Network Neutrality: A Survey of the Economic Literature

Florian Schuett

Abstract

This paper reviews the small but growing economic literature on network neutrality. It considers a number of possible departures from network neutrality, in particular termination fees, second-degree price discrimination, and vertical foreclosure.

KEYWORDS: net neutrality, two-sided markets, network management, access-tiering, foreclosure

Author Notes: I am grateful to the editor, Julian Wright, an anonymous referee, as well as Andrea Amelio, Eric van Damme, Jens Prüfer and Jasper Sluijs for helpful comments and suggestions. All errors are my own.
1 Introduction

Roughly speaking, network neutrality refers to the principle that all data packets on an information network are treated equally. While the Internet had traditionally functioned more or less according to this principle, recent technological and legal developments have opened up the possibility for network operators to distinguish between packets. The advent of broadband, enabling the transmission of high-bandwidth applications and content over the Internet, has increased the demand for differentiated traffic management. As a result, the issue of whether and to what extent net neutrality should be maintained has become the subject of much debate in policy circles.

During the first few years, the debate took place in a theoretical vacuum, at least as far as rigorous economic analysis was concerned. In recent years, economists have answered the call, making net neutrality a lively field of theoretical research. This paper surveys the small but rapidly growing literature on the topic. It tries to categorize research and to provide a critical assessment of the assumptions used and the results obtained.

One reason why the debate about net neutrality is often confusing to outsiders is that the term is used in connection with several distinct concepts. In essence, net neutrality is a theoretical benchmark with which various practices that depart from it are contrasted. The fact that network operators cannot distinguish between packets means, on the one hand, that they cannot determine their origin. It follows that they cannot charge the originator of a packet a fee for transmitting it to users. Thus, net neutrality implies a zero-price rule. On the other hand, the fact that network operators cannot distinguish between packets means they cannot discriminate in terms of price or quality of transmission depending on the type, the origin, or the destination of a data packet. Operators cannot engage in traffic management by, e.g., prioritizing traffic, favoring certain packets over others. Thus, net neutrality implies a non-discrimination rule.

Most of the economic literature on the topic can be classified according to which of the two aspects of net neutrality is considered. Section 2 presents the part of the literature dealing with net neutrality as a zero-price rule. This line of research starts from the observation that the market for Internet access fits the definition of a two-sided market, and asks whether a network operator should be allowed to charge content providers for reaching its customer base. Section 3 presents the part of the literature dealing with net neutrality as a non-discrimination rule. It considers two separate issues: first, should a network operator be allowed to offer a menu of qualities to content providers? In the simplest case, this would mean creating a priority service with faster delivery, and charging content providers a premium for this service. Second, do network operators that are vertically integrated in the
content provision sector have an incentive to degrade rival traffic? Section 4 then considers departures from net neutrality that do not fit into either of the two previous categories, and have received less attention from economists. Finally, Section 5 concludes.

2 Net neutrality as a zero-price rule

In the current system, both content providers and consumers pay their respective Internet service provider (ISP) for accessing the Internet (see Figure 1). Content providers do not, however, pay consumers’ ISP for transmitting data to its customer base, which would amount to a termination fee (Lee and Wu, 2009). Several papers evaluate the desirability of a ban on termination fees using insights from the literature on two-sided markets. In a two-sided market, a platform facilitates interaction between two different sides (e.g., buyers and merchants in the case of credit cards, or players and game developers in the case of video game consoles). The Internet fits the definition of a two-sided market because of the existence of network externalities and limits on side-payments between the two sides of the market (Rochet and Tirole, 2006). Content providers benefit from a larger number of consumers (for example, through increased advertising revenue), while consumers benefit from a larger number of content providers (through greater variety of available content). Both sides depend on ISPs to connect to the Internet and access each other. At the same time, the two sides cannot easily undo the access charges levied by ISPs through side-payments, so the structure of these charges matters.¹

The literature on two-sided markets distinguishes between membership fees (Armstrong, 2006) and usage fees (Rochet and Tirole, 2003). Economides and T˚ag (2009) use Armstrong’s membership-fee model to show that a restriction on fees on the content-provider side may increase social welfare.² They study both the case of monopoly and duopoly ISPs and show that the results are qualitatively the same. A crucial condition for a zero-price rule to be welfare-enhancing is that content providers value additional consumers more highly than consumers value additional

¹This argument applies mainly to content financed through advertising. There is also content which is financed through direct payments from users to content providers. The price for such paid content could in principle adjust to charges levied by ISPs, thereby rendering the structure of these charges irrelevant. For various reasons, however, the bulk of Internet content is at least partially financed through advertising. See Lee and Wu for a detailed discussion of this important point.

²Arguably, termination fees would more accurately be modeled as usage fees. Modeling them as membership fees may be defensible because consumers and content providers typically do not have the same ISP. Thus, if the platform in the two-sided market is taken to be the consumers’ ISP, net neutrality implies there are no fees for content providers.
content providers. This is a necessary condition for the ISP to charge positive fees to content providers in the absence of a restriction.

To understand the condition, note that the ISP may not necessarily want to charge content providers. In a two-sided market, the platform often subsidizes one of the two sides of the market because this increases its value to the other side, i.e., the platform (partly) internalizes the network externality. An ISP may want to subsidize content providers because this increases the willingness to pay of consumers. If this is the case, imposing zero pricing decreases total surplus in Economides and Tåg’s model.

Note also that even if their conditions for zero pricing to increase total surplus are satisfied, Economides and Tåg find that it hurts consumers. This is because, when allowed to charge content providers, the ISP lowers its subscription price to attract more consumers and thus increase the willingness-to-pay of content providers. In addition to yielding ambiguous results, the model by Economides and Tåg is somewhat difficult to interpret. The sufficient conditions that are needed for a zero-price rule to raise welfare impose numerous restrictions on parameters. As pointed out by Caves (2010), it is hard to assess how reasonable these restrictions

---

3Lee and Wu point out that no ISP has so far introduced termination fees, which may be seen as evidence that ISPs do not find it profitable to charge content providers, although other explanations are possible.
are because the parameters of the model lack empirical counterparts.\textsuperscript{4}

Musacchio, Schwartz, and Walrand (2009) present a model with $N$ ISPs, each of which is a local monopolist. This setup allows them to study the negative externality that termination fees impose on other ISPs: each individual ISP captures the full benefit from charging a fee to content providers – higher revenue – but does not have to bear the full cost – decreased content on the Internet, lowering consumers’ willingness to pay – which is spread over all ISPs. As a result, ISPs overcharge content providers compared to the social optimum. As $N$ increases, this effect becomes more important and leads to an expansion of the set of parameters for which a zero-price rule is socially beneficial. Musacchio et al. show that, for a given $N$, authorizing termination fees is desirable when advertising revenue is low or demand for Internet subscriptions is elastic. This result derives from the complementarity between content and the quality of the network. Low advertising revenue means that content providers have too little incentive to invest in content, while elastic demand for subscriptions means ISPs have too little incentive to invest in the quality of the network. Both problems can be overcome if (positive or negative) termination fees are allowed. Negative fees, i.e., subsidies from ISPs to content providers, can compensate for low advertising revenue, and positive fees can compensate for low margins in the subscription market.

Lee and Wu (2009) provide a number of other arguments in support of a zero-price rule. The absence of fees may contribute to creation and invention on the Internet. Since the distribution of returns to content production is skewed and the expected value low, even moderate fees may have a strong negative effect on entry. Moreover, many business models on the Internet, such as social networking sites, depend on attaining a critical mass of users. A transactions cost argument implies that these businesses might not be viable if they had to negotiate with a large number of ISPs. Another danger is that fees might lead to fragmentation, with certain content being available from some ISPs but not others (as is the case with cable television), which can be inefficient in the presence of network effects and incomplete contracting.

Lee and Wu criticize the argument according to which termination fees would raise the incentive to invest in network infrastructure, arguing that, while they would probably raise ISPs’ profits, the impact on the \textit{marginal} incentive to upgrade existing infrastructure is indeterminate. Termination fees would unambiguously encourage investment only in those areas not yet covered by any broadband infrastructure.

\textsuperscript{4}Caves calculates that the restrictions on parameters imply that the ratio between the content providers’ profits and the ISP’s profits cannot exceed 0.4. He claims that this is at odds with the real-world division of profits between ISPs and content providers.
3 Net neutrality as a non-discrimination rule

Net neutrality, when interpreted as a non-discrimination rule, prevents ISPs from prioritizing certain traffic. This has the effect of restricting ISPs to offering a single product-line. Lifting such a restriction could lead to two types of responses. ISPs could sell different qualities at different prices, allowing content providers to choose their preferred quality (second-degree price discrimination). Alternatively, they could prioritize or degrade traffic without giving content providers a say about how their traffic is treated.

3.1 Offering a menu of qualities (second-degree price discrimination)

Hermalin and Katz (2007) analyze a model in which a monopolist ISP offers internet access to consumers and charges content providers a fee for connecting with its consumers. The fee depends on the quality of the connection chosen by the content provider. Content providers differ in the attractiveness of their content. High types (those with highly attractive content) have a higher marginal valuation for connection quality. The ISP, however, does not observe a content provider’s type and offers a menu of contracts to screen them. In equilibrium, higher types purchase higher qualities. But only the highest type obtains the efficient quality, while all other types obtain a quality that is distorted downwards. In an extension of the model, Hermalin and Katz consider a duopoly of ISPs and show that the menu of qualities offered to content providers is the same as in the monopoly case, implying that the policy implications are also the same.

In Hermalin and Katz’s model, restricting the ISP to offering a single connection quality has three effects on welfare. First, it reduces the set of content providers that are active. Low types, who in the absence of a product-line restriction would have purchased a low-quality connection, no longer purchase any connection at all. Second, it reduces the efficiency of connection of high types who are forced to purchase the single, lower quality. Third, there are some intermediate types who purchase a more efficient quality (recall the downward distortion of quality for all

5This is a standard result in contract theory. The intuition is that high types have an incentive to mimic low types in order to pay less. To induce the high types to choose the contracts designed for them they need to be given an information rent. Distorting the qualities for lower types below the efficient level reduces the high types’ information rent, thereby increasing the monopolist’s profit.

6The main intuition for this result is that, when a consumer “single-homes,” i.e., buys access from only one ISP, this ISP has a monopoly over access to the consumer. Hermalin and Katz derive a condition on the cost of providing Internet access to consumers that ensures that single-homing arises in equilibrium.
but the highest type in the unrestricted case). The first and second effect decrease welfare, while the third effect increases it. The overall effect of a product-line restriction is ambiguous, but often negative, as Hermalin and Katz argue.

They obtain two other noteworthy results. First, requiring the ISP to supply the highest possible connection quality reduces the content available to consumers even more than a product-line restriction not specifying the connection quality. While it eliminates the inefficiency from high types being forced to purchase suboptimal quality, the ISP will charge a price that makes it unattractive to connect for more types than before. Second, preventing the ISP from charging content providers (i.e., a zero-price rule) leads the ISP to supply only one connection quality. This quality will be lower than efficient, and also lower than what the ISP would choose when allowed to charge content providers but restricted to offering a single quality.

Choi and Kim (forthcoming), Cheng, Bandyopadhyay, and Guo (forthcoming), and Krämer and Wiewiorra (2010) explicitly take into account network congestion. All of them use results about congestion from queuing theory in operations research to evaluate the welfare effects of discrimination. Queuing theory is the standard framework to analyze congestion in computer networks. In the strikingly similar models of Choi and Kim and Cheng et al., a monopolist ISP connects consumers with two competing content providers. The content providers are located at the ends of the standard Hotelling line and may differ in their profit margin from delivering content. Such a difference in margins could stem from differences in advertising revenue, differences in the cost of delivering content, or both. In Krämer and Wiewiorra, there is a continuum of content providers differing in the extent to which their content is sensitive to delay.

In all three papers, advertising is the content providers’ only source of revenue. They further assume that under non-discrimination, the ISP does not charge any of the content providers a fee, and traffic is delivered on a first-in first-out basis. Under a discriminatory regime, the ISP can offer a priority service. Prioritized content is delivered before non-prioritized content. Queuing theory can then be used to determine the expected waiting times for traffic from each of the content providers. In particular, the theory predicts that prioritized traffic has shorter waiting times than traffic under non-discrimination, while non-priority traffic has longer waiting times. The average waiting time for traffic is independent of whether or not there is a priority service. Importantly, as network capacity increases, the difference in waiting times between priority and non-priority traffic becomes smaller.

The difference between Choi and Kim and Cheng et al. stems from their modeling of how priority service is sold. Choi and Kim assume that the ISP sells a “fast lane” to the highest bidder. In equilibrium, it is the high-margin content provider that purchases the fast lane. As a result, this provider’s content can be
accessed faster, leading some consumers to shift from the low-margin to the high-
margin content provider. Choi and Kim show that the prioritization of traffic then
has three short-run effects. It increases the average profit margin from content pro-
vision since the high-margin firm now serves a larger share of the market. For the
same reason, however, consumers’ “transportation costs” increase (not everybody
consumes their favorite content anymore). Finally, there is an effect on waiting
costs that depends on which content provider’s traffic is more sensitive to delay. If
both are equally sensitive, prioritization has no effect on average waiting costs. The
overall welfare effect then depends on how the difference in profit margins between
the two content providers compares to the parameter measuring consumers’ trans-
portation cost. A discriminatory regime leads to higher short-run welfare if and
only if the difference in margins is large.

Cheng et al. assume that the ISP sells priority service at a fixed price to
all content providers that demand it. In equilibrium, either only the high-margin
content provider buys priority service (Case A), or both do (Case B). Cheng et al.
show that Case A arises when the difference in margins between content providers
is large, while Case B arises when the difference is small.\footnote{Case B resembles
the outcome of a prisoner’s dilemma. Choi and Kim argue that the sale
procedure adopted by Cheng et al. is less plausible than their own because it
relies on the ISP’s ability to commit to selling priority service only once. They
also examine the relative profitability of the two procedures for the ISP in an appendix.} This leads to welfare
results that differ markedly from Choi and Kim’s: in Case B, both content providers
purchase priority service, so no content is actually prioritized. The discriminatory
regime merely transfers surplus from content providers to the ISP while leaving
total surplus unchanged. In Case A, the discriminatory regime is associated with
higher welfare than non-discrimination. This is because in Cheng et al.’s setting,
Case A occurs precisely when priority service is welfare-enhancing: namely, when
the difference in profit margins is relatively large.

In the long-run, the effects of non-discrimination also depend on providers’
investments: the ISP’s investment in capacity expansion, and the content providers’
investments in the quality of their content (which is assumed to result in higher
profit margins). Capacity expansion affects the ISP’s profits through two channels:
on the one hand, by leading to faster delivery, it allows him to charge consumers
more for their internet subscription; on the other hand, it also affects the amount
that he can charge content providers for the fast lane. In a discriminatory regime,
capacity expansion decreases the value of the fast lane (recall that waiting times
become more similar). This suggests that, contrary to ISPs’ claims, the possibility
of prioritization may depress investment in network capacity. Choi and Kim are
unable to show this in general, however, because the effect on the profits from
consumer subscriptions is ambiguous. Cheng et al. arrive at similar conclusions,
but show that the region of the parameter space for which non-discrimination leads to greater investment is relatively small.

Choi and Kim show that the effect on investment by content providers is also unclear: while prioritization decreases the low-margin provider’s incentive to invest, the high-margin provider’s incentive depends on its bargaining power. In an extension, Choi and Kim depart from the assumption that waiting times are determined mechanically and investigate the ISP’s incentive to degrade low-priority traffic. By making the fast lane more attractive, degradation increases the content providers’ willingness to pay. Therefore, degradation can be profitable for an ISP even if he is not vertically integrated with (one of the) content providers.\footnote{At the same time, Choi and Kim note that the possibility of degradation might restore the ISP’s incentive to invest in capacity expansion, because it can offset the decline in the value of the fast lane.}

Krämer and Wiewiorra add to the picture the positive network externalities that content providers exert on consumers. In their model, consumer surplus increases with the variety of content available. Content providers are not in competition with each other; every user visits each content provider an exogenous number of times. The number of visits is independent of the number of active content providers and of network waiting times. A content provider’s advertising revenue, however, decreases with waiting time, and the magnitude of this effect depends on how sensitive his content is to delay. There is a threshold of sensitivity to delay above which content providers choose not to be active. Users are identical, so that the ISP extracts all the surplus, and all users connect to the network in equilibrium.

Under these assumptions, Krämer and Wiewiorra find that a discriminatory regime improves welfare both in the short run and in the long run. Because each content provider is visited equally often, average waiting times are independent of whether or not there is a priority service. As a result, the number of content providers that are active in the market is the same in the discriminatory and the non-discriminatory regime; the ISP extracts the increased surplus of content providers with priority service through the fee for the service. Since waiting times and content variety are the same, the only short-run difference between the two regimes comes through the fact that in a discriminatory regime, content that is highly sensitive to delay is delivered faster, while less sensitive content is delivered slower. This increases efficiency and short-run welfare.

In Krämer and Wiewiorra’s model, capacity expansion by the ISP has three effects. By reducing waiting times, capacity expansion decreases consumers’ waiting costs and increases the number of active content providers, both of which boost consumers’ willingness to pay for Internet access. These two effects arise regardless of whether or not discrimination is possible. A third effect arises only in the

http://www.bepress.com/rne/vol9/iss2/1
DOI: 10.2202/1446-9022.1224
discriminatory regime, where the larger number of active content providers also generates more revenue from the sale of priority service. Because of this last effect, Krämer and Wiewiorra find that the discriminatory regime is associated with greater capacity investment, and thus long-run welfare. It should be noted, however, that the effect on revenues from priority service crucially depends on their assumptions, which imply in particular that the optimal price for priority service is independent of network capacity. This contrasts with Choi and Kim’s framework, in which increased capacity decreases the value of priority service. By ignoring some of the tradeoffs involved, Krämer and Wiewiorra portray discrimination in a way that might well be overly optimistic.

Nevertheless, on the whole, second-degree price discrimination seems to be a rather benign departure from net neutrality; while its welfare effects are not entirely clear, in many scenarios they are indeed likely to be positive. Lee and Wu also view the sale of a fast lane as relatively unproblematic, as long as “basic” service remains free. They warn, however, that there is a risk that ISPs might degrade the basic service so much that it would cease to be an attractive option for any content provider. Such a situation would be equivalent to a generalized termination fee.

3.2 Degradation of traffic (vertical foreclosure)

One of the main arguments against a strict net-neutrality rule is that applications and content differ in their sensitivity to delay. Real-time, high-bandwidth applications such as Voice-over-Internet-Protocol (VoIP), video streams, and online games are more sensitive to delay than, say, email. It makes intuitive sense to manage traffic in such a way that delay is minimized for content that is most sensitive to it. This could bring about service improvements for users even without ISPs charging content providers for (priority) access to consumers. At the same time, proponents of net neutrality fear that this kind of traffic management could open the door to less innocuous practices. Many ISPs are vertically integrated in the content provision sector. For example, many of them offer their own VoIP and video-on-demand services which compete with services from independent content providers. Observers have voiced concerns that ISPs may have incentives to degrade the quality of competing services to increase the demand for their own.

This is an example of what is generally referred to as vertical foreclosure. The only paper that addresses foreclosure in the context of broadband services is Chen and Nalebuff (2006). They study a model with two complementary products of which only one is essential. The essential good is supplied by firm A, while the non-essential good is supplied by firm B and, potentially, firm A (if it chooses

9See Rey and Tirole (2007) for an overview.
to enter B’s market). Chen and Nalebuff show that when A and B compete in the market for the non-essential good, A drives down the price to marginal cost unless the value that consumers derive from the non-essential good is sufficiently large; A extracts the surplus by charging a higher price for the essential good. They go on to look at the case where A has the possibility to degrade the quality of B’s product, and find that it has no incentive to do so assuming the non-essential good is of low value. This result, however, seems to be an artifact of their low-value assumption, for which (according to their previous result) the price of the non-essential good is equal to marginal cost. In that setting, there is no profit to be earned in the B-market, and therefore no incentive to degrade the competing product.

The result that an ISP has no incentive to degrade competing services also seems implausible in the light of available empirical evidence. In 2005, the Federal Communications Commission (FCC) took action against Madison River Communication, a regional ISP that was blocking VoIP services. Many European mobile operators similarly block Skype on 3G networks. Both suggest that ISPs may find it profitable to exclude or degrade rival services. Additional evidence for foreclosure comes from the cable television industry, which shares a number of characteristics with the provision of Internet access. Much like ISPs, cable operators act as platforms, connecting program service providers with their customers. In addition, they are often vertically integrated in the program service sector. In a study of the cable TV industry in the United States, Chipty (2001) shows that cable operators that are integrated with program service providers offer fewer rival programs than non-integrated operators.

Note that degradation does not have to occur in a direct way. In fact, direct degradation or outright blocking of rival services may in many cases be a violation of antitrust or competition laws (at least when the firm in question has market power). ISPs may, however, disguise such practices by creating a two-tiered access structure and pricing the fast lane prohibitively high. Only their own downstream affiliates would then be able to profitably use the fast lane, while competitors would be left with the slow lane, which ISPs could make sure to be of inadequate quality to compete effectively with their own integrated services (Economides, 2008).

4 Other aspects

Kocsis and de Bijl (2007) argue that the possibility for differentiation will lead competing ISPs to strike exclusive deals with content providers. Thus, certain content would only be available from one ISP, while other content would only be available

---

10The FCC’s investigation was dropped under the terms of a settlement reached between the parties; see http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-543A2.pdf.
from another. Such a strategy would result in horizontal differentiation between ISPs and thus reduce competition. Economides and Täg express similar concerns. In their view, the fact that ISPs may give exclusive (priority) access to one of several competing content providers, e.g., through an auction, would decrease innovation, since small firms with financial constraints are unlikely to win such an auction.

Prüfer and Jahn (2007) point to the fact that the broadband industry faces what they call a “capacity paradox.” Many high-bandwidth applications on the Internet depend on the existence of excess capacity. ISPs, however, face tougher competition when there is excess capacity, and therefore have little incentive to keep expanding capacity as the demand for it increases. Eventually, excess capacity is bound to disappear. This would go to the benefit of ISPs, but to the detriment of content providers. If one follows their argument, allowing for network management and prioritization may be the only way to restore the quality of high-bandwidth applications, and thus incentives to develop such applications.

5 Conclusion

A general theme that emerges from the literature is that the welfare effects of net-neutrality regulation tend to be ambiguous. A zero-price rule may increase or decrease welfare, depending on the relative magnitudes of the network externalities between consumers and content providers, among other things. A non-discrimination rule may increase or decrease short-run welfare, depending on assumptions about the nature of competition between content providers, the organization of the sale of priority service, and various other parameters. Furthermore, a non-discrimination rule may increase or decrease investment in network capacity and thus long-run welfare.

The policy implication of these ambiguous results is that, given our current understanding, the case for strict net-neutrality regulation is not compelling. This is even more so considering that the only net-neutrality violation of which there have been actual examples is degradation of rival services, an issue that existing antitrust and competition laws may well be able to deal with without additional regulation. Still, this conclusion is tentative, and much work remains to be done.

Looking ahead, net neutrality promises to be a fruitful area for economic theorizing, precisely because most of the departures from net neutrality that are feared by its proponents are, so far, purely hypothetical. Economic theory can be useful to assess the likely effects of regulation aimed at preventing such departures. Since economic theory is one of the few available tools to inform policy makers about the potential benefits of regulation, its results are likely to be influential. The flip side of the counterfactual nature of the research is, of course, the absence of
empirical facts that could discipline the analysis. For this reason, net-neutrality researchers must be particularly cautious about the assumptions from which they derive their results and the conclusions they draw.

References


