1 Network Economics: Neutrality, Competition, and Service Differentiation

John Musacchio†, Galina Schwartz‡, and Jean Walrand‡
University of California, Santa Cruz† University of California, Berkeley‡

In 2007 Comcast, a cable TV and Internet Service Provider in the United States, began to selectively rate limit or "shape" the traffic from users of the peer-to-peer application Bit Torrent. The access technology that Comcast uses is asymmetric in its capacity - the "uplink" from users is much slower than the "downlink." For client-server applications like web, this asymmetry is fine, but for peer-to-peer, where home users are serving up huge files for others to download, the uplink quickly becomes congested. Comcast felt that it had to protect the rest of its users from a relatively small number of heavy peer-to-peer users that were using a disproportionate fraction of the system's capacity. In other words, peer-to-peer users were imposing a negative externality by creating congestion that harmed other users.

This negative externality reduces the welfare of the system because users act selfishly. The peer-to-peer user is going to continue to exchange movies even though this action is disrupting his neighbor's critical, work-related video conference. Comcast thought that by singling out users of peer-to-peer applications, it could limit the ill effects of this externality and keep the rest of its users (who mostly don't use peer-to-peer) happy. Instead Comcast's decision placed them in the center of the ongoing network neutrality debate. Supporters of the network neutrality concept feel that the Internet access provider ought not to be allowed to "discriminate" between traffic of different users or different applications. Said another way, the provider of a network layer service should not act on information about the application layer that generated the traffic. Some supporters of net-neutrality argue that this principal of non-discrimination is not just a good architectural principle, but it is also vital for a number of societal reasons such as maintaining free-speech, and the ability for new, innovative content providers to enter the market and have the same potential reach as entrenched incumbents with more resources.

Related to the question of whether ISPs ought to be allowed to treat traffic from different applications differently, there are the ideas of service differentiation and quality of service. Service differentiation is not a new concept in communications networks. For instance, over a decade ago the designers of the Asynchronous Transfer Mode (ATM) networking technology understood that traffic from interactive applications like telephony is much more sensitive to delay than traffic from file transfers which is also more bursty. Therefore mixing the traffic of both types together has the potential to greatly reduce the utility of the entire system by hurting the users of interactive applications. Many have argued that these problems would disappear as the capacity of the network's links grew exponentially. The argument was that the network would be so fast, that traffic of all types would see negligible delays, so the differences in traffic requirements would not matter. However as the network's capacity grew, so did the consumption of that capacity by new applications, like peer-to-peer and streaming video. When Comcast recently deployed a new voice over IP (VoIP) telephony service in their network, they indeed chose to keep that traffic protected from the congestion of other Internet traffic, and not coincidentally from traffic from other providers of VoIP service like Skype. The question of whether service differentiation ought to be included in the Internet might still be debated, but the reality is that it is being deployed today. The danger is that it will be deployed in an ad-hoc way by different ISPs, and the possible benefits of having a coherent architectural approach across ISPs will be lost. For instance the ability to independently select application service provider and Internet service provider might be lost. For example the Skype/Comcast "stack" for VoIP could be at a severe disadvantage compared to a Comcast/Comcast stack.

Another directly related set of questions is how revenue ought to be shared between the players of this "stack." What is the right way to share revenue between content providers and ISPs? In particular, should ISPs like Comcast be allowed to charge Skype a fee for offering services to Comcast subscribers? This is one of the questions we will examine in this chapter.

The potential of a better Internet is enormous. Today, the Internet delivers insufficient or inconsistent service quality for many applications. For instance, companies pay thousands of dollars for access to private networks for business grade video conferencing because the quality on the public Internet in many cases is too unreliable. As the costs of energy rise, it is not hard to envision a day when most of the workforce will not be able to commute to their jobs every day. If that were to happen, the continued productivity of our economy will depend on the workforce having access to high quality, reliable, interactive applications. They would even be willing to pay for them – a few dollars to have a reliable HDTV video conference with colleagues would be worth it if it saves having to buy a $100 tank of gasoline for instance. Unfortunately, today's Internet does not give one the option of paying a little extra for quality.

In addition to the likely growth in importance of interactive applications, another important trend is the move toward cloud computing and the related trend of service oriented architecture. With these approaches an organization can push parts of their computing and Information Technology (IT) infrastructures out of their own facilities and instead "outsource" them to specialized providers. This capability has the potential to greatly lower administrative costs, and also make an organization's IT system much more adaptable to changing needs. However, this approach requires a reliable, fast, and low latency network in order that these distributed services be as responsive for users as a more traditional approach with mostly on-site infrastructure.

As we have implied there are a large number of economic questions facing the Internet, and how they get resolved will arguably be the driving force behind the future evolution of the Internet. We cannot address all of the questions in depth in this chapter, but instead we will limit our discussion to three. First we will look at the question of revenue sharing between content providers and ISPs; an issue that we argue is core to the net-neutrality debate. Next we will discuss some fundamental modeling work that looks at the economic efficiency that results when multiple interconnected ISPs compete on price, and users seek lower prices and lower delays. Lastly, we examine some of the basic issues behind service differentiation.

Economic modeling is also central to the study of architectures that employ explicit congestion notification and charging in order to achieve fairness and utility maximization across a network of users with rate-dependent utility functions. These ideas are extremely useful both for understanding the behavior of existing protocols like TCP, and for the design of new networking technologies. In this volume Chapter 4.3 by Kelly and Raina discusses the issues behind explicit congestion notification and Chapter 3.2 by Yi and Chiang addresses the network utility maximization approach for wireless networks.

1.1 Neutrality

Today, an Internet service provider (ISP) charges both end-users who subscribe to that ISP for their last-mile Internet access as well as content providers that are directly connected to the ISP. However, an ISP generally does not charge content providers that are not directly attached to it for delivering content to end-users. One of the focal questions in the network neutrality policy debate is whether these current charging practices should continue and be mandated by law, or if ISPs ought to be allowed to charge all content providers that deliver content to the ISP's end-users. Indeed the current network neutrality debate began when the CEO of AT&T suggested that such charges be allowed [1].

To address this question, we develop a two-sided market model of the interaction of ISPs, end-users, and content providers. A more complete description of this work is available in [2]. The model is closely related to the existing two-sided markets literature as we detail later in this section. We model a "neutral" network as a regime in which ISPs are allowed to charge only content providers that buy their Internet access from them. We argue that with such a charging structure, the ISPs compete on price to attract content providers to buy their access from them, driving these prices to the ISP's costs. For simplicity we normalize the price content providers pay ISPs to be net of ISP connection cost, so in a "neutral" network only the end-users (and not the content providers) pay a positive price to the ISPs. In a “non-neutral” network, all ISPs are allowed to charge all content providers, and thus ISPs extract revenues from both content providers and end-users.

The question we address in this work is part of the larger debate on network neutrality , which includes diverse issues such as whether service differentiation should be allowed, or whether charges for content constitute an impingement of freedom of speech (see [3] and [4]). In 2006 there was a considerable divergence of opinions on the subject of net-neutrality. Indeed the issue was intensely debated by law and policy makers, and the imposition of restrictive network regulations on Internet service providers (ISPs) in order to achieve network neutrality seemed likely. In 2007 the situation began to change. In June of that year, the Federal Trade Commission (FTC) issued a report, forcefully stating the lack of FTC support for network neutrality regulatory restraints, and warning of “potentially adverse and unintended effects of regulation [5].” Similarly, on September 7, 2007, the Department of Justice issued comments “cautioning against premature regulation of the Internet [6].” However, US president Barack Obama, elected in 2008, voiced support for network neutrality, thus the debate is far from over. We do not attempt to address all of the questions in the network neutrality debate. We only study the issue of whether ISPs ought to be allowed to charge content providers for accessing the end-users.

Our model is based on the ideas of two-sided markets, and there is a large literature on the subject. For a survey of two-sided markets see for example works by Rochet and Tirole [7] and Armstrong[8]. Other work has used the ideas of two-sided markets to study network neutrality, For instance, Hermalin and Katz [9] model network neutrality as a restriction on the product space, and consider whether ISPs should be allowed to offer more than one grade of service. While Hogendorn [10] studies two-sided markets where intermediaries sit between “conduits” and content providers. In his context, net-neutrality means content has open access to conduits where an “open access” regime affords open access to the intermediaries. Weiser [11] discusses policy issues related to two-sided markets. The work we describe here is unique from other studies of network neutrality in that we develop a game-theoretic model to study investment incentives of the providers under each network regime.

1.1.1 Model

Figure 1.1 illustrates our setting. In the model, there are M content providers and N ISPs. Each ISP T_n is attached to end-users U_n $(n = 1, 2, \ldots, N)$ and charges them p_n per click. The ISP T_n has a monopoly over its end-user base U_n. Thus, the end-users are divided between the ISPs, with each ISP having $1/N$ of the entire market. This assumption reflects the market power of local ISPs. Each ISP T_n charges each content provider C_m an amount equal to q_n per click. Content provider C_m invests c_m and ISP T_n invests t_n.

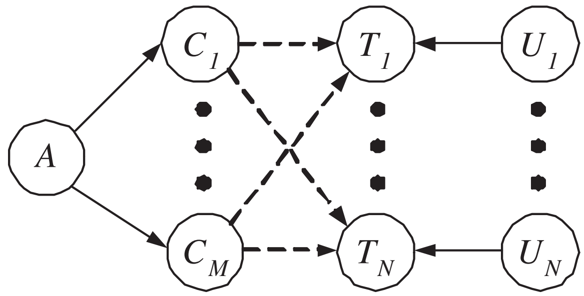

Figure 1.1 The direction of payments in the model. Each C_m is a content provider; each T_n is an ISP; and each U_n is a pool of users subscribing to ISP U_n. The A block denotes the set of advertisers that pay the content providers. The dotted lines indicate payments made only with two-sided pricing ("non-neutral").

Recall from the beginning of this section that we measure q net of content providers' access payment (for network attachment), which is set at marginal cost due to the competition amongst ISPs. Accordingly, we measure the content provider per-user charges to advertisers (which we denote as a) net of the content provider's access payment.

We characterize usage of end-users U_n by the number of "clicks" B_n they make. Since Internet advertising is most often priced per click, clicks are a natural metric for expressing advertising revenue. It is a less natural metric for expressing ISP revenue from end-users, because ISPs do not charge users per click but rather base their charges on bits. However, it is convenient to use only one metric and argue that one could approximate one metric from knowledge of the other using an appropriate scaling factor. The rate B_n of clicks of end-users U_n depends on the price p_n but also on the quality of the network, which in-turn is determined by provider investments. The rate of clicks B_n, which characterizes end-user demand, depends on the end-user access price p_n and investments as

$$B_n = \left\{ \begin{array}{l} \frac{1}{N^{1-w}}(c_1^v + \cdots + c_M^v)\times \\ \qquad\qquad \left[(1-\rho)t_n^w + \frac{\rho}{N}(t_1^w + \cdots + t_N^w)\right] \end{array} \right\} e^{-p_n/\theta} \tag{1.1}$$

where $\rho \in (0,1)$, $\theta > 0$, and $v, w \geq 0$ with $v + w < 1$. For a given network quality (the expression in the curly brackets) the rate of clicks exponentially decreases with price p_n.

The term $c_1^v + \cdots + c_M^v$ is the value of the content providers' investments as seen by a typical end-user. This expression is concave in the investments of the individual providers, and the interpretation is that each content provider adds value to the network. Also note that the structure of the expression is such that end-users value a network in which content is produced by numerous content providers higher than a network in which the content is provided by a single provider with the same cumulative investment. Our end-users' preference for content variety is similar to that of the classical monopolistic competition model by Dixit and Stiglitz [12]. In the expression (1.1), the term in square brackets reflects the value of the ISP's investments for end-users. Clearly users U_n value the investment made by their ISP, but they may also value the investments made by other ISPs. For instance, a user of one ISP might derive more value by having a better connection with users of another ISP. In our model the parameter ρ captures this spill-over effect. When $\rho = 1$, end-users U_n value investments of all ISPs equally while when $\rho = 0$, they value only the investment of their ISP. When $\rho \in (0, 1)$ end-users U_n value investment of his ISP T_n more than investments of other ISPs $T_k \neq T_n$. The term ρ reflects the direct network effect of ISPs on end-users (not between them and content providers). This effect captures a typical network externality (see [13] for a discussion of investment spill-over effects). The factor $1/N^{1-w}$ is a convenient normalization. It reflects the division of the end-user pool among N providers and it is justified as follows. Suppose there were no spill-over and each ISP were to invest t/N. The total rate of clicks should be independent of N. In our model, the total click rate is proportional to $(1/N^{1-w})(N(t/N)^w)$, which is indeed independent of N.

The rate R_{mn} of clicks from end-users U_n to C_m is given by

$$R_{mn} = \frac{c_m^v}{c_1^v + \cdots + c_M^v} B_n. \tag{1.2}$$

Thus, the total rate of clicks for content provider C_m is given by

$$D_m = \sum_n R_{mn}.$$

We assume that content providers charge a fixed amount a per click to the advertisers. Each content provider's objective is to maximize its profit which is equal to revenues from end-user clicks net of investment costs. Thus

$$\Pi_{Cm} = \sum_{n=1}^{N} (a - q_n) R_{mn} - \beta c_m$$

where the term $\beta > 1$ is the outside option (alternative use of funds c_m).

ISP T_n profit is

$$\Pi_{Tn} = \ (p_n + q_n) B_n - \alpha t_n.$$

where $\alpha > 1$ is the outside option of the ISP. We assume providers of each type are identical and we will focus on finding symmetric equilibria for both one- and two-sided pricing.

1.1.2 The Analysis of One- and Two-sided Pricing

To compare one-sided and two-sided pricing (neutral and non-neutral networks), we make the following assumptions.

(a) One-sided pricing (neutral network): In stage 1 each T_n simultaneously chooses (t_n, p_n). The price q_n charged to content providers is constrained to be 0. (Recall the discussion in section 1.1.) In stage 2 each C_m chooses c_m.

(b) Two-sided pricing (non-neutral network): In stage 1 each T_n simultaneously chooses (t_n, p_n, q_n). In stage 2 each C_m chooses c_m.

In both cases, we assume that content providers observe ISP investments, and can subsequently adjust their investments based on the ISPs' choices. We justify this assumption by the difference in time and scale of the required initial investments. The investments of ISPs tend to be longer-term investments in infrastructure, such as deploying networks of fibre-optic cable. Conversely, the investments of content-providers tend to be shorter-term and more ongoing in nature, such as development of content, making ongoing improvements to a search algorithm, or adding/replacing servers in a server farm.

1.1.2.1 Two-Sided Pricing

In a network with two-sided pricing (non-neutral network), each ISP chooses (t_n, p_n, q_n) and each content provider chooses c_m. To analyze the game we use the principle of backwards induction – analyzing the last stage of the game first and then working back in time supposing that the players in earlier stages will anticipate the play in the latter stages. A content provider C_m in the last stage of the game should choose the optimal c_m after having observed actions (t_n, p_n, q_n) from the preceding stage of the game. Because of a cancelation of terms, it turns out that content provider C_m's profit Π_{Cm} is independent of other content provider investments c_j, $j \neq m$. Therefore, each content provider's optimization is unaffected by the simultaneously made (but correctly anticipated in equilibrium) investment decisions of the other content providers. We therefore simply can find the optimal c_m as a function of (t_n, p_n, q_n), and since the ISPs should anticipate that the content providers will play their best response, we can substitute the function $c_m(t_n, p_n, q_n)$ into the expression for each ISP T_n's profit Π_{Tn}. Each ISP T_n gets to choose the investment and prices (t_n, p_n, q_n) that maximize his profit Π_{Tn}. Since the simultaneous decisions of each of the ISPs affect each other, in order to find a Nash equilibrium we need to identify a point where the best response functions intersect. By carrying out this analysis, we can find closed-form expressions for the Nash equilibrium actions and profits of the content and transit providers (see our working paper, [14] for the detailed derivations):

$$\begin{aligned}
p_n &= p = \theta - a; &(1.3)\\
q_n &= q = a - \theta\frac{v}{N(1-v)+v};\\
t_n &= t \text{ with } (Nt)^{1-v-w} = x^{1-v}y^{v}e^{-(\theta-a)/\theta};\\
c_m &= c \text{ with } c^{1-v-w} = x^{w}y^{1-w}e^{-(\theta-a)/\theta};\\
\Pi_{Cm} &= \Pi_C := \left(\frac{\theta v(1-v)}{N(1-v)+v}\right)\left[x^{w}y^{v}e^{-(\theta-a)/\theta}\right]^{\frac{1}{1-v-w}}; &(1.4)\\
\Pi_{Tn} &= \Pi_T := \left(\frac{M\theta(N(1-v)(1-w\phi)-wv)}{N(N(1-v)+v)}\right)\left[x^{w}y^{v}e^{-(\theta-a)/\theta}\right]^{\frac{1}{1-v-w}}
\end{aligned}$$

where

$$x := \frac{M\theta w}{\alpha}\frac{N\phi(1-v)+v}{N(1-v)+v};\quad y := \frac{\theta}{\beta}\frac{v^2}{N(1-v)+v}. \tag{1.5}$$

1.1.2.2 One-Sided Pricing

The analysis of the one-sided case is similar that of the two-sided pricing (non-neutral), except that $q_n = 0$ as we argued in section 1.1.1 for $n = 1, \dots, N$. We use the same backwards induction approach that we described for analyzing the two-sided case. Doing that leads to the following solution for the content providers' and ISPs' actions and profits in equilibrium (see [14] for the derivations):

$$\begin{aligned}
p_n &= p_0 := \frac{\theta N(1-v)}{N(1-v)+v};\\
q_m &= 0;\\
t_n &= t_0 \text{ where } (Nt_0)^{1-v-w} = x^{1-v}y_0^{v}e^{-p_0/\theta}\\
c_m &= c_0 \text{ where } c_0^{1-v-w} = x^{w}y_0^{1-w}e^{-p_0/\theta}\\
\Pi_{Cm} &= \Pi_{C0} := a(1-v)\left[x^{w}y_0^{v}e^{-p_0/\theta}\right]^{\frac{1}{1-v-w}} &(1.6)\\
\Pi_{Tn} &= \Pi_{T0} := \left(\frac{M\theta(N(1-v)(1-w\phi)-wv)}{N(N(1-v)+v)}\right)\left[x^{w}y_0^{v}e^{-p_0/\theta}\right]^{\frac{1}{1-v-w}}
\end{aligned}$$

where x is given in (1.5), and $y_0 := \frac{av}{\beta}$.

In section 1.1.4 we compare the profits and social welfare of the two regimes for a range of parameters.

1.1.3 User Welfare and Social Welfare

In order to compare the one-sided and two-sided regimes, we would like to find expressions for the social welfare of each regime. Social welfare is simply the sum of the net payoffs of all the players involved – content providers, ISPs, and users. The welfare of each content provider and ISP is simply their profit, but for users we need to define their welfare. To do this, we will use the concept of consumer surplus, which is the total amount of the difference between what consumers would have been willing to pay for the service they consumed and what they actually had to pay. In order to make this calculation, we need a demand function that relates end-user price to the quantity consumed, and for our model that is the expression that relates total click rate to price. We compute the consumer surplus by taking the integral of the demand function from the equilibrium price to infinity. This integral is taken with the investment levels of content providers and ISPs fixed. We find

$$W_U(\text{two-sided}) = M\theta x^{w/(1-v-w)} y^{v/(1-v-w)} e^{-\frac{\theta-a}{\theta(1-v-w)}}.$$

The expression for the one-sided case is the same, but with y exchanged for y_0 and $\theta - a$ in the exponent exchanged with p_0. The ratio of the social welfare with one- vs. two-sided pricing has the form

$$\frac{W_U(\text{one-sided}) + N\Pi_T(\text{one-sided}) + M\Pi_C(\text{one-sided})}{W_U(\text{two-sided}) + N\Pi_T(\text{two-sided}) + M\Pi_C(\text{two-sided})}.$$

1.1.4 Comparison

To compare the revenue in the one- and two-sided cases, we define the ratio

$$r(\Pi_C) := \left(\frac{\Pi_C(\text{one-sided})}{\Pi_C(\text{two-sided})}\right)^{1-v-w}$$

where $\Pi_C(\text{two-sided})$ is the profit per content provider in the two-sided case as expressed in (1.4) and $\Pi_C(\text{one-sided})$ is the profit per content provider with one-sided pricing (1.6). We define $r(\Pi_T)$ similarly. We find

$$r(\Pi_T) = \left(\frac{\delta}{\pi}\right)^{v} e^{\pi-\delta}, \quad r(\Pi_C) = \left(\frac{\delta}{\pi}\right)^{1-w} e^{\pi-\delta}. \tag{1.7}$$

where

$$\pi := \frac{v}{N(1-v)+v} \text{ and } \delta := \frac{a}{\theta}.$$

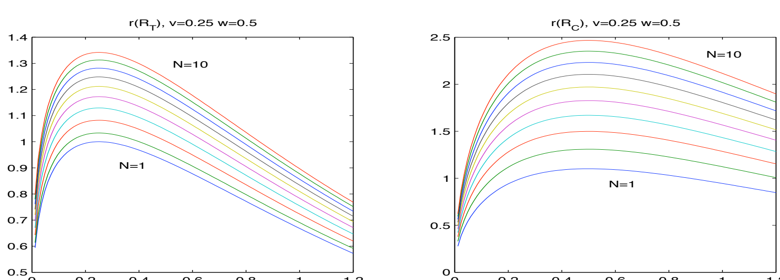

Figure 1.2 The ratios of profits ($v = 0.25, w = 0.5$) for different values of N.

Figure 1.2 shows the ratios of revenues with one- vs. two-sided pricing for both content providers and ISPs. The figures show that for small or large values of

a/θ, the ratio of advertising revenue per click to the constant characterizing price sensitivity of end-users, two-sided pricing is preferable to both content providers and ISPs. (Here we say "preferable" in that the revenues are larger, though we have seen that the rate of return on investments are the same.) For mid range values of a/θ, one-sided pricing is preferable to both, though the values of a/θ where the transition between one-sided being preferable to two-sided are not exactly the same for content providers and ISPs. Furthermore, as N, the number of ISPs increases, the range of a/θ values for which one-sided pricing is superior increases, while also the degree by which it is superior (in terms of revenues to content providers and ISPs) increases.

These results can be explained by the following reasoning. When a/θ is large, the content providers' revenues from advertising are relatively high, and the ISPs' revenue from end-users are relatively low. Because of this, the ISPs' incentives to invest are suboptimal (too low relative to the socially optimal ones), unless they can extract some of the content providers' advertising revenue by charging the content providers. Thus in the one-sided pricing case, the ISPs under invest, making the rewards for them as well as content providers less than it could have been with two-sided pricing.

It is important to note that when a/θ is larger than 1, the price p charged end-users becomes negative in the two-sided case as can be seen from (1.3). If end-users were actually paid to click, intuition suggests that they would click an unbounded amount and therefore our exponential model of demand (1.1) would not be valid in this region of a/θ. However, one could interpret price p to be net any variable costs ν to end-users - similar to how we define q. With this interpretation, the price p could be negative while the actual prices users see positive, so long as $|p| < \nu$. We therefore show numerical results in our plots for a/θ as large as 1.2.

When a/θ is very small, the content providers' advertising revenue is relatively low, and the ISP's end-user revenue is relatively high. In order to get the content providers to invest adequately, the ISPs need to pay the content providers. That is why for small enough a/θ the price q actually becomes negative, representing a per click payment from the ISPs to the content providers.

Finally, when a/θ is in the intermediate range, in between the two extremes, both content providers and ISPs have adequate incentive to invest. However another effect comes into play – ISP free riding becomes an important factor when N is large. As N increases in the two-sided pricing case there are more ISPs that levy a charge against each content provider. As the price ISPs charge content providers increases, it becomes less attractive for content providers to invest. Thus an ISP choosing the price to charge content providers is balancing the positive effect of earning more revenue per click from content providers versus the negative effect of having fewer clicks because the content providers have reduced their investment. But each ISP sees the entire gain of raising its price, but the loss is borne by all N ISPs. Consequently, the ISPs overcharge the content providers in Nash equilibrium, and the degree of this overcharging increases with N. This is analogous to the tragedy of the commons where people overexploit a public resource. Another perhaps more direct analogy is the "castles on the Rhine effect" where each castle owner is incentivized to increase transit tolls on passing traffic excessively by ignoring the fact that the resulting reduction in traffic harms not only him, but also other castle owners. When all castles do the same, the traffic on the Rhine decreases [15]. The extent of this negative externality, and hence the degree of over-charging, increases with N.

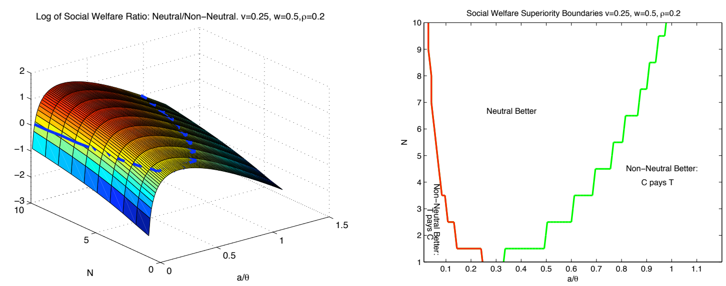

Figure 1.3 Left: The Log of Social Welfare Ratio (one-sided to two-sided). Right: Regions of Social Welfare Superiority (one-sided vs. two-sided).

Figure 1.3 shows a three dimensional plot of the ratio of social welfare for one- vs. two-sided pricing. The plot shows how the ratio changes for different N and a/θ. The second panel of Figure 1.3 is a simplified version of the first panel. It depicts only the boundaries in the parameter space where one-sided pricing is preferable to two-sided and vice versa.

It is also worthwhile to note that the spill-over parameter ρ and the number of content providers M do not appear in the expression for the ratio of content provider revenue between the two regimes nor do they appear in the ratio of revenues for ISPs (1.7). This is in spite of the fact that ρ and M do appear in the expressions for both the one-sided and two-sided pricing equilibria. This suggests that the spill-over effect and number of content providers have little or no effect on the comparative welfare of the two regimes.

1.1.5 Conclusions

Our model shows how each pricing regime affects investment incentives of transit and content providers. In particular, the model reveals how parameters such as advertising rate, end-user price sensitivity, and the number of ISPs influence whether one- or two-sided pricing achieves a higher social welfare. From our results, when the ratio of advertising rates to the constant characterizing price sensitivity is an extreme value, either large or small, two-sided pricing is preferable. If the ratio of advertising rates to the constant characterizing price sensitivity is not extreme, then an effect like the "castles on the Rhine effect" becomes more important. ISPs in a two-sided pricing regime have the potential to over charge content providers, and this effect becomes stronger as the number of ISPs increases.

Of course our model contains many simplifications. Among these simplifications is the assumption that ISPs have local monopolies over their users. We believe that if we had studied a model where each ISP is a duopolist, which better models the degree of choice most end-users have today, our results would have been qualitatively similar. However, there are a number of competing effects such a model might introduce. First, such a scenario would reduce the market power of ISPs over end-users, thus reducing the revenue the ISPs could extract from them. Since two-sided pricing provides ISPs with another source of revenue to justify their investments, this effect would tend to increase the parameter region for which two-sided pricing is social welfare superior. Second, a duopolist competing on product quality invests more than a monopolist, so this would tend to increase efficiency of one-sided pricing. Third, if the model were changed from having N to $2N$ ISPs, then the free riding or "castles on the Rhine" effect would grow, tending to reduce the welfare of the two-sided pricing case. The net effect of all these individual effects would of course depend on the detailed specifications of such a model.

Our two-sided pricing model implicitly assumes that in-bound traffic to a local ISP could be identified as originating at a particular content provider, in order for the ISP to levy the appropriate charge to the content provider. This assumption would not strictly hold if content providers had some way of reaching end-users of an ISP without paying this ISP for end-user traffic. For instance if there were a second ISP that enjoyed settlement free peering with the first ISP, the content provider could route its traffic through the second ISP and thus avoid the first ISP charge for end-user access. This strategy might be facilitated by the fact that the end-users of both ISPs send traffic to each other, and perhaps the traffic from the content provider could be masked in some way to look like traffic originating from the second ISP's end-users. However, the communication protocols of the Internet require that packets be labeled with the origin (IP) address. It seems unlikely today that a large content provider could have the origin addresses of its traffic falsified in a way that would both prevent ISPs from being able to charge the content provider while still enabling end-users to send traffic in the reverse direction back to the content provider. However, it is certainly possible that technology would be developed to enable such a strategy in the future, especially if there were an economic incentive for developing it.

1.2 Competition

The subject of communication network economics is fundamentally complex largely because there are so many interactions. For instance, the Internet is built and operated by many different providers, most of which are making pricing, capacity investment, and routing decisions in order to further their own commercial interests. Users are also acting selfishly – choosing their ISPs, when and how much to use the network, and sometimes even their traffic's routing, in order to maximize their payoff – the difference between the utility they enjoy from the network minus what they have to pay. Studying how efficiently such a complex market system can operate is fundamentally important, because it is with this understanding that we will be able to evaluate the potential of new network architectures that may change the structure of the interactions between the agents of such a market.

There has been an enormous amount of research on the fundamentals of communications networks economics, yet there are still important gaps in our basic understanding. For instance past work has looked at the effects of selfish routing by users, selfish interdomain routing between ISPs, congestion effects, ISP price competition, and ISP capacity investment decisions (There is a large amount of literature in each of these areas. For example for selfish routing by users see [16, 17, 18]; for selfish interdomain routing see [19]; for price competition see [20], [21], [22], and [23].) There are many properties that are of interest when one studies such models, but of particular importance is the efficiency loss of the system due to the selfish interactions of the agents as compared to a hypothetical, perfect, central agent making all the decisions. One way of quantifying the efficiency loss is the "price of anarchy," concept introduced by Koutsoupias and Papadimitriou [24]. The price of anarchy is simply the ratio of the optimum total utility of the system divided by the worst utility achieved in Nash equilibrium.

A large number of price of anarchy results have been shown for general models that contain various combinations of a few of the above features we listed. Yet, there are few general results (price of anarchy bounds for a class of models) when one considers more than a few of these features – even when the simplest possible models are used for each feature. For instance if one considers a model with selfish routing, elastic demand (users can send less traffic if the network is too expensive or slow), and price competition between ISPs, there are price of anarchy results for limited cases, but there is not yet a complete understanding of the problem. In the remainder of this section we look at this particular class of problems, and describe some recent results that push in the direction of getting a more general understanding.

The model we consider is based on a model first proposed and studied by Acemoglu and Ozdaglar [20] and later extended by Hayrapetyan, Tardos, and Wexler [22]. The model studies the pricing behavior of several providers competing to offer users connectivity between two nodes, and it has the features that i) a provider's link becomes less attractive as it becomes congested; and ii) that user demand is elastic – users will choose not to use any link if the sum of the price and latency of the available links is too high. In the first version of the model studied by Acemoglu and Ozdaglar, the user elasticity is modeled by assuming that all users have a single reservation utility and that if the best available price plus latency exceeds this level, users do not use any service. In this setting, the authors find that the price of anarchy – the worst case ratio of social welfare achieved by a social planner choosing prices to the social welfare arising when providers strategically choose their prices – is $(\sqrt{2}+1)/2$. (Or expressed the other way, as the ratio of welfare in Nash equilibrium to social optimum, the ratio is $2\sqrt{2}-2$.) In a later work, Ozdaglar and Acemoglu [25] extend the model to consider providers in a parallel-serial combination. Traffic chooses between several branches connected in parallel, and then for each branch, the traffic traverses the links of several providers connected serially.

Hayrapetyan et. al. [22] consider a model in which user demand is elastic, which is the form of the demand model we will study here. The topology they consider is just the simple situation of competing links between two nodes. They derived the first loose bounds on the price of anarchy for this model. Later Ozdaglar showed that the bound is actually 1.5, and furthermore that this bound is tight [21]. Ozdaglar's derivation uses techniques of mathematical programming, and is similar to the techniques used in [20]. In [23], Musacchio and Wu provide an independent derivation of the same result using an analogy to an electrical circuit where each branch represents a provider link and the current represents flow.

The work we describe next generalizes the work of [23] by considering a more general topology of providers connected in parallel and serial combinations. A more detailed description of the work is available in [26]. As in the earlier work, we use a circuit analogy to derive bounds on the price of anarchy. Furthermore, our bounds depend on a metric that is a measure of how concentrated the market power is in the network. That metric is the reciprocal of the slope of the latency function of the network branch with the least slope divided by the harmonic mean of the slopes of all the branches. In terms of the circuit analogy, this is simply the ratio between the conductance of the most conductive branch to the conductance of the whole system.

1.2.1 Model

We consider a model in which users need to send traffic from a single source to a single destination, but there are several alternative paths that the users can choose from. The paths consist of parallel-serial combinations of Internet service providers as in the example illustrated by the left panel of Figure 1.4. The traffic incurs delay as it crosses each ISP $\mathbf{i}$ that is linear in the amount of traffic that is passing through that ISP. Thus if the flow through ISP $\mathbf{i}$ is $f_{\mathbf{i}}$ we suppose the delay is $a_{\mathbf{i}} f_{\mathbf{i}}$ where $a_{\mathbf{i}}$ is a constant. Furthermore, there is a fixed latency along each main branch, as illustrated by Figure 1.4. Note that we denote the index $\mathbf{i}$ in bold because it represents a vector of indices that determine the location of that ISP in the network.

Users control over which path(s) their traffic goes. A user chooses a path for her traffic selfishly and she cares both about the delay her traffic incurs as well as the price she pays. To account for both of these preferences, we suppose that users seek to use the path with the minimum price plus delay (just as in the model of [20] and others). We term the metric price plus delay, "disutility," just as in [22]. We make the further assumption that the traffic from each user represents a negligible fraction of the total traffic, thus when a single user switches routes, that one user's switch does not change the delays along significantly. An equilibrium occurs when users can not improve their disutility by changing paths, so therefore in such an equilibrium all used paths must have the same disutility and all unused paths must have a disutility that is larger (not smaller.) This kind of equilibrium in which users select routes under the assumption that their individual choice will have a negligible change on the congestion of the route they choose is called a Wardrop equilibrium [27].

As in the model of [22] we suppose that the demand is elastic so if the disutility is too high, some users do not connect to any network or reduce the amount of traffic they send. If d is the disutility on the used paths, then we suppose that the total amount of traffic that is willing to accept that disutility is $f = U^{-1}(d)$ where $U(\cdot)$ is a decreasing function. Furthermore we suppose that $U(\cdot)$ is concave.

In summary, we suppose that given a set of prices from the ISPs (a price profile), the users selfishly select both routes and the amount of traffic to send and as a result, the network reaches a Wardrop equilibrium and furthermore the amount of traffic sent satisfies $f = U^{-1}(d)$. It turns out that for a given set of prices, there exists a unique Wardrop equilibrium that satisfies these properties. (We refer to $U(\cdot)$ as the "inverse demand" function and disutility curve interchangeably.)

Given this user behavior, we now turn to the ISPs. The ISPs play a game in which each ISP $\mathbf{i}$ chooses their price $p_{\mathbf{i}}$, and the resulting Wardrop equilibrium induced by price profile determines the amount of traffic each ISP carries. In turn, the profit each ISP earns is the product of their price times the traffic they carry. We are interested in comparing the social welfare in a Nash equilibrium of this game and comparing that to the maximum social welfare achievable if a central agent were to choose the price profile. (Of course we are also interested in whether a Nash equilibrium even exists, and if so, if it is unique.) The social welfare of the game includes the sum of the ISP profits, as well as the welfare of users.

We use the standard notion of consumer surplus as our metric of user welfare. If the total rate offered web traffic is f, the consumer surplus is found simply by taking $\int_0^f U(z)dz - fU(f)$. The integral has the following interpretation. From the disutility curve $U(\cdot)$, we see that the first ϵ units of traffic would be willing to pay a price as high as $k - \epsilon$ per unit, the next ϵ units of traffic would be willing to pay a price of $k - 2\epsilon$ per unit, and so on. Thus integrating the disutility curve from 0 to f captures the total amount the traffic is willing to pay, and then subtracting the amount it actually pays $fU(f)$, yields the surplus enjoyed by the traffic (users).

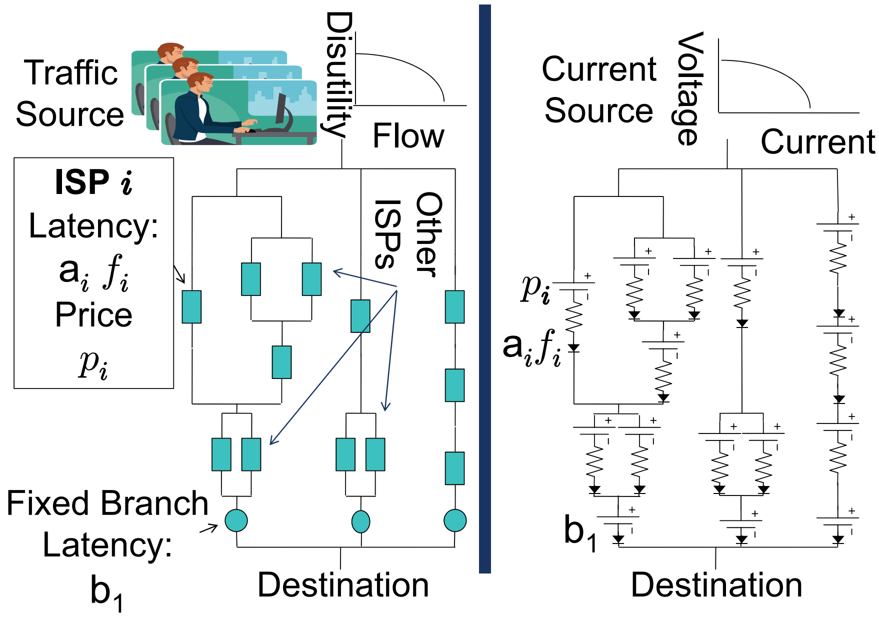

Figure 1.4 An analogous circuit for analyzing the network.

1.2.2 Circuit Analogy

In order to analyze this game we are going to draw an analogy between this game and an electric circuit as suggested by Figure 1.4. It turns out that the relationships between prices and flows in our model are analogous to the Kirchoff voltage and current law relations between the voltages and currents in the circuit pictured in the figure.

To begin analyzing the system, we start by determining the prices that induce the socially optimum flow. The well known theory of Pigovian taxes says that one way to align the objectives of individuals with that of society as a whole is to make individuals pay the cost of their externalities on others [28]. Following this notion the socially optimal pricing should price the flow so that each user bears the marginal cost to society of each additional unit of new flow.

The latency through each ISP is $a_\mathbf{i} f_\mathbf{i}^*$. However, the social cost is the latency times the amount of flow bearing that latency. Therefore the cost is $a_\mathbf{i} f_\mathbf{i}^{*2}$. Thus the marginal cost is $2a_\mathbf{i} f_\mathbf{i}^*$. The latency borne by users is $a_\mathbf{i} f_\mathbf{i}^*$. Therefore to make the disutility borne by users reflect marginal social cost, the price they pay ought to be $a_\mathbf{i} f_\mathbf{i}^*$. With such a price, each user would see the true social cost of their actions, and thus each user's selfish decisions would actually be working toward maximizing social welfare. A formal proof of the optimality of these prices is provided in [26].

Now that we have an expression for the optimal prices for each ISP $\mathbf{i}$ as a function of the flow that passes through that ISP, it is clear that the optimal assignment of flows ought to satisfy the constraint that $p_{\mathbf{i}} = a_{\mathbf{i}} f^*_{\mathbf{i}}$ for each ISP $\mathbf{i}$ in addition to the Wardrop equilibrium and the inverse demand relationship between flow and disutility. If we now turn to the circuit pictured in the right panel of Figure 1.4 and substitute a resistor of size $a_{\mathbf{i}}$ in for each voltage source that represents price $p_{\mathbf{i}}$, we have a circuit whose solution gives us the solution to the problem of finding the optimal assignment of flows (and thus prices) for all the ISPs. We call the circuit with these resistor substitutions the *optimal circuit*.

We now turn to understanding the Nash equilibrium of the game. First it is important to realize that providers connected serially can have very inefficient equilibria. Consider just two providers connected serially, each of which charges a price larger than the highest point on the disutility curve. The flow through these providers will be zero, and furthermore it is a Nash equilibrium for these providers to charge these prices. This is because one player cannot cause the flow to become positive by lowering his price, so each player is playing a best response to the other player's prices. In examples for which there are more than one provider connected serially on each branch, one can construct a Nash equilibrium in which no flow is carried, and therefore the price of anarchy across all Nash equilibria is infinite.

However these infinite price of anarchy Nash equilibria seem unlikely to be played in a real situation. Intuitively, a provider would want to charge a low enough price so that it would be at least possible that he could carry some flow if the other providers on the branch also had a low enough price. With that in mind, we would like to identify a more restricted set of Nash equilibria that seem more "reasonable" and that have a price of anarchy that can be bounded. To that end, we define the notion of a *zero-flow zero-price* equilibrium. A zero-flow zero-price equilibrium is a Nash equilibrium for which players who are carrying zero-flow must charge a zero-price. The intuition motivating this definition is that a player who is not attracting any flow will at least try lowering his price as far as possible in order to attract some flow. It indeed turns out that for this restricted set of equilibria, the price of anarchy is bounded.

Just as we obtained the the social optimum solution, we will seek to express the ratio of price to flow in Nash equilibrium. An ISP that is trying to maximize his profit must consider how much his flow is reduced when he increases his price. This could be very complicated to compute in a complex network, but fortunately we can use our circuit analogy to simplify the problem. A basic result of circuit theory is that a resistive electric circuit viewed from a port, such as the pair of circuit nodes that ISP $\mathbf{i}$ is connected to, can be reduced to an equivalent circuit containing a single resistor and voltage source [29]. Such an equivalent is known as the Thévenin equivalent. Thus we can abstract ISP $\mathbf{i}'s$ "competitive environment" into a single resistor and voltage source. The Thévenin equivalent resistance is found by adding the resistances of resistors connected in series, and taking the reciprocal of the sum of reciprocals of those connected in parallel.

There are a few more details that we need to take into account, as Thévenin equivalents hold only for linear circuits, and we have a few nonlinearities – the diodes in each branch that keep an ISPs flow from being negative, and the inverse demand function. It turns out that the inverse demand function can be linearized at the Nash equilibrium point, and its slope s can be used as a "resistor" in the computation of the Thévenin equivalent. The diodes can be taken into account by only including resistors of branches that are "on" in Nash equilibrium.

With these consideration taken in mind, suppose that the Thévenin equivalent resistance player $\mathbf{i}$ sees is $\delta_{\mathbf{i}}$. If $\mathbf{i}$ unilaterally raises his price by $+\epsilon$ the flow he carries will be reduced by $\epsilon/(\delta_{\mathbf{i}} + a_{\mathbf{i}}) - o(\epsilon^2)$. (It turns out the term $-o(\epsilon^2)$ is negative because of the concavity of the inverse demand function.) Thus we have that

$$(p_{\mathbf{i}} + \epsilon)\left(f_{\mathbf{i}} - \epsilon/(\delta_{\mathbf{i}} + a_{\mathbf{i}}) - o(\epsilon^2)\right) = p_{\mathbf{i}} f_{\mathbf{i}} + [-p_{\mathbf{i}}(\delta_{\mathbf{i}} + a_{\mathbf{i}})^{-1} + f_{\mathbf{i}}]\epsilon - o(\epsilon^2).$$

The above expression is less than the original profit $p_{\mathbf{i}} f_{\mathbf{i}}$ for all nonzero ϵ if and only if

$$\frac{p_{\mathbf{i}}}{f_{\mathbf{i}}} = \delta_{\mathbf{i}} + a_{\mathbf{i}}. \tag{1.8}$$

At this point we have shown a property of the Nash equilibrium flow and prices, but we have not shown that an equilibrium actually exists. It turns out that a Nash equilibrium does exist, and this is proved by construction in [26]. Another technical detail is that in some cases the Thévenin equivalent could be different for a small price increase than it is for a small price decrease. This is because a small price increase might induce flow in some branches that are otherwise off. For these cases, one can show that a best response must satisfy

$$\frac{p_{\mathbf{i}}}{f_{\mathbf{i}}} \in [\delta_{\mathbf{i}}^{+} + a_{\mathbf{i}}, \delta_{\mathbf{i}}^{-} + a_{\mathbf{i}}].$$

where $\delta_{\mathbf{i}}^{+}$ and $\delta_{\mathbf{i}}^{-}$ are the two different Thévenin equivalents.

Form equation (1.8) we see that the price to flow ratio in Nash equilibrium is higher than it is for the social optimum prices. As we did when we constructed the "optimal circuit" we can construct something we call a *Nash circuit* by substituting resistors of size $\delta_{\mathbf{i}} + a_{\mathbf{i}}$ in for the voltage sources representing price $p_{\mathbf{i}}$ for each ISP $\mathbf{i}$. The solution to this circuit gives us the Nash equilibrium flows and prices.

Now that we have the Social optimum and Nash equilibrium configurations of the system described by linear circuits, it is possible to derive closed form expressions that lead to bounds on the price of anarchy. For the class of networks we call "simple parallel serial" which is a network of parallel branches each containing one or more serially connected providers, we can write Kirchoff's voltage laws in matrix form, and then from that derive quadratic forms for the total profit of the providers in both the social optimum and Nash cases. This is done in detail [26].

It turns out that for any problem instance(consisting of a disutility function, a topology, and ISP latencies $\{a_i\}$) with Nash equilibrium flow f and disutility function slope $-s = U'(f)$, we can construct a new problem with a price of

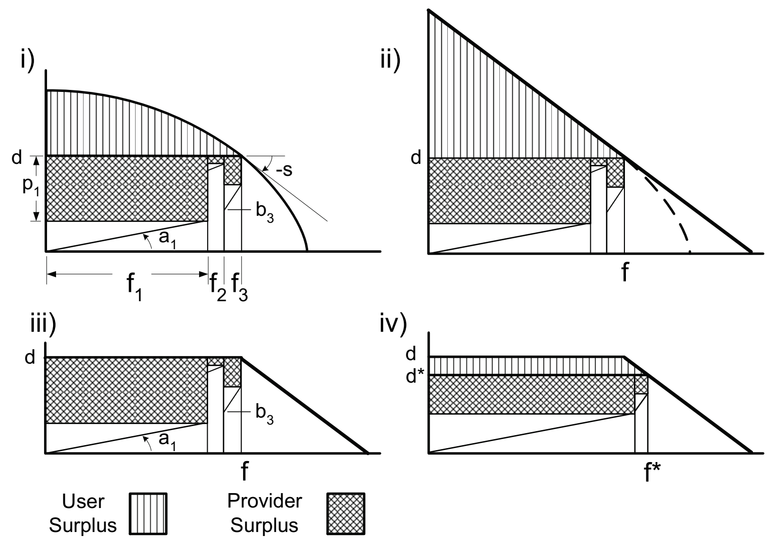

Figure 1.5 i) The Nash equilibrium of game $\mathcal{G}$. ii) The Nash equilibrium of game $\mathcal{G}_l$, where the disutility function of $\mathcal{G}$ has been linearized. iii) The Nash equilibrium of the game $\mathcal{G}_t$, where the disutility function of $\mathcal{G}$ has been linearized and "truncated." iv) The social optimum configuration of the game $\mathcal{G}_t$. Note that the flow $f^* > f$ and that link 2 is not used in the social optimum configuration of this example.

anarchy at least as high in the following way. We modify the disutility function to be flat for flows between 0 and f and then make it affine decreasing with slope $-s$ for higher flows. (This is an argument adapted from [22]). This is basically because the new problem instance has the same Nash equilibrium as the old problem, but now the user welfare is zero. The argument is illustrated by Figure 1.5. Because of this argument, we can restrict our attention to disutility functions of the shape shown in the lower part of the figure. After invoking this argument we can express the user welfare in social optimum for this shape of disutility function using a quadratic form.

It turns out that the algebra works out so that our bounds are found to be a function of a parameter y we term the *conductance ratio*. The conductance ratio is the conductance of the most conductive branch divided by the conductance of network as a whole. The conductance ratio is therefore a measure of how concentrated the capabilities of the network are in a single branch. A conductance ratio near 1 means that most of the conductance of the system is concentrated in a single branch. The smaller the conductance ratio, the more that the overall conductance of the system is distributed across multiple branches. Thus, in a sense the conductance ratio reflects the market power or concentration of the system. As one would expect, the price of anarchy bounds that we find increase as the conductance ratio approaches 1. The following Theorem comes from and is proven in [26].

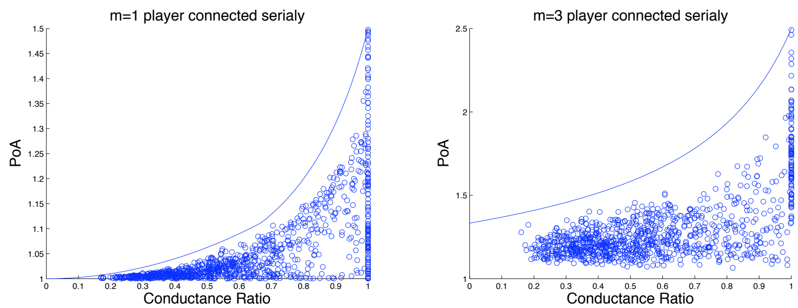

Figure 1.6 Simple parallel serial PoA bound for the cases where there are either $m = 1$ or 3 providers connected serially. The bounds are plotted as a function of the conductance ratio – the ratio of the conductance of the most conductive branch to the conductance of the whole circuit. The points are the PoA of randomly generated example topologies.

Theorem 1.1. *Consider the game with a simple parallel-serial topology. Consider the following ratio*

$$y = \frac{\max_i 1/a_i}{\sum_i 1/a_i},$$

which is the conductance of the most conductive branch divided by the overall conductance. The price of anarchy for zero-flow zero-price Nash equilibria is no more than

$$\begin{cases} \frac{1}{4}\frac{m^2+2m(1+y)+(y-1)^2}{m} & y \leq m/3 \\ \frac{m^2(2-y)+m(4-y^2-y)+2(y-1)^2}{8m-6my} & y \geq m/3 \end{cases} \tag{1.9}$$

where m is the maximum number of providers connected serially. Furthermore, the maximum of the above bound occurs when $y = 1$, and consequently the price of anarchy is mo more than

$$1 + m/2. \tag{1.10}$$

The bounds given in Theorem 1.1 are illustrated in Figure 1.6. Note how the price of anarchy falls as the conductance ratio falls, i.e., becomes less monopolistic. Also note the increase in the price of anarchy as the number of serially connected providers increases. This is an example of the well known “double marginalization” affect in which serially connected providers tend to overcharge (as compared to social optimum) because they do not consider how their price increases will hurt the profits of the other serially connected providers.

The results of Theorem 1.1 are for simple-parallel serial topologies. For general parallel serial topologies (with arbitrary groupings of ISPs connected in parallel and series), it turns out the same bounds hold, with an additional factor of 2. The argument is given in [26]. We do believe that the bound with a factor of 2 is not tight, but it remains an open problem to get a tighter bound.

1.3 Service Differentiation

As we said in the introduction of this chapter, service differentiation is an important tool for improving the Internet, and furthermore as a whole by providing quality to those that need it most. Furthermore, for architectures that separate high priority traffic from low priority traffic, that separation can increase the utility of the system by preventing the mixing of traffic of different types that do not mix well (i.e., video conferencing and peer-to-peer file sharing). In addition to increasing user utility, service differentiation has the potential of increasing provider revenues. Walrand [30] demonstrates that with a simple model, and other works have shown this as well using different models.

Another interesting set of questions comes up when one considers the combination of service differentiation and ISP competition. Musacchio and Wu [31] considers such a combination and shows that architectures that support service differentiation by offering delay sensitive traffic with a priority queue lead to competition games between ISPs that have a lower price of anarchy than if the same ISPs were to use a shared queue, no differentiation architecture. That work supposes that the delay sensitive traffic comes from applications that are from applications that are close to constant bit rate (called "voice" for short), while the delay insensitive applications generate more bursty traffic (called "web" for short).

Musacchio and Wu [31] models the network with a simple queueing model and supposes a capacity constraint for each ISP. Using this approach we derive a space of feasible regions for the vector of web and voice traffic for both the shared and differentiated architecture cases. Basically with a shared architecture, as the fraction of web traffic is increased, the ISP needs to operate at a lower utilization to meet the delay constraint. However if the ISP chooses not to provide good service to voice traffic, that ISP can operate at a much higher utilization. With a priority architecture, the ISP can operate at a high utilization regardless of the web vs. voice mix because the voice traffic is always protected. Using this approach we show that the price of anarchy can be as high as 2 for a shared architecture but only 1.5 for a priority architecture. The model in this study is still quite simplistic, so there are still many important open problems regarding the price of anarchy in differentiated services competition.

Despite the enormous potential advantages of the service differentiation there are a number of challenges. One set of challenges are in developing technical architectures to support service differentiation in a scalable way. A large amount of work has been done in this area, and many protocols and architectures have been proposed which we will not attempt to survey here. Another challenge is a coordination problem – if one ISP adopts a new architecture, the full benefit of that adoption might not be realized until other ISPs also adopt it. Several works have looked at the problem of getting new architectures adopted in the presence of an incumbent architecture (for example [32] and [33]).

Another challenge is that the move from a single service class to multiple service classes actually has the potential to cause some users to be worse off, even if in the aggregate the population of users benefits. This is because users who want mid-range quality and are happy with a single class of service might be forced to choose between a low quality service or a high-priced, high quality service in a differentiated system. Schwartz, Shetty and Walrand study this issue, and they suggest [34] that the number of users who fall into this category could be reduced, while still getting most of the benefits of service differentiation.

Another challenge is that there can be a potential instability in systems that provide service differentiation. Consider a system with a priority queue and a best effort queue. Any "cycles" spent serving the priority queue come at the expense of the best effort queue, so if a lot of traffic uses the priority queue, the best effort traffic will see lower quality. Now consider the following dynamic. Suppose some users move from the best effort queue to the more expensive priority queue in order to get better service. This shift will degrade the quality of the best effort queue, causing more users to switch to priority queue, which again in turn degrades the best effort queue. Depending on the precise utility functions of the users this process might not stop until all the users are using the priority queue. In other examples the mix might tip to everyone using best effort, or perhaps to some positive fraction using both. In summary, the situation is quite "tippy," and so a small change in modeling assumptions can lead to very different predicted outcomes.

This is a phenomenon that has been long recognized by other researchers. This phenomenon is one of the reasons that Odlyzko [35] proposes a Paris metro pricing (PMP) for Internet service classes. Like the Paris metro once did, Odlyzko's PMP scheme would offer a predetermined fraction of the resources to the different classes, but at different prices. The scheme is inherently more stable than a priority scheme, because the quality of each class depends only on the congestion of that class. However, one drawback is that one loses some statistical multiplexing gain. For instance if there is instantaneous demand for the lower class but not the upper class, such a system would not offer its full capacity to the lower class traffic.

Although the "tippiness" problem of schemes for which the congestion of one class affects the quality of the other classes makes these schemes difficult to analyze, the task is not impossible. A recent model of Schwartz, Shetty and Walrand (manuscript in submission as of 2009) includes these effects. Another work by Gibbens, Mason, and Steinberg [36] demonstrates that a different effect driven by competition between ISPs can lead to the market tipping to a single service class. Clearly, there is potential for future work in the area to better understand these phenomena.

1.4 Acknowledgement

The authors gratefully acknowledge the National Science Foundation for supporting this work under grants NeTS-FIND 0626397 and 0627161.

References

[1] E. Whitacre, At SBC, it's all about 'scale and scope'. *Business Week*, Interview by R. O. Crockett, (November 7, 2005).

[2] J. Musacchio, G. Schwartz and J. Walrand, A two-sided market analysis of provider investment incentives with an application to the net neutrality issue. *Review of Network Economics*, **8**:1 (2009), 22-39.

[3] A. Odlyzko, Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets. *Review of Network Economics*, **8** (2009), 40–60.

[4] Chong, The 31 flavors of the net neutrality debate: Beware the trojan horse. *Advanced Communications Law and Policy Institute, Scholarship Series*, New York Law School, (December 2007).

[5] Federal Trade Commission, Broadband connectivity competition policy. Report (June 2007).

[6] Department of Justice, Comments on network neutrality in federal communications commission proceeding. Press Release, (September 7, 2007).

[7] J.-C. Rochet and J. Tirole, Two-sided markets: A progress report. *RAND Journal of Economics*, **37**:3 (2006), 655–667.

[8] M. Armstrong, Competition in two sided markets. *RAND Journal of Economics*, **37**:3 (2006), 668–691.

[9] B. Hermalin and M. Katz, The economics of product-line restrictions with an application to the network neutrality controversy. *Information Economics and Policy*, **19** (2007), 215–48.

[10] C. Hogendorn, Broadband internet: Net neutrality versus open access. *International Economics and Economic Policy*, **4** (2007), 185–208.

[11] P. Weiser, Report from the center for the new west putting network neutrality in perspective. Center for the New West Discussion Paper (January 2007).

[12] A. Dixit and J. Stiglitz, Monopolistic competition and optimum product diversity. *The American Economic Review*, **67**:3 (1977), 297–308.

[13] J. Thijssen, *Investment Under Uncertainty, Coalition Spillovers and Market Evolution in a Game Theoretic Perspective*, (New York: Springer, 2004).

[14] J. Musacchio, G. Schwartz and J. Walrand, *A Two-Sided Market Analysis of Provider Investment Incentives With an Application to the Net-Neutrality Issue*, School of Engineering, Universitiy of California, Santa Cruz, UCSC-SOE-09-08 (2009).

[15] J. A. Kay, Tax policy: A survey. *The Economic Journal*, **100**:399 (1990), 18–75.

[16] T. Roughgarden, Stackelberg scheduling strategies. In *Proceedings of the 33rd Annual Symposium ACM Symposium on Theory of Computing*, (2001), 104–13.

[17] T. Roughgarden, How bad is selfish routing?. *Journal of ACM*, **49**:2 (2002), 236–59.
[18] T. Roughgarden, The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, **67**:2 (2003), 341–64.
[19] J. Feigenbaum, R. Sami and S. Shenker, Mechanism design for policy routing. *Distributed Computing*, **18** (2006), 293–305.
[20] D. Acemoglu and A. Ozdaglar, Competition and efficiency in congested markets. *Mathematics of Operations Research*, **32**:1 (2007), 1–31.
[21] A. Ozdaglar, Price competition with elastic traffic. *Networks*, **52**:3 (2008), 141–55.
[22] A. Hayrapetyan, E. Tardos and T. Wexler, A network pricing game for selfish traffic. *Distributed Computing*, **19**:4 (2007), 255–66.
[23] J. Musacchio and S. Wu, The price of anarchy in a network pricing game. In *Proceedings of the 45th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, (2007), 882–91.
[24] E. Koutsoupias and C. H. Papadimitriou, Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science*, Trier, Germany, (1999), 404–13.
[25] D. Acemoglu and A. Ozdaglar, Competition in parlallel-serial networks. *IEEE Journal on Selected Areas in Communications*, **25**:6 (2007), 1180–92.
[26] J. Musacchio, *The Price of Anarchy in Parallel - Serial Competition with Elastic Demand*, School of Engineering, Universitiy of California, Santa Cruz, UCSC-SOE-09-20 (2009).
[27] J. Wardrop, Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers, Part II*, **1**:36 (1952), 352–62.
[28] A. C. Pigou, *The Economics of Welfare*, (Macmillan, 1920).
[29] L. Chua, C. Desoer and E. Kuh, *Linear and Nonlinear Circuits*, (New York: McGraw-Hill, 1987).
[30] J. Walrand, Economic models of communication networks. In *Performance Modeling and Engineering*, ed. Xia and Liu. (Springer, 2008), pp.57–90.
[31] J. Musacchio and S. Wu, The price of anarchy in differentiated services networks. In *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, (2008), 615–22.
[32] J. Musacchio, J. Walrand and S. Wu, A game theoretic model for network upgrade decisions. In *Proceedings of the 44th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, (2006), 191–200.
[33] Y. Jin, S. Sen, R. Guerin, K. Hosanagar and Z.-L. Zhang, Dynamics of competition between incumbent and emerging network technologies. In *Proceedings of the ACM NetEcon08: The Workshop on the Economics of Networks, Systems, and Computation*, Seattle, WA, (2008), 49–54.
[34] G. Schwartz, N. Shetty and J. Walrand, Network neutrality: Avoiding the extremes. In *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, (2008), 1086–93.
[35] A. Odlyzko, Paris metro pricing for the internet. In *Proceedings of the 1st ACM conference on Electronic Commerce*, Denver, CO, (1999), 140–47.
[36] R. Gibbens, R. Mason and R. Steinberg, Internet service classes under competition. *IEEE Journal on Selected Areas in Communications*, **18**:12 (2000), 2490–98.

Subject index