t_F + t_G)^3}{t_F t_G}. \tag{12}$$

Proof. See Appendix.

We assume condition (12) to hold throughout the paper. The condition says that the advertising revenue should be sufficiently high: in that case, firm G can take advantage of any extra expected click due to prioritized traffic and that will induce the ISP to increase the corresponding premium fee, which also ensures that the fringe firms find it too costly to opt for priority.

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

$$\Pi_{ISP}^{PP} = P^{PP} + f_H - \mu^{PP}.$$

---

[24]Instead, no one in the fringe is pivotal when firm G chooses priority, and this is why each fringe member simply compares ap_L and $ap_H - f_H$, as described by (10). Also we rule out the possibility of mergers among small CPs, e.g., due to transaction costs and differences in their core businesses.

Finally, the ISP sets the subscription fee P to extract all surplus (5) from final users. Under network neutrality this implies

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

where $W(x_F, x_G)$ is given by (4). In case PP is allowed, the charge to final users is

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

where $\overline{W}(x_H, x_L)$ is given by (7) and, under (12), $v_L = v_F$ and $v_H = v_G$.

In what follows, we concentrate only on the more interesting case where the ISP finds it optimal to supply both the fringe and firm G, instead of extracting all the surplus only from firm G while neglecting the fringe: this is ensured by having the consumers' preference for variety which is strong enough. The following condition guarantees that this is always the case.

Assumption 1 *Users' evaluation of content is such that*

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

Assumption 1 has a simple interpretation: value, for a certain type of content and for a given waiting time, exceeds the marginal development cost and the marginal cost to supply capacity needed to accommodate for the extra traffic generated.

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

Proposition 1 *(i) The equilibrium average congestion is always the same under both regimes.*

(ii) The fringe always supplies more content under NN, while, for δ sufficiently small, firm G supplies more content under PP if and only if

$$a < \overline{a}_G = 8\frac{t_G^2}{t_F}.$$

(iii) For δ sufficiently small, PP leads both to a higher capacity investment and to more total content than NN if and only if

$$a < \widehat{a} = \frac{27}{8}\frac{t_G^2}{t_F} < \overline{a}_G. \tag{16}$$

14

Proof. See Appendix.

The first part of the proposition is independent of any assumption on the effects of the distribution of ad clicks and revenues. As the end users only care about average expected congestion when buying access to the Internet, the neutrality regime has no bearing on the average waiting time. The regime instead changes the amount of content provided and the expected traffic generated. As both NN and PP lead to the same waiting time, the expected capacity has to adjust to the expected traffic.

The second part of the proposition focuses on the content provided in equilibrium by different types of firms. The content decision of the fringe is determined only by expected clicks and, hence, advertising revenues: NN thus implies an increase in the participation at the edge. The matter is more intricate when dealing with the large CP. To illustrate the intuition behind the result, we need to focus on the best response functions. Under NN, firm G's profit maximization leads to a best response that is non-monotonic. It first increases in x_F, reaches a maximum when $x_G = x_F$, and then it decreases. Notice that this is a feature shared by all models of imperfectly discriminating contests *à la* Tullock (see, e.g., Congleton, 1984; Dixit, 1987). In that literature, the "effort" of a player first increases when outside competition is weak and then decreases continuously. In our setting, supplying more content plays the same role as effort: they are both costly, and the whole distribution of content determines the probability of attracting clicks with the associated revenues, exactly as the effort distribution determines the probability of winning in a rent-seeking contest.

To see this more precisely, the profit for firm G under NN is $ax_G/(x_F+x_G)-t_Gx_G$, which has the same form as a firm competing in a Tullock contest for a 'prize' of size a by putting costly effort x_G against rivals who put an aggregate effort x_F. It is the size of the prize, relative to the effort costs, that ultimately dictates if efforts/contents are strategic substitutes or complements. More formally, the best response of firm G is given by $x_G = \sqrt{ax_F/t_G} - x_F$ which increases (resp., decreases) in x_F if a/t_G is greater (resp., smaller) than $4x_F$. This explains the central role played by the advertising rate, relative to the content development costs, in our analysis.

Figure 1 plots the best response functions[25] and the equilibria under both NN (continuous lines) and PP (dashed lines). Panel (a) illustrates the case of a relatively low value of a. Equilibria take place below the 45° diagonal, where the reaction functions are downward sloping. In that case, the reduction of content of the fringe

---

[25]To be more precise, for the fringe we draw the free-entry condition.

under PP leads to an increase in the content supplied by firm G, as dictated by strategic substitutability. According to the result in the proposition, this is generally the case if the unit advertising rate is $a < \bar{a}_G$. Figure 1 (b), instead, illustrates a case in which the equilibria take place above the main diagonal. As before, priority implies a decrease in the content of the fringe but this now has the opposite effect on their large opponent: strategic complementarity, in fact, leads also firm G to *decrease* its supply in equilibrium. The proposition establishes that this is the case if the advertising rate is sufficiently high, i.e., $a > \bar{a}_G$.

[FIGURE 1 (a) and (b) HERE]

The third part of the proposition focuses on the ISP's investment in capacity and on the CPs' total supply of content. As established in part (i), the waiting time is constant; hence, investment in the network adjusts to the total content supply. The results in part (ii) help us to understand this result. Clearly, NN must lead to higher content and investment in presence of strategic complementarity: a deviation from NN implies in that case a decrease in both types of content. In presence of strategic substitutability, instead, the overall content supply and investment will depend on whether the decrease of fringe content under PP is more or less than compensated by the increase in the supply of firm G. The result in the proposition provides the conditions under which a regime leads to higher overall supply and investment: PP is beneficial if the advertising revenue is sufficiently low, i.e., $a < \hat{a}$ and, intuitively, the latter condition is more restrictive than the corresponding one in part (ii).

Finally, it is worth noticing that the ISP's incentive to invest in capacity under PP does *not* depend on the premium fee: the latter is just used to extract firm G's rent but does not have a bearing on the equilibrium traffic on the Internet. The fee, in fact, does not affect either the average congestion on the network, as shown in part (i), or the content choices of CPs and, consequently, the overall supply of the sector.

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

Proposition 2 *For δ sufficiently small, the comparison between the equilibrium variables under the NN and PP regimes implies:*

$$(i) \ \Pi_F^{NN} \ > \ \Pi_F^{PP},$$
$$\pi_G^{NN} \ < \ \pi_G^{PP};$$

16

$$(ii) \ W_H < W(x_F, x_G) = \overline{W}(x_L, x_H) < W_L;$$

$$(iii) \ P^{NN} \ < \ P^{PP} \ iff \ a < \overline{a}_P = \frac{t_G^2}{8t_F} \frac{(4v_G - v_F)^3}{v_G^3},$$

$$\Pi_{ISP}^{NN} \ < \ \Pi_{ISP}^{PP} \ iff \ a < \overline{a}_{ISP} = \frac{t_G^2}{t_F} \frac{(4v_G - v_F - 3\lambda - 3t_G)^3}{(2v_G - 2\lambda - 3t_G)^3};$$

(iv) If $v_G \ \simeq \ v_F$ and both are sufficiently large, the ISP prefers the regime that leads to higher investment.

Proof. See Appendix.

The results of Proposition 2 suggest that preserving the current "status quo", NN, has important redistributive effects on the sector that go beyond investment in infrastructure. The first part of the proposition focuses on the profits of CPs. As NN implies an increase in entry at the edge, this also translates into higher aggregate profits for the fringe companies. Moving towards a regime of PP kills part of the innovation done at the edge by the fringe, as small providers get reduced expected clicks: this damages advertising revenue and overall profits.[26] Firm G's profits, on the contrary, are enhanced if NN is abandoned: the large firm is better off as priority increases the appeal of its content and the likelihood of the associated adverts to be clicked. The ISP appropriates a chunk of those extra rents by charging the premium fee f_H; however, the fixed fee does not extract the entire amount, as it is designed by the ISP on the basis of a unilateral deviation of firm G.

The second part of the proposition confirms that the equilibrium reflects the properties of the M/M/1 system that we adopted to model internet congestion. As average congestion is unaffected by the regime, the only effect of PP is to create inequality in the waiting time faced to access the content of firm G as compared to the fringe providers.

The third part of the proposition focuses on users' fees and ISP profits. End users always have their consumer surplus completely extracted both with and without NN, but the prices they pay differ. If the fringe content is not too valuable relative to firm G's ($v_F < 4v_G$) then ad rates exist such that priority increases the subscription fees ($a < \overline{a}_P$). In this case, the result reflects the higher benefits that users enjoy from more available applications and content: as PP generates both more content and

---

[26]This result goes beyond the validity of Proposition 2 as it holds true for any value of δ.

value, consumers are charged more, as their surplus is extracted by the ISP. However, the results change if the contents of the fringe firms are very important to users: for example, newly launched but virally captivating social networks, revolutionary online services, useful apps or particularly entertaining new videogames are often highly valued. In those cases, NN leads to a higher fee independently of the values of the advertising rate: content supply under NN is relatively skewed towards fringe providers, and best meets users' preferences.

A similar logic applies to the ISP's profits. If established content is relatively more valuable, then there exist ad rates such that priority also increases the ISP profits ($a < \bar{a}_{ISP}$): intuitively, in those cases the ISP can extract more surplus from both users and firm G, exploiting the fact that the latter is supplying more content under PP. For higher values of the advertising rate ($a > \bar{a}_{ISP}$), the countervailing effect due to the reduction of firm G's supply under PP dominates the extraction of the surplus related to the high valuation of users.[27] In that case, if it is the ISP that chooses the priority regime, NN would not need to be mandated by a regulator, as it also constitutes the privately optimal choice. In the next section we make welfare assessments more precise.

Finally, in case both types of content are similarly and highly valued by end users, the most profitable regime for the ISP is exactly the one that leads to higher investment in the network: as surplus extraction from consumers is dominant under both regimes, the relevant effect is the one linked to the provision of content and this ultimately drives the ISP's choice.

To summarize, the main effect of NN regulation is therefore to direct clicks and, consequently, advertising resources towards the fringe. The result is to induce more entry of new CPs in the fringe or, in other words, to spur innovation at the edge. However, there are reasonable circumstances in which it reduces both content innovation by large CPs and investment by the ISP. If content of any type is really "king", the ISP always chooses the regime that induces more generation of aggregate content from CPs, and matches it with increased investment in capacity.

---

[27]If PP leads to a reduction of firm G's supply, ISP's profits are negatively affected in two ways: first, through a reduced price to end users, and second, through a lower fixed fee to G. However, there is one extra effect that goes in the opposite direction: the lower cost of investing in capacity. The balance of these three effects also determines the size of the area in which PP is more profitable than NN. The area may be more or less extended than the one in which the price to subscribers is higher. The thresholds $\bar{a}_P$ and $\bar{a}_{ISP}$ coincide only when content values are very large.

4.1 Welfare effects of NN regulation

We can now consider the welfare properties of the equilibrium. As prices and fees are simple transfers, the expressions for social welfare are

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

under NN and PP respectively.

These expressions implicitly assume that all advertising is informative and it increases the social value of the industry. Part of online advertising, however, may constitute a nuisance for final users, distracting them from the main reason they are surfing. Such a feature, though, would not affect the results presented below in any fashion. In fact, in our model, the priority regime has no bearing on the overall amount of advertising resources available on the Internet but only on their distribution: this can be seen immediately as $a(p_F x_F + p_G x_G) = a(p_L x_L + p_H x_H) = a \cdot 1$. Thus our welfare results (on the comparison between NN and PP) do not depend in any way on the stance taken towards advertising. If, for some reason, PP allows for more effective advertising practices, such that the overall resources in the sector may increase, then there would be a different argument in favor of PP.

The following proposition establishes what regime is socially preferable when allocations are chosen by an unregulated ISP.

Proposition 3 *For δ sufficiently small, the privately chosen allocation under PP is socially preferable to NN iff:*

$$
a < \overline{a}_{SW} = \frac{t_G^2}{t_F} \frac{(4v_G - v_F - 3\lambda - 4t_G)^3}{(2v_G - 2\lambda - 3t_G)^3} < \overline{a}_{ISP}.
$$

If $v_G \simeq v_F$ and both are sufficiently large, the social planner prefers the regime that leads to higher investment.

Proof. See Appendix.

If the content of firm G is valuable and PP guarantees more of it, then priority is the preferred regime from a welfare viewpoint. Clearly, PP leads to more content

when the advertising rate is sufficiently low and that explains the condition $a < \overline{a}_{SW}$. On the other hand, if it is NN that leads to more content, particularly the one very valuable to users, then a ban on prioritization may lead to welfare superior outcomes. Although the mechanisms driving profit and welfare results are similar, the exact forces that drive them are slightly different. This becomes clear comparing the expression of social welfare with the ISP's profits, i.e., $\Pi_{ISP}^{PP} = P + f_H - \mu$ under PP. Suppose a is sufficiently high and the overall content decreases under PP. As the price extracts the entire consumer surplus, less content has a negative effect of the same magnitude as identified in the previous section. Similarly, the effect of less content under PP identically affects the cost of investing in network capacity, which is relevant for both profits and welfare. The third effect in the case of profits is related to the fixed fee: in the case of social welfare, the latter is replaced by a further positive effect linked to the reduction of the content development costs. The reduction of both the costs of investment and development dilutes the first negative effect on prices; however, if a is sufficiently high, PP still loses its appeal from a welfare perspective.

It is easy to see that the area of welfare dominance of PP is *less* extended than the area in which it is more profitable for the ISP to adopt it ($\overline{a}_{SW} < \overline{a}_{ISP}$). As a consequence, combinations of parameters exist for which the ISP's regime preference contrasts with social welfare: particularly, if $\overline{a}_{SW} < a < \overline{a}_{ISP}$ the ISP would prefer a prioritization regime although this would be damaging overall welfare. In those situations, the ISP's choice of regime is sub-optimal and a regulatory intervention to impose NN would be necessary to re-establish the desirable outcome.

Importantly, in the special case in which both types of contents are similar and highly valued, the socially preferred regime coincides with the most profitable one, and it is the one that leads to higher content and investment in the network.[28]

5   Robustness

5.1   Investment in capacity costs

In this section we discuss the robustness of our results to the assumption on the costs of network capacity, $I(\mu) = \mu$. We show below that investment in capacity and

---

[28] We discussed in this section the welfare implications of the selection of the neutrality regime. In each regime, however, content and investment can also differ from the first best.

total content, as well as results on welfare, are *exactly* the same, independently of the specification of the cost function. The main difference is that the average waiting time now changes in the two regimes, and we discuss how it only depends on the concavity/convexity of the cost function.

To show this, we now allow for a general investment cost $I(\mu)$. Clearly, only the ISP's profit is affected by this change. Under PP, for instance, it is

$$\Pi_{ISP}^{PP} = v_F x_L + v_G x_H - s\overline{W} + f_H(x_H, x_L) - I(\mu(\overline{W})),$$

where $\mu(\overline{W}) = \frac{1}{\overline{W}} + \lambda(x_L + x_H)$. Since choosing μ determines $\overline{W}$, we can operate a change of variable by which $\mu = \frac{1}{\overline{W}} + \lambda(x_L + x_H)$.

Concentrating only on the terms that depend on waiting time, we have the FOC

$$\frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} = -s - I'\frac{\partial \mu}{\partial \overline{W}} = -s + \frac{I'}{\overline{W}^2} = 0,$$

which we rewrite as

$$F \equiv \overline{W}^{-2}I' - s = 0.$$

We assume that the SOC is always satisfied. This requires that $\partial F/\partial \overline{W} < 0$, or $-(2\overline{W}I' + I'') < 0$, which is always true if $I'' > 0$, but it is also satisfied if $I'' < 0$ but not too large. Recall that, in Proposition 1, the average waiting time did not change in the two regimes. Under a more general specification, we simply need to check how the average waiting time is affected by δ, as $\delta = 0$ corresponds to the NN regime. We apply implicit differentiation to get

$$\frac{\partial \overline{W}}{\partial \delta} = -\frac{\partial F/\partial \delta}{\partial F/\partial \overline{W}}.$$

The denominator is negative by the SOC. Hence the sign is the same as the sign of the numerator. From $\mu(\overline{W}; \delta) = \frac{1}{\overline{W}} + \lambda[x_L(\delta) + x_H(\delta)]$, this depends only on

$$\frac{\partial F}{\partial \delta} = -I''\frac{\lambda}{\overline{W}^2}\frac{\partial[x_L(\delta) + x_H(\delta)]}{\partial \delta}.$$

Around $\delta = 0$, the sign of the last term $(\frac{\partial[x_L(\delta) + x_H(\delta)]}{\partial \delta})$ has been established in Proposition 1 (ii), as the overall content does not depend on the average waiting time. Thus we obtain a very clear and general result. If investment costs are linear in μ, $I'' = 0$, we confirm that the priority regime has no impact on the average waiting time. If instead investment is convex in μ, then the average waiting time follows the

opposite behavior as total content; e.g., from Proposition 1 (iii), if $a < \hat{a}$, then PP leads to higher content *and* lower waiting time. The converse result is true if costs are concave.

We then turn to the impact of the priority regime, δ, on investment μ. From $\mu(\overline{W}; \delta) = \frac{1}{\overline{W}} + \lambda[x_L(\delta) + x_H(\delta)]$, we obtain after simplifications

$$\frac{\partial \mu}{\partial \delta} = -\frac{1}{\overline{W}^2} \frac{\partial \overline{W}}{\partial \delta} + \lambda \frac{\partial[x_L(\delta) + x_H(\delta)]}{\partial \delta} = \frac{2\lambda \overline{W} I'}{2\overline{W} I' + I''} \frac{\partial[x_L(\delta) + x_H(\delta)]}{\partial \delta}.$$

The sign of the denominator is positive as a result of the SOC. Thus: $sign\left[\frac{\partial \mu}{\partial \delta}\right] = sign\left[\frac{\partial[x_L(\delta) + x_H(\delta)]}{\partial \delta}\right]$. Therefore a central result of Proposition 1, namely that investment at the infrastructure core follows total investment in content, is still valid in general, independently of the shape of the investment function.

Finally, it is immediate to show that also the main welfare result given in Proposition 3 is still valid generally. This is because total welfare depends on $\overline{W}$, as it affects both the ISP and the users. The monopolist ISP extracts all users' surplus and thus perfectly internalizes social welfare as far as waiting time is concerned. Hence, we can apply the envelope theorem so that local changes of $\overline{W}$ do not have any first-order effect on profits and therefore also no local effect on social welfare. The threshold derived in Proposition 3 is still valid for *any* investment cost function.

5.2 Content spillover effects

In our model we allow users to have heterogeneous valuations for different types of content, but yet all contents are non-rival for users and enter additively in their utility function. As heterogeneity may also lead to asymmetry in the demand of contents, with a possible effect on ad-revenues, we extended the model to show how "spillover effects" on the demand of content can affect the results of our model.

More specifically, imagine in each time period, normalized to one, a different application/content becomes available and users stay online for that entire period. However, content may be more or less popular, implying that users may actually stay online a bit more, or a bit less, than the expected unit period. We denote this extra time by r, with $r \gtrless 0$: in particular, the users' demand of every unit of content by firm G is affected by an extra $+r$, and the demand for every unit of content by each firm F by an extra $-r$. The parameter r could be made positive or negative depending on the relative value of the content of firm G relative to the fringe. In

other words, if some content is more popular than other, then consumption of that content is magnified, but this brings a negative spillover to the other less popular content which is then consumed less. The case $r = 0$ corresponds to our benchmark model in Section 3, allowing for meaningful comparisons and robustness checks.

In the Online Appendix, we first show that extending the model to allow for content spillovers does not affect the equilibrium waiting time, which is identical under both NN and PP. The reason is that the specification preserves the $M/M/1$ formulation of internet traffic: the spillover effect is simply a multiplier of traffic (positive or negative) but does not affect the properties of the system.

Second, a positive spillover on the content of firm G, $r > 0$, negatively affects the equilibrium content of the fringe: the effect is proportionally more relevant under PP, as the benchmark content is lower in that case. The effect on the content of firm G is, instead, not clear *a priori*: under both regimes, the spillover has a direct positive effect on the time users spend on the content but also a negative effect, similar to the *cannibalization* effect of new content on the probability of an ad being clicked. In the Online Appendix we show that spillovers, r, have a quantitative impact on the results, whereas qualitatively all conclusions are robust to such an extension. The impact of the spillover is to affect the value of the various thresholds for a in Proposition 1 and 3. In particular, the thresholds decrease in r. Hence, if it is firm G to be popular and hence clicked more often ($r > 0$), then the thresholds shifts to the left and it becomes *less* likely that PP leads to higher content, more investment, and higher welfare. The opposite happens when $r < 0$, that is, the fringe firms' ads are clicked more often. The intuition is as follows. Take the case $r > 0$. The content of the fringe decreases because of the negative spillover, especially so under PP when ad revenues further decrease. As for firm G, the factor r *de facto* increases the effectiveness of ad revenues. Thus it is "as if" the ad rate goes up. We know from our previous analysis that, for high values of a, there is strategic complementarity and also firm G decreases its supply of content in equilibrium. Thus, for $r > 0$, it is more likely than in the benchmark case that PP leads of a reduction of total content.

5.3 Larger deviations from NN

The main results presented in Section 4 are based on the analysis conducted around $\delta = 0$ and hence they hold when prioritization implies small deviations from NN. But how small does δ need to be? We conducted some analysis to evaluate the sensitivity

of our findings to larger values of δ (see the Online Appendix for full details).

We know from the earlier Propositions that a central role is played by the advertising rate per click a. First, we considered relatively "low" values of a, that is, quite near to the lowest boundary $\widetilde{a}$ that ensures self-selection (Lemma 1). We found, as expected, that the content of the fringe monotonically decreases in δ. Priority pricing, instead, increases the content of firm G compared to NN. However, as the effect of prioritization becomes stronger, the difference in content for G between the two regimes starts to decrease. In other words, our result on overall content and, consequently, investment in the network is, in fact, sensitive to δ. The expected result that PP leads to a higher overall content than NN still holds but for values of δ that are not very large. In case prioritization changes the probabilities more sharply, then the "best responses" plotted in Figure 1 (a) would also shift sharply and lead to an equilibrium in which the overall content decreases. This is due to the fact that the increase in content of firm G does not compensate sufficiently for the decrease in the overall entry of fringe providers.

As far as the other variables are concerned, both the overall profits of the fringe and the profits of firm G change monotonically in δ. In particular, the overall fringe profits decrease while the profits of firm G increase when PP has a stronger impact (larger δ). Both the ISP's profits and social welfare are expected to rise under PP, according to Proposition 2 and Proposition 3: our analysis suggests that the results hold true as δ grows above zero, unless PP has a major impact on the probabilities.

Turning to the case of relatively "large" values of the advertising rate a, our results appear to be very robust. Both the overall, fringe and firm G's contents are negatively affected by PP. The results are in fact only sharpened as δ becomes larger. The results on profits are also fully confirmed. Finally, for high values of the advertising rate, both social welfare and the ISP's profits are harmed by PP; more and more so as the impact of prioritization sharpens.

6   Discussion

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. ISPs can access a large amount of information about data packets and discriminate between them at a very low cost, essentially questioning the net neutrality principles. This has triggered a heated

debate, recently exacerbated by the attempts to regulate the Internet sector in the US and in the EU and the resistance opposed by telecom operators.

In order to frame the debate and study the incentives of the main parties involved, we contribute to the economics literature on net neutrality by analyzing a model that captures the asymmetry between different types of content providers. In our approach, a monopolist ISP allows final users to use apps and browse content from two types of providers: one large and established Internet CP, like Google or Facebook, and a number of small firms, the "edge of the network". This two-sided market model allows us to tackle one of the most controversial issues: the alleged tension between investment incentives at the "core", i.e., ISPs' maintenance and upgrade of their networks, and innovation at the "edge", i.e., the ease of entry of new and innovative Internet start-ups. Indeed, increasing traffic, congestion and the connection speed are crucial elements in today's Internet: evidence suggests that even small delays in loading a webpage affect the likelihood that an advertising banner is clicked. We propose a model that formalizes prioritization as a tool that stands at the interface between Operations Management and Marketing, especially in the context of clickstream tracking. Broadband network intelligence allows the ISP both to reduce waiting time of particular applications and content (which is directly enjoyed by end users) and, contextually, redirect to these providers advertising resources. Prioritization, in fact, makes it more likely that ads on a particular content is clicked with a consequent increase in the expected advertising revenues. On the other hand, our model emphasizes that the large CP may not have incentives to introduce more content, due to a "cannibalization" effect. Unambiguously, instead, prioritization takes away resources from other stakeholders (the small CPs) who cannot afford to pay the priority fee. We identify when each one of these effects dominates and provide conditions under which prioritization increases the large CP's provision, the overall content available, and the investment in the network capacity. The results are crucially related to the advertising revenue per click: under prioritization the latter determines the reaction of the larger CP to the reduction in content supplied by the fringe. If the unit advertising revenue is not too high, the reaction is positive and content increases. However, as the ad revenue increases, the "cannibalization" effect kicks in leading to a decrease in the content supply. These effects impact on the total content supply and, in turn, on investment in the network. Moreover, we discuss the regime preferences of both the ISP and society as a whole and we find that the ISP's choice

of prioritization can be excessive. We identify situations in which the privately and socially optimal regimes do not coincide: these happen for intermediate values of the advertising rate and are related to the fact that the ISP does not fully internalize the content development costs. Although practically the identification of such instances may not prove easy, the finding detects circumstances when a regulatory imposition avoiding undesirable profit squeezing via prioritization is soundly justified.

These results have policy implications in the light of the current debate. First, concerns about prioritization and innovation at the "edge" seem grounded on the basis of our model. Intuitively, as start-ups have usually limited financial capabilities, they may be cut-off from "Internet superhighways" they cannot access and, as such, suffer more acutely from increasing congestion. If society has preferences strongly tilted in favor of these small CPs, then there is a clear case for not allowing departures from net neutrality. This view, however, ignores the impact on other stakeholders. Second, in fact, the large established CPs can profit from prioritization since it brings a higher likelihood that their ads are clicked. This holds true despite the ISP will charge a fee for priority. Third, higher profitability may not necessarily be reflected in more content supplied by the large CPs: for policy purposes, higher profits of established providers cannot be taken to proxy a wider content supply to web users. Fourth, content supply is crucially related to the advertising revenues per click. The internet sector has grown dramatically since its commercial inception and, with it, the amount of advertising resources invested online: digital advertising expenditure may soon overtake television globally.[29] The widespread use of cookies, tracking and profiling technologies and the increasing availability of consumers' data is likely to further improve the effectiveness of online advertising and, hence, to boost further the advertising revenue per click in large economic areas like the US and the EU. In that case, our results suggest that preserving net neutrality may be the right policy to guarantee the largest availability of content and to provide incentives to invest and improve the existing networks. Interestingly, our results show that the ISP should have interests aligned with the policy maker when ad revenues are high enough. On the other hand, the concerns raised by the unbounded commercial exploitation of personal data and the current worldwide debate about the right to privacy may lead to regulations that hinder the technological effectiveness of advertising with the possible

---

[29]http://www.nytimes.com/2015/12/07/business/media/digital-ad-spending-expected-to-soon-surpass-tv.html?_r=0.

consequence of reducing advertising resources.[30] In this case, our results indicate that net neutrality does not necessarily guarantee a higher level of overall online content and a switch to a regime allowing priority may both induce large providers to boost their content provision and also the ISP's incentives to invest in the network. Thus, an upshot of our findings is that a tougher privacy regime can be accompanied by a more lenient approach towards prioritization policies. However, we also show that when the effectiveness of ad revenues is reduced, there is an excessive tendency for the ISP to adopt prioritization compared to its social desirability. Only in the special case when "content is king" (i.e., all types of content are highly valuable to users) our analysis suggests that investment, profitability and welfare are all aligned and the privately and socially preferred regime coincide.

References

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

[2] Bourreau M., Kourandi F. and T. Valletti (2015), Net Neutrality with Competing Internet Platforms, Journal of Industrial Economics, 63(1), 30-73.

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

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

[5] Choi J.P., D.S. Jeon and B. Kim (2015), Network Neutrality, Business Models, and Internet Interconnection, American Economic Journal: Microeconomics, 7(3), 104-141.

[6] Congleton R.D. (1984), Committees and rent-seeking effort, Journal of Public Economics, 25(1-2), 197-209.

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

[8] D'Annunzio A. and A. Russo (2015), Net neutrality and internet fragmentation: the role of online advertising, International Journal of Industrial Organization, 43 (2015), 30-47.

---

[30] Goldfarb and Tucker (2011) find that even moderate levels of privacy regulation in the EU reduce the effectiveness of online advertising.

[9] Dixit A. (1987), Strategic behavior in contests, American Economic Review, 77(5), 891-898.

[10] Economides N. and J. Tag (2012), Net Neutrality on the Internet: A Two-sided Market Analysis, Information Economics & Policy, 24(2), 91-104.

[11] Economides N. and B.E. Hermalin (2012), The Economics of Net Neutrality on the Internet, Rand Journal of Economics, 43(4), 602-629.

[12] Gans, J. (2015), Weak Versus Strong Net Neutrality, Journal of Regulatory Economics, 47(2), 183-200.

[13] Goldfarb A. and C. Tucker (2011), Privacy Regulation and Online Advertising, Management Science, 57 (1), 57-71.

[14] Greenstein, S., Peitz M. and T. Valletti (2016), Net Neutrality: A Fast Lane to Understanding the Tradeoffs, Journal of Economic Perspectives (forthcoming).

[15] Jullien, B. and W. Sand-Zantman (2015), Internet Regulation, Two-sided Pricing, and Sponsored Data, TSE Working Paper, 12-327.

[16] Kourandi, F., Kramer, J. and T. Valletti (2015), Net neutrality, exclusivity contracts and Internet fragmentation, Information Systems Research, 26(2), 320-338.

[17] Kramer J., Wiewiorra L. and C. Weinhardt (2013), Net Neutrality: A Progress Report, Telecommunications Policy, 37(9), 794–813.

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

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

[20] 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.

[21] Musacchio J., Schwartz G. and J. Walrand (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.

[22] Njoroge P., Ozdaglar A., Stier-Moses N.E. and G.Y. Weintraub (2013), Investment in two sided markets and the NN debate, Review of Network Economics, 12(4), 355-402.

[23] Peitz M. and F. Schuett (2015), Net neutrality and inflation of traffic, mimeo.

[24] 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.

[25] Schuett, F. (2010), Network neutrality: A survey of the economic literature, Review of Network Economics 9(2), 1-13.

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

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

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

A Appendix

Proof of Lemma 1. The ISP's profit maximization implies that firm G's self-selection condition (11) is satisfied with the equality sign. Substituting (11) into (10) gives

$$\frac{ax_L \left((1+\delta)x_H - (1-\delta)x_G^D\right) - 2a\delta \left(x_L + x_G^D\right)}{\left(x_L + x_G^D\right)\left[(1-\delta)x_L + (1+\delta)x_H\right]} - t_G \left(x_H - x_G^D\right) \geq 0. \quad (17)$$

Note that, if firm G refuses to pay for priority, it could still use the free lane, in which case it would offer a different amount of content. From G's deviation profits, we obtain $x_G^D = \sqrt{ax_L/t_G} - x_L$. Further, x_H and x_L are calculated in the PP regime (see (20) and (19) below). After substitution, and simplifying the derivative of the l.h.s. of (17), evaluated at $\delta = 0$, the latter inequality holds for sufficiently high values of a, i.e., $a \geq \widetilde{a} = \frac{(t_F + t_G)^3}{t_F t_G}$.

Q.E.D.

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

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

Notice that, for a given W, the capacity marginal cost when the 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

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

The first-order conditions in the two regimes are

$$\frac{\partial \Pi_{ISP}^{NN}}{\partial W} = -s - \frac{\partial \mu}{\partial W} = 0, \quad (18)$$
$$\frac{\partial \Pi_{ISP}^{PP}}{\partial \overline{W}} = -s - \frac{\partial \mu}{\partial \overline{W}} = 0.$$

These conditions are identical and thus determine the same average waiting time.[31]

---

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

(ii) Under PP, the free-entry condition for the fringe (2) allows to find the fringe content as

$$x_L = \sqrt{\frac{x_H^2 (1+\delta)^2}{4(1-\delta)^2} + \frac{a}{t_F}} - \frac{1+\delta}{1-\delta}\frac{x_H}{2} \tag{19}$$

while firm G's best response to x_L is

$$x_H = \sqrt{x_L \frac{a}{t_G}\frac{1-\delta}{1+\delta}} - \frac{1-\delta}{1+\delta}x_L. \tag{20}$$

Solving simultaneously give

$$x_L^* = \sqrt[3]{a\frac{t_G}{t_F^2}\frac{(1-\delta)}{(1+\delta)}}, \qquad x_H^* = \frac{t_F}{t_G}\sqrt[3]{\left(a\frac{t_G}{t_F^2}\frac{(1-\delta)}{(1+\delta)}\right)^2} - \sqrt[3]{a\frac{t_G}{t_F^2}\frac{(1-\delta)}{(1+\delta)}}\frac{1-\delta}{1+\delta}.$$

The equilibrium contents under NN, x_F^* and x_G^*, are obtained for $\delta = 0$. Clearly, $x_L^* < x_F^*$ for any positive δ while it is easy to show that $\frac{\partial x_H^*}{\partial \delta}|_{\delta=0} > 0$ iff $a < \bar{a}_G = 8\frac{t_G^2}{t_F}$.

(iii) The equality of the average waiting times implies that, in equilibrium, $\mu^{NN} - \lambda(x_F^* + x_G^*)$ under NN must equal $\mu^{PP} - \lambda(x_L^* + x_H^*)$ under PP. The level of capacity therefore depends on the comparison of the total expected traffic, which is generated by total content:

$$\mu^{PP} > \mu^{NN} \text{ iff } x_L^* + x_H^* > x_F^* + x_G^*. \tag{21}$$

For δ small and under assumption (12), (21) boils down to: $\frac{\partial(x_L^* + x_H^*)}{\partial \delta}|_{\delta=0} > 0$; the latter requires $a < \hat{a} = \frac{27}{8}\frac{t_G^2}{t_F}$.

Q.E.D.

Proof of Proposition 2. From the proof of Proposition 1 we know x_L^* and x_H^* and we can calculate $\overline{W} = 1/\sqrt{s}$, and thus $\mu - \lambda(x_L + x_H) = \sqrt{s}$ under PP. The equilibrium under NN is simply found imposing $\delta = 0$. The characterization of the PP equilibrium is then completed by:

$$\mu = \lambda\left(\frac{t_F}{t_G}\sqrt[3]{\left(a\frac{t_G}{t_F^2}\frac{(1-\delta)}{(1+\delta)}\right)^2} + \sqrt[3]{a\frac{t_G}{t_F^2}\frac{(1-\delta)}{(1+\delta)}\frac{2\delta}{1+\delta}}\right) + \sqrt{s},$$

$$f_H = 2\sqrt{at_G x_L^*} - \sqrt[3]{\frac{at_F t_G(1-\delta)}{(1+\delta)}}x_L^* - \frac{2\delta}{1+\delta}t_G x_L^* - \frac{t_F^2 x_L^{*3}}{\sqrt[3]{\frac{at_F t_G(1-\delta)}{(1+\delta)}}},$$

31

$$\Pi_F = \sqrt[3]{\frac{a^2 t_G^2}{8 t_F} \frac{(1-\delta)^2}{(1+\delta)^2}}, \quad \pi_G = \left(\sqrt{a} - \sqrt{t_G}\sqrt[6]{a\frac{t_G(1-\delta)}{t_F^2(1+\delta)}}\right)^2,$$

$$P = \left(v_G\left(\sqrt[3]{\frac{a t_F(1-\delta)}{t_G^2(1+\delta)}} - \frac{1-\delta}{1+\delta}\right) + v_F\right)\sqrt[3]{\frac{a t_G}{t_F^2}\frac{1-\delta}{1+\delta}} - \sqrt{s},$$

$$\Pi_{ISP} = P + f_H - \mu.$$

(i) First, the effect of priority on the total profits of the fringe is immediate: $\delta > 0$ ensures that $\Pi_F^{NN} > \Pi_F^{PP}$. The result on firm G's profits follows by evaluating the derivative of the equilibrium expression around $\delta = 0$. To check whether the profits increase or decrease under PP, $\frac{\partial \pi_G^{PP}}{\partial \delta}|_{\delta=0}$ is computed and it turns out that it is always positive under the maintained assumption $a > \widetilde{a}$.

(ii) 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.

(iii) As before, the results follow by evaluating the derivative of the equilibrium expressions around $\delta = 0$. To check whether the prices to end users increase or decrease under PP, $\frac{\partial P^{PP}}{\partial \delta}|_{\delta=0}$ is computed. The latter is positive iff $a < \overline{a}_P = \frac{t_G^2}{8 t_F}\frac{(4v_G - v_F)^3}{v_G^3}$. Turning to the profits of the ISP, it is

$$\frac{\partial \Pi_{ISP}^{PP}}{\partial \delta}|_{\delta=0} > 0 \text{ iff } a < \overline{a}_{ISP} = \frac{t_G^2}{t_F}\frac{(4v_G - v_F - 3\lambda - 3t_G)^3}{(2v_G - 2\lambda - 3t_G)^3}. \tag{22}$$

implying $\Pi_{ISP}^{PP} > \Pi_{ISP}^{NN}$.

(iv) If $v_G \simeq v_F$ and both are large compared to λ and t_G, then $\overline{a}_{ISP}$ tends to $\widehat{a} = \frac{27}{8}\frac{t_G^2}{t_F}$, which is the same threshold obtained in Proposition 1, part (iii).

Q.E.D.

Proof of Proposition 3. Under PP, the social welfare associated to the priority regime under the privately optimal allocation can be written as:

$$SW^{PP} = v_G x_H + v_F x_L - s\overline{W} + a - t_G x_H - t_F\frac{x_L^2}{2} - \frac{1}{\overline{W}} - \lambda(x_L + x_H).$$

Substituting the private equilibrium values, taking the derivative with respect to δ, evaluated around $\delta = 0$, leads to conclude that, under (12), $SW^{PP} > SW^{NN}$ iff $a < \overline{a}_{SW} = \frac{t_G^2}{t_F}\frac{(4v_G - v_F - 3\lambda - 4t_G)^3}{(2v_G - 2\lambda - 3t_G)^3}$. If $v_G \simeq v_F$ and both are large compared to λ and t_G, then $\overline{a}_{SW}$ tends to $\widehat{a} = \frac{27}{8}\frac{t_G^2}{t_F}$, which is again the same threshold obtained in Proposition 1, part (iii).

Q.E.D.
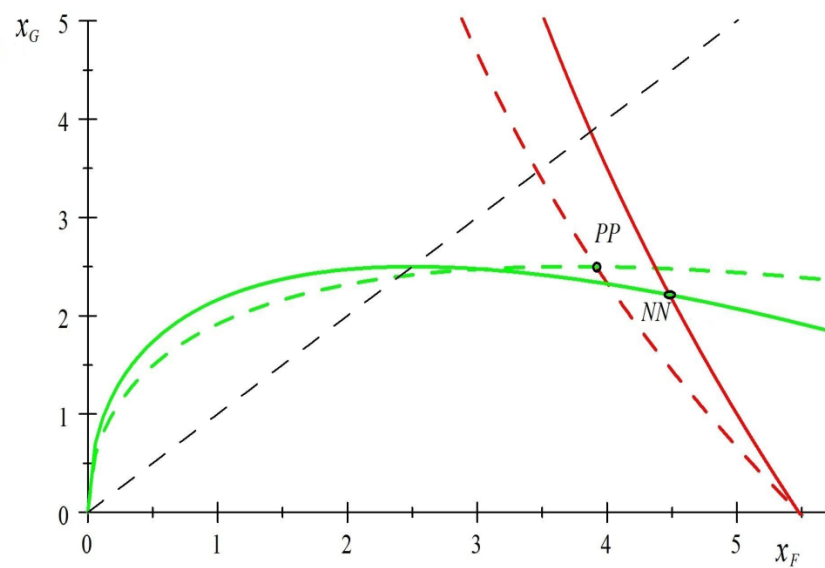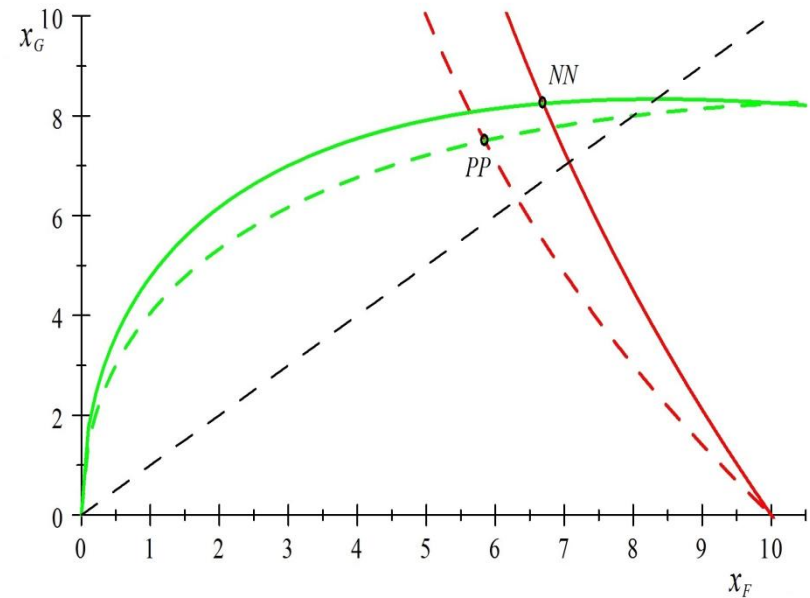
Fig. 1 Content choices with and without Net Neutrality.

The figure illustrates the best response functions of firm G and the zero profit conditions determining the entrance of the fringe firms. Firm G's best responses are in *green* and the fringe's zero profit conditions are in *red*. The *continuous* lines represent the Net Neutrality regime and the *dashed* lines the Priority Pricing regime. The points labelled *NN* and *PP* show the equilibrium content choices in the two regimes respectively. Panel **(a)** illustrates a case in which contents are *strategic substitutes*; Panel **(b)** a case in which contents are *strategic complements*. The parameter values are $t_G = 3$, $t_F = 1$, $\delta = 0.2$, as well as $a = 30$ (panel **(a)**) and $a = 100$ (panel **(b)**).