

Matters are quite different for the *New York Times*. Its loyal customers are high income readers who dislike advertising but are willing to pay more for the paper's content than marginal readers who are less sensitive to advertising.<sup>1</sup> Thus the *Times* fails to internalize loyal readers' distaste for advertising, leading to potentially excessive advertising as a result of *below* optimal pricing to advertisers despite market power. Increases in the costs of distribution that reduce the number of subscribers will tend to *reduce* advertisements as the paper internalizes the costs to its wealthier readers. While intuitive in many markets, this opposite case has been assumed away by previous theoretical and empirical analysis of two-sided markets.

This paper shows that both of these are special cases of a simplified reformulation of the Rochet and Tirole, henceforth RT, (2006) model of monopoly in two-sided markets. The crucial difference between cases is the source of user heterogeneity. While credit card users primarily differ in the *interaction* (or usage) value they take from merchants accepting cards, newspaper readers differ most importantly in their *membership* value from reading the paper's content. This distinction is crucial because participation on one side of the market effectively determines the quality of the platform on the other side. Therefore, like any monopolist who must choose a single quality as well as quantity, the platform internalizes network effects to marginal rather than average participating users (Spence 1975).

The diversity of possibilities in two-sided markets does not eliminate the theory's predictive power. Because the distortions caused by market power (Section II) are linked to observable comparative statics (Section IV) through user heterogeneity, both intuition about the sources of this heterogeneity and empirical data can help calibrate the model in applications. Further restrictions may then be placed on the models (Section V) with a clear sense of how special assumptions increase predictive and prescriptive power. Together these results inform policy analysis in two-sided markets (Section VI), providing a general measure of market power and helping to predict the effects of regulation and mergers.

My analysis rests on a view of the platform's problem as choosing participation rates on the two sides rather than the prices supporting this allocation. This approach (Section I) is justified by an appropriate platform pricing strategy, the *insulating tariff*, that avoids potential coordination failures, thereby allowing the platform to achieve any desired allocation as a unique equilibrium. It applies, along with much of my analysis, more generally (Section III) than the RT (2006) model, allowing a simple approach to analyzing network industries with arbitrary heterogeneous utility, network effects, and any number of sides. I thereby answer perhaps the oldest open question in the theory of network industries (S. J. Liebowitz and Stephen E. Margolis 1994): does a monopolistic platform internalize and therefore neutralize network effects? The answer is yes, but imperfectly given the Spence distortion.

Of course this article is only a first pass at a general analysis of network pricing. Section VII therefore concludes by discussing directions for future research. Longer and less instructive proofs are collected into an Appendix available at <http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.4.1642>.

<sup>1</sup> Note that the intuitive stories I tell about these industries are not intended to be specific prescriptions about policy in these industries, but rather concrete instances of general theoretical possibilities. It is the mapping between the story's assumptions and the conclusions about policy, rather than the validity of the assumptions in a particular industry, that I am interested in here. Thus I do not have any empirical evidence substantiating my stories, evidence that would be highly desirable before reaching definite policy conclusions. For example, Ulrich Kaiser and Minjae Song (2009) argues that users do not actually dislike advertising, as my story assumes. Furthermore, in some cases at least, content may be viewed as an interaction rather than membership benefit if it is tightly tailored to accompany an ad, though I doubt this is the case for newspapers.

## I. Framework

The definition of two-sided markets is controversial (RT 2006; Marc Rysman 2009). For me, the phrase denotes a style of industrial organization modeling<sup>2</sup> developed by, among others, Bernard Caillaud and Bruno Jullien (2001; Caillaud and Jullien (2003; Rochet and Tirole (2003); Simon P. Anderson and Stephen Coate (2005); Mark Armstrong (2006); and Rochet and Tirole (2006). These models tend to emphasize three features.<sup>3</sup>

- i) *Multi-product firm*: A platform provides distinct services to two sides of the market, which can be explicitly charged different prices.
- ii) *Cross network effects*: Users' benefits from participation depend on the extent of user participation on the other side of the market, which varies with market conditions.
- iii) *Bilateral market power*: Platforms are price setters (monopolistic or oligopolistic) on both sides of the market and typically set uniform prices.

The failure of any of these conditions makes simpler and better-understood models more appropriate. If a platform does not explicitly charge different prices to different groups of users, it is best viewed as a standard, one-sided network. When participation does not vary on both sides a vertical monopolies model fits better. An absence of market power allows us to model the firm as a distributor. However, many industries<sup>4</sup> relevant to industrial policy exhibit all of the above characteristics. RT (2006) introduce a "canonical model," in their words, of monopoly capturing these three features in a simple manner that still generalizes the two most influential models in the literature, those of RT (2003) and Armstrong (2006).

### A. The RT (2006) model

Before describing it more formally, I highlight a few key assumptions of the model, on top of the familiar notions of monopoly and constant marginal cost.

- i) User valuations are taken as exogenous to any direct interactions between users on the two sides. Thus the RT (2006) model takes a "macro" approach, in the terminology of Nicholas Economides (1996). While "micro" models that directly incorporate such interactions (Julian Wright 2004; Andrei Hagiu 2006; Graeme Guthrie and Julian Wright 2007) have similar positive forms, their normative implications can be quite different.
- ii) Network effects are all *across*, not *within* the two sides. This rules out, for example, negative within-side effects from competition among software creators or positive collaboration effects among operating system users.

<sup>2</sup> These can either be viewed as models aimed at capturing important features of some true class of "two-sided markets" or they can be viewed as a style of modeling that captures some elements of "two-sidedness" that are more or less important in different industries. I lean towards the second view.

<sup>3</sup> I am grateful to Bruno Jullien and Patrick Rey for helping me refine these criteria.

<sup>4</sup> For example, credit cards, newspapers, operating systems, Internet service providers and others discussed by RT (2003) and Armstrong.

- iii) Users on each side interact with either all or a random subset of users on the other side, price discrimination<sup>5</sup> within a particular side based on the number of such interactions is impossible, and user values are affine in the number of users on the other side. This does not rule out some users interacting with a larger random and unpriced sample of users on the other side; this will just magnify their interaction valuations.
- iv) Finally, it assumes that users on each side of the market are of equal value to those on the other side. This rules out, for example, high advertising-value readers of a newspaper and video games that are especially valued by gamers (Robin S. Lee 2009).

Few of these assumptions are necessary for my analysis; in fact, assumptions ii–iv can be substantially relaxed or eliminated entirely. However, doing so complicates the exposition. Furthermore, given its greater parsimony, I suspect the RT (2006) model will continue to be the most attractive framework in many applications, including those I focus on: the payments and newspaper industries. Most importantly, its assumptions fit many industries quite well.

Consider the case of the newspaper industry as an example. The ways in which advertisers gain from readers viewing their ads, or why exactly users dislike ads (Gary S. Becker and Kevin M. Murphy 1993), seems fairly exogenous to industrial policy in the newspaper industry. While advertisements sometimes compete within a paper for user attention, it seems fairly reasonable to assume that advertisers are close to indifferent as to the number of other advertisements included in a paper, and readers are indifferent to the number of other readers of the paper. Readers usually read all or a fairly random selection of advertisements in a paper, and certainly it seems difficult to charge users (or advertisers) differentially based on the number of advertisements viewed. Finally, some advertisements are certainly more annoying than others and some readers more valuable than others to advertisers. However, I follow many top past applied papers (Stephen T. Berry and Joel Waldfogel 1999; Matthew Gentzkow and Jesse Shapiro 2010; Yin Fan 2010) on industries with advertising in viewing this as of second-order importance.

Therefore I develop most of my analysis in the context of the RT (2006) application, treating the general case only in Section III. There I show that my basic message in the RT (2006) applies generally. Therefore little is lost by focusing on the RT (2006) model, and Section III will likely be of most interest to theoretically inclined readers.

### B. User Preferences and Heterogeneity

There is a continuum of potential users on each side  $\mathcal{I} = \mathcal{A}, \mathcal{B}$  of the market, with mass normalized to 1. Thus the number of users participating on each side represents the fraction of potential participants choosing to do so. All quantities are scaled accordingly as discussed below.  $\mathcal{I}$  refers to a generic side of the market and  $\mathcal{A}$  and  $\mathcal{B}$  to refer to specific sides in examples.

A typical user  $i$  on side  $\mathcal{I}$  has an inherent *membership* benefit or cost  $B_i^{\mathcal{I}}$  from participating in the service if no users participate on the other side. For example, developers must pay fixed costs even if no users own the operating system the software runs on. Given my normalization of a unit mass of users,  $B_i^{\mathcal{I}}$  must be measured in terms of the total value all users on side  $\mathcal{I}$  would derive if they participated given that they have the same preferences as user  $i$ . Suppose a town has

<sup>5</sup> As in all models with market power, the impossibility of price discrimination plays a crucial role in normative conclusions. I believe price discrimination is probably neither systematically easier nor more difficult in two-sided markets than in standard markets. Even when some discrimination is possible, I believe the discrimination-free model gives some insight, as long as the discrimination is imperfect.

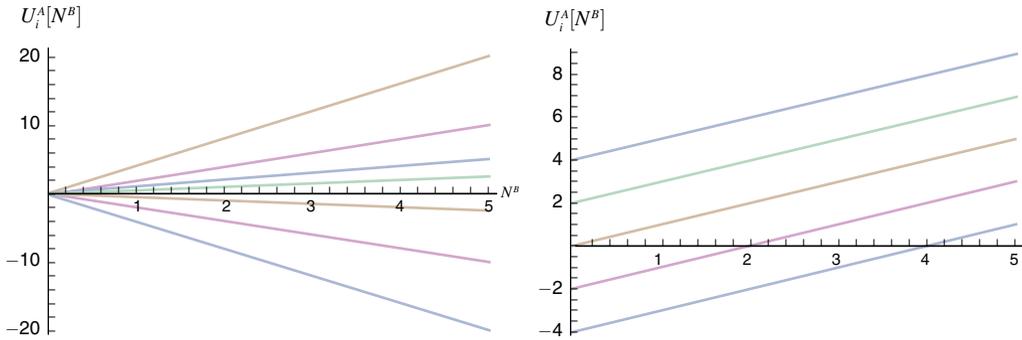


FIGURE 1.  $U_i^A(N^B)$  FOR VARIOUS RT (2003) (LEFT) AND ARMSTRONG (RIGHT) PREFERENCES.

Note: This Illustrates the Two Dimensions of Heterogeneity: Interaction and Membership Values, Respectively

100 possible newspaper subscribers and user  $i$  values reading her paper at \$500 a year; then her membership benefit would be  $B_i^T = \$50,000$ .

Each user also derives an *interaction*<sup>6</sup> benefit or cost of participation  $b_i^T$  for every user that participates on the other side. Again this must be appropriately scaled.<sup>7</sup> One of 1,000 credit card holders who makes 100 purchases every year deriving a 50 cent surplus from credit over cash would have a interaction benefit of \$50,000 per year, as this would be the value to all users on her side if all cards were accepted and all users on her side had the same preferences as she does. If there were 1,000 possible advertisers and 100 readers in a town and the disutility to a user  $i$  of each advertiser being included in a local newspaper were ten cents, then the interaction cost of that user would be  $b_i^T = \$10,000$ . I follow most of the industrial organization literature in assuming that user utility is quasi-linear in money.

Formally the utility derived by user  $i$  on side  $\mathcal{I}$  from participating is

$$U_i^T = B_i^T + b_i^T N^J - P^T(N^J)$$

where  $N^J$  is the number of users participating on side  $\mathcal{J} = -\mathcal{I}$ , the other side than  $\mathcal{I}$ .  $P^T(N^J)$  is the tariff set by the platform (independence of  $i$  disallows price discrimination), prescribing how much users must pay (or will be paid) to participate conditional on a given size of the platform on side  $\mathcal{J}$ . Users on each side can therefore be heterogeneous along two dimensions: interaction and membership values. Two natural special cases involve only one dimension of heterogeneity. RT (2003) assume that  $B_i^T \equiv 0$  and that users have heterogeneous interaction values. Armstrong (2006) assumes homogeneous interaction values ( $b_i^T \equiv b^T$ ) and allows heterogeneous membership values. Figure 1 shows the difference between these specifications. Utility is graphed as a function of participation on the other side of the market for various RT (2003) preferences (left) and Armstrong preferences (right). When, in general, there are both dimensions of heterogeneity, even fixing  $N^J$  and  $P^T$ , many different types of users may be just on the margin between participating and not (have  $U_i^T = 0$ ): some may have high interaction benefits but large membership costs; others may have low interaction benefits and no membership costs. This is pictured in Figure 2, where all users lying along the lines are marginal. The implications of these different

<sup>6</sup> RT (2006) refers to this as the user's usage valuation; I eschew this terminology to avoid confusion, as users have no choice over how intensively to use the service in the RT (2006) model.

<sup>7</sup> Of course these scales can be renormalized as suits a given application, so long as this is done consistently.

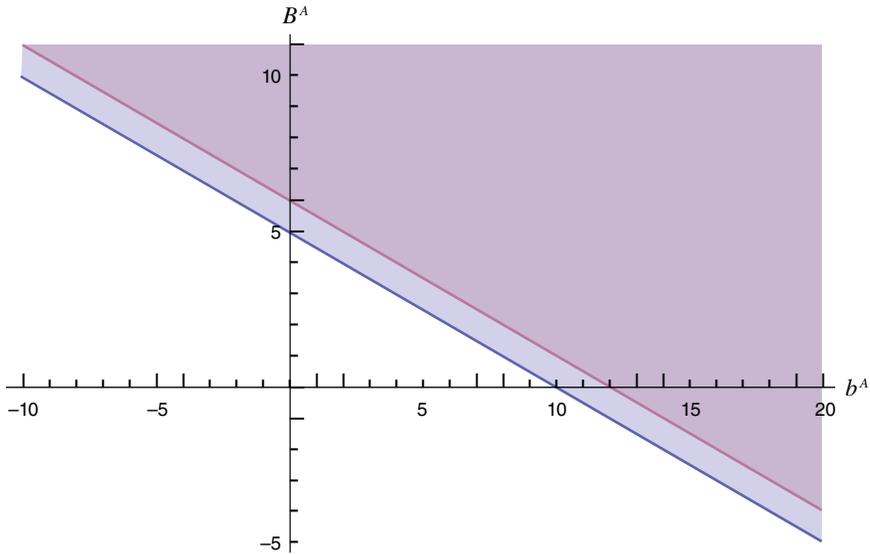


FIGURE 2. THE SET OF USERS PARTICIPATING ON SIDE  $\mathcal{A}$  WHEN HALF OF USERS PARTICIPATE ON SIDE  $\mathcal{B}$  AND  $p^{\mathcal{A}} = 5$  AND  $6$  RESPECTIVELY

sources of user heterogeneity are the primary focus of this paper. Formally I assume that the user parameters are distributed according to some massless, twice continuously differentiable<sup>8</sup> bi-variate distribution with probability density function  $f^{\mathcal{I}}(B_i^{\mathcal{I}}, b_i^{\mathcal{I}})$  and a full support.

C. Coordination and Insulating Tariffs

Once the tariff is set, users on the two sides of the market play a game. A user  $i$  on side  $\mathcal{A}$  will choose to participate if and only if

$$B_i^{\mathcal{A}} + b_i^{\mathcal{A}}N^{\mathcal{B}} > P^{\mathcal{A}}(N^{\mathcal{B}}).$$

However, this typically depends on the decisions of user on side  $\mathcal{B}$ , which  $\mathcal{A}$  users take as given. There may be multiple equilibria for some tariffs and distributions of user tastes. For example, suppose  $b_i^{\mathcal{I}} \equiv 1$ ,  $B_i^{\mathcal{I}} \equiv 0$  and  $P^{\mathcal{I}} \equiv 1/2$  for  $\mathcal{I} = \mathcal{A}, \mathcal{B}$ .<sup>9</sup> Then it is clearly an equilibrium for either all or none of the users to participate. In the former case the utility of participation on either side (taking the other as given) is  $1/2$ ; in the latter it is  $-1/2$ . This is the classic “chicken and egg” problem in two-sided markets (Caillaud and Jullien 2003).

Yet, in a sense, equilibrium multiplicity is inessential to the analysis of two-sided markets. A given pair of participation rates  $\widetilde{N}^{\mathcal{A}}$  and  $\widetilde{N}^{\mathcal{B}}$  leads to a unique profit and social welfare. To see this, note that, given a fixed side  $\mathcal{B}$  participation rate  $\widetilde{N}^{\mathcal{B}}$ , there is a well-defined demand function determining the number of users who participate on side  $\mathcal{A}$  as a function of  $P^{\mathcal{A}}$ , the equilibrium price to side  $\mathcal{A}$ . Visually, this is depicted in Figure 2, where the set of side  $\mathcal{A}$  users participating

<sup>8</sup>Note that the assumption of twice-continuous differentiability actually rules out all of the one-dimensional cases. As the online Appendix shows, the only assumption needed for the analysis is smoothness with regard to movements of the boundary of a set; that is sufficient conditions for the Leibnitz Integral Rule to apply.

<sup>9</sup>Note this example does not fit my full support and massless assumption, but an analogous example that does can be constructed by perturbing it.

when  $N^B = 0.5$  and  $P^A = 5$  or 6 is shaded. Clearly participation on side  $\mathcal{A}$ , holding fixed side  $\mathcal{B}$  participation, declines in  $P^A$ . Formally for either  $\mathcal{I}$

$$(1) \quad N^{\mathcal{I}}(P^{\mathcal{I}}, \widetilde{N}^{\mathcal{J}}) \equiv \int_{-\infty}^{\infty} \int_{P^{\mathcal{I}} - b^{\mathcal{I}} \widetilde{N}^{\mathcal{J}}}^{\infty} f^{\mathcal{I}}(B^{\mathcal{I}}, b^{\mathcal{I}}) dB^{\mathcal{I}} db^{\mathcal{I}}.$$

Clearly<sup>10</sup>  $N^{\mathcal{I}} \equiv \partial N^{\mathcal{I}} / \partial P^{\mathcal{I}} < 0$ . Therefore inverting it with respect to its first argument yields a well-defined function  $P^{\mathcal{I}}(N^{\mathcal{I}}, N^{\mathcal{J}})$ . Thus there is a unique pair of prices, and therefore profit and welfare,  $P^A(\widetilde{N}^A, \widetilde{N}^B)$  and  $P^B(\widetilde{N}^B, \widetilde{N}^A)$  consistent with  $\widetilde{N}^A$  users participating on side  $\mathcal{A}$ ,  $\widetilde{N}^B$  users participating on side  $\mathcal{B}$ , and uniform pricing.

Thus multiplicity plays no role if one thinks of the platform as simply choosing an allocation<sup>11</sup> to maximize some objective function. The only concern is that the platform may struggle to consistently implement its desired allocation; there could be a “failure to launch” as a result of a “critical mass problem,” in the terminology of David S. Evans and Richard Schmalensee (2009).

This can be avoided, however, by a conscientious platform. The platform can lower (raise) its price<sup>12</sup> on side  $\mathcal{I}$  when hoped-for (undesirable) side  $\mathcal{J}$  users that are valuable (harmful) to marginal users on side  $\mathcal{I}$  fail to show up. This insulates the platform’s allocation on side  $\mathcal{I}$  from the influence of side  $\mathcal{J}$  participation. At the logical extreme the platform can ensure that  $\widetilde{N}^{\mathcal{I}}$  users participate on side  $\mathcal{I}$  regardless of side  $\mathcal{J}$  decisions if it charges the *insulating tariff*<sup>13</sup>  $P^{\mathcal{I}}(N^{\mathcal{J}}) \equiv P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, N^{\mathcal{J}})$ . Then the unique equilibrium is the platform’s target allocation.<sup>14</sup>

Insulating tariffs are intuitive in many applications. With homogeneous interactions values (Armstrong), the insulating tariff is an insurance scheme, as originally proposed by Phillip H. Dybvig and Chester S. Spatt (1983) for a one-sided market. The platform charges users a price  $b^{\mathcal{I}}$  per user on side  $\mathcal{J}$  and charges an “hedonic” price (in Armstrong’s language) which determines participation. Therefore side  $\mathcal{I}$  users are indifferent to  $N^{\mathcal{J}}$ . With no membership values (RT 2003), the insulating tariff is a pure interaction price  $p^{\mathcal{I}}$  so that any side  $\mathcal{I}$  user earns utility  $(b_i^{\mathcal{I}} - p^{\mathcal{I}})N^{\mathcal{J}}$  from participating. Thus participating side  $\mathcal{I}$  users prefer high side  $\mathcal{J}$  participation and thus are not insured but, because the sign of their utility is independent of side  $\mathcal{J}$  participation, still choose to participate independent of the decisions of side  $\mathcal{J}$  users. In general, the composition, but not level, of participants may shift with participation on the other side: a rise in  $N^{\mathcal{J}}$  selects  $\mathcal{I}$  users with high interaction values.

<sup>10</sup> This follows from my assumption of full support.

<sup>11</sup> This approach, which is the key method used throughout the paper to simplify the complexities of pricing in two-sided markets, was first suggested to me in the context of the RT (2003) model by Jeremy Bulow, to whom I am tremendously grateful. Because of the single dimensionality of user heterogeneity in that model, the allocation approach is not much simpler than the price approach there. This led me, much to my later regret, to ignore Jeremy’s advice until after having wasted months trying to implement the price approach. On this, as many other matters, I have come round to seeing the elegance of his perspective. My approach was also inspired by the fulfilled expectations equilibrium of Michael L. Katz and Carl Shapiro (1985) and more broadly by the contract theory literature, starting with Roger B. Myerson (1981). It is surprising that, given the long history of the allocation approach in related literatures, it has not to my knowledge been applied previously to the general analysis of network industries.

<sup>12</sup> A further assumption of the RT (2006) model, not discussed extensively above, is that a (uniform) price can be set at any desired level on both sides of the market costlessly. This assumption fails in broadcast media, as argued by Anderson and Coate (2005). They explore, in a Hybrid model (see Section V) with RT (2003) preferences for advertisers and Armstrong preferences for consumers, the consequences of this price rigidity. A more general analysis of restrictions on pricing is, as discussed in Subsection VIA, an important direction for future research.

<sup>13</sup> I am grateful to Bruno Jullien for helping to guide me towards this name.

<sup>14</sup> Note that the platform can charge an insulating tariff on just one side  $\mathcal{I}$  of the market and achieve the same guarantee, as this assures that any equilibrium must have  $\widetilde{N}^{\mathcal{I}}$  users, removing expectations from the decision making of  $\mathcal{J}$  users. This is what makes possible Anderson and Coate’s (2005) analysis: they assume the platform chooses quantity, rather than price, to advertisers (effectively assuming an insulating tariff to one side). See footnote 27 for a more general discussion.

Schemes resembling insulating tariffs are used explicitly in many industries: Web site ad rates are typically per click and credit card fees/incentives per transaction. In fact, in broadcasting, as Anderson and Coate (2005) argue, the structure of programming often allows platforms to commit to a quantity of advertising directly. However, even when such explicit schemes are not used, the static RT (2006) model can reasonably be thought of as a reduced form for a dynamic model, in the spirit of Evans and Schmalensee (2009). In this case insulating tariffs simply require that the platform provides subsidies at early stages of product development which it recoups once its desired allocation is achieved. This pattern is commonly observed in video games, operating systems and Web sites.

However, there may be some circumstances under which firms would refrain or be constrained from employing them; see Section VII for further discussion. In these cases the critical mass problem binds and the coordination problems considered by Glenn Ellison and Drew Fudenberg (2003); Atilla Ambrus and Rossella Argenziano (2009); and Evans and Schmalensee (2009) become important.<sup>15</sup> However, in most mature industries, the focus of the RT (2006) model, price flexibility is sufficient to avoid these problems.<sup>16</sup>

Many other tariffs achieve the platform’s desired allocation, even uniquely. In fact, as argued by RT (2003), RT (2006), and Armstrong, any tariff with  $P^I(\widetilde{N}^I) = P^I(\widetilde{N}^I, \widetilde{N}^J)$  for both  $I$  has the pair  $(\widetilde{N}^A, \widetilde{N}^B)$  as an equilibrium. Thus none of my analysis, except a brief discussion of competition in Subsection VIC, assumes any particular tariff. Rather, this subsection is meant to justify my approach of ignoring the specifics of tariffs and coordination and to show, perhaps surprisingly, that adding optimization *simplifies* the analysis. Thus even a reader skeptical of the possibility of insulating tariffs but willing to focus, exogenously, on a given equilibrium, should accept my analysis in the monopoly case I focus on.

## II. Pricing

Industrial policy typically aims to alleviate the social harms caused by market power. The first step towards formulating such policy is therefore understanding the nature of those harms. Towards that goal, this section develops and compares the socially optimal and profit maximizing allocation rules, emphasizing the prices that support these allocations.

### A. Pigouvian Pricing

The value created by the platform is the benefits it brings to users less the costs of providing the service. RT (2006) assumes marginal costs constant in both participation rates, taking the other participation rate as given. Thus there may be two types of cost: membership costs  $C^I N^I$  and interaction costs  $cN^A N^B$ . The benefits the platform brings to users on side  $I$  are

$$(2) \quad V^I(N^I, N^J) = \int_{-\infty}^{\infty} \int_{P^I(N^I, N^J) - b^I N^J}^{\infty} [B^I + b^I N^J] f^I(B^I, b^I) dB^I db^I.$$

<sup>15</sup> However, I suspect that explicitly modeling why critical mass problems exist in a given application, ignored in previous work on this problem, would be crucial to understanding their welfare implications. For example, if imperfect information is the cause, platforms and social welfare might actually be harmed by attempts to “solve” the deliberately designed critical mass problem!

<sup>16</sup> An identical argument clearly applies to one-sided networks, or any coordination game. I thus believe that the importance of coordination has been exaggerated in situations when an optimizing agent with the ability to make transfers can regulate coordination. Even in the cases when it is relevant, I believe it is more a choice than a constraint. However, this is obviously a controversial view. See, for example, Joseph Farrell and Paul D. Klemperer (2007) for a well-argued contrasting view and the last paragraph of the paper for further discussion.

Thus the total social value of the platform is

$$(3) \quad V(N^A, N^B) = V^A(N^A, N^B) + V^B(N^B, N^A) - C^A N^A - C^B N^B - c N^A N^B.$$

A benevolent social planner equates marginal social benefits to their marginal social costs:

$$V_1^I + V_2^J = C^I + cN^J$$

where  $V_i^I$  is the derivative of  $V^I$  with respect to the  $i$ th argument.  $V_1^I = P^I$  as the user added on side  $I$  must be marginal and therefore earn zero net surplus from participating.  $V_2^J$  is the value an additional user on side  $I$  brings to users on side  $J$ :  $b^J N^J$ , where

$$\overline{b^J} = \frac{\int_{-\infty}^{\infty} \int_{P^J(N^J, N^I) - b^J N^I}^{\infty} b^J f^I(B^J, b^J) dB^J db^J}{\int_{-\infty}^{\infty} \int_{P^J(N^J, N^I) - b^J N^I}^{\infty} f^I(B^J, b^J) dB^J db^J}$$

is the *average interaction value of participating users* on side  $J$ . Thus the optimal price is

$$(4) \quad P^I = \underbrace{C^I + cN^J}_{\text{marginal private cost}} - \underbrace{\overline{b^J} N^J}_{\text{marginal external benefit}}$$

This is the standard Pigouvian<sup>17</sup> condition: the price of an activity should equal its private cost less any external benefits. This last term is the essential difference between optimal pricing in two-sided markets and standard multiproduct pricing: because network effects are external to individual decisions, price should diverge from cost. Thus positive network effects should be subsidized and negative ones taxed.

Newspapers offer a simple example. Optimal pricing calls for readers to be subsidized, below the cost of providing the news by the value they bring to advertisers, and for advertisers to be taxed, above the cost of printing their ads, by the amount readers dislike them.

In the Armstrong model, interaction values are homogeneous ( $b_i^I \equiv b^I$ ) and interaction costs are disallowed ( $c = 0$ ) so (4) becomes Alex Gaudeul and Bruno Jullien’s (2008) formula

$$P^I = C^I - b^J N^J.$$

RT (2003) rule out membership values/costs so user prices and surplus are all from interactions. Letting  $p^I \equiv P^I/N^I$  be the *per-interaction price* and  $\overline{s^I} \equiv (V^I/N^J) - p^I$  the average per-interaction surplus on side  $I$  gives Bedre-Defolie and Calvano’s (2010) and Weyl’s (2009b) optimal pricing rule

$$p^A + p^B - c = -\overline{s^A} = -\overline{s^B}.$$

I now compare this classical rule to that which a profit-maximizing monopolist would adopt.

<sup>17</sup> First-best pricing has traditionally been known in the literature as Lindahl pricing (Özlem Bedre-Defolie and Emilio Calvano 2010; Weyl 2009b). However, because price discrimination is ruled out in the RT (2006) model, pricing follows Pigou (1920) rather than Lindahl (1919).

B. Profit-Maximizing Pricing

Often the operators of platforms are concerned with their profits rather than with social welfare. Because price discrimination is typically imperfect, these differ. To make as clear as possible the distortions introduced by imperfect price discrimination it is useful to take them to their logical extreme, ruling out all discrimination.<sup>18</sup> Profits are then

$$(5) \quad \pi(N^A, N^B) = (P^A[N^A, N^B] - C^A)N^A + (P^B[N^B, N^A] - C^B)N^B - cN^AN^B.$$

A profit-maximizing monopolist equates marginal revenues of participation to marginal cost:

$$\underbrace{P^I + P_1^I N^I + P_2^J N^J}_{\text{marginal revenue}} = \underbrace{C^I + cN^J}_{\text{marginal cost}}$$

The first two terms of marginal revenue are classical: price minus the inverse hazard rate of demand (or *market power*)  $\mu^I \equiv -P_1^I N^I = P^I/\epsilon^I$  where  $\epsilon^I$  is the elasticity of demand. The final term is special to two-sided markets: it is the revenue that can be extracted from side  $J$  by adding an additional user on side  $I$ . Letting  $\tilde{b}^J$  be the *average interaction value of marginal users* (AIVMU) on side  $J$ , by the implicit function theorem and equation (1)

$$P_2^J = -\frac{N_2^J}{N_1^J} = \frac{\int_{-\infty}^{\infty} b^J f^J(P^J[N^J, N^I] - b^J N^I, b^J) db^J}{\int_{-\infty}^{\infty} f^J(P^J[N^J, N^I] - b^J N^I, b^J) db^J} \equiv \tilde{b}^J.$$

The platform can extract only the value *marginal users* on side  $J$  place on an additional side  $I$  user joining. This is an example of the general tendency, emphasized by Spence (1975) and discussed extensively below, of monopolists to serve the preferences of marginal, rather than all participating, users. The platform’s side  $J$  marginal revenue from a side  $I$  user is therefore  $\tilde{b}^J N^J$ . Privately optimal pricing follows a simple extension of Lerner’s formula<sup>19</sup>

$$(6) \quad \frac{P^I - (C^I + cN^J - \tilde{b}^J N^J)}{P^I} = \frac{1}{\epsilon^I}.$$

In the Armstrong case this immediately simplifies to Armstrong’s pricing condition

$$P^I = C^I - b^J N^J + \mu^I.$$

In RT (2003), only interaction benefits exist so  $\tilde{b}^J = p^J$ . Therefore the pricing condition is

$$p^A + p^B - c = m^A = m^B$$

where  $m^I \equiv \mu^I/N^I$ . This is the formula that RT (2003) derives.

<sup>18</sup> For an analysis of distortions that can arise even under perfect price discrimination and with a single group of homogeneous users, when there are externalities to nonparticipating consumers or other frictions, see Ilya Segal (1999).

<sup>19</sup> RT (2006) states the general condition for optimal two-product pricing in terms of derivatives of  $N^A$  and  $N^B$ , determined as fixed points of an equilibrium among users. However, as a function of the allocation  $(N^A, N^B)$ , profits are just the simple explicit function above. This is what allows me to express the first-order condition for optimal in terms of the primitive properties of preferences in two-sided markets.

Comparing private to socially optimal pricing,

$$(7) \quad P^{\mathcal{I}} = \underbrace{C^{\mathcal{I}} + cN^{\mathcal{J}} - \bar{b}^{\mathcal{J}}N^{\mathcal{J}}}_{\text{socially optimal price}} + \underbrace{\mu^{\mathcal{I}}}_{\text{classical market power distortion}} + \underbrace{(\bar{b}^{\mathcal{J}} - \widetilde{b}^{\mathcal{J}})N^{\mathcal{J}}}_{\text{Spence distortion}}$$

Thus there are two distortions in a two-sided market. First, classical marginal revenue lies below price by the amount of the market power  $\mu^{\mathcal{I}}$ . Second, if  $\bar{b}^{\mathcal{J}} \neq \widetilde{b}^{\mathcal{J}}$ , the average interaction values of marginal users differ from those of loyal users; the platform will either over- or undersubsidize (tax) users on side  $\mathcal{I}$ . Like the classical market power distortion, this *Spence distortion* is a consequence of the platform's inability to price discriminate. The platform internalizes network externalities but does so imperfectly (see Subsection III C).

The tendency to truckle to marginal users is familiar to anyone living in, or observant while visiting,<sup>20</sup> a tourist destination: the city government and businesses tend to cater to mobile tourists rather than to locked-in residents.<sup>21</sup> This Spence distortion is likely more important in two-sided markets than the contexts for which it was originally conceived. A platform is unlikely to partially ameliorate inefficiency (while introducing other distortions) by offering multiple products (Michael Mussa and Sherwin Rosen 1978; Mary O'Keeffe, W. Kip Viscusi, and Richard J. Zeckhauser 1984; David Besanko, Shabtai Donnenfeld, and Lawrence J. White 1987) as this would require inefficiently wasting potential interactions.<sup>22</sup> Once "quality" is provided to some users on one side of the market, it is free to provide to others.

The existence and sign of the Spence distortion depend crucially on the source of user heterogeneity.<sup>23</sup>  $\bar{b}^{\mathcal{I}}$  will tend to exceed  $\widetilde{b}^{\mathcal{I}}$  if users differ primarily in their interaction values. For example, in the extreme case of only interaction heterogeneity (RT 2003), the Spence distortion is exactly the per-interaction surplus on side  $\mathcal{I}$ , while when there is only membership heterogeneity (Armstrong), there is no Spence distortion. The Spence distortion may even be downward, as in the newspaper example above. If heterogeneity in willingness to pay for content dominates and is correlated with willingness to pay to avoid advertising, then loyal users dislike advertising more than marginals, and the Spence distortion is downwards.

<sup>20</sup> Of course in real life, as in the RT (2006) model, marginal users (tourists) are a heterogeneous bunch, and many, including the author, have preferences more similar to natives than to those of other tourists.

<sup>21</sup> Readers living in less-frequented climes may find a joke instructive. I tell a variant of a classic Israeli joke, first told to me by David Hariton, to whom I am grateful. In the original joke, Smith is replaced by David Ben-Gurion.

Adam Smith dies and, for his service to economics, is given a choice of where to spend eternity. He requests to see each option before deciding. First he is shown Hell, which, full of decadent food, French wines and beautiful women, seems a merry way to spend the rest of time. Heaven, on the other hand, is an unending stream of presentations of leading research in economics and philosophy. Having spent his life in contemplation on these topics, Smith decides he has earned a bit of relaxation in the afterlife and opts for Hell. Immediately he is thrown onto the rack, whipped, water-boarded, and subjected to other "enhanced" methods of entertainment. Astonished, he says, "I was just here a few minutes ago and things were so much nicer. What happened?" Lucifer replies, "Then you were a tourist."

<sup>22</sup> If the incentive for price discrimination is sufficiently large the platform might "throw away" quality. While such strategies are common in standard markets, in two-sided markets they seem to occur only when justified by other concerns outside this model, such as optimal matching (e.g., targeted ads). However, this is an important question for future research.

<sup>23</sup> Another, perhaps more general way to put this follows the language of Spence more closely. Spence argued quality would be undersupplied ( $P^{\mathcal{J}}$  distorted upward) when  $P_{12}^{\mathcal{I}} < 0$  and oversupplied ( $P^{\mathcal{J}}$  distorted downward) when  $P_{12}^{\mathcal{I}} > 0$ . If, as in the RT (2006) model, each user can participate at most once, the former is equivalent to users with high utility (lower reservation values) having high sensitivity to quality and users with low utility (high reservation values) being less so; the latter conversely. Note that  $P_{12}^{\mathcal{I}} = -\mu_b^{\mathcal{I}}/N^{\mathcal{I}}$ , the measure of local interaction heterogeneity I develop in Section IV. Thus there is a one-to-one correspondence between Spence's cross-partial of the price function and my focus on user heterogeneity.

Thus the harms of market power depend crucially on the source of heterogeneity. If, as is typically assumed, the costs of price distortions are convex, then market power over card accepters is particularly pernicious as it compounds the Spence distortion from cardholders. However, it may actually be beneficial that the *Times* has market power over advertisers, as this offsets the Spence distortion potentially leading to a better second-best level of advertising. Even with market power, therefore, it is possible ad rates may be too low.

C. Ramsey Pricing

Achieving first-best prices may be infeasible in practice as it would require subsidies whose granting, given the cost of raising public funds, political economy constraints, and imperfect information, would be more costly than the monopoly distortions they seek to address. When granting subsidies is infeasible, second-best pricing requires maximizing social welfare subject to some constraint, such as allowing the firm a rate of return (possibly 0) on its variable or fixed costs. Because of the externalities in two-sided markets, this Ramsey solution must be extended as proposed by Tae Hoon Oum and Michael M. Tretheway (1988) to take these into account.

I consider three formulations of the Ramsey problem, all of which are equivalent if the required level of profit is 0. First, in the text, I consider the classic Ramsey problem: social welfare is maximized subject to achieving a minimum absolute profit. In the Appendix, I consider a modified version of the Ramsey problem that RT (2003)<sup>24</sup> uses in a two-sided market where the rate of return is required on variable costs. As I argued in Weyl (2009b), there are two possible social objectives: maximizing user or social surplus subject to the rate-of-return constraint. The first approach addresses externalities more completely, while the second comes closer to the monopolist’s constrained goals.

**THEOREM 1:** *Interior Ramsey prices maximizing user or social surplus subject to the constraint that the platform makes a profit of at least K must solve*

$$(8) \quad \frac{\overbrace{P^I - (C^I + cN^J - \widetilde{b}^J N^J)}^{\text{private marginal opportunity cost}} - \overbrace{[1 - \lambda]}^{\text{Lagrangian weighting}} \overbrace{[\overline{b}^J - \widetilde{b}^J]}^{\text{Spence distortion}}}{P^I} = \lambda \frac{1}{\epsilon^I}$$

where

$$\lambda \equiv \frac{\overbrace{K}^{\text{target profit}} + \overbrace{(b^A + b^B - c)N^A N^B}^{\text{subsidy required for (local) Pigouvian prices}}}{\underbrace{N^A \mu^A + N^B \mu^B + (\overline{b}^A + \overline{b}^A - \widetilde{b}^A - \widetilde{b}^B)N^A N^B}_{\text{(local) profit gain moving to monopoly from Pigouvian prices}}}$$

<sup>24</sup> Rochet and Tirole use this modified Ramsey set-up to consider whether firms distort the “balance” of prices as separate from their level, a major focus of mine in Weyl (2009c).

## PROOF:

See the Ramsey Pricing portion of the Appendix.

Thus the Ramsey pricing condition is just a simple weighted average of the Pigouvian and profit-maximizing prices. These, again, diverge in their attention to both the Spence and classical market power distortions. Prices are closer to profit maximization i) the higher is the target profit, ii) the larger is the subsidy called for by Pigouvian prices and iii) the further one must move towards monopoly to achieve a given gain in net profits. Just as first-best prices take a classic Pigouvian form, Ramsey prices take OT's Pigou-Ramsey form.

### III. Generalization

The primary aim of this paper is to understand the price theory of and proper policy towards industries such as payment cards and newspapers. After a brief interlude in this section, I continue toward this goal in Section IV, to which a casual reader may wish to skip directly. However, the general character of my basic ideas thus far suggests they may help analyze a broader class of models than that RT (2006) specifically adapted to those industries. In fact with any number of groups of users and essentially arbitrary heterogeneous preferences, the same principles developed above apply. Insulating tariffs exist, allowing a simple analysis of the platform's choice of allocation showing in general that the Spence distortion is the key element added by network externalities. This section considers such a generalization.

I maintain four important assumptions of the RT (2006) model:

- i) (Quasi-linear) user preferences are taken as exogenous (RT 2006 assumption 1).
- ii) All groups of users can be explicitly (third-degree) price discriminated and all users within each group differ only in their preferences.<sup>25</sup>
- iii) No price discrimination is possible, but prices to any given group can take any positive or negative value. Users interact with an exogenous collection of other users (in their own and other groups); any marginal price for such interactions is exogenous to the model and enters only to the extent that it determines preferences.
- iv) Externalities are only to participating users.<sup>26</sup>

#### A. The Model

There are  $M$  groups  $\mathcal{I} = \mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$  and users may value participation by members not only of other groups, but of their own. A typical user  $i$  on side  $\mathcal{I}$  is characterized by a vector  $\theta_i^{\mathcal{I}}$  of characteristics drawn according to a smooth and massless distribution with probability density function

<sup>25</sup> Andre Veiga and Weyl (2010) have made significant progress in relating this assumption.

<sup>26</sup> Unlike the others, this assumption can easily be dispensed with. This generalizes Segal's (1999) basic model of contracting with externalities to allow asymmetric information (on reservation values) and asymmetric agents ("sides of the market"). However, the increase in notational complexity and distance from a realistic model of network industries (most nonparticipant externalities arise in contracting, rather than uniform pricing, settings) led me to this assumption. The intuition of that model should be clear from Segal and the general model here: the social planner internalizes all externalities, while a profit-maximizer internalizes the *reverse sign* of externalities to *marginal* nonparticipating consumers, scaled by the number of participating users. Effectively, to a profit maximizer, negative out-group externalities are equivalent to positive in-group externalities, while to a social planner they are opposite. Details are available on request. This extension and the more general connections between the theory of multi-sided platforms (network industries) and contracting with externalities are promising areas for future research.

$f^{\mathcal{I}}$  with full support on  $R^{K^{\mathcal{I}}}$  where  $K^{\mathcal{I}} \in N$ . Let  $\mathbf{N} \equiv (N^A, N^B, N^C, \dots)$  be an *allocation*, a vector of participation rates on each side.<sup>27</sup> The utility of user  $i$  on side  $\mathcal{I}$  from participating is

$$U_i^{\mathcal{I}} = u^{\mathcal{I}}(\mathbf{N}; \theta_i^{\mathcal{I}}) - P^{\mathcal{I}}$$

where  $P^{\mathcal{I}}$  is the price a user on side  $\mathcal{I}$  must pay to participate. I assume that  $u^{\mathcal{I}}$  is smooth in and finite for all values of the allocation and characteristics.

Note that this general model has a few special cases of particular interest:

- i)  $M = 1$  is a one-sided network with arbitrary utility and heterogeneity. I do not believe this model has ever been considered, but given the substantial interest in one-sided network monopolists (Economides 1996) it seems a natural general model.
- ii)  $M = 2$  is RT (2006) with arbitrary heterogeneous utilities and within-side network effects.
- iii) Suppose  $M$  sides can be split into two groups  $\mathbf{A}$  and  $\mathbf{B}$  such that  $u^{\mathcal{I}}$  is independent of  $N^{\mathcal{J}}$  whenever either  $\mathcal{I}, \mathcal{J} \in \mathbf{A}$  or  $\mathcal{I}, \mathcal{J} \in \mathbf{B}$ . This is (ii) without within-side effects but with groups of discriminable, heterogeneously valuable users on each side.

For a particular allocation  $\mathbf{N}$  and price  $P^{\mathcal{I}}$  the set of  $\mathcal{I}$  users weakly benefiting from participating is  $\overline{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}) \equiv \{\theta^{\mathcal{I}} : u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) \geq P^{\mathcal{I}}\}$  and the set of all marginal users is  $\widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}) \equiv \{\theta^{\mathcal{I}} : u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) = P^{\mathcal{I}}\}$ . Then the fraction of users interested in participating on side  $\mathcal{I}$  given an allocation  $\mathbf{N}$  and a price  $P^{\mathcal{I}}$  is

$$\widetilde{N}^{\mathcal{I}}(P^{\mathcal{I}}, \mathbf{N}) = \int_{\overline{\Theta}^{\mathcal{I}}(P^{\mathcal{I}}, \mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}.$$

Because the set  $\overline{\Theta}^{\mathcal{I}}$  is clearly contracting in  $P^{\mathcal{I}}$ ,  $\widetilde{N}^{\mathcal{I}} < 0$  and<sup>28</sup>  $\widetilde{N}^{\mathcal{I}}$  can be inverted to yield  $P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, \mathbf{N})$ , the price needed to attract  $\widetilde{N}^{\mathcal{I}}$  of users who anticipate allocation  $\mathbf{N}$ .

Note that the potential multiplicity problem here is far worse than in the two-sided case, as utility functions have arbitrary structure, and there can be an arbitrary number of sides. However, this enormous coordination problem can again be avoided by careful tariffs. In particular the platform may charge an insulating tariff, which is here a price to side  $\mathcal{I}$  depending on the full realized allocation that ensures the desired allocation is realized. Formally the insulating tariff for a desired participation rate  $\widetilde{N}^{\mathcal{I}}$  is  $P^{\mathcal{I}}(\mathbf{N}) \equiv P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, \mathbf{N})$ . As with RT (2006), if the platform charges the insulating tariff associated with its desired allocation on all sides, then the unique equilibrium is its desired allocation.<sup>29</sup> Thus once again the platform’s problem can be viewed as

<sup>27</sup> In particular, I assume every utility level is achieved by same type, given any  $\mathbf{N}$ .

<sup>28</sup> When participation is positive, but not total, from my assumption of smooth  $f^{\mathcal{I}}$ s and full support.

<sup>29</sup> Here, again, insulating *every side from every side* can be dispensed with. Imagine drawing a graph where each node represents a side of the market and a directed edge is drawn between each side and those sides whose participation affects their utility, but against whose participation they are not insulated. I would conjecture, but have only the sketch of a proof that, so long as this graph is acyclic there is a unique equilibrium. Intuitively if the graph is acyclic, one can trace back from its sinks to tie down the unique participation rate on each side. Furthermore other tariffs than the insulating tariff may do the trick for particular (distributions of) user preferences. However, I believe that the “simplest” approach to “robustly” ensuring uniqueness is fully insulating every side of the market from every other side. A formal analysis of all this will likely appear in joint work in progress with Alex White, as also referred to in footnote 53.

one of choosing an allocation  $\tilde{\mathbf{N}}$  to maximize some objective, eliminating the need to consider derivatives of complex, multi-sided fixed points.

B. Pricing

Let  $P^{\mathcal{I}}(\mathbf{N}) = P^{\mathcal{I}}(N^{\mathcal{I}}, \mathbf{N})$ , where  $N^{\mathcal{I}}$  is the  $\mathcal{I}$ th entry of  $\mathbf{N}$ ,  $\overline{\Theta}^{\mathcal{I}}(\mathbf{N}) \equiv \overline{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}[\mathbf{N}])$  and  $\widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}) \equiv \widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}[\mathbf{N}])$ . The gross value created on side  $\mathcal{I}$  by an allocation  $\mathbf{N}$  is simply

$$V^{\mathcal{I}}(\mathbf{N}) = \int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}.$$

I allow for arbitrary smooth, positive cost functions  $C(\mathbf{N})$ . Thus the (net) surplus created by the service as a function of the allocation is

$$V(\mathbf{N}) = \sum_{\mathcal{I}} V^{\mathcal{I}}(\mathbf{N}) - C(\mathbf{N}).$$

Maximizing the surplus created by the service requires equating marginal social value to marginal cost. Let  $X_{\mathcal{J}} \equiv \partial X / \partial N^{\mathcal{J}}$ . A socially optimal allocation then requires that for each  $\mathcal{I}$

$$\sum_{\mathcal{J}} V_{\mathcal{I}}^{\mathcal{J}} = C_{\mathcal{I}}.$$

The following theorem states that these conditions can again be written in a Pigouvian form.

**THEOREM 2:** *The first-order conditions for a socially optimal allocation are*

$$(9) \quad P^{\mathcal{I}} = \underbrace{C_{\mathcal{I}}}_{\text{marginal cost}} - \underbrace{\sum_{\mathcal{J}} \overline{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}}}_{\text{marginal externalities}}$$

where the average marginal interaction value of participating side  $\mathcal{I}$  users for side  $\mathcal{J}$  users is

$$\overline{u}_{\mathcal{J}}^{\mathcal{I}}(\mathbf{N}) \equiv \frac{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u_{\mathcal{J}}^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}.$$

**PROOF:**

See the Generalization portion of the Appendix.

Thus the Pigouvian formula (4) extends in the most natural way possible: interaction values are replaced by the *marginal* value of users who have potentially nonlinear utility and all externalities, to those within side  $\mathcal{I}$  and on other sides  $\mathcal{J} \neq \mathcal{I}$ , are included.

On the other hand revenues on side  $\mathcal{I}$  are  $R^{\mathcal{I}}(\mathbf{N}) = P^{\mathcal{I}}(\mathbf{N})N^{\mathcal{I}}$  and profits

$$\pi(\mathbf{N}) = \sum_{\mathcal{I}} R^{\mathcal{I}}(\mathbf{N}) - C(\mathbf{N}).$$

Profit maximization requires equating marginal revenue of an additional side  $\mathcal{I}$  user, from all sides of the market, to the marginal cost of serving that user:

$$\sum_{\mathcal{J}} R_{\mathcal{I}}^{\mathcal{J}} = C_{\mathcal{I}}.$$

This yields a similarly intuitive extension of the RT (2006) profit maximizing pricing.

**THEOREM 3:** *The first-order conditions for a profit maximizing allocation are*

$$(10) \quad \frac{P^{\mathcal{I}} - (C_{\mathcal{I}} - \sum_{\mathcal{J}} \widetilde{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}})}{P^{\mathcal{I}}} = \frac{1}{\epsilon^{\mathcal{I}}}$$

or equivalently

$$(11) \quad P^{\mathcal{I}} = \underbrace{C_{\mathcal{I}} - \sum_{\mathcal{J}} \overline{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}}}_{\text{socially optimal price}} + \underbrace{\mu^{\mathcal{I}}}_{\text{classical market power distortion}} + \underbrace{\sum_{\mathcal{J}} (\overline{u}_{\mathcal{I}}^{\mathcal{J}} - \widetilde{u}_{\mathcal{I}}^{\mathcal{J}}) N^{\mathcal{J}}}_{\text{Spence distortion(s)}}$$

where the average marginal interaction value of marginal side  $\mathcal{I}$  users for side  $\mathcal{J}$  users is

$$\widetilde{u}_{\mathcal{I}}^{\mathcal{J}}(\mathbf{N}) \equiv \frac{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u_{\mathcal{I}}^{\mathcal{J}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}.$$

**PROOF:**

See the Appendix.

Thus again profit maximization distorts the allocation in two ways. First it raises prices (lowers participation) as classical marginal revenue falls below price. Second it imperfectly internalizes network externalities, as preferences of marginal rather than all participating users determine the marginal revenues generated by an additional side  $\mathcal{I}$  user. Thus there are now  $M$  classical distortions and  $M^2$  Spence distortions.

### C. Discussion

Conceptually little has changed from the RT (2006) model. Insulating tariffs exist and the platform can therefore achieve any desired allocation. The platform thus maximizes its objectives over possible allocations, making its problem simple. Profit maximization leads to classical and Spence distortions. The nature of these latter distortions depends on how the preferences of loyal and marginal users diverge, that is, on the source of user heterogeneity.

This suggests three interesting conclusions. First, while most of this paper focuses on affine user preferences, this is without significant loss of generality. While affine preferences allow only two dimensions of heterogeneity, these are two dimensions which generally matter. This extends even to my comparative statics analysis below, as none of the terms governing these include the curvature of utility (none involve  $V_{\mathcal{J}\mathcal{J}}^{\mathcal{I}}$ ). Of course the irrelevance of further dimensions of heterogeneity depends crucially on the impossibility of price discrimination. If user utility is not affine, platforms may use a marginal price, such as differential charges for viewing certain Web sites, to discriminate among users. In this case social value and profits depend not only on

participation rates, but also on marginal prices. This important and largely open<sup>30</sup> problem is well beyond the scope of this paper.

Second, it provides a simple and general strategy for analyzing monopoly networks: the allocation approach. While my results here constitute only the most superficial of first passes, having no comparative static or policy analysis, they suggest a path for future research.

Finally, it answers perhaps the oldest open question in network economics: the general validity of the (Liebowitz and Margolis 1994) conjecture that optimizing networks will internalize (and thereby neutralize) network externalities. Liebowitz and Margolis were partially correct, but only up to a point. While platforms do internalize externalities, they do so imperfectly as they take into account only the preferences of marginal users. This Spence distortion will be smallest, and therefore network externalities most nearly neutralized, when loyal and marginal users place a similar value on the participation of a marginal network user. In this case, the only distortions caused by market power are the classic, familiar ones of any multiproduct monopolist. In this case actions of users influence the welfare of other users only indirectly, through platform optimization (see Subsection IVB). On other hand, when loyal and marginal users have sharply different preferences, network monopolies have much more complex distortions with large direct network externalities persisting in equilibrium.

#### IV. Comparative Statics

A primary motivation for the theory of two-sided markets is that conditions on each side affect participation and welfare on the other. As with any comparative statics exercise, understanding these indirect cross-effects relies on the second-order conditions for optimization (Paul A. Samuelson 1941) and specifically, because of the multiproduct (Weyl 2009a) monopoly (Weyl and Michal Fabinger 2009) context, on pass-through rates and the cross partial of the allocation in profits. I begin by formally developing these closely related concepts.

The *pass-through rate* on side  $\mathcal{I}$ ,

$$\rho^{\mathcal{I}} \equiv \left. \frac{dP^{\mathcal{I}}}{dC^{\mathcal{I}}} \right|_{N^{\mathcal{J}}} = - \frac{\mu^{\mathcal{I}}}{N^{\mathcal{I}} \frac{\partial^2 \pi}{\partial N^{\mathcal{I}2}}}$$

is the amount a private platform finds it optimal to increase  $P^{\mathcal{I}}$  in response to an increase in  $C^{\mathcal{I}}$  if  $N^{\mathcal{J}}$  is held fixed. The cross partial

$$\chi \equiv \frac{\partial^2 \pi}{\partial N^{\mathcal{A}} \partial N^{\mathcal{B}}}$$

measures the complementarity/substitutability (if positive/negative) of participation rates.

For traditional comparative static analysis, it is important that the first-order conditions used actually represent the optimal allocation for the platform. To ensure this, a convenient assumption is that the platform's profit function is concave. However, it is important to avoid overly restrictive assumptions that are sufficient, but unnecessary, for the purpose as these can bias analysis; log-concavity is a typical such assumption. To add tractability without undue restrictions, I

<sup>30</sup> See Bedre-Defolie and Calvano (2008) for a first pass, in the context of the RT (2003) model.

propose a “weak” (in a sense formalized by Theorem 4) second-order condition. As far as I know this *two-sided contraction* (2SC) is the first second-order condition to be proposed for the general RT (2006) model.

If for all  $(N^A, N^B) \in (0, 1)^2, \rho^A, \rho^B > (\geq) 0$  and  $(\mu^A \mu^B / N^A N^B) > (\geq) \rho^A \rho^B \chi^2$ , I will say that  $f$  satisfies strict (weak) *two-sided contraction* (2SC) given interaction cost  $c$ .

**THEOREM 4:** *If  $f^A, f^B$ , and  $c$  exhibit strict 2SC then for any  $C^A$  and  $C^B$  a solution to equation (6) for both  $I$  is the unique platform’s optimal price. If they violate weak 2SC then there exists a pair  $(C^A, C^B)$  for which there is a solution to equation (6) which is not an optimum.*

**PROOF:**

See Appendix.

In the RT (2003) case,<sup>31</sup>  $\chi = \mu^I / N^J$  for both  $I$  so the condition becomes my (Weyl 2009c) “cross-subsidy contraction” condition  $\rho^A \rho^B < 1$ . The Comparative Statics portion of the Appendix gives Pigouvian SOCs. These could be extended to the general model of Section III by deriving conditions for the Hessian matrix of cross partials of profits with respect to the allocation to be negative definite.

### A. Complements versus Substitutes

The most famous, supposedly robust result on the comparative statics of two-sided markets is what RT (2006) calls the “simple ‘seesaw principle’: a factor that is conducive to a high price on one side, to the extent that it raises the platform’s margin on that side, tends also to call for a low price on the other side as attracting members on that other side becomes more profitable.” While intuitive, this result faces two challenges. First, the appropriate notion of “price” is unclear. In the only (RT 2003) model where the seesaw principle has been demonstrated formally (RT 2003; Weyl 2009c), the price is per interaction. In other models this price has little special significance. However, as discussed above, holding fixed the number of users on side  $\mathcal{I}$  the price (in any sense) on side  $\mathcal{J}$  is decreasing in the number of users on side  $\mathcal{J}$ . Therefore RT (2006)’s seesaw principle can be reformulated as follows: factors leading the platform to choose higher  $N^I$  lead it to choose lower  $N^J$ . That is  $\partial^2 \pi / \partial N^A \partial N^B < 0$  or participation on the two sides are substitutes *for the platform*.<sup>32</sup> In RT (2003), this holds, and the two formulations are equivalent. However, this formulation can be examined beyond the context of the RT (2003) model.

The more serious challenge to the seesaw principle is that this broader formulation is not generally true but instead depends on the source of user heterogeneity.

To see this formally, it is useful to construct a general measure of the local importance of the two dimensions of heterogeneity. A natural such measure is how interaction and membership benefits of marginal users increase with price. Price is, by definition, always equal to the total value of marginal users. It is therefore natural to decompose increases in price into changes in interaction and membership values. From Subsection IIB  $P_1^I = -\mu^I / N^I$ ; but the total gross utility of a marginal user is  $\widetilde{B}_1^I + \widetilde{b}_1^I N^J$  so

$$\widetilde{B}_1^I + \widetilde{b}_1^I N^J = -\frac{\mu^I}{N^I} .$$

<sup>31</sup> See below. Also see the online Appendix for the Armstrong special case.

<sup>32</sup> Note that the “demand system” does not necessarily exhibit either complements or substitutes: Slutsky symmetry is not obeyed ( $\widetilde{b}^I = P_2^I \neq P_2^J = \widetilde{b}^J$ ) and may even be violated in signs, despite quasi-linearity, because of the externalities between the sides.

We can therefore define natural measures of local heterogeneity along the two dimensions as the projection of market power onto each dimension.  $\mu_{\tilde{b}^{\mathcal{I}}} \equiv -\tilde{b}_1^{\mathcal{I}} N^{\mathcal{I}}$  is the *membership market power* and  $\mu_{\tilde{b}}^{\mathcal{I}} \equiv -\tilde{b}_1^{\mathcal{I}} N^{\mathcal{A}} N^{\mathcal{B}}$  is the *interaction market power*. The cross partial

$$(12) \quad \chi = \tilde{b}^{\mathcal{A}} + \tilde{b}^{\mathcal{B}} - c - \frac{\mu_{\tilde{b}}^{\mathcal{A}}}{N^{\mathcal{B}}} - \frac{\mu_{\tilde{b}}^{\mathcal{B}}}{N^{\mathcal{A}}}$$

as the effect of side  $\mathcal{J}$  participation on  $\mathcal{I}$  marginal revenue is the difference between its effect on price  $P^{\mathcal{I}}$ ,  $\tilde{b}^{\mathcal{I}}$ , and market power  $\mu^{\mathcal{I}}$ , as shown formally in the proof of Theorem 5 below. Intuitively, interaction benefits favor complementarity: the value of a side  $\mathcal{A}$  user is proportional to the number of users she interacts with on side  $\mathcal{B}$ . Thus an increase in side  $\mathcal{B}$  users makes it more attractive to recruit side  $\mathcal{A}$  users. Offsetting this is the fact that when interaction benefits are the main dimension of heterogeneity, increasing the participation on side  $\mathcal{B}$  requires recruiting low interaction benefit users. Thus increased side  $\mathcal{B}$  participation reduces the AIVMU, eroding the cross-subsidy to, and therefore participation by, side  $\mathcal{A}$ .

Thus the sign of the cross partial is determined by how the surplus created by marginal interaction benefits compares to their heterogeneity. Perhaps the sharpest way to express this is in terms of the relative importance of interaction benefits in profits compared to their relative importance in heterogeneity. Platform profits are just the sum of (twice) marginal interaction surplus  $\tilde{b} N^{\mathcal{A}} N^{\mathcal{B}} \equiv (\tilde{b}^{\mathcal{A}} + \tilde{b}^{\mathcal{B}} - c) N^{\mathcal{A}} N^{\mathcal{B}}$  and marginal membership surplus,  $\sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} (\tilde{b}^{\mathcal{I}} - C^{\mathcal{I}}) N^{\mathcal{I}}$ . It is therefore natural to consider the fraction of profits arising from marginal interaction surplus, the *interaction surplus ratio*  $\alpha \equiv \tilde{b} N^{\mathcal{A}} N^{\mathcal{B}} / \pi$ . Similarly the *interaction heterogeneity ratio*  $\beta \equiv \mu_{\tilde{b}} / \mu_{\tilde{b}^{\mathcal{I}}}$ , where  $\mu_{\tilde{b}} = \sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} N^{\mathcal{I}} \mu_{\tilde{b}}^{\mathcal{I}}$  and  $\mu_{\tilde{b}^{\mathcal{I}}} = \sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} N^{\mathcal{I}} \mu_{\tilde{b}^{\mathcal{I}}}^{\mathcal{I}}$ , measures the relative *aggregate* importance of interaction heterogeneity.

**THEOREM 5:** *Participation on the two sides of the market are complements if  $\mu_{\tilde{b}} > 0$  and  $\alpha > \beta$ , substitutes if either  $\mu_{\tilde{b}} \leq 0$  or  $\beta > \alpha$ , and independent if  $\mu_{\tilde{b}} > 0$  and  $\alpha = \beta$ .*

**PROOF:**

See the Scale-Income Model portion of the Appendix.

Thus user heterogeneity ties the Spence distortion to the cross partial of participation rates. Because observing the cross partial requires only marginal shocks to market conditions, it may be easier to study empirically than the Spence distortion directly. Therefore one might measure basic features of user heterogeneity by the sign of the cross-participation effect, for example by observing the effect of a shock to one membership cost. Sadly, this is a coarse instrument, unable to distinguish which side of the market generates the interaction heterogeneity nor anything beyond its magnitude relative to the interaction surplus ratio. Measuring these finer properties requires richer data<sup>33</sup> or stronger assumptions.

The theorem makes clear the source of the seesaw effect in the RT (2003) model: there is no membership heterogeneity, so  $\beta = \infty$ , implying substitutes. By contrast, in Armstrong’s model  $\beta = 0$  as there is no interaction heterogeneity, and  $\alpha > 0$  as otherwise the firm would end the

<sup>33</sup> A companion paper under preparation, Weyl (2009a), treats identification in multiproduct monopoly, with a focus on two-sided markets. I show that first-order instruments for participation rates reveal elasticities and the AIVMU, while quantitatively observable cost shocks reveal pass-through rates and the cross partial. Some tests of general multiproduct monopoly are also possible, but many of the finer normative features, and tests of the RT (2006) model specifically, require stronger assumptions or higher-order variation.

two-sidedness, separately providing services to the two sides. Thus the Armstrong model always has complements, showing that the seesaw principle is far from general.

### B. Welfare Effects

In Section I, I argue that cross-group externalities *in the absence of transfers* are a defining feature of two-sided markets. However, others take the view (Hagiu 2007; Hagiu 2009; Rysman 2009) that two-sided markets are ones where, holding fixed some notion of price, each group's welfare depends on the other's participation and thereby *indirectly* (Jeffrey Church and Neil Gandal 1992; Michael Katz and Carl Shapiro 1994) on its own.

Such views are difficult to parse in multi-sided network models because the welfare-irrelevant details of pricing in these models still lead to very different indirect network effects holding fixed prices. For example, if insulating tariffs are charged to both sides then, by construction, such indirect network effects *never* exist. Thus, unless these authors think the canonical models miss the "essential nature" of two-sided markets, which I view as largely being defined by these models, it is difficult to see how such a test can be applied.<sup>34</sup>

Perhaps a more reasonable line of inquiry is therefore the nature of network effects in equilibrium. Suppose that participation on one side of a market rises for a reason, such as membership costs falling or membership values uniformly shifting up, that has no direct effect on the platform's incentives on the other side. I call the effect of such an exogenous increase in side  $\mathcal{A}$  participation on side  $\mathcal{B}$  welfare the equilibrium network effect.

**THEOREM 6:** *The equilibrium network effect from side  $\mathcal{J}$  to side  $\mathcal{I}$  has the same sign as*

$$(13) \quad \overline{b^{\mathcal{I}}} - \widetilde{b^{\mathcal{I}}} + \rho^{\mathcal{I}}\chi.$$

**PROOF:**

See the Scale-Income Model portion of the Appendix.

The first term of expression (13) is the direct effect of  $\mathcal{J}$  participation on  $\mathcal{I}$  welfare: the Spence distortion from  $\mathcal{I}$  to  $\mathcal{J}$ . Only the distortion matters:  $\widetilde{b^{\mathcal{I}}}$  is internalized by the platform as a higher price to side  $\mathcal{I}$ . One might view this direct effect as the equilibrium network *externality*. The second term is an indirect effect through the platform's optimization: the pass-through of the cross-side pricing effect. For example, if  $\chi < 0$  (participation rates are substitutes as in RT 2003), side  $\mathcal{I}$  will tend to benefit from an increase in prices on side  $\mathcal{J}$  as this will incentivize the firm to obtain greater participation in side  $\mathcal{I}$  by reducing prices.

Interaction heterogeneity both enlarges the Spence distortion and makes  $\chi$  negative, while membership heterogeneity eliminates it or even reverses its sign but makes  $\chi$  positive. Thus the source of heterogeneity has an ambiguous effect on expression (externalities). However, the first effect is fundamentally inframarginal, while the second depends only on local properties.

For example, in Armstrong's model, which has no Spence distortion, complementarity implies positive equilibrium network effects.<sup>35</sup> In the RT (2003) model, as discussed in Section II, the Spence distortion from side  $\mathcal{I}$  is  $\overline{s^{\mathcal{I}}}$ , the user surplus on side  $\mathcal{I}$ , and the cross partial can be shown

<sup>34</sup> Liebowitz and Margolis (1994) discuss the dangers of abusing the concept of network externalities.

<sup>35</sup> On the other hand, if interaction surplus is negative and participation rates are substitutes (I do not know of any simple example of this), equilibrium network effects are negative. These conditions do not have any consistent relationship to the *primitive* externalities, the level of interaction values on the two sides.

to be the negative of per-interaction market power on either side of the market  $m^{\mathcal{I}}$ . Furthermore in Weyl and Fabinger (2009) we show that  $\bar{s} = \bar{\rho}m$  where  $\bar{\rho}$  is an average of pass-through rates over prices above the equilibrium level, as pass-through measures the log-curvature of demand. Expression (13) therefore becomes, in the RT (2003) case,

$$-m^{\mathcal{I}}(\bar{\rho}^{\mathcal{I}} - \rho^{\mathcal{I}}),$$

whose sign is determined by the slope of  $\rho^{\mathcal{I}}$  with respect to cost/price: increasing pass-through implies average pass-through exceeds local pass-through, decreasing pass-through the reverse. Thus the *third* derivative of log-demand determines equilibrium network effects.

It may seem immediate that an increase in costs on side  $\mathcal{I}$  harms side  $\mathcal{I}$  users, but in Weyl (2009a) I showed that in the RT (2003) model the average user on one side of the market may actually want her prices increased to encourage a reduction in prices to users on the other side. The following corollary provides general conditions for this counterintuitive result.

**COROLLARY 7:**  $dV^{\mathcal{I}}/dC^{\mathcal{I}}$  has the sign of

$$(14) \quad -\left(\frac{\mu^A \mu^B}{N^A N^B} + \chi \rho^{\mathcal{J}} \left[ \bar{b}^{\mathcal{I}} - \tilde{b}^{\mathcal{I}} \right] \right).$$

Thus the counterintuitive effect can occur at either extreme of heterogeneity. When interaction heterogeneity dominates,  $\chi$  is negative while interaction surplus is strongly positive, so average users on side  $\mathcal{I}$  may benefit from higher prices which encourage the platform to bring in more customers on side  $\mathcal{J}$ . For example in the RT (2003) model, expression (14) becomes  $\bar{\rho}^{\mathcal{I}} \rho^{\mathcal{J}} - 1$ ; second-order conditions require  $\rho^{\mathcal{I}} \rho^{\mathcal{J}} < 1$ , as shown above, so beneficial own-cost increases require a  $\rho^{\mathcal{I}}$  increasing rapidly in price, yielding a large Spence distortion.

On the other hand, when membership heterogeneity is strong enough to give negative interaction heterogeneity, interaction surplus is negative but  $\chi > 0$  and average newspaper readers<sup>36</sup> may actually want higher prices to force firms to internalize their distaste for advertising and reduce its quantity. In intermediate cases, such as the Armstrong model, with small Spence distortions, own-cost effects are always negative.

### C. Arbitrary Comparative Statics

Effects of local shocks to the market can always be expressed as a combination of direct externalities<sup>37</sup> and indirect effects through optimally chosen participation rates. The former can be analyzed through a partial derivative holding fixed firm actions; the second is equivalent to some combination of changes in the firm's (opportunity) cost on each side of the market. Therefore knowing  $dN^{\mathcal{I}}/dC^{\mathcal{J}}, N_1^{\mathcal{I}}, N_2^{\mathcal{I}}, V_1^{\mathcal{I}}, V_2^{\mathcal{I}}$  for both  $I, J = A, B$  is sufficient to compute arbitrary comparative statics; expressions for these are given in the text and Appendix. The same approach may be taken in the more general model proposed in Section III, though explicit expressions of the relevant derivatives do not appear in this paper.

<sup>36</sup> This makes clear that all the reasoning about surplus is about the *total* user surplus on each side of the market: it integrates over all users. Clearly marginal or near-marginal users are harmed by any increase in prices, even if these benefit loyal users. In some settings we may care about such distributional consequences (is ritzy readers' distaste for advertising reason enough to exclude poorer marginal readers?), but that analysis is beyond the scope of this paper and in fact most standard industrial organization.

<sup>37</sup> An earlier draft of this paper, available on request, provided a variety of such explicit comparative statics.

## V. An Example: the Scale-Income Model

A primary contribution of this paper is to simplify the platform's problem to analyze, for the first time, the effect of multiple dimensions of user heterogeneity. The ability to analyze these more general models does not, however, eliminate all motivation for unidimensional models. As shown above, multidimensional heterogeneity leaves substantial ambiguities about the direction of various distortions and comparative statics. In cases when most heterogeneity plausibly lies along a single dimension, making this assumption explicit can help resolve these. Furthermore, from an empirical perspective it may be difficult to identify a two-dimensional model without parametric assumptions; restricting heterogeneity to a single dimension may be a simple and transparent way to impose the necessary additional structure.

Unfortunately, the source of heterogeneity in the most commonly applied model (Rysman 2004; Kaiser and Julian Wright 2006; Elena Argentesi and Lapo Filistrucchi 2007), Armstrong's, seems implausible in most settings where it is applied. A primary dimension of heterogeneity for at least one side of the market is almost certainly the value derived from the other side. The RT (2003) model focuses on this source of heterogeneity but has the unfortunate feature that it rules out any membership costs or benefits, making it implausible in many industries. However, my foregoing analysis emphasized that most results in the RT (2003) model are due to the source of heterogeneity rather than the absence of membership costs and benefits. Thus most of the results of the RT (2003) model extend to a Generalized RT (2003) (henceforth GRT 2003) model that allows for (homogeneous) membership costs and benefits.<sup>38</sup>

However, the GRT (2003) model still seems to fit many markets poorly. Newspaper readers and software producers, to name a few, clearly differ substantially in their membership benefits and costs, respectively, of participating in a platform. One reasonable model<sup>39</sup> of such settings (Anderson and Coate 2005) has GRT (2003)-like users on one side and Armstrong-like users on the other. In this section I propose an alternative that I think is likely to be most fruitful in applications: *Scale-Income* (SI) model. It offers a useful rule of thumb for thinking about sources of heterogeneity, making analysis a bit more concrete.

Users on each side agree on the relative size of membership and interaction values but differ in scale. All newspaper readers (side  $\mathcal{A}$ ) lose a fraction  $-\beta^{\mathcal{A}}N^{\mathcal{B}}$  of the value they take from reading if a fraction  $N^{\mathcal{B}}$  of advertisers participate; however, they may differ in their total utility. Intuitively, higher income users have greater willingness to pay to gain the utility of reading the newspaper and avoid the disutility of advertising. Advertisers have the same value of circulation as a fraction of the fixed cost  $-B_i^{\mathcal{B}}$  they expend to establish a relationship with the newspaper, but differ in the scale of both of these depending on their business size. Thus  $b_i^{\mathcal{I}}/B_i^{\mathcal{I}} = \beta^{\mathcal{I}}$  for all  $i, \mathcal{I}$ , but users differ in the scale of their utility. They are heterogeneous vertically (Spence 1976; Eytan Sheshinski 1976) rather than horizontally (Harold Hotelling 1929).

I believe this model provides a better approximation to many two-sided markets than any of the other unidimensional models.<sup>40</sup> It seems to me a fairly good fit to software platforms<sup>41</sup> (operating

<sup>38</sup> This model was analyzed extensively in a previous draft of this paper and, while omitted here for brevity, is available on request.

<sup>39</sup> This *Hybrid* model was extensively analyzed in a previous draft of this paper, available on request.

<sup>40</sup> Note that the RT (2003) model is the special case of the SI model where  $\beta^{\mathcal{I}} = \infty$ . An interesting potential extension of the SI model is to extend this in the way the GRT (2003) model extends the RT (2003) model: allow users to lie along any line in  $R^2$ .

<sup>41</sup> As an example, I will go into a bit more detail on this case. Users typically derive some value from the platform itself and some proportional to the media (games or programs) on the platform. It seems reasonable to assume that the ratio between these is probably quite homogeneous in the population. Similarly software producers have development costs and average per user (unit profits multiplied by the probability of a purchase). At least in expected terms, this ratio is likely quite homogeneous, as software producers that expend larger fixed costs for the same variable benefit as another

systems, video games etc.), dating clubs, commercial intermediation (supermarkets, stock markets, eBay, etc.) and Internet service provision.

For concreteness, I focus here on a version of the model adapted to newspapers or other advertising platforms.  $\beta^{\mathcal{I}} < 0$  on both sides of the market: readers on side  $\mathcal{A}$  have positive membership values from reading the paper but negative interaction values from viewing advertising, while side  $\mathcal{B}$  advertisers have positive interaction value of circulation but membership costs associated with establishing relationships with the paper. Because

$$P^{\mathcal{I}} = \widetilde{B}^{\mathcal{I}} + \widetilde{b}^{\mathcal{I}}N^{\mathcal{J}} = \widetilde{b}^{\mathcal{I}}\left(N^{\mathcal{J}} + \frac{1}{\beta^{\mathcal{I}}}\right)$$

we have that  $\widetilde{b}^{\mathcal{I}} = P^{\mathcal{I}}/(N^{\mathcal{J}} + (1/\beta^{\mathcal{I}})) = P^{\mathcal{I}}\nu^{\mathcal{I}}$  where  $\nu^{\mathcal{I}} \equiv 1/(N^{\mathcal{J}} + (1/\beta^{\mathcal{I}}))$ . The Spence distortion from side  $\mathcal{I}$  is now  $\mu^{\mathcal{I}}\rho^{\mathcal{I}}N^{\mathcal{I}}$  as interaction surplus is just interaction market power multiplied by the average pass-through of its distribution (see Subsection IIIB above). Rather than the sign of the Spence distortion's being dictated directly by the model or left entirely ambiguous, it is given in an intuitive way by market conditions that can be reflected upon or econometrically measured. If prices on side  $\mathcal{B}$  (advertisers) have the same sign as interaction benefits on that side, then loyal users tend to have higher (or less negative, in the case of negative prices) interaction benefits than marginal users and therefore prices on side  $\mathcal{A}$  (readers) are distorted upwards. On the other hand if interaction benefits on side  $\mathcal{A}$  have the opposite sign of price, as with a high-quality newspaper whose readers dislike advertising, then loyal users tend to have more negative (or less positive when prices are negative) interaction values than marginal readers and therefore prices on side  $\mathcal{B}$  are distorted *downward* (advertisers).

Note that the crucial difference here is *not just the sign of interaction values, but how these compare to the sign of price*. Free tabloids essentially have a negative price, given their aggressive marketing in public transport hubs, and therefore have low scale-income, advertising-insensitive loyal readers, implying an upward Spence distortion *despite negative interaction benefits*.<sup>42</sup> Thus the SI model would have very different predictions about the behavior of tabloids versus high-quality papers as the marginal readers of both desert for the Internet: tabloids will become further laden with advertising and market more aggressively, while quality papers will pare back advertising and raise subscription fees.

Comparative statics are similarly dictated by the market conditions. The Scale-Income Model portion of the Appendix shows participation on the two sides are complements (substitutes) if and only if

$$\sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} \left( \nu^{\mathcal{I}} \left[ \widetilde{b}^{\mathcal{J}}N^{\mathcal{J}} - C^{\mathcal{I}} - cN^{\mathcal{J}} \right] \right) - c > (<)0.$$

For quality newspapers  $\nu^{\mathcal{A}} < 0 < \nu^{\mathcal{B}}$ . Assuming subscribers are net profitable even in the absence of subscription fees (advertisers are obviously unprofitable in this sense),  $\{\widetilde{b}^{\mathcal{B}}N^{\mathcal{B}} - C^{\mathcal{A}} - cN^{\mathcal{B}} > 0$ . So long as these effects are large enough to outweigh interaction costs, participation rates are complements. Also intuitively the equilibrium network effect from readers to advertisers is positive by complementarity, but the sign of equilibrium network effect of advertiser participation to

producer will be driven out of the market. However, some games and software are clearly much more prominent and higher impact than others, having larger fixed costs and variable benefits. Thus the SI model seems a sensible fit.

<sup>42</sup> Similarly if programs for an operating system are subsidized, as with Macintosh in the 1990s, low scale programs will tend to be served and thus there will be a downward Spence distortion (potentially underpriced operating systems) despite positive interaction benefits.

readers is ambiguous (the harm to loyal readers may outweigh the benefits to marginal readers or not).

Empirical data become particularly useful in the SI model as it has substantial identifying power even when little can be observed,<sup>43</sup> especially when explicit links can be made to observable income or size distributions,<sup>44</sup> as is common in structural empirical work (Berry, James Levinsohn, and Ariel Pakes 1995). Thus in cases where the source of user heterogeneity is unknown, relevant policy implications are theoretically ambiguous, and empirical data for full identification are missing, the SI model provides a reasonable way to structure policy intuitions. Furthermore, it can easily be extended to the general model of Section III:  $u^T(\mathbf{N}; \theta^T) \equiv \theta^T u^T(\mathbf{N})$  where  $u^T(\cdot)$  is an arbitrary smooth function of the allocation.

## VI. Applications

This section briefly discusses three policy-oriented applications of my results, designed to demonstrate how the tools developed help address longstanding applied questions.

### A. Measuring Market Power and Predation

In applied antitrust analysis, price-cost margins are used to measure market power or as a screen for predatory pricing. It has long been argued (David S. Evans 2003; Wright 2004) that pricing below cost is not indicative of predatory behavior<sup>45</sup> in a two-sided market as users may be subsidized on one side to reflect the benefits of users on the other side. Similarly pricing significantly above cost need not indicate large market power, as users on one side may be taxed if users on the other side have interaction costs. Measuring market power and predation in two-sided markets is therefore an old open question. My framework provides a simple answer: a general Lerner index for two-sided markets, which encompasses and unifies previous Lerner indices proposed for special models, such as Armstrong's and RT (2003).

One approach to such a Lerner index is to construct them for each side of the market individually, in which case they are given by equation (6) and require a measurement of the AIVMU, as well as costs.<sup>46</sup> Measuring the AIVMU may be difficult, but it's not much harder than observing costs. These measures can then be used, as any Lerner index, as a test for market power<sup>47</sup> and predation. Because prices are often near or below zero in two-sided markets, absolute market power  $\mu^T$ , perhaps normalized by something other than price, may be a more attractive metric as it is guaranteed to be positive and finite for a statically optimizing firm. This may be calculated

<sup>43</sup> A decomposition of price into interaction and membership benefits and identification of market power, which is feasible simply based on first-order instruments for participation or price on both sides of the market, suffice to identify interaction market power.

<sup>44</sup> These predict higher-order properties of demand, allowing pass-through rates and cross partials to be predicted, and the size of interaction surplus, and therefore normative comparative statics, to be estimated.

<sup>45</sup> While there is much dislike about requiring below true cost pricing as a necessary condition for predation (Aaron S. Edlin 2002), most legal (Frank H. Easterbrook 1981; Brooke Group Ltd. 1993) and economic (Patrick Bolton, Joseph F. Brodley, and Michael H. Riordan 2000) doctrine holds that allegations of predation must establish in Easterbrook's words "a sacrifice of today's profits for tomorrow's." This means that, in practice, to the extent predation cases arise at all in two-sided markets, the argument that below-cost pricing does not establish that prices are below "true" costs is likely to be a potent one.

<sup>46</sup> Market power may also be estimated structurally (Argentesi and Filistrucchi 2007). I discuss this approach, which also allows costs to be estimated rather than observed, extensively in Weyl (2009a).

<sup>47</sup> It is not immediately clear why market power, and not market power combined with the Spence distortion, is the right thing to measure. For the purposes of my discussion here, I just take as given the policymakers' interest in measuring market power.

just as easily:  $P^{\mathcal{I}} - C^{\mathcal{I}} - cN^{\mathcal{J}} + \widetilde{b}^{\mathcal{J}}N^{\mathcal{J}}$ . Then a natural test for predation on one side individually is that this be negative.

If, instead, an aggregate measure of market power is desired, weighting by participation on the two sides is a natural way to aggregate. The aggregate Lerner index is then

$$N^A\mu^A + \frac{N^B\mu^B}{N^AP^A} + N^BP^B = (1 + \alpha)\frac{\pi}{R}$$

where  $R$  is revenue and  $\alpha$  is the interaction surplus ratio of Subsection IVA. Intuitively if two-sidedness makes up a large part of profits, one should expect relatively low prices for a given amount of market power, as the platform will tend to subsidize users for participation. Therefore even a small profit to revenue ratio indicates significant market power if two-sidedness is a main source of profits. The test for predation is the natural extension of the standard test: profits are negative if and only if the aggregate Lerner index is. My formulae, side-specific or aggregated, extend intuitively to the general multi-sided model of Section III.

### B. Regulation

Regulation of two-sided markets has been a topic of substantial recent interest. Two prominent examples are the policy debates over interchange fee caps on prices to card-accepting merchants, and net neutrality regulations, interpreted variously as price caps on fees Internet service providers (ISPs) can charge Web sites or a limit on their price discrimination. As with merger analysis, designing regulation in two-sided markets is beyond the scope of this paper. However, I believe the paper does provide three issues for future research to consider.

First it emphasizes that, to the extent that regulation aims to emulate the optimal benchmarks of Pigouvian or Ramsey pricing, it should solve distortions on both sides. In fact Pigouvian and Ramsey pricing require solving a constant fraction of distortions on each side, rather than only one side, as with net neutrality or interchange fee regulations. In considering the size of these distortions it suggests two factors are crucial: the size of classical market power and the Spence distortion *on the other side of the market*.

Thus the novel element in two-sided markets is that regulators should focus most on reducing price opposite a side with a large Spence distortion. Thus regulators of ISPs should focus on limiting prices to Web sites (net neutrality) if there is more (interaction) surplus among loyal users than among highly profitable Web sites. But if the situation is reversed, forcing ISPs to reduce prices and build more line to consumer homes may be a higher priority.

Second, implementing Ramsey-type regulation requires a detailed knowledge of demand<sup>48</sup> that may not be available to a regulator. If so it may be more attractive to regulate only one side of the market, especially if market power is thought to particularly distort that side's prices. However a price cap on side  $\mathcal{I}$  can create further distortions, especially with positive interaction benefits, as the platform can lower side  $\mathcal{I}$ 's price either by increasing participation on side  $\mathcal{I}$  (which the regulator wants) or by decreasing participation on side  $\mathcal{J}$  (which she likely does not want). Thus Sheshinski's (1976) argument that price regulation tends to reduce quality provision is even stronger. In two-sided markets "quality reduction" comes from further distorting prices charged to users on the other side of the market. Of course, when interaction benefits are negative, especially if the Spence distortion is upward, this may be desirable: price caps on newspaper readers

<sup>48</sup> When information is more limited, the appropriate response is to explicitly incorporate these informational constraints into a model of policy design (David P. Baron and Myerson 1982; Jean-Jacques Laffont and Jean Tirole 1993). This is an important open problem in two-sided markets and is certainly beyond the scope of this paper.

may lead to more ads,<sup>49</sup> but this could well be an efficient counterbalance to their market power over advertisers especially if, as with tabloids, loyal readers dislike advertising less than marginal readers.

In the positive interaction benefit and Spence distortion case, when price regulation is particularly unattractive, Sheshinski's suggestion of quantity regulation may be more attractive as it does not change pricing incentives on the other side of the market. The simplest way to see this<sup>50</sup> is to note that the privately optimal pricing condition on side  $\mathcal{I}$  takes as given participation on side  $\mathcal{J}$ , and thus the first-order condition on side  $\mathcal{I}$  is not (directly) affected by a constraint on participation on side  $\mathcal{J}$ . A regulator might require an ISP to have a certain fraction of Web sites available on its service, rather than prohibiting the charging of Web sites. This might well encourage the recruitment of more Internet users, as a natural way to increase Web site participation without lowering price is to increase the number of subscribers. Of course, as in any market where quantity regulation is proposed, implementation would require care, to ensure that the ISP does not cheat by signing up the smallest Web sites. Given the diversity of Web sites, the practical complexity of implementing such a policy may outweigh its theoretical benefits. Furthermore, even at a theoretical level, more detailed analysis would be needed to establish the cases in which, once all indirect effects are taken into account, participation regulations are truly preferable to price regulations, and for whom. Nonetheless, such *allocation regulation* at least merits further investigation in multi-sided networks.

Finally, the analysis above seems to provide further rationale for allowing price discrimination in two-sided markets, at least when Spence distortions are positive. In this case discrimination has the additional benefit (Weyl 2009c; Rysman 2009) of increasing the subsidy to users on the other side of the market, helping ameliorate both the market power (on the other side) and Spence (on the discriminated side) distortions. Because Spence distortions are likely upward among Web sites (incumbents like Google make greater profits from a marginal surfer than entrants), this seems to lean in favor of allowing price discrimination,<sup>51</sup> that is, repealing net neutrality. However, if the Spence distortion is negative, as among *Times* readers, price discrimination may be more harmful than usual as it may lead to higher advertiser prices exacerbating market power.<sup>52</sup> Again, more detailed analysis of price discrimination would be required to formalize such arguments.

### C. Mergers

Merger analysis requires a general model of competition, which is beyond the scope of this paper. Nonetheless my results make three small contributions towards this goal.

First, the approach taken here is likely to be useful in analyzing such merger models. To illustrate this, I show in the Applications portion of the Appendix how it can be used to analyze mergers in a nonparametric, market-expanding version of Armstrong's model of symmetrically differentiated single-homing duopoly, the *generalized Armstrong single-homing model* (GASH). A companion paper (Weyl 2008) uses the same techniques to analyze mergers between

<sup>49</sup> Some of these issues are analyzed, under particular assumptions about user heterogeneity (see footnote 12) by Anderson and Coate (2005).

<sup>50</sup> A formalization is available on request.

<sup>51</sup> Of course, as in any vertical moral hazard/double marginalization problem, transferring incentives to the platform is not all good; this may hold up Web sites, extracting surplus from their investment in producing quality content if contracts are not sufficiently rich (Bengt Holmström 1982). For an analysis that emphasizes the effects on Web site investment see Jay Pil Choi and Byung-Cheol Kim (forthcoming).

<sup>52</sup> Of course this depends on whether prices are initially too high or too low to advertisers; in the latter case, the effect is ambiguous.

a platform and a non-two-sided firm producing a good that is a substitute (broadcast TV merger with advertising-free cable) or a complement (operating systems and hardware manufacturers) for users on one side of the market, considering the second case in substantial detail.

Second, the insulating tariff offers an approach to overcoming a problem<sup>53</sup> plaguing the analysis of competition in multi-sided networks. As Armstrong points out, a tremendous multiplicity of equilibria are possible in competition between platforms depending on the tariffs  $P^A(\cdot)$  and  $P^B(\cdot)$  the other firm uses *at participation levels other than the equilibrium*. For example, if one payment card firm uses a fixed fee, this will encourage the other to steal its cardholders as a means of attracting merchants who now have fewer partners, while if it uses a negative fixed fee and a large per-interaction fee this softens competition as stealing cardholders actually discourages merchant participation.

However, if one assumes firms choose insulating tariffs, these cross-side participation stealing effects are reduced and, in the GASH case, entirely eliminated. This *insulated equilibrium* greatly simplifies the analysis of competition. It also seems at least as reasonable an assumption as the more basic Nash-in-prices (Bertrand) assumption universal in the multi-sided networks literature,<sup>54</sup> given that this tariff is both intuitive and plausible, as well as robustly ensuring good equilibria are uniquely selected. It is a simple extension of the common assumption in price-quality competition that firms take as given other firms' choice of quality (Avner Shaked and John Sutton 1982) when choosing price, as the number of users participating on side  $\mathcal{I}$  is effectively the quality of the platform's "product" on side  $\mathcal{I}$ .

Third, my results suggest that in any model of competition, the source of user heterogeneity will be central to determining the positive and normative effects of mergers. Mergers largely affect firm market power, and potentially the size of network effects, both of which act to shift platform (opportunity) costs (Farrell and Carl Shapiro 2008). Because the positive and normative effects of changes in costs and network effects are determined by the sources of user heterogeneity, so too will be the effects of mergers. Furthermore, whether market power is more or less harmful in a two-sided market depends on the source of heterogeneity.

This is confirmed by the two most prominent existing models of competition in two-sided markets. In Weyl (2009c) I show that a merger (with no efficiencies) in the RT (2003) model of competition is tantamount to an increase in market power on one or more sides of the market. It will therefore *increase* participation, and potentially benefit users, on one side if competition is much more intense on one side (participation on the two sides are substitutes). On the other hand, in the online Appendix I show that, at least when competitors use insulating tariffs and regardless of the relative intensity of competition, a merger (without efficiencies) in the GASH model increases market power and therefore reduces participation and welfare on both sides, as

<sup>53</sup> An alternative approach to making a specific assumption about conduct, as I suggest here, is to search for results that are robust across various solution concepts or to attempt to explicitly identify the solution concept. The first approach seems reasonable, if challenging, and is an interesting direction for future research. A simple example of this strategy was a result, included in a previous version of this paper and available on request, that in many reasonable cases, even without an insulating tariff, mergers from GASH lead to lower participation on both sides. The second approach is in the spirit of the classic contributions of Timothy F. Bresnahan (1982) but has proven difficult to implement empirically given its data demands (Aviv Nevo 1998). Nonetheless there has been some recent interest in identifying solution concepts in other contexts, such as vertical relations (Sofia Villas-Boas and Rebecca Hellerstein 2006), so asking how one would go about identifying the two-sided markets solution concept (what sort of price schedules do firms take as given) would be an interesting topic for future research. Finally, one might use demand uncertainty to tie down a unique optimal tariff (Klemperer and Margaret A. Meyer 1989), though this approach has proved challenging to implement in applications in the simpler context of one-sided supply function equilibrium. Nonetheless I think operationalizing uncertainty-based refinements of oligopoly equilibria is an exciting direction for future research.

<sup>54</sup> Bruno Jullien proposed to me, in a private conversation, a model of undifferentiated Cournot-style competition. However, this model has symmetric equilibria only when there is a single dimension of user heterogeneity, making it difficult to analyze more generally. A proof is available on request.

participation rates are complements. Thus merger models in two-sided markets must show care in their assumptions about the sources of user heterogeneity.

## VII. Conclusion

This paper makes two contributions. First, by formulating the platform's problem in terms of its choice of allocation, rather than prices, I simplify and generalize the analysis of network industries. Second, I show that the key normative properties and comparative statics of two-sided markets depend on the source of user heterogeneity, which previous analysis has restricted. The modesty of these contributions makes clear the early stage of the literature. I therefore conclude by discussing directions for future research.

On the empirical side a number of questions are suggested quite directly by my arguments above. Does the SI model fit well in some market where *ex ante* the sources of user heterogeneity seem unclear? How well do the predictions of the RT (2003) model fit actual payment card data? Do newspapers actually exhibit complements? Comparing market power to the Spence distortions, are there overall too many or too few ads in most papers? Applications will largely be driven by the data available, so I will not dwell on them excessively here.

On the theoretical side, much remains to be done to understand pricing in networks more generally. For example, my approach so far allows only extremely stylized models of competition of limited direct empirical relevance. I consider formulating a workhorse, general empirical model<sup>55</sup> of two-sided markets, and practical means for identifying it, to be the most important open question in this area. For regulatory policy the monopoly model is likely to be of greater use, but a more careful analysis of price discrimination and regulatory design (Baron and Myerson 1982; Laffont and Tirole 1993) are needed.

A number of fundamental theoretical problems remain open, three of which I will mention.

First, the exploding literature on matching design, surveyed by Alvin E. Roth (2002), has thus far had limited interaction with the literature on pricing in two-sided markets; see Glenn Ellison, Drew Fudenberg and Markus Möbius (2004); Susan Athey and Glenn Ellison (2008); Ettore Damiano and Hao Li (2008); Andrei Hagiu and Bruno Jullien (2008); Weyl and Tirole (2010) for notable, if early, exceptions. These literatures have much in common, though market design has largely focused on efficiency and paid little attention to prices, while the two-sided markets literature largely ignores, as this paper does, the possibility of designing platforms to increase surplus. I suspect optimal pricing interacts importantly with platform design and therefore that such "revenue maximizing matching" is a fruitful direction for future research.

Finally, as my discussions in Subsection IC and IIIA emphasize, the coordination problems that have long been thought central to networks can generally be overcome by appropriate tariffs. This does not seem to always occur in practice, however. Insulating tariffs might be difficult to implement if demand is not known exactly to the platform; they might, in fact, be unprofitable if demand is uncertain as the critical mass problem might be an effective screen for high demand states. Standard capital market imperfections could also play a role in limiting the platform's ability to borrow, which might be necessary in the true dynamic process of network formation swept under the static model here. A platform might signal to its financiers that it knows it will succeed by overcoming the critical mass problem without subsidies. These are all interesting topics for future theoretical research.

Regardless of the precise explanation for imperfect insulation, my discussion suggests that coordination problems may be a choice, rather than a constraint. If correct, this would imply, for

<sup>55</sup> Alex White and Weyl (2010) make a first attempt at this, extending the insulating tariff to oligopoly.

example, that coordination is, on its own, an important basic source of market power and possible coordination failures are not a reasonable rationale for a merger or collusion. More careful evaluation of this controversial claim is an important theoretical challenge.

## REFERENCES

- Ambrus, Attila, and Rossella Argenziano.** 2009. "Asymmetric Networks in Two-Sided Markets." *American Economic Journal: Microeconomics*, 1(1): 17–52.
- Anderson, Simon P., and Stephen Coate.** 2005. "Market Provision of Broadcasting: A Welfare Analysis." *Review of Economic Studies*, 72(4): 947–72.
- Argentesi, Elena, and Lapo Filistrucchi.** 2007. "Estimating Market Power in a Two-Sided Market: The Case of Newspapers." *Journal of Applied Econometrics*, 22(7): 1247–66.
- Armstrong, Mark.** 2006. "Competition in Two-Sided Markets." *RAND Journal of Economics*, 37(3): 668–91.
- Athey, Susan, and Glenn Ellison.** 2008. "Position Auctions and Consumer Search." <http://kuznets.harvard.edu/~athey/position.pdf>.
- Baron, David P., and Roger B. Myerson.** 1982. "Regulating a Monopolist with Unknown Costs." *Econometrica*, 50(4): 911–30.
- Becker, Gary S., and Kevin M. Murphy.** 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics*, 108(4): 941–64.
- Bedre-Defolie, Özlem, and Emilio Calvano.** 2010. "Pricing Payment Cards." Unpublished.
- Berry, Steven T., and Joel Waldfoegel.** 1999. "Free Entry and Social Inefficiency in Radio Broadcasting." *RAND Journal of Economics*, 30(3): 397–420.
- Berry, Steven T., James Levinsohn, and Ariel Pakes.** 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841–90.
- Besanko, David, Shabtai Donnenfeld, and Lawrence J. White.** 1987. "Monopoly and Quality Distortion: Effects and Remedies." *Quarterly Journal of Economics*, 102(4): 743–67.
- Bolton, Patrick, Joseph F. Brodley, and Michael H. Riordan.** 2000. "Predatory Pricing: Strategic Theory and Legal Policy." *Georgetown Law Journal*, 88(8): 2239–330.
- Bresnahan, Timothy F.** 1982. "The Oligopoly Solution Concept Is Identified." *Economics Letters*, 10(1–2): 87–92.
- Caillaud, Bernard, and Bruno Jullien.** 2001. "Competing Cybermediaries." *European Economic Review*, 45(4–6): 797–808.
- Caillaud, Bernard, and Bruno Jullien.** 2003. "Chicken & Egg: Competition among Intermediation Service Providers." *RAND Journal of Economics*, 34(2): 309–28.
- Choi, Jay Pil, and Byung-Cheol Kim.** Forthcoming. "Net Neutrality and Investment Incentives." *RAND Journal of Economics*.
- Church, Jeffrey, and Neil Gandal.** 1992. "Network Effects, Software Provision, and Standardization." *Journal of Industrial Economics*, 40(1): 85–103.
- Damiano, Ettore, and Hao Li.** 2008. "Competing Matchmaking." *Journal of the European Economic Association*, 6(4): 789–818.
- Dybvig, Philip H., and Chester S. Spatt.** 1983. "Adoption Externalities as Public Goods." *Journal of Public Economics*, 20(2): 231–47.
- Easterbrook, Frank H.** 1981. "Predatory Strategies and Counterstrategies." *University of Chicago Law Review*, 48(2): 263–337.
- Economides, Nicholas.** 1996. "The Economics of Networks." *International Journal of Industrial Organization*, 14(6): 673–99.
- Edlin, Aaron S.** 2002. "Stopping above-Cost Predatory Pricing." *Yale Law Journal*, 111(4): 941–91.
- Ellison, Glenn, and Drew Fudenberg.** 2003. "Knife-Edge or Plateau: When Do Market Models Tip?" *Quarterly Journal of Economics*, 118(4): 1249–78.
- Ellison, Glenn, Drew Fudenberg, and Markus Mobius.** 2004. "Competing Auctions." *Journal of the European Economic Association*, 2(1): 30–66.
- Evans, David S.** 2003. "The Antitrust Economics of Multi-sided Platform Markets." *Yale Journal on Regulation*, 20(2): 325–81.
- Evans, David S., and Richard Schmalensee.** 2009. "Failure to Launch: Critical Mass in Platform Businesses." <http://www.interic.org/Conference/Schmalensee.pdf>.
- Fan, Ying.** 2010. "Ownership Consolidation and Product Quality: A Study of the U.S. Daily Newspaper Market." [http://www.personal.umich.edu/~yingfan/DailyNewspaper\\_Fan.pdf](http://www.personal.umich.edu/~yingfan/DailyNewspaper_Fan.pdf).

- Farrell, Joseph, and Paul Klemperer.** 2007. "Coordination and Lock-in: Competition with Switching Costs and Network Effects." In *Handbook of Industrial Organization*, Vol. 3, ed. Mark Armstrong and Robert H. Porter, 1967–2072. Amsterdam: North-Holland.
- Farrell, Joseph, and Carl Shapiro.** 2008. "Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition." <http://faculty.haas.berkeley.edu/shapiro/alternative.pdf>.
- Gaudeul, Alex, and Bruno Jullien.** 2008. "E-Commerce, Two-Sided Markets and Info-Mediation." In *Internet and Digital Economics: Principles, Methods and Applications*, ed. Eric Brousseau and Nicolas Curien, 268–90. Cambridge, UK: Cambridge University Press.
- Gentzkow, Matthew, and Jesse Shapiro.** 2010. "What Drives Media Slant? Evidence from U.S. Newspapers." *Econometrica*, 78(1): 35–71.
- Guthrie, Graeme, and Julian Wright.** 2007. "Competing Payment Schemes." *Journal of Industrial Economics*, 55(1): 37–67.
- Hagiu, Andrei.** 2006. "Pricing and Commitment by Two-Sided Platforms." *RAND Journal of Economics*, 37(3): 720–37.
- Hagiu, Andrei.** 2007. "Merchant or Two-Sided Platform?" *Review of Network Economics*, 6(2): 115–33.
- Hagiu, Andrei.** 2009. "Two-Sided Platforms: Product Variety and Pricing Structures." *Journal of Economics and Management Strategy*, 18(4): 1011–43.
- Hagiu, Andrei, and Bruno Jullien.** 2008. "Why Do Intermediaries Divert Search?" <http://www.people.hbs.edu/ahagiu/research.html>.
- Holmstrom, Bengt.** 1982. "Moral Hazard in Teams." *Bell Journal of Economics*, 13(2): 324–40.
- Hotelling, Harold.** 1929. "Stability in Competition." *Economic Journal*, 39(153): 41–57.
- Kaiser, Ulrich, and Minjae Song.** 2009. "Do Media Consumers Really Dislike Advertising?: An Empirical Assessment of the Role of Advertising in Print Media Markets." *International Journal of Industrial Organization*, 27(2): 292–301.
- Kaiser, Ulrich, and Julian Wright.** 2006. "Price Structure in Two-Sided Markets: Evidence from the Magazine Industry." *International Journal of Industrial Organization*, 24(1): 1–28.
- Katz, Michael L., and Carl Shapiro.** 1985. "Network Externalities, Competition, and Compatibility." *American Economic Review*, 75(3): 424–40.
- Katz, Michael L., and Carl Shapiro.** 1994. "Systems Competition and Network Effects." *Journal of Economic Perspectives*, 8(2): 93–115.
- Klemperer, Paul D., and Margaret A. Meyer.** 1989. "Supply Function Equilibria in Oligopoly under Uncertainty." *Econometrica*, 57(6): 1243–77.
- Laffont, Jean-Jacques, and Jean Tirole.** 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.
- Lee, Robin S.** 2009. "Vertical Integration and Exclusivity in Platform and Two-Sided Markets." <http://pages.stern.nyu.edu/~rslee/>.
- Liebowitz, S. J., and Stephen E. Margolis.** 1994. "Network Externality: An Uncommon Tragedy." *Journal of Economic Perspectives*, 8(2): 133–50.
- Lindhal, Erik.** 1919. "Positive Lösung." *Die Gerechtigkeit der Besteuerung*.
- Ltd., Brooke Group.** 1993. "Brooke Group Ltd. v. Brown & Williamson Tobacco Corp." (92–466), 509 U.S.(209).
- Mussa, Michael, and Sherwin Rosen.** 1978. "Monopoly and Product Quality." *Journal of Economic Theory*, 18(2): 301–17.
- Myerson, Roger B.** 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73.
- Nevo, Aviv.** 1998. "Identification of the Oligopoly Solution Concept in a Differentiated-Products Industry." *Economics Letters*, 59(3): 391–95.
- O’Keefe, Mary, W. Kip Viscusi, and Richard J. Zeckhauser.** 1984. "Economic Contests: Comparative Reward Schemes." *Journal of Labor Economics*, 2(1): 27–56.
- Oum, Tae Hoon, and Michael W. Tretheway.** 1988. "Ramsey Pricing in the Presence of Externality Costs." *Journal of Transport Economics and Policy*, 22(3): 307–17.
- Pigou, Arthur.** 1920. *The Economics of Welfare*. London: MacMillan.
- Rochet, Jean-Charles, and Jean Tirole.** 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association*, 1(4): 990–1029.
- Rochet, Jean-Charles, and Jean Tirole.** 2006. "Two-Sided Markets: A Progress Report." *RAND Journal of Economics*, 37(3): 645–67.
- Roth, Alvin E.** 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." *Econometrica*, 70(4): 1341–78.
- Rysman, Marc.** 2004. "Competition between Networks: A Study of the Market for Yellow Pages." *Review of Economic Studies*, 71(2): 483–512.

- Rysman, Marc.** 2009. "The Economics of Two-Sided Markets." *Journal of Economic Perspectives*, 23(3): 125–43.
- Samuelson, Paul A.** 1941. "The Stability of Equilibrium: Comparative Statics and Dynamics." *Econometrica*, 9(2): 97–120.
- Segal, Ilya.** 1999. "Contracting with Externalities." *Quarterly Journal of Economics*, 114(2): 337–88.
- Shaked, Avner, and John Sutton.** 1982. "Relaxing Price Competition through Product Differentiation." *Review of Economic Studies*, 49(1): 3–13.
- Sheshinski, Eytan.** 1976. "Price, Quality and Quantity Regulation in Monopoly Situations." *Economica*, 43(17): 127–37.
- Spence, A. Michael.** 1975. "Monopoly, Quality, and Regulation." *Bell Journal of Economics*, 6(2): 417–29.
- Spence, A. Michael.** 1976. "Product Differentiation and Welfare." *American Economic Review*, 66(2): 407–14.
- Veiga, Andre, and E. Glen Weyl.** 2010. "Platforms with Heterogeneous Externalities." Unpublished.
- Villas-Boas, Sofia, and Rebecca Hellerstein.** 2006. "Identification of Supply Models of Retailer and Manufacturer Oligopoly Pricing." *Economics Letters*, 90(1): 132–40.
- Weyl, E. Glen.** 2008. "Double Marginalization in Two-Sided Markets." [http://www.people.fas.harvard.edu/~Eweyl/Genius\\_JMP.pdf](http://www.people.fas.harvard.edu/~Eweyl/Genius_JMP.pdf).
- Weyl, E. Glen.** 2009a. "Slutsky Meets Marschak: First-Order Identification of Multi-Product Production." [http://www.people.fas.harvard.edu/~weyl/SlutMars\\_12\\_09\\_pdf](http://www.people.fas.harvard.edu/~weyl/SlutMars_12_09_pdf).
- Weyl, E. Glen.** 2009b. "Monopoly, Ramsey and Lindahl in Rochet and Tirole (2003)." *Economics Letters*, 103(2): 99–100.
- Weyl, E. Glen.** 2009c. "The Price Theory of Two-Sided Markets." [http://www.people.fas.harvard.edu/~weyl/pt2sms\\_3\\_09.pdf](http://www.people.fas.harvard.edu/~weyl/pt2sms_3_09.pdf).
- Weyl, E. Glen, and Michal Fabinger.** 2009. "Pass-Through as an Economic Tool." [http://www.people.fas.harvard.edu/~weyl/Pass-through\\_10\\_09.pdf](http://www.people.fas.harvard.edu/~weyl/Pass-through_10_09.pdf).
- White, Alexander, and E. Glen Weyl.** 2010. "Imperfect Platform Competition: A General Framework." Unpublished.
- Wright, Julian.** 2004. "The Determinants of Optimal Interchange Fees in Payment Systems." *Journal of Industrial Economics*, 52(1): 1–26.

## A Price Theory of Multi-Sided Platforms

By E. GLEN WEYL\*

*I develop a general theory of monopoly pricing of networks. Platforms use insulating tariffs to avoid coordination failure, implementing any desired allocation. Profit maximization distorts in the spirit of A. Michael Spence (1975) by internalizing only network externalities to marginal users. Thus the empirical and prescriptive content of the popular Jean-Charles Rochet and Jean Tirole (2006) model of two-sided markets turns on the nature of user heterogeneity. I propose a more plausible, yet equally tractable, model of heterogeneity in which users differ in their income or scale. My approach provides a general measure of market power and helps predict the effects of price regulation and mergers. (JEL D42, D85, L14)*

The pricing problems of payment and advertising platforms have much in common. Both seek to attract two distinct groups of users: AmEx needs cardholders and merchants, while the *New York Times* recruits readers and advertisers. Because the value each group takes from using these services depends on the size of the *other side of the market*, the platform's pricing and marketing strategies to each group are closely linked. Therefore policy directed at alleviating distortions caused by market power in these industries must take account of how interventions on one side affect welfare and platform behavior on the other.

Yet despite credit cards and newspapers both being canonical *two-sided markets*, the economics of these industries seem intuitively quite different. Consumers most likely to carry AmEx are those who most value the opportunity to use the card. These loyal cardholders therefore value the participation of merchants more than those indifferent between AmEx and another payment form do. Given its limited ability to price discriminate, AmEx fails to fully internalize the preferences of loyal users, putting too little effort into attracting merchants and charging them a higher price than would be socially optimal. However, when the costs of attracting cardholders rise and therefore cardholder incentives fall, AmEx will tend to serve only users who value merchant participation more strongly, leading them to attract more merchants with lower fees. This logic is the basis of the burgeoning literature on two-sided markets pioneered by Rochet and Tirole (2003).

\*Harvard Society of Fellows and Toulouse School of Economics: 78 Mount Auburn Street, Cambridge, MA 02138 (e-mail: weyl@fas.harvard.edu). This paper was split off as one part of a previous working paper "Monopolies in Two-Sided Markets: Comparative Statics and Identification." I am grateful to el Ministerio de Hacienda de Chile and the Centro de Investigación Económica at the Instituto Tecnológico Autonomo de México (CIE-ITAM), which hosted me on a visit while I conducted parts of this research. The Harvard Milton Fund supported the last stage of this project and financed the excellent research assistance provided by Will Weingarten, Alex White, and especially Stephanie Lo. Ahmed Jaber and Dan Sacks also acted as research assistants with the permission of their employer Amy Finkelstein, to whom I am grateful for her generosity. I also appreciate the helpful comments and advice on this research supplied by Mark Armstrong, David Evans, Jeremy Fox, Andrea Hawkey, Alisha Holland, Scott Kominers, Luis Rayo, Bill Rogerson, Dick Schmalensee, Jesse Shapiro, and especially Maxim Engers, and seminar participants at the Banco de México, the Barcelona Graduate School of Economics, CIE-ITAM, Harvard University, IESE Business School, Northwestern University, Télécom ParisTech Conference on the Economics of Information and Communications Technology, the Toulouse School of Economics (TSE), University of Chicago, and University of California at Berkeley. I owe a special debt to Pai-Ling Yin for detailed discussions of the paper. All confusion and errors are my own.

Matters are quite different for the *New York Times*. Its loyal customers are high income readers who dislike advertising but are willing to pay more for the paper's content than marginal readers who are less sensitive to advertising.<sup>1</sup> Thus the *Times* fails to internalize loyal readers' distaste for advertising, leading to potentially excessive advertising as a result of *below* optimal pricing to advertisers despite market power. Increases in the costs of distribution that reduce the number of subscribers will tend to *reduce* advertisements as the paper internalizes the costs to its wealthier readers. While intuitive in many markets, this opposite case has been assumed away by previous theoretical and empirical analysis of two-sided markets.

This paper shows that both of these are special cases of a simplified reformulation of the Rochet and Tirole, henceforth RT, (2006) model of monopoly in two-sided markets. The crucial difference between cases is the source of user heterogeneity. While credit card users primarily differ in the *interaction* (or usage) value they take from merchants accepting cards, newspaper readers differ most importantly in their *membership* value from reading the paper's content. This distinction is crucial because participation on one side of the market effectively determines the quality of the platform on the other side. Therefore, like any monopolist who must choose a single quality as well as quantity, the platform internalizes network effects to marginal rather than average participating users (Spence 1975).

The diversity of possibilities in two-sided markets does not eliminate the theory's predictive power. Because the distortions caused by market power (Section II) are linked to observable comparative statics (Section IV) through user heterogeneity, both intuition about the sources of this heterogeneity and empirical data can help calibrate the model in applications. Further restrictions may then be placed on the models (Section V) with a clear sense of how special assumptions increase predictive and prescriptive power. Together these results inform policy analysis in two-sided markets (Section VI), providing a general measure of market power and helping to predict the effects of regulation and mergers.

My analysis rests on a view of the platform's problem as choosing participation rates on the two sides rather than the prices supporting this allocation. This approach (Section I) is justified by an appropriate platform pricing strategy, the *insulating tariff*, that avoids potential coordination failures, thereby allowing the platform to achieve any desired allocation as a unique equilibrium. It applies, along with much of my analysis, more generally (Section III) than the RT (2006) model, allowing a simple approach to analyzing network industries with arbitrary heterogeneous utility, network effects, and any number of sides. I thereby answer perhaps the oldest open question in the theory of network industries (S. J. Liebowitz and Stephen E. Margolis 1994): does a monopolistic platform internalize and therefore neutralize network effects? The answer is yes, but imperfectly given the Spence distortion.

Of course this article is only a first pass at a general analysis of network pricing. Section VII therefore concludes by discussing directions for future research. Longer and less instructive proofs are collected into an Appendix available at <http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.4.1642>.

<sup>1</sup> Note that the intuitive stories I tell about these industries are not intended to be specific prescriptions about policy in these industries, but rather concrete instances of general theoretical possibilities. It is the mapping between the story's assumptions and the conclusions about policy, rather than the validity of the assumptions in a particular industry, that I am interested in here. Thus I do not have any empirical evidence substantiating my stories, evidence that would be highly desirable before reaching definite policy conclusions. For example, Ulrich Kaiser and Minjae Song (2009) argues that users do not actually dislike advertising, as my story assumes. Furthermore, in some cases at least, content may be viewed as an interaction rather than membership benefit if it is tightly tailored to accompany an ad, though I doubt this is the case for newspapers.

## I. Framework

The definition of two-sided markets is controversial (RT 2006; Marc Rysman 2009). For me, the phrase denotes a style of industrial organization modeling<sup>2</sup> developed by, among others, Bernard Caillaud and Bruno Jullien (2001; Caillaud and Jullien (2003; Rochet and Tirole (2003); Simon P. Anderson and Stephen Coate (2005); Mark Armstrong (2006); and Rochet and Tirole (2006). These models tend to emphasize three features.<sup>3</sup>

- i) *Multi-product firm*: A platform provides distinct services to two sides of the market, which can be explicitly charged different prices.
- ii) *Cross network effects*: Users' benefits from participation depend on the extent of user participation on the other side of the market, which varies with market conditions.
- iii) *Bilateral market power*: Platforms are price setters (monopolistic or oligopolistic) on both sides of the market and typically set uniform prices.

The failure of any of these conditions makes simpler and better-understood models more appropriate. If a platform does not explicitly charge different prices to different groups of users, it is best viewed as a standard, one-sided network. When participation does not vary on both sides a vertical monopolies model fits better. An absence of market power allows us to model the firm as a distributor. However, many industries<sup>4</sup> relevant to industrial policy exhibit all of the above characteristics. RT (2006) introduce a "canonical model," in their words, of monopoly capturing these three features in a simple manner that still generalizes the two most influential models in the literature, those of RT (2003) and Armstrong (2006).

### A. The RT (2006) model

Before describing it more formally, I highlight a few key assumptions of the model, on top of the familiar notions of monopoly and constant marginal cost.

- i) User valuations are taken as exogenous to any direct interactions between users on the two sides. Thus the RT (2006) model takes a "macro" approach, in the terminology of Nicholas Economides (1996). While "micro" models that directly incorporate such interactions (Julian Wright 2004; Andrei Hagiu 2006; Graeme Guthrie and Julian Wright 2007) have similar positive forms, their normative implications can be quite different.
- ii) Network effects are all *across*, not *within* the two sides. This rules out, for example, negative within-side effects from competition among software creators or positive collaboration effects among operating system users.

<sup>2</sup> These can either be viewed as models aimed at capturing important features of some true class of "two-sided markets" or they can be viewed as a style of modeling that captures some elements of "two-sidedness" that are more or less important in different industries. I lean towards the second view.

<sup>3</sup> I am grateful to Bruno Jullien and Patrick Rey for helping me refine these criteria.

<sup>4</sup> For example, credit cards, newspapers, operating systems, Internet service providers and others discussed by RT (2003) and Armstrong.

- iii) Users on each side interact with either all or a random subset of users on the other side, price discrimination<sup>5</sup> within a particular side based on the number of such interactions is impossible, and user values are affine in the number of users on the other side. This does not rule out some users interacting with a larger random and unpriced sample of users on the other side; this will just magnify their interaction valuations.
- iv) Finally, it assumes that users on each side of the market are of equal value to those on the other side. This rules out, for example, high advertising-value readers of a newspaper and video games that are especially valued by gamers (Robin S. Lee 2009).

Few of these assumptions are necessary for my analysis; in fact, assumptions ii–iv can be substantially relaxed or eliminated entirely. However, doing so complicates the exposition. Furthermore, given its greater parsimony, I suspect the RT (2006) model will continue to be the most attractive framework in many applications, including those I focus on: the payments and newspaper industries. Most importantly, its assumptions fit many industries quite well.

Consider the case of the newspaper industry as an example. The ways in which advertisers gain from readers viewing their ads, or why exactly users dislike ads (Gary S. Becker and Kevin M. Murphy 1993), seems fairly exogenous to industrial policy in the newspaper industry. While advertisements sometimes compete within a paper for user attention, it seems fairly reasonable to assume that advertisers are close to indifferent as to the number of other advertisements included in a paper, and readers are indifferent to the number of other readers of the paper. Readers usually read all or a fairly random selection of advertisements in a paper, and certainly it seems difficult to charge users (or advertisers) differentially based on the number of advertisements viewed. Finally, some advertisements are certainly more annoying than others and some readers more valuable than others to advertisers. However, I follow many top past applied papers (Stephen T. Berry and Joel Waldfogel 1999; Matthew Gentzkow and Jesse Shapiro 2010; Yin Fan 2010) on industries with advertising in viewing this as of second-order importance.

Therefore I develop most of my analysis in the context of the RT (2006) application, treating the general case only in Section III. There I show that my basic message in the RT (2006) applies generally. Therefore little is lost by focusing on the RT (2006) model, and Section III will likely be of most interest to theoretically inclined readers.

### B. User Preferences and Heterogeneity

There is a continuum of potential users on each side  $\mathcal{I} = \mathcal{A}, \mathcal{B}$  of the market, with mass normalized to 1. Thus the number of users participating on each side represents the fraction of potential participants choosing to do so. All quantities are scaled accordingly as discussed below.  $\mathcal{I}$  refers to a generic side of the market and  $\mathcal{A}$  and  $\mathcal{B}$  to refer to specific sides in examples.

A typical user  $i$  on side  $\mathcal{I}$  has an inherent *membership* benefit or cost  $B_i^{\mathcal{I}}$  from participating in the service if no users participate on the other side. For example, developers must pay fixed costs even if no users own the operating system the software runs on. Given my normalization of a unit mass of users,  $B_i^{\mathcal{I}}$  must be measured in terms of the total value all users on side  $\mathcal{I}$  would derive if they participated given that they have the same preferences as user  $i$ . Suppose a town has

<sup>5</sup> As in all models with market power, the impossibility of price discrimination plays a crucial role in normative conclusions. I believe price discrimination is probably neither systematically easier nor more difficult in two-sided markets than in standard markets. Even when some discrimination is possible, I believe the discrimination-free model gives some insight, as long as the discrimination is imperfect.

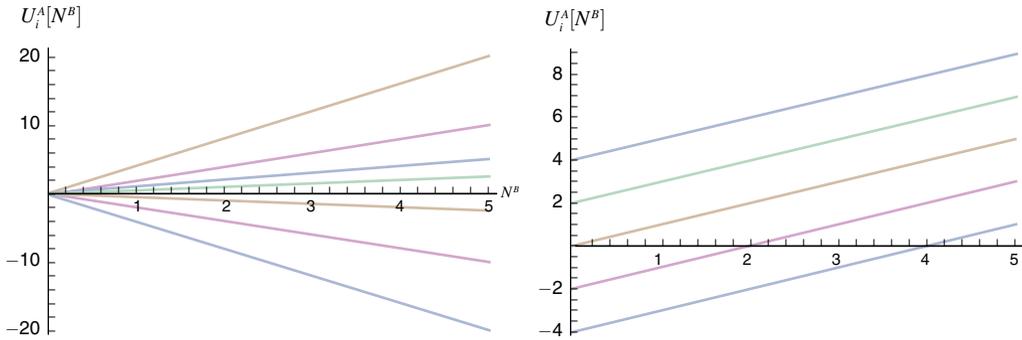


FIGURE 1.  $U_i^A(N^B)$  FOR VARIOUS RT (2003) (LEFT) AND ARMSTRONG (RIGHT) PREFERENCES.

Note: This Illustrates the Two Dimensions of Heterogeneity: Interaction and Membership Values, Respectively

100 possible newspaper subscribers and user  $i$  values reading her paper at \$500 a year; then her membership benefit would be  $B_i^T = \$50,000$ .

Each user also derives an *interaction*<sup>6</sup> benefit or cost of participation  $b_i^T$  for every user that participates on the other side. Again this must be appropriately scaled.<sup>7</sup> One of 1,000 credit card holders who makes 100 purchases every year deriving a 50 cent surplus from credit over cash would have a interaction benefit of \$50,000 per year, as this would be the value to all users on her side if all cards were accepted and all users on her side had the same preferences as she does. If there were 1,000 possible advertisers and 100 readers in a town and the disutility to a user  $i$  of each advertiser being included in a local newspaper were ten cents, then the interaction cost of that user would be  $b_i^T = \$10,000$ . I follow most of the industrial organization literature in assuming that user utility is quasi-linear in money.

Formally the utility derived by user  $i$  on side  $\mathcal{I}$  from participating is

$$U_i^T = B_i^T + b_i^T N^J - P^T(N^J)$$

where  $N^J$  is the number of users participating on side  $\mathcal{J} = -\mathcal{I}$ , the other side than  $\mathcal{I}$ .  $P^T(N^J)$  is the tariff set by the platform (independence of  $i$  disallows price discrimination), prescribing how much users must pay (or will be paid) to participate conditional on a given size of the platform on side  $\mathcal{J}$ . Users on each side can therefore be heterogeneous along two dimensions: interaction and membership values. Two natural special cases involve only one dimension of heterogeneity. RT (2003) assume that  $B_i^T \equiv 0$  and that users have heterogeneous interaction values. Armstrong (2006) assumes homogeneous interaction values ( $b_i^T \equiv b^T$ ) and allows heterogeneous membership values. Figure 1 shows the difference between these specifications. Utility is graphed as a function of participation on the other side of the market for various RT (2003) preferences (left) and Armstrong preferences (right). When, in general, there are both dimensions of heterogeneity, even fixing  $N^J$  and  $P^T$ , many different types of users may be just on the margin between participating and not (have  $U_i^T = 0$ ): some may have high interaction benefits but large membership costs; others may have low interaction benefits and no membership costs. This is pictured in Figure 2, where all users lying along the lines are marginal. The implications of these different

<sup>6</sup> RT (2006) refers to this as the user’s usage valuation; I eschew this terminology to avoid confusion, as users have no choice over how intensively to use the service in the RT (2006) model.

<sup>7</sup> Of course these scales can be renormalized as suits a given application, so long as this is done consistently.

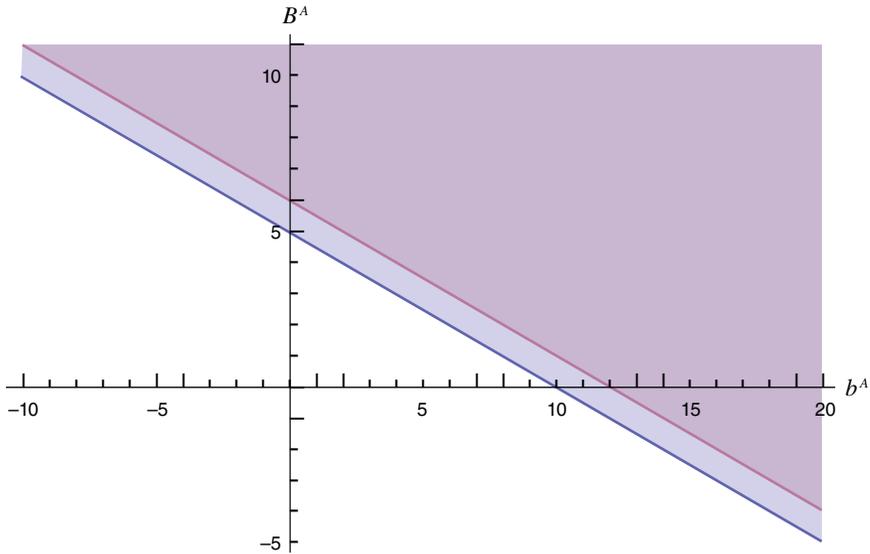


FIGURE 2. THE SET OF USERS PARTICIPATING ON SIDE  $\mathcal{A}$  WHEN HALF OF USERS PARTICIPATE ON SIDE  $\mathcal{B}$  AND  $p^{\mathcal{A}} = 5$  AND  $6$  RESPECTIVELY

sources of user heterogeneity are the primary focus of this paper. Formally I assume that the user parameters are distributed according to some massless, twice continuously differentiable<sup>8</sup> bi-variate distribution with probability density function  $f^{\mathcal{I}}(B_i^{\mathcal{I}}, b_i^{\mathcal{I}})$  and a full support.

C. Coordination and Insulating Tariffs

Once the tariff is set, users on the two sides of the market play a game. A user  $i$  on side  $\mathcal{A}$  will choose to participate if and only if

$$B_i^{\mathcal{A}} + b_i^{\mathcal{A}}N^{\mathcal{B}} > P^{\mathcal{A}}(N^{\mathcal{B}}).$$

However, this typically depends on the decisions of user on side  $\mathcal{B}$ , which  $\mathcal{A}$  users take as given. There may be multiple equilibria for some tariffs and distributions of user tastes. For example, suppose  $b_i^{\mathcal{I}} \equiv 1$ ,  $B_i^{\mathcal{I}} \equiv 0$  and  $P^{\mathcal{I}} \equiv 1/2$  for  $\mathcal{I} = \mathcal{A}, \mathcal{B}$ .<sup>9</sup> Then it is clearly an equilibrium for either all or none of the users to participate. In the former case the utility of participation on either side (taking the other as given) is  $1/2$ ; in the latter it is  $-1/2$ . This is the classic “chicken and egg” problem in two-sided markets (Caillaud and Jullien 2003).

Yet, in a sense, equilibrium multiplicity is inessential to the analysis of two-sided markets. A given pair of participation rates  $\widetilde{N}^{\mathcal{A}}$  and  $\widetilde{N}^{\mathcal{B}}$  leads to a unique profit and social welfare. To see this, note that, given a fixed side  $\mathcal{B}$  participation rate  $\widetilde{N}^{\mathcal{B}}$ , there is a well-defined demand function determining the number of users who participate on side  $\mathcal{A}$  as a function of  $P^{\mathcal{A}}$ , the equilibrium price to side  $\mathcal{A}$ . Visually, this is depicted in Figure 2, where the set of side  $\mathcal{A}$  users participating

<sup>8</sup>Note that the assumption of twice-continuous differentiability actually rules out all of the one-dimensional cases. As the online Appendix shows, the only assumption needed for the analysis is smoothness with regard to movements of the boundary of a set; that is sufficient conditions for the Leibnitz Integral Rule to apply.

<sup>9</sup>Note this example does not fit my full support and massless assumption, but an analogous example that does can be constructed by perturbing it.

when  $N^B = 0.5$  and  $P^A = 5$  or 6 is shaded. Clearly participation on side  $\mathcal{A}$ , holding fixed side  $\mathcal{B}$  participation, declines in  $P^A$ . Formally for either  $\mathcal{I}$

$$(1) \quad N^{\mathcal{I}}(P^{\mathcal{I}}, \widetilde{N}^{\mathcal{J}}) \equiv \int_{-\infty}^{\infty} \int_{P^{\mathcal{I}} - b^{\mathcal{I}} \widetilde{N}^{\mathcal{J}}}^{\infty} f^{\mathcal{I}}(B^{\mathcal{I}}, b^{\mathcal{I}}) dB^{\mathcal{I}} db^{\mathcal{I}}.$$

Clearly<sup>10</sup>  $N^{\mathcal{I}} \equiv \partial N^{\mathcal{I}} / \partial P^{\mathcal{I}} < 0$ . Therefore inverting it with respect to its first argument yields a well-defined function  $P^{\mathcal{I}}(N^{\mathcal{I}}, N^{\mathcal{J}})$ . Thus there is a unique pair of prices, and therefore profit and welfare,  $P^A(\widetilde{N}^A, \widetilde{N}^B)$  and  $P^B(\widetilde{N}^B, \widetilde{N}^A)$  consistent with  $\widetilde{N}^A$  users participating on side  $\mathcal{A}$ ,  $\widetilde{N}^B$  users participating on side  $\mathcal{B}$ , and uniform pricing.

Thus multiplicity plays no role if one thinks of the platform as simply choosing an allocation<sup>11</sup> to maximize some objective function. The only concern is that the platform may struggle to consistently implement its desired allocation; there could be a “failure to launch” as a result of a “critical mass problem,” in the terminology of David S. Evans and Richard Schmalensee (2009).

This can be avoided, however, by a conscientious platform. The platform can lower (raise) its price<sup>12</sup> on side  $\mathcal{I}$  when hoped-for (undesirable) side  $\mathcal{J}$  users that are valuable (harmful) to marginal users on side  $\mathcal{I}$  fail to show up. This insulates the platform’s allocation on side  $\mathcal{I}$  from the influence of side  $\mathcal{J}$  participation. At the logical extreme the platform can ensure that  $\widetilde{N}^{\mathcal{I}}$  users participate on side  $\mathcal{I}$  regardless of side  $\mathcal{J}$  decisions if it charges the *insulating tariff*<sup>13</sup>  $P^{\mathcal{I}}(N^{\mathcal{J}}) \equiv P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, N^{\mathcal{J}})$ . Then the unique equilibrium is the platform’s target allocation.<sup>14</sup>

Insulating tariffs are intuitive in many applications. With homogeneous interactions values (Armstrong), the insulating tariff is an insurance scheme, as originally proposed by Phillip H. Dybvig and Chester S. Spatt (1983) for a one-sided market. The platform charges users a price  $b^{\mathcal{I}}$  per user on side  $\mathcal{J}$  and charges an “hedonic” price (in Armstrong’s language) which determines participation. Therefore side  $\mathcal{I}$  users are indifferent to  $N^{\mathcal{J}}$ . With no membership values (RT 2003), the insulating tariff is a pure interaction price  $p^{\mathcal{I}}$  so that any side  $\mathcal{I}$  user earns utility  $(b_i^{\mathcal{I}} - p^{\mathcal{I}})N^{\mathcal{J}}$  from participating. Thus participating side  $\mathcal{I}$  users prefer high side  $\mathcal{J}$  participation and thus are not insured but, because the sign of their utility is independent of side  $\mathcal{J}$  participation, still choose to participate independent of the decisions of side  $\mathcal{J}$  users. In general, the composition, but not level, of participants may shift with participation on the other side: a rise in  $N^{\mathcal{J}}$  selects  $\mathcal{I}$  users with high interaction values.

<sup>10</sup> This follows from my assumption of full support.

<sup>11</sup> This approach, which is the key method used throughout the paper to simplify the complexities of pricing in two-sided markets, was first suggested to me in the context of the RT (2003) model by Jeremy Bulow, to whom I am tremendously grateful. Because of the single dimensionality of user heterogeneity in that model, the allocation approach is not much simpler than the price approach there. This led me, much to my later regret, to ignore Jeremy’s advice until after having wasted months trying to implement the price approach. On this, as many other matters, I have come round to seeing the elegance of his perspective. My approach was also inspired by the fulfilled expectations equilibrium of Michael L. Katz and Carl Shapiro (1985) and more broadly by the contract theory literature, starting with Roger B. Myerson (1981). It is surprising that, given the long history of the allocation approach in related literatures, it has not to my knowledge been applied previously to the general analysis of network industries.

<sup>12</sup> A further assumption of the RT (2006) model, not discussed extensively above, is that a (uniform) price can be set at any desired level on both sides of the market costlessly. This assumption fails in broadcast media, as argued by Anderson and Coate (2005). They explore, in a Hybrid model (see Section V) with RT (2003) preferences for advertisers and Armstrong preferences for consumers, the consequences of this price rigidity. A more general analysis of restrictions on pricing is, as discussed in Subsection VIA, an important direction for future research.

<sup>13</sup> I am grateful to Bruno Jullien for helping to guide me towards this name.

<sup>14</sup> Note that the platform can charge an insulating tariff on just one side  $\mathcal{I}$  of the market and achieve the same guarantee, as this assures that any equilibrium must have  $\widetilde{N}^{\mathcal{I}}$  users, removing expectations from the decision making of  $\mathcal{J}$  users. This is what makes possible Anderson and Coate’s (2005) analysis: they assume the platform chooses quantity, rather than price, to advertisers (effectively assuming an insulating tariff to one side). See footnote 27 for a more general discussion.

Schemes resembling insulating tariffs are used explicitly in many industries: Web site ad rates are typically per click and credit card fees/incentives per transaction. In fact, in broadcasting, as Anderson and Coate (2005) argue, the structure of programming often allows platforms to commit to a quantity of advertising directly. However, even when such explicit schemes are not used, the static RT (2006) model can reasonably be thought of as a reduced form for a dynamic model, in the spirit of Evans and Schmalensee (2009). In this case insulating tariffs simply require that the platform provides subsidies at early stages of product development which it recoups once its desired allocation is achieved. This pattern is commonly observed in video games, operating systems and Web sites.

However, there may be some circumstances under which firms would refrain or be constrained from employing them; see Section VII for further discussion. In these cases the critical mass problem binds and the coordination problems considered by Glenn Ellison and Drew Fudenberg (2003); Atilla Ambrus and Rossella Argenziano (2009); and Evans and Schmalensee (2009) become important.<sup>15</sup> However, in most mature industries, the focus of the RT (2006) model, price flexibility is sufficient to avoid these problems.<sup>16</sup>

Many other tariffs achieve the platform’s desired allocation, even uniquely. In fact, as argued by RT (2003), RT (2006), and Armstrong, any tariff with  $P^I(\widetilde{N}^I) = P^I(\widetilde{N}^I, \widetilde{N}^J)$  for both  $I$  has the pair  $(\widetilde{N}^A, \widetilde{N}^B)$  as an equilibrium. Thus none of my analysis, except a brief discussion of competition in Subsection VIC, assumes any particular tariff. Rather, this subsection is meant to justify my approach of ignoring the specifics of tariffs and coordination and to show, perhaps surprisingly, that adding optimization *simplifies* the analysis. Thus even a reader skeptical of the possibility of insulating tariffs but willing to focus, exogenously, on a given equilibrium, should accept my analysis in the monopoly case I focus on.

## II. Pricing

Industrial policy typically aims to alleviate the social harms caused by market power. The first step towards formulating such policy is therefore understanding the nature of those harms. Towards that goal, this section develops and compares the socially optimal and profit maximizing allocation rules, emphasizing the prices that support these allocations.

### A. Pigouvian Pricing

The value created by the platform is the benefits it brings to users less the costs of providing the service. RT (2006) assumes marginal costs constant in both participation rates, taking the other participation rate as given. Thus there may be two types of cost: membership costs  $C^I N^I$  and interaction costs  $cN^A N^B$ . The benefits the platform brings to users on side  $I$  are

$$(2) \quad V^I(N^I, N^J) = \int_{-\infty}^{\infty} \int_{P^I(N^I, N^J) - b^I N^J}^{\infty} [B^I + b^I N^J] f^I(B^I, b^I) dB^I db^I.$$

<sup>15</sup> However, I suspect that explicitly modeling why critical mass problems exist in a given application, ignored in previous work on this problem, would be crucial to understanding their welfare implications. For example, if imperfect information is the cause, platforms and social welfare might actually be harmed by attempts to “solve” the deliberately designed critical mass problem!

<sup>16</sup> An identical argument clearly applies to one-sided networks, or any coordination game. I thus believe that the importance of coordination has been exaggerated in situations when an optimizing agent with the ability to make transfers can regulate coordination. Even in the cases when it is relevant, I believe it is more a choice than a constraint. However, this is obviously a controversial view. See, for example, Joseph Farrell and Paul D. Klemperer (2007) for a well-argued contrasting view and the last paragraph of the paper for further discussion.

Thus the total social value of the platform is

$$(3) \quad V(N^A, N^B) = V^A(N^A, N^B) + V^B(N^B, N^A) - C^A N^A - C^B N^B - c N^A N^B.$$

A benevolent social planner equates marginal social benefits to their marginal social costs:

$$V_1^{\mathcal{I}} + V_2^{\mathcal{J}} = C^{\mathcal{I}} + cN^{\mathcal{J}}$$

where  $V_i^{\mathcal{I}}$  is the derivative of  $V^{\mathcal{I}}$  with respect to the  $i$ th argument.  $V_1^{\mathcal{I}} = P^{\mathcal{I}}$  as the user added on side  $\mathcal{I}$  must be marginal and therefore earn zero net surplus from participating.  $V_2^{\mathcal{J}}$  is the value an additional user on side  $\mathcal{I}$  brings to users on side  $\mathcal{J}$ :  $b^{\mathcal{J}}N^{\mathcal{J}}$ , where

$$\overline{b^{\mathcal{J}}} = \frac{\int_{-\infty}^{\infty} \int_{P^{\mathcal{J}}(N^{\mathcal{J}}, N^{\mathcal{I}}) - b^{\mathcal{J}}N^{\mathcal{I}}}^{\infty} b^{\mathcal{J}} f^{\mathcal{I}}(B^{\mathcal{J}}, b^{\mathcal{J}}) dB^{\mathcal{J}} db^{\mathcal{J}}}{\int_{-\infty}^{\infty} \int_{P^{\mathcal{J}}(N^{\mathcal{J}}, N^{\mathcal{I}}) - b^{\mathcal{J}}N^{\mathcal{I}}}^{\infty} f^{\mathcal{I}}(B^{\mathcal{J}}, b^{\mathcal{J}}) dB^{\mathcal{J}} db^{\mathcal{J}}}$$

is the *average interaction value of participating users* on side  $\mathcal{J}$ . Thus the optimal price is

$$(4) \quad P^{\mathcal{I}} = \underbrace{C^{\mathcal{I}} + cN^{\mathcal{J}}}_{\text{marginal private cost}} - \underbrace{\overline{b^{\mathcal{J}}N^{\mathcal{J}}}}_{\text{marginal external benefit}}$$

This is the standard Pigouvian<sup>17</sup> condition: the price of an activity should equal its private cost less any external benefits. This last term is the essential difference between optimal pricing in two-sided markets and standard multiproduct pricing: because network effects are external to individual decisions, price should diverge from cost. Thus positive network effects should be subsidized and negative ones taxed.

Newspapers offer a simple example. Optimal pricing calls for readers to be subsidized, below the cost of providing the news by the value they bring to advertisers, and for advertisers to be taxed, above the cost of printing their ads, by the amount readers dislike them.

In the Armstrong model, interaction values are homogeneous ( $b_i^{\mathcal{I}} \equiv b^{\mathcal{I}}$ ) and interaction costs are disallowed ( $c = 0$ ) so (4) becomes Alex Gaudeul and Bruno Jullien’s (2008) formula

$$P^{\mathcal{I}} = C^{\mathcal{I}} - b^{\mathcal{J}}N^{\mathcal{J}}.$$

RT (2003) rule out membership values/costs so user prices and surplus are all from interactions. Letting  $p^{\mathcal{I}} \equiv P^{\mathcal{I}}/N^{\mathcal{I}}$  be the *per-interaction price* and  $\overline{s^{\mathcal{I}}} \equiv (V^{\mathcal{I}}/N^{\mathcal{J}}) - p^{\mathcal{I}}$  the average per-interaction surplus on side  $\mathcal{I}$  gives Bedre-Defolie and Calvano’s (2010) and Weyl’s (2009b) optimal pricing rule

$$p^A + p^B - c = -\overline{s^A} = -\overline{s^B}.$$

I now compare this classical rule to that which a profit-maximizing monopolist would adopt.

<sup>17</sup> First-best pricing has traditionally been known in the literature as Lindahl pricing (Özlem Bedre-Defolie and Emilio Calvano 2010; Weyl 2009b). However, because price discrimination is ruled out in the RT (2006) model, pricing follows Pigou (1920) rather than Lindahl (1919).

B. Profit-Maximizing Pricing

Often the operators of platforms are concerned with their profits rather than with social welfare. Because price discrimination is typically imperfect, these differ. To make as clear as possible the distortions introduced by imperfect price discrimination it is useful to take them to their logical extreme, ruling out all discrimination.<sup>18</sup> Profits are then

$$(5) \quad \pi(N^A, N^B) = (P^A[N^A, N^B] - C^A)N^A + (P^B[N^B, N^A] - C^B)N^B - cN^AN^B.$$

A profit-maximizing monopolist equates marginal revenues of participation to marginal cost:

$$\underbrace{P^I + P_1^I N^I + P_2^J N^J}_{\text{marginal revenue}} = \underbrace{C^I + cN^J}_{\text{marginal cost}}$$

The first two terms of marginal revenue are classical: price minus the inverse hazard rate of demand (or *market power*)  $\mu^I \equiv -P_1^I N^I = P^I/\epsilon^I$  where  $\epsilon^I$  is the elasticity of demand. The final term is special to two-sided markets: it is the revenue that can be extracted from side  $J$  by adding an additional user on side  $I$ . Letting  $\tilde{b}^J$  be the *average interaction value of marginal users* (AIVMU) on side  $J$ , by the implicit function theorem and equation (1)

$$P_2^J = -\frac{N_2^J}{N_1^J} = \frac{\int_{-\infty}^{\infty} b^J f^J(P^J[N^J, N^I] - b^J N^I, b^J) db^J}{\int_{-\infty}^{\infty} f^J(P^J[N^J, N^I] - b^J N^I, b^J) db^J} \equiv \tilde{b}^J.$$

The platform can extract only the value *marginal* users on side  $J$  place on an additional side  $I$  user joining. This is an example of the general tendency, emphasized by Spence (1975) and discussed extensively below, of monopolists to serve the preferences of marginal, rather than all participating, users. The platform’s side  $J$  marginal revenue from a side  $I$  user is therefore  $\tilde{b}^J N^J$ . Privately optimal pricing follows a simple extension of Lerner’s formula<sup>19</sup>

$$(6) \quad \frac{P^I - (C^I + cN^J - \tilde{b}^J N^J)}{P^I} = \frac{1}{\epsilon^I}.$$

In the Armstrong case this immediately simplifies to Armstrong’s pricing condition

$$P^I = C^I - b^J N^J + \mu^I.$$

In RT (2003), only interaction benefits exist so  $\tilde{b}^J = p^J$ . Therefore the pricing condition is

$$p^A + p^B - c = m^A = m^B$$

where  $m^I \equiv \mu^I/N^I$ . This is the formula that RT (2003) derives.

<sup>18</sup> For an analysis of distortions that can arise even under perfect price discrimination and with a single group of homogeneous users, when there are externalities to nonparticipating consumers or other frictions, see Ilya Segal (1999).

<sup>19</sup> RT (2006) states the general condition for optimal two-product pricing in terms of derivatives of  $N^A$  and  $N^B$ , determined as fixed points of an equilibrium among users. However, as a function of the allocation  $(N^A, N^B)$ , profits are just the simple explicit function above. This is what allows me to express the first-order condition for optimal in terms of the primitive properties of preferences in two-sided markets.

Comparing private to socially optimal pricing,

$$(7) \quad P^{\mathcal{I}} = \underbrace{C^{\mathcal{I}} + cN^{\mathcal{J}} - \bar{b}^{\mathcal{J}}N^{\mathcal{J}}}_{\text{socially optimal price}} + \underbrace{\mu^{\mathcal{I}}}_{\text{classical market power distortion}} + \underbrace{(\bar{b}^{\mathcal{J}} - \widetilde{b}^{\mathcal{J}})N^{\mathcal{J}}}_{\text{Spence distortion}}$$

Thus there are two distortions in a two-sided market. First, classical marginal revenue lies below price by the amount of the market power  $\mu^{\mathcal{I}}$ . Second, if  $\bar{b}^{\mathcal{J}} \neq \widetilde{b}^{\mathcal{J}}$ , the average interaction values of marginal users differ from those of loyal users; the platform will either over- or undersubsidize (tax) users on side  $\mathcal{I}$ . Like the classical market power distortion, this *Spence distortion* is a consequence of the platform's inability to price discriminate. The platform internalizes network externalities but does so imperfectly (see Subsection IIIC).

The tendency to truckle to marginal users is familiar to anyone living in, or observant while visiting,<sup>20</sup> a tourist destination: the city government and businesses tend to cater to mobile tourists rather than to locked-in residents.<sup>21</sup> This Spence distortion is likely more important in two-sided markets than the contexts for which it was originally conceived. A platform is unlikely to partially ameliorate inefficiency (while introducing other distortions) by offering multiple products (Michael Mussa and Sherwin Rosen 1978; Mary O'Keeffe, W. Kip Viscusi, and Richard J. Zeckhauser 1984; David Besanko, Shabtai Donnenfeld, and Lawrence J. White 1987) as this would require inefficiently wasting potential interactions.<sup>22</sup> Once "quality" is provided to some users on one side of the market, it is free to provide to others.

The existence and sign of the Spence distortion depend crucially on the source of user heterogeneity.<sup>23</sup>  $\bar{b}^{\mathcal{I}}$  will tend to exceed  $\widetilde{b}^{\mathcal{I}}$  if users differ primarily in their interaction values. For example, in the extreme case of only interaction heterogeneity (RT 2003), the Spence distortion is exactly the per-interaction surplus on side  $\mathcal{I}$ , while when there is only membership heterogeneity (Armstrong), there is no Spence distortion. The Spence distortion may even be downward, as in the newspaper example above. If heterogeneity in willingness to pay for content dominates and is correlated with willingness to pay to avoid advertising, then loyal users dislike advertising more than marginals, and the Spence distortion is downwards.

<sup>20</sup> Of course in real life, as in the RT (2006) model, marginal users (tourists) are a heterogeneous bunch, and many, including the author, have preferences more similar to natives than to those of other tourists.

<sup>21</sup> Readers living in less-frequented climes may find a joke instructive. I tell a variant of a classic Israeli joke, first told to me by David Hariton, to whom I am grateful. In the original joke, Smith is replaced by David Ben-Gurion.

Adam Smith dies and, for his service to economics, is given a choice of where to spend eternity. He requests to see each option before deciding. First he is shown Hell, which, full of decadent food, French wines and beautiful women, seems a merry way to spend the rest of time. Heaven, on the other hand, is an unending stream of presentations of leading research in economics and philosophy. Having spent his life in contemplation on these topics, Smith decides he has earned a bit of relaxation in the afterlife and opts for Hell. Immediately he is thrown onto the rack, whipped, water-boarded, and subjected to other "enhanced" methods of entertainment. Astonished, he says, "I was just here a few minutes ago and things were so much nicer. What happened?" Lucifer replies, "Then you were a tourist."

<sup>22</sup> If the incentive for price discrimination is sufficiently large the platform might "throw away" quality. While such strategies are common in standard markets, in two-sided markets they seem to occur only when justified by other concerns outside this model, such as optimal matching (e.g., targeted ads). However, this is an important question for future research.

<sup>23</sup> Another, perhaps more general way to put this follows the language of Spence more closely. Spence argued quality would be undersupplied ( $P^{\mathcal{J}}$  distorted upward) when  $P_{12}^{\mathcal{I}} < 0$  and oversupplied ( $P^{\mathcal{J}}$  distorted downward) when  $P_{12}^{\mathcal{I}} > 0$ . If, as in the RT (2006) model, each user can participate at most once, the former is equivalent to users with high utility (lower reservation values) having high sensitivity to quality and users with low utility (high reservation values) being less so; the latter conversely. Note that  $P_{12}^{\mathcal{I}} = -\mu_b^{\mathcal{I}}/N^{\mathcal{I}}$ , the measure of local interaction heterogeneity I develop in Section IV. Thus there is a one-to-one correspondence between Spence's cross-partial of the price function and my focus on user heterogeneity.

Thus the harms of market power depend crucially on the source of heterogeneity. If, as is typically assumed, the costs of price distortions are convex, then market power over card accepters is particularly pernicious as it compounds the Spence distortion from cardholders. However, it may actually be beneficial that the *Times* has market power over advertisers, as this offsets the Spence distortion potentially leading to a better second-best level of advertising. Even with market power, therefore, it is possible ad rates may be too low.

C. Ramsey Pricing

Achieving first-best prices may be infeasible in practice as it would require subsidies whose granting, given the cost of raising public funds, political economy constraints, and imperfect information, would be more costly than the monopoly distortions they seek to address. When granting subsidies is infeasible, second-best pricing requires maximizing social welfare subject to some constraint, such as allowing the firm a rate of return (possibly 0) on its variable or fixed costs. Because of the externalities in two-sided markets, this Ramsey solution must be extended as proposed by Tae Hoon Oum and Michael M. Tretheway (1988) to take these into account.

I consider three formulations of the Ramsey problem, all of which are equivalent if the required level of profit is 0. First, in the text, I consider the classic Ramsey problem: social welfare is maximized subject to achieving a minimum absolute profit. In the Appendix, I consider a modified version of the Ramsey problem that RT (2003)<sup>24</sup> uses in a two-sided market where the rate of return is required on variable costs. As I argued in Weyl (2009b), there are two possible social objectives: maximizing user or social surplus subject to the rate-of-return constraint. The first approach addresses externalities more completely, while the second comes closer to the monopolist’s constrained goals.

**THEOREM 1:** *Interior Ramsey prices maximizing user or social surplus subject to the constraint that the platform makes a profit of at least K must solve*

$$(8) \quad \frac{\overbrace{P^I - (C^I + cN^J - \widetilde{b}^J N^J)}^{\text{private marginal opportunity cost}} - \overbrace{[1 - \lambda]}^{\text{Lagrangian weighting}} \overbrace{[\overline{b}^J - \widetilde{b}^J]}^{\text{Spence distortion}}}{P^I} = \lambda \frac{1}{\epsilon^I}$$

where

$$\lambda \equiv \frac{\overbrace{K}^{\text{target profit}} + \overbrace{(b^A + b^B - c)N^A N^B}^{\text{subsidy required for (local) Pigouvian prices}}}{\underbrace{N^A \mu^A + N^B \mu^B + (\overline{b}^A + \overline{b}^A - \widetilde{b}^A - \widetilde{b}^B)N^A N^B}_{\text{(local) profit gain moving to monopoly from Pigouvian prices}}}$$

<sup>24</sup> Rochet and Tirole use this modified Ramsey set-up to consider whether firms distort the “balance” of prices as separate from their level, a major focus of mine in Weyl (2009c).

## PROOF:

See the Ramsey Pricing portion of the Appendix.

Thus the Ramsey pricing condition is just a simple weighted average of the Pigouvian and profit-maximizing prices. These, again, diverge in their attention to both the Spence and classical market power distortions. Prices are closer to profit maximization i) the higher is the target profit, ii) the larger is the subsidy called for by Pigouvian prices and iii) the further one must move towards monopoly to achieve a given gain in net profits. Just as first-best prices take a classic Pigouvian form, Ramsey prices take OT's Pigou-Ramsey form.

### III. Generalization

The primary aim of this paper is to understand the price theory of and proper policy towards industries such as payment cards and newspapers. After a brief interlude in this section, I continue toward this goal in Section IV, to which a casual reader may wish to skip directly. However, the general character of my basic ideas thus far suggests they may help analyze a broader class of models than that RT (2006) specifically adapted to those industries. In fact with any number of groups of users and essentially arbitrary heterogeneous preferences, the same principles developed above apply. Insulating tariffs exist, allowing a simple analysis of the platform's choice of allocation showing in general that the Spence distortion is the key element added by network externalities. This section considers such a generalization.

I maintain four important assumptions of the RT (2006) model:

- i) (Quasi-linear) user preferences are taken as exogenous (RT 2006 assumption 1).
- ii) All groups of users can be explicitly (third-degree) price discriminated and all users within each group differ only in their preferences.<sup>25</sup>
- iii) No price discrimination is possible, but prices to any given group can take any positive or negative value. Users interact with an exogenous collection of other users (in their own and other groups); any marginal price for such interactions is exogenous to the model and enters only to the extent that it determines preferences.
- iv) Externalities are only to participating users.<sup>26</sup>

#### A. The Model

There are  $M$  groups  $\mathcal{I} = \mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$  and users may value participation by members not only of other groups, but of their own. A typical user  $i$  on side  $\mathcal{I}$  is characterized by a vector  $\theta_i^{\mathcal{I}}$  of characteristics drawn according to a smooth and massless distribution with probability density function

<sup>25</sup> Andre Veiga and Weyl (2010) have made significant progress in relating this assumption.

<sup>26</sup> Unlike the others, this assumption can easily be dispensed with. This generalizes Segal's (1999) basic model of contracting with externalities to allow asymmetric information (on reservation values) and asymmetric agents ("sides of the market"). However, the increase in notational complexity and distance from a realistic model of network industries (most nonparticipant externalities arise in contracting, rather than uniform pricing, settings) led me to this assumption. The intuition of that model should be clear from Segal and the general model here: the social planner internalizes all externalities, while a profit-maximizer internalizes the *reverse sign* of externalities to *marginal* nonparticipating consumers, scaled by the number of participating users. Effectively, to a profit maximizer, negative out-group externalities are equivalent to positive in-group externalities, while to a social planner they are opposite. Details are available on request. This extension and the more general connections between the theory of multi-sided platforms (network industries) and contracting with externalities are promising areas for future research.

$f^{\mathcal{I}}$  with full support on  $R^{K^{\mathcal{I}}}$  where  $K^{\mathcal{I}} \in N$ . Let  $\mathbf{N} \equiv (N^A, N^B, N^C, \dots)$  be an *allocation*, a vector of participation rates on each side.<sup>27</sup> The utility of user  $i$  on side  $\mathcal{I}$  from participating is

$$U_i^{\mathcal{I}} = u^{\mathcal{I}}(\mathbf{N}; \theta_i^{\mathcal{I}}) - P^{\mathcal{I}}$$

where  $P^{\mathcal{I}}$  is the price a user on side  $\mathcal{I}$  must pay to participate. I assume that  $u^{\mathcal{I}}$  is smooth in and finite for all values of the allocation and characteristics.

Note that this general model has a few special cases of particular interest:

- i)  $M = 1$  is a one-sided network with arbitrary utility and heterogeneity. I do not believe this model has ever been considered, but given the substantial interest in one-sided network monopolists (Economides 1996) it seems a natural general model.
- ii)  $M = 2$  is RT (2006) with arbitrary heterogeneous utilities and within-side network effects.
- iii) Suppose  $M$  sides can be split into two groups  $\mathbf{A}$  and  $\mathbf{B}$  such that  $u^{\mathcal{I}}$  is independent of  $N^{\mathcal{J}}$  whenever either  $\mathcal{I}, \mathcal{J} \in \mathbf{A}$  or  $\mathcal{I}, \mathcal{J} \in \mathbf{B}$ . This is (ii) without within-side effects but with groups of discriminable, heterogeneously valuable users on each side.

For a particular allocation  $\mathbf{N}$  and price  $P^{\mathcal{I}}$  the set of  $\mathcal{I}$  users weakly benefiting from participating is  $\overline{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}) \equiv \{\theta^{\mathcal{I}} : u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) \geq P^{\mathcal{I}}\}$  and the set of all marginal users is  $\widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}) \equiv \{\theta^{\mathcal{I}} : u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) = P^{\mathcal{I}}\}$ . Then the fraction of users interested in participating on side  $\mathcal{I}$  given an allocation  $\mathbf{N}$  and a price  $P^{\mathcal{I}}$  is

$$\widetilde{N}^{\mathcal{I}}(P^{\mathcal{I}}, \mathbf{N}) = \int_{\overline{\Theta}^{\mathcal{I}}(P^{\mathcal{I}}, \mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}.$$

Because the set  $\overline{\Theta}^{\mathcal{I}}$  is clearly contracting in  $P^{\mathcal{I}}$ ,  $\widetilde{N}^{\mathcal{I}} < 0$  and<sup>28</sup>  $\widetilde{N}^{\mathcal{I}}$  can be inverted to yield  $P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, \mathbf{N})$ , the price needed to attract  $\widetilde{N}^{\mathcal{I}}$  of users who anticipate allocation  $\mathbf{N}$ .

Note that the potential multiplicity problem here is far worse than in the two-sided case, as utility functions have arbitrary structure, and there can be an arbitrary number of sides. However, this enormous coordination problem can again be avoided by careful tariffs. In particular the platform may charge an insulating tariff, which is here a price to side  $\mathcal{I}$  depending on the full realized allocation that ensures the desired allocation is realized. Formally the insulating tariff for a desired participation rate  $\widetilde{N}^{\mathcal{I}}$  is  $P^{\mathcal{I}}(\mathbf{N}) \equiv P^{\mathcal{I}}(\widetilde{N}^{\mathcal{I}}, \mathbf{N})$ . As with RT (2006), if the platform charges the insulating tariff associated with its desired allocation on all sides, then the unique equilibrium is its desired allocation.<sup>29</sup> Thus once again the platform’s problem can be viewed as

<sup>27</sup> In particular, I assume every utility level is achieved by same type, given any  $\mathbf{N}$ .

<sup>28</sup> When participation is positive, but not total, from my assumption of smooth  $f^{\mathcal{I}}$ s and full support.

<sup>29</sup> Here, again, insulating *every side from every side* can be dispensed with. Imagine drawing a graph where each node represents a side of the market and a directed edge is drawn between each side and those sides whose participation affects their utility, but against whose participation they are not insulated. I would conjecture, but have only the sketch of a proof that, so long as this graph is acyclic there is a unique equilibrium. Intuitively if the graph is acyclic, one can trace back from its sinks to tie down the unique participation rate on each side. Furthermore other tariffs than the insulating tariff may do the trick for particular (distributions of) user preferences. However, I believe that the “simplest” approach to “robustly” ensuring uniqueness is fully insulating every side of the market from every other side. A formal analysis of all this will likely appear in joint work in progress with Alex White, as also referred to in footnote 53.

one of choosing an allocation  $\tilde{\mathbf{N}}$  to maximize some objective, eliminating the need to consider derivatives of complex, multi-sided fixed points.

B. Pricing

Let  $P^{\mathcal{I}}(\mathbf{N}) = P^{\mathcal{I}}(N^{\mathcal{I}}, \mathbf{N})$ , where  $N^{\mathcal{I}}$  is the  $\mathcal{I}$ th entry of  $\mathbf{N}$ ,  $\overline{\Theta}^{\mathcal{I}}(\mathbf{N}) \equiv \overline{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}[\mathbf{N}])$  and  $\widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}) \equiv \widetilde{\Theta}^{\mathcal{I}}(\mathbf{N}, P^{\mathcal{I}}[\mathbf{N}])$ . The gross value created on side  $\mathcal{I}$  by an allocation  $\mathbf{N}$  is simply

$$V^{\mathcal{I}}(\mathbf{N}) = \int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}.$$

I allow for arbitrary smooth, positive cost functions  $C(\mathbf{N})$ . Thus the (net) surplus created by the service as a function of the allocation is

$$V(\mathbf{N}) = \sum_{\mathcal{I}} V^{\mathcal{I}}(\mathbf{N}) - C(\mathbf{N}).$$

Maximizing the surplus created by the service requires equating marginal social value to marginal cost. Let  $X_{\mathcal{J}} \equiv \partial X / \partial N^{\mathcal{J}}$ . A socially optimal allocation then requires that for each  $\mathcal{I}$

$$\sum_{\mathcal{J}} V_{\mathcal{I}}^{\mathcal{J}} = C_{\mathcal{I}}.$$

The following theorem states that these conditions can again be written in a Pigouvian form.

**THEOREM 2:** *The first-order conditions for a socially optimal allocation are*

$$(9) \quad P^{\mathcal{I}} = \underbrace{C_{\mathcal{I}}}_{\text{marginal cost}} - \underbrace{\sum_{\mathcal{J}} \overline{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}}}_{\text{marginal externalities}}$$

where the average marginal interaction value of participating side  $\mathcal{I}$  users for side  $\mathcal{J}$  users is

$$\overline{u}_{\mathcal{J}}^{\mathcal{I}}(\mathbf{N}) \equiv \frac{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u_{\mathcal{J}}^{\mathcal{I}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}.$$

**PROOF:**

See the Generalization portion of the Appendix.

Thus the Pigouvian formula (4) extends in the most natural way possible: interaction values are replaced by the *marginal* value of users who have potentially nonlinear utility and all externalities, to those within side  $\mathcal{I}$  and on other sides  $\mathcal{J} \neq \mathcal{I}$ , are included.

On the other hand revenues on side  $\mathcal{I}$  are  $R^{\mathcal{I}}(\mathbf{N}) = P^{\mathcal{I}}(\mathbf{N})N^{\mathcal{I}}$  and profits

$$\pi(\mathbf{N}) = \sum_{\mathcal{I}} R^{\mathcal{I}}(\mathbf{N}) - C(\mathbf{N}).$$

Profit maximization requires equating marginal revenue of an additional side  $\mathcal{I}$  user, from all sides of the market, to the marginal cost of serving that user:

$$\sum_{\mathcal{J}} R_{\mathcal{I}}^{\mathcal{J}} = C_{\mathcal{I}}.$$

This yields a similarly intuitive extension of the RT (2006) profit maximizing pricing.

**THEOREM 3:** *The first-order conditions for a profit maximizing allocation are*

$$(10) \quad \frac{P^{\mathcal{I}} - (C_{\mathcal{I}} - \sum_{\mathcal{J}} \widetilde{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}})}{P^{\mathcal{I}}} = \frac{1}{\epsilon^{\mathcal{I}}}$$

or equivalently

$$(11) \quad P^{\mathcal{I}} = \underbrace{C_{\mathcal{I}} - \sum_{\mathcal{J}} \overline{u}_{\mathcal{I}}^{\mathcal{J}} N^{\mathcal{J}}}_{\text{socially optimal price}} + \underbrace{\mu^{\mathcal{I}}}_{\text{classical market power distortion}} + \underbrace{\sum_{\mathcal{J}} (\overline{u}_{\mathcal{I}}^{\mathcal{J}} - \widetilde{u}_{\mathcal{I}}^{\mathcal{J}}) N^{\mathcal{J}}}_{\text{Spence distortion(s)}}$$

where the average marginal interaction value of marginal side  $\mathcal{I}$  users for side  $\mathcal{J}$  users is

$$\widetilde{u}_{\mathcal{I}}^{\mathcal{J}}(\mathbf{N}) \equiv \frac{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} u_{\mathcal{I}}^{\mathcal{J}}(\mathbf{N}; \theta^{\mathcal{I}}) f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}{\int_{\theta^{\mathcal{I}} \in \overline{\Theta}^{\mathcal{I}}(\mathbf{N})} f^{\mathcal{I}}(\theta^{\mathcal{I}}) d\theta^{\mathcal{I}}}.$$

**PROOF:**

See the Appendix.

Thus again profit maximization distorts the allocation in two ways. First it raises prices (lowers participation) as classical marginal revenue falls below price. Second it imperfectly internalizes network externalities, as preferences of marginal rather than all participating users determine the marginal revenues generated by an additional side  $\mathcal{I}$  user. Thus there are now  $M$  classical distortions and  $M^2$  Spence distortions.

### C. Discussion

Conceptually little has changed from the RT (2006) model. Insulating tariffs exist and the platform can therefore achieve any desired allocation. The platform thus maximizes its objectives over possible allocations, making its problem simple. Profit maximization leads to classical and Spence distortions. The nature of these latter distortions depends on how the preferences of loyal and marginal users diverge, that is, on the source of user heterogeneity.

This suggests three interesting conclusions. First, while most of this paper focuses on affine user preferences, this is without significant loss of generality. While affine preferences allow only two dimensions of heterogeneity, these are two dimensions which generally matter. This extends even to my comparative statics analysis below, as none of the terms governing these include the curvature of utility (none involve  $V_{\mathcal{J}\mathcal{J}}^{\mathcal{I}}$ ). Of course the irrelevance of further dimensions of heterogeneity depends crucially on the impossibility of price discrimination. If user utility is not affine, platforms may use a marginal price, such as differential charges for viewing certain Web sites, to discriminate among users. In this case social value and profits depend not only on

participation rates, but also on marginal prices. This important and largely open<sup>30</sup> problem is well beyond the scope of this paper.

Second, it provides a simple and general strategy for analyzing monopoly networks: the allocation approach. While my results here constitute only the most superficial of first passes, having no comparative static or policy analysis, they suggest a path for future research.

Finally, it answers perhaps the oldest open question in network economics: the general validity of the (Liebowitz and Margolis 1994) conjecture that optimizing networks will internalize (and thereby neutralize) network externalities. Liebowitz and Margolis were partially correct, but only up to a point. While platforms do internalize externalities, they do so imperfectly as they take into account only the preferences of marginal users. This Spence distortion will be smallest, and therefore network externalities most nearly neutralized, when loyal and marginal users place a similar value on the participation of a marginal network user. In this case, the only distortions caused by market power are the classic, familiar ones of any multiproduct monopolist. In this case actions of users influence the welfare of other users only indirectly, through platform optimization (see Subsection IVB). On other hand, when loyal and marginal users have sharply different preferences, network monopolies have much more complex distortions with large direct network externalities persisting in equilibrium.

#### IV. Comparative Statics

A primary motivation for the theory of two-sided markets is that conditions on each side affect participation and welfare on the other. As with any comparative statics exercise, understanding these indirect cross-effects relies on the second-order conditions for optimization (Paul A. Samuelson 1941) and specifically, because of the multiproduct (Weyl 2009a) monopoly (Weyl and Michal Fabinger 2009) context, on pass-through rates and the cross partial of the allocation in profits. I begin by formally developing these closely related concepts.

The *pass-through rate* on side  $\mathcal{I}$ ,

$$\rho^{\mathcal{I}} \equiv \left. \frac{dP^{\mathcal{I}}}{dC^{\mathcal{I}}} \right|_{N^{\mathcal{J}}} = - \frac{\mu^{\mathcal{I}}}{N^{\mathcal{I}} \frac{\partial^2 \pi}{\partial N^{\mathcal{I}^2}}}$$

is the amount a private platform finds it optimal to increase  $P^{\mathcal{I}}$  in response to an increase in  $C^{\mathcal{I}}$  if  $N^{\mathcal{J}}$  is held fixed. The cross partial

$$\chi \equiv \frac{\partial^2 \pi}{\partial N^{\mathcal{A}} \partial N^{\mathcal{B}}}$$

measures the complementarity/substitutability (if positive/negative) of participation rates.

For traditional comparative static analysis, it is important that the first-order conditions used actually represent the optimal allocation for the platform. To ensure this, a convenient assumption is that the platform's profit function is concave. However, it is important to avoid overly restrictive assumptions that are sufficient, but unnecessary, for the purpose as these can bias analysis; log-concavity is a typical such assumption. To add tractability without undue restrictions, I

<sup>30</sup> See Bedre-Defolie and Calvano (2008) for a first pass, in the context of the RT (2003) model.

propose a “weak” (in a sense formalized by Theorem 4) second-order condition. As far as I know this *two-sided contraction* (2SC) is the first second-order condition to be proposed for the general RT (2006) model.

If for all  $(N^A, N^B) \in (0, 1)^2, \rho^A, \rho^B > (\geq) 0$  and  $(\mu^A \mu^B / N^A N^B) > (\geq) \rho^A \rho^B \chi^2$ , I will say that  $f$  satisfies strict (weak) *two-sided contraction* (2SC) given interaction cost  $c$ .

**THEOREM 4:** *If  $f^A, f^B$ , and  $c$  exhibit strict 2SC then for any  $C^A$  and  $C^B$  a solution to equation (6) for both  $I$  is the unique platform’s optimal price. If they violate weak 2SC then there exists a pair  $(C^A, C^B)$  for which there is a solution to equation (6) which is not an optimum.*

**PROOF:**

See Appendix.

In the RT (2003) case,<sup>31</sup>  $\chi = \mu^I / N^J$  for both  $I$  so the condition becomes my (Weyl 2009c) “cross-subsidy contraction” condition  $\rho^A \rho^B < 1$ . The Comparative Statics portion of the Appendix gives Pigouvian SOCs. These could be extended to the general model of Section III by deriving conditions for the Hessian matrix of cross partials of profits with respect to the allocation to be negative definite.

### A. Complements versus Substitutes

The most famous, supposedly robust result on the comparative statics of two-sided markets is what RT (2006) calls the “simple ‘seesaw principle’: a factor that is conducive to a high price on one side, to the extent that it raises the platform’s margin on that side, tends also to call for a low price on the other side as attracting members on that other side becomes more profitable.” While intuitive, this result faces two challenges. First, the appropriate notion of “price” is unclear. In the only (RT 2003) model where the seesaw principle has been demonstrated formally (RT 2003; Weyl 2009c), the price is per interaction. In other models this price has little special significance. However, as discussed above, holding fixed the number of users on side  $\mathcal{I}$  the price (in any sense) on side  $\mathcal{J}$  is decreasing in the number of users on side  $\mathcal{J}$ . Therefore RT (2006)’s seesaw principle can be reformulated as follows: factors leading the platform to choose higher  $N^I$  lead it to choose lower  $N^J$ . That is  $\partial^2 \pi / \partial N^A \partial N^B < 0$  or participation on the two sides are substitutes *for the platform*.<sup>32</sup> In RT (2003), this holds, and the two formulations are equivalent. However, this formulation can be examined beyond the context of the RT (2003) model.

The more serious challenge to the seesaw principle is that this broader formulation is not generally true but instead depends on the source of user heterogeneity.

To see this formally, it is useful to construct a general measure of the local importance of the two dimensions of heterogeneity. A natural such measure is how interaction and membership benefits of marginal users increase with price. Price is, by definition, always equal to the total value of marginal users. It is therefore natural to decompose increases in price into changes in interaction and membership values. From Subsection IIB  $P_1^I = -\mu^I / N^I$ ; but the total gross utility of a marginal user is  $\widetilde{B}_1^I + \widetilde{b}_1^I N^J$  so

$$\widetilde{B}_1^I + \widetilde{b}_1^I N^J = -\frac{\mu^I}{N^I} .$$

<sup>31</sup> See below. Also see the online Appendix for the Armstrong special case.

<sup>32</sup> Note that the “demand system” does not necessarily exhibit either complements or substitutes: Slutsky symmetry is not obeyed ( $\widetilde{b}_1^I = P_2^I \neq P_2^J = \widetilde{b}_1^J$ ) and may even be violated in signs, despite quasi-linearity, because of the externalities between the sides.

We can therefore define natural measures of local heterogeneity along the two dimensions as the projection of market power onto each dimension.  $\mu_{\tilde{b}^{\mathcal{I}}} \equiv -\tilde{b}_1^{\mathcal{I}} N^{\mathcal{I}}$  is the *membership market power* and  $\mu_{\tilde{b}}^{\mathcal{I}} \equiv -\tilde{b}_1^{\mathcal{I}} N^{\mathcal{A}} N^{\mathcal{B}}$  is the *interaction market power*. The cross partial

$$(12) \quad \chi = \tilde{b}^{\mathcal{A}} + \tilde{b}^{\mathcal{B}} - c - \frac{\mu_{\tilde{b}}^{\mathcal{A}}}{N^{\mathcal{B}}} - \frac{\mu_{\tilde{b}}^{\mathcal{B}}}{N^{\mathcal{A}}}$$

as the effect of side  $\mathcal{J}$  participation on  $\mathcal{I}$  marginal revenue is the difference between its effect on price  $P^{\mathcal{I}}$ ,  $\tilde{b}^{\mathcal{I}}$ , and market power  $\mu^{\mathcal{I}}$ , as shown formally in the proof of Theorem 5 below. Intuitively, interaction benefits favor complementarity: the value of a side  $\mathcal{A}$  user is proportional to the number of users she interacts with on side  $\mathcal{B}$ . Thus an increase in side  $\mathcal{B}$  users makes it more attractive to recruit side  $\mathcal{A}$  users. Offsetting this is the fact that when interaction benefits are the main dimension of heterogeneity, increasing the participation on side  $\mathcal{B}$  requires recruiting low interaction benefit users. Thus increased side  $\mathcal{B}$  participation reduces the AIVMU, eroding the cross-subsidy to, and therefore participation by, side  $\mathcal{A}$ .

Thus the sign of the cross partial is determined by how the surplus created by marginal interaction benefits compares to their heterogeneity. Perhaps the sharpest way to express this is in terms of the relative importance of interaction benefits in profits compared to their relative importance in heterogeneity. Platform profits are just the sum of (twice) marginal interaction surplus  $\tilde{b} N^{\mathcal{A}} N^{\mathcal{B}} \equiv (\tilde{b}^{\mathcal{A}} + \tilde{b}^{\mathcal{B}} - c) N^{\mathcal{A}} N^{\mathcal{B}}$  and marginal membership surplus,  $\sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} (\tilde{b}^{\mathcal{I}} - C^{\mathcal{I}}) N^{\mathcal{I}}$ . It is therefore natural to consider the fraction of profits arising from marginal interaction surplus, the *interaction surplus ratio*  $\alpha \equiv \tilde{b} N^{\mathcal{A}} N^{\mathcal{B}} / \pi$ . Similarly the *interaction heterogeneity ratio*  $\beta \equiv \mu_{\tilde{b}} / \mu_{\tilde{b}^{\mathcal{I}}}$ , where  $\mu_{\tilde{b}} = \sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} N^{\mathcal{I}} \mu_{\tilde{b}}^{\mathcal{I}}$  and  $\mu_{\tilde{b}^{\mathcal{I}}} = \sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} N^{\mathcal{I}} \mu_{\tilde{b}^{\mathcal{I}}}^{\mathcal{I}}$ , measures the relative *aggregate* importance of interaction heterogeneity.

**THEOREM 5:** *Participation on the two sides of the market are complements if  $\mu_{\tilde{b}} > 0$  and  $\alpha > \beta$ , substitutes if either  $\mu_{\tilde{b}} \leq 0$  or  $\beta > \alpha$ , and independent if  $\mu_{\tilde{b}} > 0$  and  $\alpha = \beta$ .*

**PROOF:**

See the Scale-Income Model portion of the Appendix.

Thus user heterogeneity ties the Spence distortion to the cross partial of participation rates. Because observing the cross partial requires only marginal shocks to market conditions, it may be easier to study empirically than the Spence distortion directly. Therefore one might measure basic features of user heterogeneity by the sign of the cross-participation effect, for example by observing the effect of a shock to one membership cost. Sadly, this is a coarse instrument, unable to distinguish which side of the market generates the interaction heterogeneity nor anything beyond its magnitude relative to the interaction surplus ratio. Measuring these finer properties requires richer data<sup>33</sup> or stronger assumptions.

The theorem makes clear the source of the seesaw effect in the RT (2003) model: there is no membership heterogeneity, so  $\beta = \infty$ , implying substitutes. By contrast, in Armstrong’s model  $\beta = 0$  as there is no interaction heterogeneity, and  $\alpha > 0$  as otherwise the firm would end the

<sup>33</sup> A companion paper under preparation, Weyl (2009a), treats identification in multiproduct monopoly, with a focus on two-sided markets. I show that first-order instruments for participation rates reveal elasticities and the AIVMU, while quantitatively observable cost shocks reveal pass-through rates and the cross partial. Some tests of general multiproduct monopoly are also possible, but many of the finer normative features, and tests of the RT (2006) model specifically, require stronger assumptions or higher-order variation.

two-sidedness, separately providing services to the two sides. Thus the Armstrong model always has complements, showing that the seesaw principle is far from general.

### B. Welfare Effects

In Section I, I argue that cross-group externalities *in the absence of transfers* are a defining feature of two-sided markets. However, others take the view (Hagiu 2007; Hagiu 2009; Rysman 2009) that two-sided markets are ones where, holding fixed some notion of price, each group's welfare depends on the other's participation and thereby *indirectly* (Jeffrey Church and Neil Gandal 1992; Michael Katz and Carl Shapiro 1994) on its own.

Such views are difficult to parse in multi-sided network models because the welfare-irrelevant details of pricing in these models still lead to very different indirect network effects holding fixed prices. For example, if insulating tariffs are charged to both sides then, by construction, such indirect network effects *never* exist. Thus, unless these authors think the canonical models miss the "essential nature" of two-sided markets, which I view as largely being defined by these models, it is difficult to see how such a test can be applied.<sup>34</sup>

Perhaps a more reasonable line of inquiry is therefore the nature of network effects in equilibrium. Suppose that participation on one side of a market rises for a reason, such as membership costs falling or membership values uniformly shifting up, that has no direct effect on the platform's incentives on the other side. I call the effect of such an exogenous increase in side  $\mathcal{A}$  participation on side  $\mathcal{B}$  welfare the equilibrium network effect.

**THEOREM 6:** *The equilibrium network effect from side  $\mathcal{J}$  to side  $\mathcal{I}$  has the same sign as*

$$(13) \quad \overline{b}^{\mathcal{I}} - \widetilde{b}^{\mathcal{I}} + \rho^{\mathcal{I}}\chi.$$

**PROOF:**

See the Scale-Income Model portion of the Appendix.

The first term of expression (13) is the direct effect of  $\mathcal{J}$  participation on  $\mathcal{I}$  welfare: the Spence distortion from  $\mathcal{I}$  to  $\mathcal{J}$ . Only the distortion matters:  $\widetilde{b}^{\mathcal{I}}$  is internalized by the platform as a higher price to side  $\mathcal{I}$ . One might view this direct effect as the equilibrium network *externality*. The second term is an indirect effect through the platform's optimization: the pass-through of the cross-side pricing effect. For example, if  $\chi < 0$  (participation rates are substitutes as in RT 2003), side  $\mathcal{I}$  will tend to benefit from an increase in prices on side  $\mathcal{J}$  as this will incentivize the firm to obtain greater participation in side  $\mathcal{I}$  by reducing prices.

Interaction heterogeneity both enlarges the Spence distortion and makes  $\chi$  negative, while membership heterogeneity eliminates it or even reverses its sign but makes  $\chi$  positive. Thus the source of heterogeneity has an ambiguous effect on expression (externalities). However, the first effect is fundamentally inframarginal, while the second depends only on local properties.

For example, in Armstrong's model, which has no Spence distortion, complementarity implies positive equilibrium network effects.<sup>35</sup> In the RT (2003) model, as discussed in Section II, the Spence distortion from side  $\mathcal{I}$  is  $\overline{s}^{\mathcal{I}}$ , the user surplus on side  $\mathcal{I}$ , and the cross partial can be shown

<sup>34</sup> Liebowitz and Margolis (1994) discuss the dangers of abusing the concept of network externalities.

<sup>35</sup> On the other hand, if interaction surplus is negative and participation rates are substitutes (I do not know of any simple example of this), equilibrium network effects are negative. These conditions do not have any consistent relationship to the *primitive* externalities, the level of interaction values on the two sides.

to be the negative of per-interaction market power on either side of the market  $m^{\mathcal{I}}$ . Furthermore in Weyl and Fabinger (2009) we show that  $\bar{s} = \bar{\rho}m$  where  $\bar{\rho}$  is an average of pass-through rates over prices above the equilibrium level, as pass-through measures the log-curvature of demand. Expression (13) therefore becomes, in the RT (2003) case,

$$-m^{\mathcal{I}}(\bar{\rho}^{\mathcal{I}} - \rho^{\mathcal{I}}),$$

whose sign is determined by the slope of  $\rho^{\mathcal{I}}$  with respect to cost/price: increasing pass-through implies average pass-through exceeds local pass-through, decreasing pass-through the reverse. Thus the *third* derivative of log-demand determines equilibrium network effects.

It may seem immediate that an increase in costs on side  $\mathcal{I}$  harms side  $\mathcal{I}$  users, but in Weyl (2009a) I showed that in the RT (2003) model the average user on one side of the market may actually want her prices increased to encourage a reduction in prices to users on the other side. The following corollary provides general conditions for this counterintuitive result.

**COROLLARY 7:**  $dV^{\mathcal{I}}/dC^{\mathcal{I}}$  has the sign of

$$(14) \quad -\left(\frac{\mu^A \mu^B}{N^A N^B} + \chi \rho^{\mathcal{J}} \left[ \bar{b}^{\mathcal{I}} - \tilde{b}^{\mathcal{I}} \right] \right).$$

Thus the counterintuitive effect can occur at either extreme of heterogeneity. When interaction heterogeneity dominates,  $\chi$  is negative while interaction surplus is strongly positive, so average users on side  $\mathcal{I}$  may benefit from higher prices which encourage the platform to bring in more customers on side  $\mathcal{J}$ . For example in the RT (2003) model, expression (14) becomes  $\bar{\rho}^{\mathcal{I}} \rho^{\mathcal{J}} - 1$ ; second-order conditions require  $\rho^{\mathcal{I}} \rho^{\mathcal{J}} < 1$ , as shown above, so beneficial own-cost increases require a  $\rho^{\mathcal{I}}$  increasing rapidly in price, yielding a large Spence distortion.

On the other hand, when membership heterogeneity is strong enough to give negative interaction heterogeneity, interaction surplus is negative but  $\chi > 0$  and average newspaper readers<sup>36</sup> may actually want higher prices to force firms to internalize their distaste for advertising and reduce its quantity. In intermediate cases, such as the Armstrong model, with small Spence distortions, own-cost effects are always negative.

### C. Arbitrary Comparative Statics

Effects of local shocks to the market can always be expressed as a combination of direct externalities<sup>37</sup> and indirect effects through optimally chosen participation rates. The former can be analyzed through a partial derivative holding fixed firm actions; the second is equivalent to some combination of changes in the firm's (opportunity) cost on each side of the market. Therefore knowing  $dN^{\mathcal{I}}/dC^{\mathcal{J}}, N_1^{\mathcal{I}}, N_2^{\mathcal{I}}, V_1^{\mathcal{I}}, V_2^{\mathcal{I}}$  for both  $I, J = A, B$  is sufficient to compute arbitrary comparative statics; expressions for these are given in the text and Appendix. The same approach may be taken in the more general model proposed in Section III, though explicit expressions of the relevant derivatives do not appear in this paper.

<sup>36</sup> This makes clear that all the reasoning about surplus is about the *total* user surplus on each side of the market: it integrates over all users. Clearly marginal or near-marginal users are harmed by any increase in prices, even if these benefit loyal users. In some settings we may care about such distributional consequences (is ritzy readers' distaste for advertising reason enough to exclude poorer marginal readers?), but that analysis is beyond the scope of this paper and in fact most standard industrial organization.

<sup>37</sup> An earlier draft of this paper, available on request, provided a variety of such explicit comparative statics.

## V. An Example: the Scale-Income Model

A primary contribution of this paper is to simplify the platform's problem to analyze, for the first time, the effect of multiple dimensions of user heterogeneity. The ability to analyze these more general models does not, however, eliminate all motivation for unidimensional models. As shown above, multidimensional heterogeneity leaves substantial ambiguities about the direction of various distortions and comparative statics. In cases when most heterogeneity plausibly lies along a single dimension, making this assumption explicit can help resolve these. Furthermore, from an empirical perspective it may be difficult to identify a two-dimensional model without parametric assumptions; restricting heterogeneity to a single dimension may be a simple and transparent way to impose the necessary additional structure.

Unfortunately, the source of heterogeneity in the most commonly applied model (Rysman 2004; Kaiser and Julian Wright 2006; Elena Argentesi and Lapo Filistrucchi 2007), Armstrong's, seems implausible in most settings where it is applied. A primary dimension of heterogeneity for at least one side of the market is almost certainly the value derived from the other side. The RT (2003) model focuses on this source of heterogeneity but has the unfortunate feature that it rules out any membership costs or benefits, making it implausible in many industries. However, my foregoing analysis emphasized that most results in the RT (2003) model are due to the source of heterogeneity rather than the absence of membership costs and benefits. Thus most of the results of the RT (2003) model extend to a Generalized RT (2003) (henceforth GRT 2003) model that allows for (homogeneous) membership costs and benefits.<sup>38</sup>

However, the GRT (2003) model still seems to fit many markets poorly. Newspaper readers and software producers, to name a few, clearly differ substantially in their membership benefits and costs, respectively, of participating in a platform. One reasonable model<sup>39</sup> of such settings (Anderson and Coate 2005) has GRT (2003)-like users on one side and Armstrong-like users on the other. In this section I propose an alternative that I think is likely to be most fruitful in applications: *Scale-Income* (SI) model. It offers a useful rule of thumb for thinking about sources of heterogeneity, making analysis a bit more concrete.

Users on each side agree on the relative size of membership and interaction values but differ in scale. All newspaper readers (side  $\mathcal{A}$ ) lose a fraction  $-\beta^{\mathcal{A}}N^{\mathcal{B}}$  of the value they take from reading if a fraction  $N^{\mathcal{B}}$  of advertisers participate; however, they may differ in their total utility. Intuitively, higher income users have greater willingness to pay to gain the utility of reading the newspaper and avoid the disutility of advertising. Advertisers have the same value of circulation as a fraction of the fixed cost  $-B_i^{\mathcal{B}}$  they expend to establish a relationship with the newspaper, but differ in the scale of both of these depending on their business size. Thus  $b_i^{\mathcal{I}}/B_i^{\mathcal{I}} = \beta^{\mathcal{I}}$  for all  $i, \mathcal{I}$ , but users differ in the scale of their utility. They are heterogeneous vertically (Spence 1976; Eytan Sheshinski 1976) rather than horizontally (Harold Hotelling 1929).

I believe this model provides a better approximation to many two-sided markets than any of the other unidimensional models.<sup>40</sup> It seems to me a fairly good fit to software platforms<sup>41</sup> (operating

<sup>38</sup> This model was analyzed extensively in a previous draft of this paper and, while omitted here for brevity, is available on request.

<sup>39</sup> This *Hybrid* model was extensively analyzed in a previous draft of this paper, available on request.

<sup>40</sup> Note that the RT (2003) model is the special case of the SI model where  $\beta^{\mathcal{I}} = \infty$ . An interesting potential extension of the SI model is to extend this in the way the GRT (2003) model extends the RT (2003) model: allow users to lie along any line in  $R^2$ .

<sup>41</sup> As an example, I will go into a bit more detail on this case. Users typically derive some value from the platform itself and some proportional to the media (games or programs) on the platform. It seems reasonable to assume that the ratio between these is probably quite homogeneous in the population. Similarly software producers have development costs and average per user (unit profits multiplied by the probability of a purchase). At least in expected terms, this ratio is likely quite homogeneous, as software producers that expend larger fixed costs for the same variable benefit as another

systems, video games etc.), dating clubs, commercial intermediation (supermarkets, stock markets, eBay, etc.) and Internet service provision.

For concreteness, I focus here on a version of the model adapted to newspapers or other advertising platforms.  $\beta^{\mathcal{I}} < 0$  on both sides of the market: readers on side  $\mathcal{A}$  have positive membership values from reading the paper but negative interaction values from viewing advertising, while side  $\mathcal{B}$  advertisers have positive interaction value of circulation but membership costs associated with establishing relationships with the paper. Because

$$P^{\mathcal{I}} = \widetilde{B}^{\mathcal{I}} + \widetilde{b}^{\mathcal{I}}N^{\mathcal{J}} = \widetilde{b}^{\mathcal{I}}\left(N^{\mathcal{J}} + \frac{1}{\beta^{\mathcal{I}}}\right)$$

we have that  $\widetilde{b}^{\mathcal{I}} = P^{\mathcal{I}}/(N^{\mathcal{J}} + (1/\beta^{\mathcal{I}})) = P^{\mathcal{I}}\nu^{\mathcal{I}}$  where  $\nu^{\mathcal{I}} \equiv 1/(N^{\mathcal{J}} + (1/\beta^{\mathcal{I}}))$ . The Spence distortion from side  $\mathcal{I}$  is now  $\mu^{\mathcal{I}}\rho^{\mathcal{I}}N^{\mathcal{I}}$  as interaction surplus is just interaction market power multiplied by the average pass-through of its distribution (see Subsection IIIB above). Rather than the sign of the Spence distortion's being dictated directly by the model or left entirely ambiguous, it is given in an intuitive way by market conditions that can be reflected upon or econometrically measured. If prices on side  $\mathcal{B}$  (advertisers) have the same sign as interaction benefits on that side, then loyal users tend to have higher (or less negative, in the case of negative prices) interaction benefits than marginal users and therefore prices on side  $\mathcal{A}$  (readers) are distorted upwards. On the other hand if interaction benefits on side  $\mathcal{A}$  have the opposite sign of price, as with a high-quality newspaper whose readers dislike advertising, then loyal users tend to have more negative (or less positive when prices are negative) interaction values than marginal readers and therefore prices on side  $\mathcal{B}$  are distorted *downward* (advertisers).

Note that the crucial difference here is *not just the sign of interaction values, but how these compare to the sign of price*. Free tabloids essentially have a negative price, given their aggressive marketing in public transport hubs, and therefore have low scale-income, advertising-insensitive loyal readers, implying an upward Spence distortion *despite negative interaction benefits*.<sup>42</sup> Thus the SI model would have very different predictions about the behavior of tabloids versus high-quality papers as the marginal readers of both desert for the Internet: tabloids will become further laden with advertising and market more aggressively, while quality papers will pare back advertising and raise subscription fees.

Comparative statics are similarly dictated by the market conditions. The Scale-Income Model portion of the Appendix shows participation on the two sides are complements (substitutes) if and only if

$$\sum_{\mathcal{I}=\mathcal{A},\mathcal{B}} \left( \nu^{\mathcal{I}} \left[ \widetilde{b}^{\mathcal{J}}N^{\mathcal{J}} - C^{\mathcal{I}} - cN^{\mathcal{J}} \right] \right) - c > (<)0.$$

For quality newspapers  $\nu^{\mathcal{A}} < 0 < \nu^{\mathcal{B}}$ . Assuming subscribers are net profitable even in the absence of subscription fees (advertisers are obviously unprofitable in this sense),  $\{\widetilde{b}^{\mathcal{B}}N^{\mathcal{B}} - C^{\mathcal{A}} - cN^{\mathcal{B}} > 0$ . So long as these effects are large enough to outweigh interaction costs, participation rates are complements. Also intuitively the equilibrium network effect from readers to advertisers is positive by complementarity, but the sign of equilibrium network effect of advertiser participation to

producer will be driven out of the market. However, some games and software are clearly much more prominent and higher impact than others, having larger fixed costs and variable benefits. Thus the SI model seems a sensible fit.

<sup>42</sup> Similarly if programs for an operating system are subsidized, as with Macintosh in the 1990s, low scale programs will tend to be served and thus there will be a downward Spence distortion (potentially underpriced operating systems) despite positive interaction benefits.

readers is ambiguous (the harm to loyal readers may outweigh the benefits to marginal readers or not).

Empirical data become particularly useful in the SI model as it has substantial identifying power even when little can be observed,<sup>43</sup> especially when explicit links can be made to observable income or size distributions,<sup>44</sup> as is common in structural empirical work (Berry, James Levinsohn, and Ariel Pakes 1995). Thus in cases where the source of user heterogeneity is unknown, relevant policy implications are theoretically ambiguous, and empirical data for full identification are missing, the SI model provides a reasonable way to structure policy intuitions. Furthermore, it can easily be extended to the general model of Section III:  $u^T(\mathbf{N}; \theta^T) \equiv \theta^T u^T(\mathbf{N})$  where  $u^T(\cdot)$  is an arbitrary smooth function of the allocation.

## VI. Applications

This section briefly discusses three policy-oriented applications of my results, designed to demonstrate how the tools developed help address longstanding applied questions.

### A. Measuring Market Power and Predation

In applied antitrust analysis, price-cost margins are used to measure market power or as a screen for predatory pricing. It has long been argued (David S. Evans 2003; Wright 2004) that pricing below cost is not indicative of predatory behavior<sup>45</sup> in a two-sided market as users may be subsidized on one side to reflect the benefits of users on the other side. Similarly pricing significantly above cost need not indicate large market power, as users on one side may be taxed if users on the other side have interaction costs. Measuring market power and predation in two-sided markets is therefore an old open question. My framework provides a simple answer: a general Lerner index for two-sided markets, which encompasses and unifies previous Lerner indices proposed for special models, such as Armstrong's and RT (2003).

One approach to such a Lerner index is to construct them for each side of the market individually, in which case they are given by equation (6) and require a measurement of the AIVMU, as well as costs.<sup>46</sup> Measuring the AIVMU may be difficult, but it's not much harder than observing costs. These measures can then be used, as any Lerner index, as a test for market power<sup>47</sup> and predation. Because prices are often near or below zero in two-sided markets, absolute market power  $\mu^T$ , perhaps normalized by something other than price, may be a more attractive metric as it is guaranteed to be positive and finite for a statically optimizing firm. This may be calculated

<sup>43</sup> A decomposition of price into interaction and membership benefits and identification of market power, which is feasible simply based on first-order instruments for participation or price on both sides of the market, suffice to identify interaction market power.

<sup>44</sup> These predict higher-order properties of demand, allowing pass-through rates and cross partials to be predicted, and the size of interaction surplus, and therefore normative comparative statics, to be estimated.

<sup>45</sup> While there is much dislike about requiring below true cost pricing as a necessary condition for predation (Aaron S. Edlin 2002), most legal (Frank H. Easterbrook 1981; Brooke Group Ltd. 1993) and economic (Patrick Bolton, Joseph F. Brodley, and Michael H. Riordan 2000) doctrine holds that allegations of predation must establish in Easterbrook's words "a sacrifice of today's profits for tomorrow's." This means that, in practice, to the extent predation cases arise at all in two-sided markets, the argument that below-cost pricing does not establish that prices are below "true" costs is likely to be a potent one.

<sup>46</sup> Market power may also be estimated structurally (Argentesi and Filistrucchi 2007). I discuss this approach, which also allows costs to be estimated rather than observed, extensively in Weyl (2009a).

<sup>47</sup> It is not immediately clear why market power, and not market power combined with the Spence distortion, is the right thing to measure. For the purposes of my discussion here, I just take as given the policymakers' interest in measuring market power.

just as easily:  $P^{\mathcal{I}} - C^{\mathcal{I}} - cN^{\mathcal{J}} + \widetilde{b}^{\mathcal{J}}N^{\mathcal{J}}$ . Then a natural test for predation on one side individually is that this be negative.

If, instead, an aggregate measure of market power is desired, weighting by participation on the two sides is a natural way to aggregate. The aggregate Lerner index is then

$$N^A\mu^A + \frac{N^B\mu^B}{N^AP^A} + N^BP^B = (1 + \alpha)\frac{\pi}{R}$$

where  $R$  is revenue and  $\alpha$  is the interaction surplus ratio of Subsection IVA. Intuitively if two-sidedness makes up a large part of profits, one should expect relatively low prices for a given amount of market power, as the platform will tend to subsidize users for participation. Therefore even a small profit to revenue ratio indicates significant market power if two-sidedness is a main source of profits. The test for predation is the natural extension of the standard test: profits are negative if and only if the aggregate Lerner index is. My formulae, side-specific or aggregated, extend intuitively to the general multi-sided model of Section III.

### B. Regulation

Regulation of two-sided markets has been a topic of substantial recent interest. Two prominent examples are the policy debates over interchange fee caps on prices to card-accepting merchants, and net neutrality regulations, interpreted variously as price caps on fees Internet service providers (ISPs) can charge Web sites or a limit on their price discrimination. As with merger analysis, designing regulation in two-sided markets is beyond the scope of this paper. However, I believe the paper does provide three issues for future research to consider.

First it emphasizes that, to the extent that regulation aims to emulate the optimal benchmarks of Pigouvian or Ramsey pricing, it should solve distortions on both sides. In fact Pigouvian and Ramsey pricing require solving a constant fraction of distortions on each side, rather than only one side, as with net neutrality or interchange fee regulations. In considering the size of these distortions it suggests two factors are crucial: the size of classical market power and the Spence distortion *on the other side of the market*.

Thus the novel element in two-sided markets is that regulators should focus most on reducing price opposite a side with a large Spence distortion. Thus regulators of ISPs should focus on limiting prices to Web sites (net neutrality) if there is more (interaction) surplus among loyal users than among highly profitable Web sites. But if the situation is reversed, forcing ISPs to reduce prices and build more line to consumer homes may be a higher priority.

Second, implementing Ramsey-type regulation requires a detailed knowledge of demand<sup>48</sup> that may not be available to a regulator. If so it may be more attractive to regulate only one side of the market, especially if market power is thought to particularly distort that side's prices. However a price cap on side  $\mathcal{I}$  can create further distortions, especially with positive interaction benefits, as the platform can lower side  $\mathcal{I}$ 's price either by increasing participation on side  $\mathcal{I}$  (which the regulator wants) or by decreasing participation on side  $\mathcal{J}$  (which she likely does not want). Thus Sheshinski's (1976) argument that price regulation tends to reduce quality provision is even stronger. In two-sided markets "quality reduction" comes from further distorting prices charged to users on the other side of the market. Of course, when interaction benefits are negative, especially if the Spence distortion is upward, this may be desirable: price caps on newspaper readers

<sup>48</sup> When information is more limited, the appropriate response is to explicitly incorporate these informational constraints into a model of policy design (David P. Baron and Myerson 1982; Jean-Jacques Laffont and Jean Tirole 1993). This is an important open problem in two-sided markets and is certainly beyond the scope of this paper.

may lead to more ads,<sup>49</sup> but this could well be an efficient counterbalance to their market power over advertisers especially if, as with tabloids, loyal readers dislike advertising less than marginal readers.

In the positive interaction benefit and Spence distortion case, when price regulation is particularly unattractive, Sheshinski's suggestion of quantity regulation may be more attractive as it does not change pricing incentives on the other side of the market. The simplest way to see this<sup>50</sup> is to note that the privately optimal pricing condition on side  $\mathcal{I}$  takes as given participation on side  $\mathcal{J}$ , and thus the first-order condition on side  $\mathcal{I}$  is not (directly) affected by a constraint on participation on side  $\mathcal{J}$ . A regulator might require an ISP to have a certain fraction of Web sites available on its service, rather than prohibiting the charging of Web sites. This might well encourage the recruitment of more Internet users, as a natural way to increase Web site participation without lowering price is to increase the number of subscribers. Of course, as in any market where quantity regulation is proposed, implementation would require care, to ensure that the ISP does not cheat by signing up the smallest Web sites. Given the diversity of Web sites, the practical complexity of implementing such a policy may outweigh its theoretical benefits. Furthermore, even at a theoretical level, more detailed analysis would be needed to establish the cases in which, once all indirect effects are taken into account, participation regulations are truly preferable to price regulations, and for whom. Nonetheless, such *allocation regulation* at least merits further investigation in multi-sided networks.

Finally, the analysis above seems to provide further rationale for allowing price discrimination in two-sided markets, at least when Spence distortions are positive. In this case discrimination has the additional benefit (Weyl 2009c; Rysman 2009) of increasing the subsidy to users on the other side of the market, helping ameliorate both the market power (on the other side) and Spence (on the discriminated side) distortions. Because Spence distortions are likely upward among Web sites (incumbents like Google make greater profits from a marginal surfer than entrants), this seems to lean in favor of allowing price discrimination,<sup>51</sup> that is, repealing net neutrality. However, if the Spence distortion is negative, as among *Times* readers, price discrimination may be more harmful than usual as it may lead to higher advertiser prices exacerbating market power.<sup>52</sup> Again, more detailed analysis of price discrimination would be required to formalize such arguments.

### C. Mergers

Merger analysis requires a general model of competition, which is beyond the scope of this paper. Nonetheless my results make three small contributions towards this goal.

First, the approach taken here is likely to be useful in analyzing such merger models. To illustrate this, I show in the Applications portion of the Appendix how it can be used to analyze mergers in a nonparametric, market-expanding version of Armstrong's model of symmetrically differentiated single-homing duopoly, the *generalized Armstrong single-homing model* (GASH). A companion paper (Weyl 2008) uses the same techniques to analyze mergers between

<sup>49</sup> Some of these issues are analyzed, under particular assumptions about user heterogeneity (see footnote 12) by Anderson and Coate (2005).

<sup>50</sup> A formalization is available on request.

<sup>51</sup> Of course, as in any vertical moral hazard/double marginalization problem, transferring incentives to the platform is not all good; this may hold up Web sites, extracting surplus from their investment in producing quality content if contracts are not sufficiently rich (Bengt Holmström 1982). For an analysis that emphasizes the effects on Web site investment see Jay Pil Choi and Byung-Cheol Kim (forthcoming).

<sup>52</sup> Of course this depends on whether prices are initially too high or too low to advertisers; in the latter case, the effect is ambiguous.

a platform and a non-two-sided firm producing a good that is a substitute (broadcast TV merger with advertising-free cable) or a complement (operating systems and hardware manufacturers) for users on one side of the market, considering the second case in substantial detail.

Second, the insulating tariff offers an approach to overcoming a problem<sup>53</sup> plaguing the analysis of competition in multi-sided networks. As Armstrong points out, a tremendous multiplicity of equilibria are possible in competition between platforms depending on the tariffs  $P^A(\cdot)$  and  $P^B(\cdot)$  the other firm uses *at participation levels other than the equilibrium*. For example, if one payment card firm uses a fixed fee, this will encourage the other to steal its cardholders as a means of attracting merchants who now have fewer partners, while if it uses a negative fixed fee and a large per-interaction fee this softens competition as stealing cardholders actually discourages merchant participation.

However, if one assumes firms choose insulating tariffs, these cross-side participation stealing effects are reduced and, in the GASH case, entirely eliminated. This *insulated equilibrium* greatly simplifies the analysis of competition. It also seems at least as reasonable an assumption as the more basic Nash-in-prices (Bertrand) assumption universal in the multi-sided networks literature,<sup>54</sup> given that this tariff is both intuitive and plausible, as well as robustly ensuring good equilibria are uniquely selected. It is a simple extension of the common assumption in price-quality competition that firms take as given other firms' choice of quality (Avner Shaked and John Sutton 1982) when choosing price, as the number of users participating on side  $\mathcal{I}$  is effectively the quality of the platform's "product" on side  $\mathcal{I}$ .

Third, my results suggest that in any model of competition, the source of user heterogeneity will be central to determining the positive and normative effects of mergers. Mergers largely affect firm market power, and potentially the size of network effects, both of which act to shift platform (opportunity) costs (Farrell and Carl Shapiro 2008). Because the positive and normative effects of changes in costs and network effects are determined by the sources of user heterogeneity, so too will be the effects of mergers. Furthermore, whether market power is more or less harmful in a two-sided market depends on the source of heterogeneity.

This is confirmed by the two most prominent existing models of competition in two-sided markets. In Weyl (2009c) I show that a merger (with no efficiencies) in the RT (2003) model of competition is tantamount to an increase in market power on one or more sides of the market. It will therefore *increase* participation, and potentially benefit users, on one side if competition is much more intense on one side (participation on the two sides are substitutes). On the other hand, in the online Appendix I show that, at least when competitors use insulating tariffs and regardless of the relative intensity of competition, a merger (without efficiencies) in the GASH model increases market power and therefore reduces participation and welfare on both sides, as

<sup>53</sup> An alternative approach to making a specific assumption about conduct, as I suggest here, is to search for results that are robust across various solution concepts or to attempt to explicitly identify the solution concept. The first approach seems reasonable, if challenging, and is an interesting direction for future research. A simple example of this strategy was a result, included in a previous version of this paper and available on request, that in many reasonable cases, even without an insulating tariff, mergers from GASH lead to lower participation on both sides. The second approach is in the spirit of the classic contributions of Timothy F. Bresnahan (1982) but has proven difficult to implement empirically given its data demands (Aviv Nevo 1998). Nonetheless there has been some recent interest in identifying solution concepts in other contexts, such as vertical relations (Sofia Villas-Boas and Rebecca Hellerstein 2006), so asking how one would go about identifying the two-sided markets solution concept (what sort of price schedules do firms take as given) would be an interesting topic for future research. Finally, one might use demand uncertainty to tie down a unique optimal tariff (Klemperer and Margaret A. Meyer 1989), though this approach has proved challenging to implement in applications in the simpler context of one-sided supply function equilibrium. Nonetheless I think operationalizing uncertainty-based refinements of oligopoly equilibria is an exciting direction for future research.

<sup>54</sup> Bruno Jullien proposed to me, in a private conversation, a model of undifferentiated Cournot-style competition. However, this model has symmetric equilibria only when there is a single dimension of user heterogeneity, making it difficult to analyze more generally. A proof is available on request.

participation rates are complements. Thus merger models in two-sided markets must show care in their assumptions about the sources of user heterogeneity.

## VII. Conclusion

This paper makes two contributions. First, by formulating the platform's problem in terms of its choice of allocation, rather than prices, I simplify and generalize the analysis of network industries. Second, I show that the key normative properties and comparative statics of two-sided markets depend on the source of user heterogeneity, which previous analysis has restricted. The modesty of these contributions makes clear the early stage of the literature. I therefore conclude by discussing directions for future research.

On the empirical side a number of questions are suggested quite directly by my arguments above. Does the SI model fit well in some market where *ex ante* the sources of user heterogeneity seem unclear? How well do the predictions of the RT (2003) model fit actual payment card data? Do newspapers actually exhibit complements? Comparing market power to the Spence distortions, are there overall too many or too few ads in most papers? Applications will largely be driven by the data available, so I will not dwell on them excessively here.

On the theoretical side, much remains to be done to understand pricing in networks more generally. For example, my approach so far allows only extremely stylized models of competition of limited direct empirical relevance. I consider formulating a workhorse, general empirical model<sup>55</sup> of two-sided markets, and practical means for identifying it, to be the most important open question in this area. For regulatory policy the monopoly model is likely to be of greater use, but a more careful analysis of price discrimination and regulatory design (Baron and Myerson 1982; Laffont and Tirole 1993) are needed.

A number of fundamental theoretical problems remain open, three of which I will mention.

First, the exploding literature on matching design, surveyed by Alvin E. Roth (2002), has thus far had limited interaction with the literature on pricing in two-sided markets; see Glenn Ellison, Drew Fudenberg and Markus Möbius (2004); Susan Athey and Glenn Ellison (2008); Ettore Damiano and Hao Li (2008); Andrei Hagiu and Bruno Jullien (2008); Weyl and Tirole (2010) for notable, if early, exceptions. These literatures have much in common, though market design has largely focused on efficiency and paid little attention to prices, while the two-sided markets literature largely ignores, as this paper does, the possibility of designing platforms to increase surplus. I suspect optimal pricing interacts importantly with platform design and therefore that such "revenue maximizing matching" is a fruitful direction for future research.

Finally, as my discussions in Subsection IC and IIIA emphasize, the coordination problems that have long been thought central to networks can generally be overcome by appropriate tariffs. This does not seem to always occur in practice, however. Insulating tariffs might be difficult to implement if demand is not known exactly to the platform; they might, in fact, be unprofitable if demand is uncertain as the critical mass problem might be an effective screen for high demand states. Standard capital market imperfections could also play a role in limiting the platform's ability to borrow, which might be necessary in the true dynamic process of network formation swept under the static model here. A platform might signal to its financiers that it knows it will succeed by overcoming the critical mass problem without subsidies. These are all interesting topics for future theoretical research.

Regardless of the precise explanation for imperfect insulation, my discussion suggests that coordination problems may be a choice, rather than a constraint. If correct, this would imply, for

<sup>55</sup> Alex White and Weyl (2010) make a first attempt at this, extending the insulating tariff to oligopoly.

example, that coordination is, on its own, an important basic source of market power and possible coordination failures are not a reasonable rationale for a merger or collusion. More careful evaluation of this controversial claim is an important theoretical challenge.

## REFERENCES

- Ambrus, Attila, and Rossella Argenziano.** 2009. "Asymmetric Networks in Two-Sided Markets." *American Economic Journal: Microeconomics*, 1(1): 17–52.
- Anderson, Simon P., and Stephen Coate.** 2005. "Market Provision of Broadcasting: A Welfare Analysis." *Review of Economic Studies*, 72(4): 947–72.
- Argentesi, Elena, and Lapo Filistrucchi.** 2007. "Estimating Market Power in a Two-Sided Market: The Case of Newspapers." *Journal of Applied Econometrics*, 22(7): 1247–66.
- Armstrong, Mark.** 2006. "Competition in Two-Sided Markets." *RAND Journal of Economics*, 37(3): 668–91.
- Athey, Susan, and Glenn Ellison.** 2008. "Position Auctions and Consumer Search." <http://kuznets.harvard.edu/~athey/position.pdf>.
- Baron, David P., and Roger B. Myerson.** 1982. "Regulating a Monopolist with Unknown Costs." *Econometrica*, 50(4): 911–30.
- Becker, Gary S., and Kevin M. Murphy.** 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics*, 108(4): 941–64.
- Bedre-Defolie, Özlem, and Emilio Calvano.** 2010. "Pricing Payment Cards." Unpublished.
- Berry, Steven T., and Joel Waldfogel.** 1999. "Free Entry and Social Inefficiency in Radio Broadcasting." *RAND Journal of Economics*, 30(3): 397–420.
- Berry, Steven T., James Levinsohn, and Ariel Pakes.** 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, 63(4): 841–90.
- Besanko, David, Shabtai Donnenfeld, and Lawrence J. White.** 1987. "Monopoly and Quality Distortion: Effects and Remedies." *Quarterly Journal of Economics*, 102(4): 743–67.
- Bolton, Patrick, Joseph F. Brodley, and Michael H. Riordan.** 2000. "Predatory Pricing: Strategic Theory and Legal Policy." *Georgetown Law Journal*, 88(8): 2239–330.
- Bresnahan, Timothy F.** 1982. "The Oligopoly Solution Concept Is Identified." *Economics Letters*, 10(1–2): 87–92.
- Caillaud, Bernard, and Bruno Jullien.** 2001. "Competing Cybermediaries." *European Economic Review*, 45(4–6): 797–808.
- Caillaud, Bernard, and Bruno Jullien.** 2003. "Chicken & Egg: Competition among Intermediation Service Providers." *RAND Journal of Economics*, 34(2): 309–28.
- Choi, Jay Pil, and Byung-Cheol Kim.** Forthcoming. "Net Neutrality and Investment Incentives." *RAND Journal of Economics*.
- Church, Jeffrey, and Neil Gandal.** 1992. "Network Effects, Software Provision, and Standardization." *Journal of Industrial Economics*, 40(1): 85–103.
- Damiano, Ettore, and Hao Li.** 2008. "Competing Matchmaking." *Journal of the European Economic Association*, 6(4): 789–818.
- Dybvig, Philip H., and Chester S. Spatt.** 1983. "Adoption Externalities as Public Goods." *Journal of Public Economics*, 20(2): 231–47.
- Easterbrook, Frank H.** 1981. "Predatory Strategies and Counterstrategies." *University of Chicago Law Review*, 48(2): 263–337.
- Economides, Nicholas.** 1996. "The Economics of Networks." *International Journal of Industrial Organization*, 14(6): 673–99.
- Edlin, Aaron S.** 2002. "Stopping above-Cost Predatory Pricing." *Yale Law Journal*, 111(4): 941–91.
- Ellison, Glenn, and Drew Fudenberg.** 2003. "Knife-Edge or Plateau: When Do Market Models Tip?" *Quarterly Journal of Economics*, 118(4): 1249–78.
- Ellison, Glenn, Drew Fudenberg, and Markus Mobius.** 2004. "Competing Auctions." *Journal of the European Economic Association*, 2(1): 30–66.
- Evans, David S.** 2003. "The Antitrust Economics of Multi-sided Platform Markets." *Yale Journal on Regulation*, 20(2): 325–81.
- Evans, David S., and Richard Schmalensee.** 2009. "Failure to Launch: Critical Mass in Platform Businesses." <http://www.interic.org/Conference/Schmalensee.pdf>.
- Fan, Ying.** 2010. "Ownership Consolidation and Product Quality: A Study of the U.S. Daily Newspaper Market." [http://www.personal.umich.edu/~yingfan/DailyNewspaper\\_Fan.pdf](http://www.personal.umich.edu/~yingfan/DailyNewspaper_Fan.pdf).

- Farrell, Joseph, and Paul Klemperer.** 2007. "Coordination and Lock-in: Competition with Switching Costs and Network Effects." In *Handbook of Industrial Organization*, Vol. 3, ed. Mark Armstrong and Robert H. Porter, 1967–2072. Amsterdam: North-Holland.
- Farrell, Joseph, and Carl Shapiro.** 2008. "Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition." <http://faculty.haas.berkeley.edu/shapiro/alternative.pdf>.
- Gaudeul, Alex, and Bruno Jullien.** 2008. "E-Commerce, Two-Sided Markets and Info-Mediation." In *Internet and Digital Economics: Principles, Methods and Applications*, ed. Eric Brousseau and Nicolas Curien, 268–90. Cambridge, UK: Cambridge University Press.
- Gentzkow, Matthew, and Jesse Shapiro.** 2010. "What Drives Media Slant? Evidence from U.S. Newspapers." *Econometrica*, 78(1): 35–71.
- Guthrie, Graeme, and Julian Wright.** 2007. "Competing Payment Schemes." *Journal of Industrial Economics*, 55(1): 37–67.
- Hagiu, Andrei.** 2006. "Pricing and Commitment by Two-Sided Platforms." *RAND Journal of Economics*, 37(3): 720–37.
- Hagiu, Andrei.** 2007. "Merchant or Two-Sided Platform?" *Review of Network Economics*, 6(2): 115–33.
- Hagiu, Andrei.** 2009. "Two-Sided Platforms: Product Variety and Pricing Structures." *Journal of Economics and Management Strategy*, 18(4): 1011–43.
- Hagiu, Andrei, and Bruno Jullien.** 2008. "Why Do Intermediaries Divert Search?" <http://www.people.hbs.edu/ahagiu/research.html>.
- Holmstrom, Bengt.** 1982. "Moral Hazard in Teams." *Bell Journal of Economics*, 13(2): 324–40.
- Hotelling, Harold.** 1929. "Stability in Competition." *Economic Journal*, 39(153): 41–57.
- Kaiser, Ulrich, and Minjae Song.** 2009. "Do Media Consumers Really Dislike Advertising?: An Empirical Assessment of the Role of Advertising in Print Media Markets." *International Journal of Industrial Organization*, 27(2): 292–301.
- Kaiser, Ulrich, and Julian Wright.** 2006. "Price Structure in Two-Sided Markets: Evidence from the Magazine Industry." *International Journal of Industrial Organization*, 24(1): 1–28.
- Katz, Michael L., and Carl Shapiro.** 1985. "Network Externalities, Competition, and Compatibility." *American Economic Review*, 75(3): 424–40.
- Katz, Michael L., and Carl Shapiro.** 1994. "Systems Competition and Network Effects." *Journal of Economic Perspectives*, 8(2): 93–115.
- Klemperer, Paul D., and Margaret A. Meyer.** 1989. "Supply Function Equilibria in Oligopoly under Uncertainty." *Econometrica*, 57(6): 1243–77.
- Laffont, Jean-Jacques, and Jean Tirole.** 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.
- Lee, Robin S.** 2009. "Vertical Integration and Exclusivity in Platform and Two-Sided Markets." <http://pages.stern.nyu.edu/~rslee/>.
- Liebowitz, S. J., and Stephen E. Margolis.** 1994. "Network Externality: An Uncommon Tragedy." *Journal of Economic Perspectives*, 8(2): 133–50.
- Lindhal, Erik.** 1919. "Positive Lösung." *Die Gerechtigkeit der Besteuerung*.
- Ltd., Brooke Group.** 1993. "Brooke Group Ltd. v. Brown & Williamson Tobacco Corp." (92–466), 509 U.S.(209).
- Mussa, Michael, and Sherwin Rosen.** 1978. "Monopoly and Product Quality." *Journal of Economic Theory*, 18(2): 301–17.
- Myerson, Roger B.** 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73.
- Nevo, Aviv.** 1998. "Identification of the Oligopoly Solution Concept in a Differentiated-Products Industry." *Economics Letters*, 59(3): 391–95.
- O’Keeffe, Mary, W. Kip Viscusi, and Richard J. Zeckhauser.** 1984. "Economic Contests: Comparative Reward Schemes." *Journal of Labor Economics*, 2(1): 27–56.
- Oum, Tae Hoon, and Michael W. Tretheway.** 1988. "Ramsey Pricing in the Presence of Externality Costs." *Journal of Transport Economics and Policy*, 22(3): 307–17.
- Pigou, Arthur.** 1920. *The Economics of Welfare*. London: MacMillan.
- Rochet, Jean-Charles, and Jean Tirole.** 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association*, 1(4): 990–1029.
- Rochet, Jean-Charles, and Jean Tirole.** 2006. "Two-Sided Markets: A Progress Report." *RAND Journal of Economics*, 37(3): 645–67.
- Roth, Alvin E.** 2002. "The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics." *Econometrica*, 70(4): 1341–78.
- Rysman, Marc.** 2004. "Competition between Networks: A Study of the Market for Yellow Pages." *Review of Economic Studies*, 71(2): 483–512.

- Rysman, Marc.** 2009. "The Economics of Two-Sided Markets." *Journal of Economic Perspectives*, 23(3): 125–43.
- Samuelson, Paul A.** 1941. "The Stability of Equilibrium: Comparative Statics and Dynamics." *Econometrica*, 9(2): 97–120.
- Segal, Ilya.** 1999. "Contracting with Externalities." *Quarterly Journal of Economics*, 114(2): 337–88.
- Shaked, Avner, and John Sutton.** 1982. "Relaxing Price Competition through Product Differentiation." *Review of Economic Studies*, 49(1): 3–13.
- Sheshinski, Eytan.** 1976. "Price, Quality and Quantity Regulation in Monopoly Situations." *Economica*, 43(17): 127–37.
- Spence, A. Michael.** 1975. "Monopoly, Quality, and Regulation." *Bell Journal of Economics*, 6(2): 417–29.
- Spence, A. Michael.** 1976. "Product Differentiation and Welfare." *American Economic Review*, 66(2): 407–14.
- Veiga, Andre, and E. Glen Weyl.** 2010. "Platforms with Heterogeneous Externalities." Unpublished.
- Villas-Boas, Sofia, and Rebecca Hellerstein.** 2006. "Identification of Supply Models of Retailer and Manufacturer Oligopoly Pricing." *Economics Letters*, 90(1): 132–40.
- Weyl, E. Glen.** 2008. "Double Marginalization in Two-Sided Markets." [http://www.people.fas.harvard.edu/~Eweyl/Genius\\_JMP.pdf](http://www.people.fas.harvard.edu/~Eweyl/Genius_JMP.pdf).
- Weyl, E. Glen.** 2009a. "Slutsky Meets Marschak: First-Order Identification of Multi-Product Production." [http://www.people.fas.harvard.edu/~weyl/SlutMars\\_12\\_09\\_pdf](http://www.people.fas.harvard.edu/~weyl/SlutMars_12_09_pdf).
- Weyl, E. Glen.** 2009b. "Monopoly, Ramsey and Lindahl in Rochet and Tirole (2003)." *Economics Letters*, 103(2): 99–100.
- Weyl, E. Glen.** 2009c. "The Price Theory of Two-Sided Markets." [http://www.people.fas.harvard.edu/~weyl/pt2sms\\_3\\_09.pdf](http://www.people.fas.harvard.edu/~weyl/pt2sms_3_09.pdf).
- Weyl, E. Glen, and Michal Fabinger.** 2009. "Pass-Through as an Economic Tool." [http://www.people.fas.harvard.edu/~weyl/Pass-through\\_10\\_09.pdf](http://www.people.fas.harvard.edu/~weyl/Pass-through_10_09.pdf).
- White, Alexander, and E. Glen Weyl.** 2010. "Imperfect Platform Competition: A General Framework." Unpublished.
- Wright, Julian.** 2004. "The Determinants of Optimal Interchange Fees in Payment Systems." *Journal of Industrial Economics*, 52(1): 1–26.

**This article has been cited by:**

1. Chokri Aloui, Khaïreddine Jebzi. 2010. Optimal pricing of a two-sided monopoly platform with a one-sided congestion effect. *International Review of Economics* . [[CrossRef](#)]

# Net neutrality and investment incentives

Jay Pil Choi\*

and

Byung-Cheol Kim\*\*

*This article analyzes the effects of net neutrality regulation on investment incentives for Internet service providers (ISPs) and content providers (CPs), and their implications for social welfare. Concerning the ISPs' investment incentives, we find that capacity expansion decreases the sale price of the priority right under the discriminatory regime. Thus, contrary to ISPs' claims that net neutrality regulations would have a chilling effect on their incentive to invest, we cannot dismiss the possibility of the opposite. A discriminatory regime can also weaken CPs' investment incentives because of CPs' concern that the ISP would expropriate some of the investment benefits.*

## 1. Introduction

■ This article analyzes the effects of net neutrality regulation on investment incentives for various players in the Internet market. Since the Internet's inception, one of the governing principles in its operation has been nondiscrimination requirements in all relevant performance dimensions, as has been true for traditional telecommunication services such as the telephone network. In 2005, however, the Federal Communications Commission (FCC) changed the classification of Internet transmissions from the category of "telecommunications services" to the category of "information services." As a result, Internet service providers (ISPs) are no longer subject to nondiscrimination restrictions. In fact, major telephone and cable operators, which together control about 98% of broadband service in the United States (as of December 2005),<sup>1</sup> recently expressed an interest in providing multitier Internet service, charging content providers (CPs) premium prices for preferential access to broadband transmission service. In response, a coalition of content providers merged forces in an effort to maintain the nondiscrimination status

---

\*University of New South Wales, Australia; choijay@gmail.com.

\*\*Georgia Institute of Technology; byung-cheol.kim@econ.gatech.edu.

We are grateful to the coeditor and two anonymous referees for their valuable comments and suggestions, which greatly improved the article. We also thank Paul de Bijl, Nick Economides, and participants in various conferences and seminars for helpful comments, and Jane Choi for her editorial assistance. The first draft of this article was completed during Choi's research visit to the ISCR and the School of Economics and Finance at Victoria University of Wellington, whose hospitality is greatly appreciated. This article received the Public Utility Research Center Prize for the best paper in regulatory economics, which was awarded at the Seventh Annual International Industrial Organization Conference. We gratefully acknowledge financial support from the NET Institute for this research. All errors, if any, are ours.

<sup>1</sup> FCC Form 477 data.

quo. Their intensive lobbying efforts led to a hot debate—known as the net neutrality debate—in Washington, along with initiatives to legislate a mandate to prevent creating a multitier Internet service. Although efforts to enact net neutrality regulations have stalled for now, the issue is expected to continuously arise in the future.<sup>2</sup>

On October 19, 2007, for instance, the Associated Press (AP) reported that Comcast, the United States' largest cable TV operator and second largest Internet provider, had interfered with users' access to file-sharing sites such as BitTorrent.<sup>3</sup> This practice is an example of discrimination in which ISPs intended to slow down some forms of traffic while giving others priority. Comcast may have had a benign reason for this practice—so called traffic shaping—to prevent file-sharing traffic from using up too much bandwidth and affecting the Internet speeds of other subscribers.<sup>4</sup> This interference, however, was certainly a move against the tradition of treating all types of Internet traffic equally—the principle of “net neutrality.” As one person's upload is another's download in file-sharing networks, this type of traffic management can have a series of repercussions in the network of file sharers. As a result, the incident received nationwide attention and stirred uproar from users of file-sharing applications who were adversely affected.

To inform this important policy debate, we analyze economic issues associated with net neutrality regulation. Considering that the Internet is a vital medium of communication, information, and commercial activities, it is of paramount importance to maintain competition and promote innovation in this market. Policymakers thus need to act with care and make an informed decision based on rigorous analysis to provide a market environment in which the right investment signals are given.

Reflecting the importance of the Internet as a main driver of economic growth and prosperity in the global economy, one of the main issues surrounding the net neutrality debate is the innovation and investment incentive for various parties involved in the market. For instance, ISPs such as Verizon, Comcast, and AT&T oppose network neutrality regulations and claim that such regulations would discourage investment in broadband networks. The logic is that they would have no incentive to invest in network capacity unless content providers supporting bandwidth-intensive multimedia applications pay a premium for heavy Internet traffic. By contrast, proponents of network neutrality regulations (comprising mostly consumer rights groups and large Internet content companies such as Google, Yahoo, and eBay) note that the Internet has operated according to the nondiscriminatory neutrality principle since its earliest days. To support their claim that net neutrality has been the main driver of the growth and innovative applications of the Internet, they rely on the so-called end-to-end design principle. Under this design principle, decisions are made “to allow the control and intelligence functions to reside largely with users at the ‘edges’ of the network, rather than in the core of the network itself” (Cerf, 2006). According to them, this creates an environment that does not require users to seek permission from the network owners and thus promotes innovations in Internet applications.

To assess the validity of conflicting claims made by opposing parties, we set up a model based on the queuing theory because scarce bandwidth and the potential need for rationing (due to substantial increases in multimedia usage of the Internet) are the root causes of the debate.<sup>5</sup> With a monopolistic network operator and two application providers, we provide a formal economic analysis of the effects of net neutrality regulations on investment incentives for ISPs and CPs, and

---

<sup>2</sup> The Obama administration has expressed support for net neutrality. On October 22, 2009, the FCC released the Net Neutrality NPRM (Notice of Proposed Rulemaking). In particular, Paragraph 106 of the NPRM states that “a broadband Internet access service provider may not charge a content, application, or service provider for enhanced or prioritized access to the subscribers of the broadband Internet access service provider.” For detailed explanation and discussion of institutional differences between the European Union and United States concerning net neutrality regulation, see Chirico, Van de Haar, and Larouche (2007).

<sup>3</sup> For more details, see “Comcast Blocks Some Internet Traffic,” October 19, 2007, by Peter Svensson, AP.

<sup>4</sup> Peer-to-peer file-sharing applications reportedly account for about 50–90% of overall Internet traffic according to a survey in 2007 by ipoque GmbH, a German traffic-management equipment vendor.

<sup>5</sup> For an economic analysis of traffic congestion at the interface between backbones in the commercial Internet, see Cremer, Rey, and Tirole (2000).

their implications for social welfare. There is no universally accepted definition of net neutrality. In principle, net neutrality means that all packets that traverse through the Internet are treated equally on the basis of first come, first served. Any routing practice deviating from this principle is a violation of net neutrality, for example, port blocking, quality degradation, and access tiering. In our article, we focus on prioritization in delivery speed for particular packets as the main mode of nonneutrality. Without net neutrality regulation, it is possible for a content provider to pay for preferential delivery while best efforts are assured for the rest without targeted degrading or purposeful content blocking.<sup>6</sup>

We compare the market equilibrium in which the monopolistic ISP is allowed to provide a two-tiered service by selling the “fast lane” to only one content provider with the equilibrium in which it cannot discriminate the delivery speed of content. This comparison of two short-run equilibria yields two major findings. First, in a discriminatory network, both content providers may engage in a Prisoners’ Dilemma type of game to receive the first priority in the delivery of content and be worse off. The ISP’s decision of whether or not it will prefer the discriminatory regime to the neutral network depends on a potential tradeoff between its network access fee from end users and the revenue from CPs through the trade of the first priority. Second, the short-run effect of net neutrality regulation on social welfare depends on the relative magnitudes of content providers’ cost/quality asymmetry and the degree of content differentiation. In particular, we show that social welfare is higher under the discriminatory regime if the asymmetry across content providers is sufficiently large.

Additionally and more importantly, we study the long-run effects of net neutrality regulation on the ISPs’ investment incentives. We find that there are two channels through which net neutrality regulation can have impacts on the ISPs’ investment incentives: the *network access fee effect* and the *rent extraction effect*. In the network with net neutrality, capacity expansion speeds up the delivery of content uniformly, thereby enabling the ISP to charge more for access. Similarly, in the discriminatory network, capacity expansion also increases the delivery speed of content and thus allows the ISP to charge a higher network fee. However, because such an effect occurs asymmetrically across different priority classes, we cannot tell unambiguously under which regime the effect of capacity extension is larger. Capacity expansion also affects the sale price of the priority right under the discriminatory regime. Because the relative merit of the first priority, and thus its value, becomes relatively small for higher capacity levels, the ISP’s incentive to invest on capacity under a discriminatory network is smaller than that under a neutral regime, where such rent extraction effects do not exist. As a result, the ISP’s investment incentive hinges upon the relative magnitudes of these two potentially opposing effects. Contrary to ISPs’ claims that net neutrality regulations would have a chilling effect on their incentive to invest, we cannot dismiss the possibility of the opposite.

We also study the effects of net neutrality regulation on application/content providers’ incentives to invest in cost reduction/quality enhancement. Because the monopolistic ISP can expropriate some of the investment benefits made by content providers through the trade of first-priority delivery in the discriminatory network, content providers’ investment incentives can be higher under the net neutrality regime. This implies that the ISP’s payoff is not necessarily increasing in its ability to extract rents from CPs when the adverse effects on CPs’ investment incentives are taken into account. As a result, the ISP may wish to limit its ability to extract rent, if such a commitment mechanism is available, to mitigate the countervailing dynamic effect on innovation incentives for CPs.

We thus find that the relationship between net neutrality regulation and investment incentives for network operators and application/content providers is subtle, and it is not easy to draw

---

<sup>6</sup> The reason for assuming best efforts for basic service is to reflect the current climate in the debate on net neutrality. No ISPs contemplate engaging in such a deliberate degradation of service, perhaps in fear of regulatory backlash. We discuss the possibility of deliberate degradation in Section 7. See Kocsis and de Bijl (2007) for various types of violations of net neutrality.

general unambiguous conclusions. However, our model informs policymakers and regulators by identifying important effects that are expected both in the short run and long run and showing the mechanism through which such effects interact.

The remainder of the article is organized in the following way. The next section offers a brief literature review of articles addressing net neutrality issues. Section 3 sets up a basic model of network markets to analyze the effects of net neutrality regulation on competition and social welfare. Section 4 analyzes the short-run equilibrium with the capacity fixed and studies the effects of net neutrality on ISPs, CPs, consumer surplus, and social welfare. Sections 5 and 6 analyze the effects of net neutrality regulations on investment incentives of ISPs and CPs, respectively. In Section 7, we provide a brief analysis with discussion about various issues around the debate of net neutrality such as heterogeneity in delay costs across content, quality degradation of information packets, and vertical integration between ISPs and CPs. Section 8 closes this article with concluding remarks, along with suggestions for further possible extensions of our basic analysis. The proofs of lemmas and propositions are relegated to Appendix A.

## 2. Related literature

■ Net neutrality regulations have been a hotly debated topic discussed with passion by both proponents and opponents alike. The discussion so far, however, has been rich in rhetoric but short on rigorous economic analysis. There are several notable exceptions.<sup>7</sup>

Hermalin and Katz (2007) examine a situation in which ISPs serve as a platform to connect content providers with end consumers in a framework of two-sided markets. They consider heterogeneous content providers whose products are vertically differentiated to analyze the effects of net neutrality regulation. Without any restrictions, ISPs can potentially offer a continuum of vertically differentiated services, although the ISP is required to provide only one service (a single tier of Internet service) with net neutrality regulation. They compare the single-service level equilibrium with the multiservice level equilibrium and show that net neutrality regulation has the following effects. Content providers who would otherwise have purchased a low-quality service are excluded from the market. That is, content providers at the bottom of the market—the ones that a single-product restriction is typically intended to aid—are almost always harmed by the restriction. Content providers in the “middle” of the market utilize more efficient and higher-quality service, which favors net neutrality regulation. Content providers at the top of the market utilize less efficient and lower-quality service than the one that would have been used in the absence of regulation, which obviously favors the discriminatory network. The overall welfare effects of such regulation can be ambiguous, but they argue that the effects are often negative. Our research complements the analysis of Hermalin and Katz (2007) in that we consider the congestion effect in the provision of Internet service and, more importantly, investigate investment incentives of ISPs and CPs; both are not addressed in theirs.

In terms of policy questions and basic framework, our research is closest to Cheng, Bandyopadhyay, and Guo (2009), who develop a game-theoretic model of competition between two content providers in a Hotelling framework. They investigate the effects of net neutrality regulation on ISPs’ incentives to expand capacity as well as study who gains and who loses as a consequence of the regulation. We build upon their framework and extend their analysis in several directions. However, there are several major differences between the two articles. First, our study goes one step further by analyzing the effects of the regulation on content providers’ incentives to provide innovative services. We find that the holdup problem can prevail under a discriminatory regime, and thus *ex ante* the ISP might prefer not to extract the entire rent available from CPs. Second, our model adopts a different pricing scheme in the sale of the priority. They assume that the ISP deals with the two competing CPs nonexclusively and charges both for priority regardless

<sup>7</sup> See also Economides (2007) and Kocsis and de Bijl (2007). In addition, there is an extensive discussion of net neutrality by lawyers. See, for instance, Wu (2003), Yoo (2006), and van Schewick (2007) and references cited therein.

of what the rival CP does. In such a setup, it is possible that both CPs pay the price for priority, but end up being back where they started because no CP would have an advantage over the other. Our model, by contrast, assumes that the ISP deals with CPs exclusively in that only one CP can be given priority. This different consideration about the pricing scheme leads to more general results with new insights. For instance, Cheng et al. find that the ISP's incentive to expand its capacity is unambiguously higher under net neutrality, but such an outcome is just one possibility in our model. In fact, we discover that it is not easy to draw general clear-cut conclusions about the relationship between net neutrality regulation and innovation incentives of either ISPs or CPs. In addition, they find that if the principle of net neutrality is abandoned, the ISPs definitely stand to gain from extracting fees for preferential treatment from the content providers. In our model, however, such a relationship is not always derived, because the discrimination may decrease the ISP's revenue from network access fees for consumers.

Economides and Tåg (2007) also provide an economic analysis of net neutrality in a two-sided market framework, but their research focus differs from ours. They are particularly interested in how net neutrality regulation affects pricing schemes on both sides of the market and social welfare in the short run; thus, our study strongly complements theirs.

Finally, Valletti and Cambini (2005) analyze network operators' incentives to invest in networks with different quality levels, as in our article. They show that quality has an impact on all calls initiated by customers (destined both on-net and off-net) and that "tacit collusion" takes place even in a symmetric model with two-part pricing, because firms tend to underinvest in quality. Their focus is on the impact of *two-way access charges* on the investment incentives in communication networks that require interconnection for off-net traffic, whereas our analysis concerns the impacts of net neutrality regulation on investment incentives of a network operator that serves as a platform for two-sided markets.

### 3. A model of net neutrality

■ We consider a situation in which online content providers deliver their content to end consumers through a broadband network provided by a monopolistic Internet service provider. For instance, we can envision a specific geographic market in which Comcast is a monopolistic ISP and content providers such as Yahoo and Google deliver their contents at the end users' requests. As Economides (2008) argues, the monopolistic ISP assumption is reasonable in the United States when we consider limited choices of broadband Internet access available to residential consumers as well as significant switching costs in changing ISPs (see Kocsis and de Bijl, 2007 for competition between oligopolistic network operators).

To understand our notion of net neutrality and the issues we address, it is useful to note that the Internet is an interconnected "network of networks." None of the network operators has the capacity to provide complete end-to-end routing between content providers and end consumers. Typically, the network operator who provides hosting services to a content provider would be different from the ISP who provides Internet connection to an end consumer. Thus, when a consumer requests specific content from a content provider, it needs to traverse several different networks. When the content is delivered, even packets comprising the same web page can travel different routes before they are assembled at the client's computer. The transit between networks is governed by a variety of peering agreements between networks. Tier 1 networks constitute the Internet backbone and have direct connections to the Internet. Tier 1 network operators interconnect with each other without purchasing transit or paying settlements. Tier 2 and 3 networks are relatively small players and purchase at least some transit from other networks to reach the Internet. We take these existing peering agreements between networks as given.

Our analysis focuses on the local market in which a local monopolistic ISP provides the "last-mile" connection service to consumers. In particular, we study the effects of the local ISP's discretion to discriminate content once the packets from content providers have arrived at the local ISP's switch box through the existing transit agreement between networks. In other words, we

consider the possibility of two-tier service at the local ISP level with the ISP extracting additional payment from content providers as a price for “priority” delivery in the absence of net neutrality regulation.

□ **The basic model.** The monopolistic ISP sells its network connection to end users at price  $a$ . There are two content providers who compete to deliver content to end users. Under net neutrality, the ISP cannot discriminate between content providers in the delivery speed of contents and does not charge content providers for sending information over its network to end users (Hahn and Wallsten, 2006).<sup>8</sup> However, note that this does not mean that CPs receive access to the Internet for free under net neutrality. We envision a typical situation in which the network operators that provide hosting service to CPs are different from the local ISP that provides access to end users. Under net neutrality, CPs pay access fees to the hosting network operator only once at the origin and are not required to pay additionally for “transit,” which is already covered in the existing traffic-routing arrangement governed by peering agreements between networks. Thus, the local ISPs that provide the last-mile transit to end consumers are not allowed to demand additional compensation from CPs. By contrast, without net neutrality regulation, preferential treatment for a particular content provider is no longer prohibited. Then, the ISP can sell the first priority, the right to be served ahead of the other, to either of the two content providers.

We assume that each consumer has demands for one of two CPs. The content request rate of each consumer follows a Poisson process with hazard rate  $\lambda$ , which represents the demand intensity. The network capacity is denoted by  $\mu$ . The service time taken to deliver the content from CPs to end users is exponentially distributed with its mean of  $1/\mu$ . Larger network capacity implies a shorter service time. In the short-run analysis, the capacity  $\mu$  is assumed to be fixed. In the long-run analysis in which investment incentives are investigated, it is endogenously derived.

Consumers, whose mass is normalized to one, are heterogeneous with respect to their preferences toward two content services in the Hotelling manner. Setting CP1 and CP2 located at the left and right ends of a line segment whose length is normalized to one, a consumer located at  $x$  pays the transport cost of  $tx$  and  $t(1-x)$  to consume CP1’s and CP2’s services, respectively. The transport cost per unit distance,  $t$ , can represent the degree of product differentiation. As in Mendelson (1985), we assume that each consumer with content request rate  $\lambda$  derives a gross utility of  $u(\lambda) = v$  from either content service, and that  $v$  is sufficiently large so that the market is fully covered in both regimes of networks.<sup>9</sup>

As in Cheng et al. (2009), Choi (2010), and Economides and Tåg (2007), we assume that content providers adopt a business model that offers their services without any direct charge to consumers but generates revenues through advertisements. Each content provider  $i$  earns a revenue stream  $r_i$  from advertisers for each consumer’s content request (“click-throughs”) it serves. The asymmetry in  $r_i$  may reflect differences in CPs’ capabilities to match advertisers and consumers. The cost of serving each consumer’s request is given by  $c_i$ . Content provider  $i$ ’s markup per each consumer’s click-through is given by  $m_i = r_i - c_i$ , where  $m_1 \geq m_2 \geq 0$  without loss of generality. Thus, the sources of asymmetry in CPs’ click-through margins can be either from the revenue side, the cost side, or a combination of both. The corresponding CP’s profit is measured by  $m_i \lambda \sigma_i$ , where  $\sigma_i$  denotes the market share for content provider  $i$ .

The sequence of the players’ choices is as follows. In the discriminatory network regime, the ISP can first sell the priority service through a trading process to only one content provider exclusively; in the neutral network, this stage does not apply. Then, the ISP posts a network access fee,  $a$ , to end users. Given the allocation of the priority classes and the network access fee, end users choose one of the content providers. As usual, the analysis for this game proceeds by using

<sup>8</sup> To quote AT&T CEO Edward Whitacre, content providers “use my lines for free.” See “Rewired and Ready for Combat,” *Business Week Online*, November 7, 2005.

<sup>9</sup> The content request rate  $\lambda$  can depend on the delivery speed of content to end users in a more general model. For instance, it is possible that end users may abort content requests in the face of long delays and leave the queue. We treat such a possibility as a second-order effect and ignore it.

backward induction, and the equilibrium concept employed here is that of subgame-perfect Nash equilibrium.

□ **Preliminaries: congestion in the M/M/1 queuing system.** To model congestion in the network, we adopt the standard framework of the M/M/1 queuing system that has been widely used by many scholars in operations research to study congestion problems and priority pricing (see Naor, 1969; Balachandran, 1972; Edelson and Hilderbrand, 1975; Mendelson and Whang, 1990). The reason for this modelling choice is two-fold. First, the root causes of the net neutrality debate are scarce bandwidth and the potential need for rationing due to substantial increases in bandwidth-intensive multimedia applications over the Internet. Second, this setup is well known to be a very good approximation for the arrival process in real systems, in which the number of customers is sufficiently large so that the impact of a single customer on the performance of the system is very small, and all customers' decisions to use the system are independent of other users'. Furthermore, this microfoundation yields nice properties with which we can analyze without any ad hoc assumptions.

In the neutral network where all packets are treated equally without any priority classes, each consumer has the expected waiting time of

$$w = \frac{1}{\mu - \lambda}, \quad (1)$$

where  $\lambda$  denotes the gross content request rate at the network (with the normalization of consumer mass to one) with the network capacity  $\mu > \lambda$ . The waiting time increases in  $\lambda$ , but decreases in  $\mu$ . If we normalize the delay cost per unit time to one, then the waiting time in (1) equals each consumer's expected waiting cost. In the basic model, we assume that all content has the same delay cost per unit time; in Section 7, we extend our analysis by considering heterogeneity in delay costs across content and applications.

On the other hand, in the discriminatory network with two priority classes, consumers' waiting costs depend on the priority classes to which their packets are designated. In the nonpreemptive discriminatory network, a consumer who requests content designated to the first-priority class has an expected waiting time of

$$w_1 = \frac{1}{\mu - \lambda_1}, \quad (2)$$

where  $\lambda_1$  is the total amount of traffic from consumers who request the content with the first priority.<sup>10</sup> By contrast, the consumer who requests content without the first priority faces the expected waiting time of

$$w_2 = \frac{\mu}{\mu - \lambda} w_1 = \frac{\mu}{\mu - \lambda} \frac{1}{\mu - \lambda_1}. \quad (3)$$

Based on these standard results in the queuing theory,<sup>11</sup> we can infer that a consumer will face a higher waiting cost by requesting the nonprioritized content instead of the prioritized one, that is,

$$\text{Property 1. } w_2 > w > w_1 \text{ for } \mu > \lambda.$$

Property 1 is easily established by examining the relative ratio of  $w_2$  to  $w_1$ , that is,  $w_2/w_1 = \mu/(\mu - \lambda) > 1$ . As a related property, we can also notice that the relative ratio of  $w_2$  to  $w_1$  is a constant, regardless of the distribution of the total traffic across different priority classes. In

<sup>10</sup> See Gross and Harris (1998) and references therein for more on the queuing theory and for the detailed derivation for waiting costs in different types of networks.

<sup>11</sup> In discriminatory networks, there are two possible priority schemes: preemptive and nonpreemptive. In the preemptive scheme, the customer request with the priority is allowed to be serviced immediately, even if another without priority is already present in service. In the nonpreemptive scheme, the customer request with the priority simply goes to the head of the queue to wait its turn without interrupting the service of a customer request already in progress.

addition, we find that the quality difference measured in waiting costs becomes smaller as the network capacity increases, that is,

$$\text{Property 2. } \frac{\partial}{\partial \mu}(w_2 - w_1) < 0.$$

This is because the marginal reduction in waiting time for the fast lane from capacity expansion decreases as the capacity level becomes high.

We adopt the M/M/1 system for our analysis because it is a standard framework to model congestion in computer networks. However, the same qualitative results can be derived with a more general framework as long as the two properties above are satisfied.

#### 4. Net neutrality and short-run analysis

□ **Equilibrium in the neutral network.** In a neutral network, end users choose one of the two content providers who provides higher net surplus, knowing that the waiting cost is given by (1). In the Hotelling model of end users, the marginal consumer  $x^*$  who is indifferent between two content providers in the neutral network is defined as<sup>12</sup>

$$v - \frac{1}{\mu - \lambda} - tx^* - a = v - \frac{1}{\mu - \lambda} - t(1 - x^*) - a, \tag{4}$$

where consumers whose preferences are represented by  $x < x^*$  choose CP1 and those by  $x > x^*$  choose CP2. With two symmetrically positioned content providers, the market for content provision is equally split between the two firms, with each content provider serving half of the market, that is,  $x^* = 1/2$ . We further consider a scenario in which each consumer's taste parameter  $x$  is fixed, which implies that it is only the middle consumer ( $x = 1/2$ ) whose participation constraint is binding.<sup>13</sup>

The ISP's profit maximization problem is thus given by

$$\max_a \pi_m = a \quad \text{s.t.} \quad v - \frac{1}{\mu - \lambda} - tx^* - a \geq 0, \tag{5}$$

where the constraint is needed to ensure that the market is covered. Then, we can derive the equilibrium network access fee and each content provider's profit as

$$\pi_m^* = a^* = v - \frac{1}{\mu - \lambda} - \frac{t}{2}; \quad \pi_i^* = \frac{m_i}{2}\lambda \quad \text{for } i = 1, 2. \tag{6}$$

□ **Equilibrium in the discriminatory network.** If the ISP is allowed to charge content providers for the higher priority class, consumers will face different expected waiting times depending on their choice of content service, as derived in (2) and (3). Let us assume that the high-margin (more efficient) content provider, CP1, obtains the first priority, with its content being entitled to be served ahead of CP2's. Later, we demonstrate that this scenario arises as an equilibrium outcome regardless of the trading mechanism. The marginal consumer at  $\tilde{x}$ , who is indifferent between the premium service and the basic service, is characterized by the following equality:

$$v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x} - a = v - \frac{\mu}{\mu - \lambda} \frac{1}{\mu - \tilde{x}\lambda} - t(1 - \tilde{x}) - a, \tag{7}$$

<sup>12</sup> The following equality is based on the assumption that there is no direct payment from end users to content providers, which simplifies the analysis. The exploration for the implications of direct payment will be an important extension of this basic model, as explained in Section 8.

<sup>13</sup> Alternatively, we can imagine a situation in which each consumer has a random arrival rate of demands for CP service, and for each demand, their taste parameter  $x$  is i.i.d. rather than constant for all time. In such a specification, each consumer would have *ex ante* the same preference and thus the same expected surplus from using the ISP. This alternative specification yields qualitatively the same results as in our specification.

where we use a tilde to denote variables associated with the discriminatory regime.<sup>14</sup> By comparing (4) and (7), we can derive an intuitive result that the content provider with the first priority has a larger market share than the one without it, that is,  $\tilde{x} \geq x^* = 1/2$ , due to the difference in waiting costs.

Note that as more consumers switch from CP2 to CP1, the delivery speed of CP1's premium service deteriorates, but the delivery speed of CP2's basic service worsens more.

$$\frac{\partial w_2}{\partial \lambda_1} = \frac{\mu}{\mu - \lambda} \frac{1}{(\mu - \lambda_1)^2} > \frac{1}{(\mu - \lambda_1)^2} = \frac{\partial w_1}{\partial \lambda_1} > 0 \tag{8}$$

The gap in waiting time between the premium and basic services widens as more consumers switch from CP2 to CP1, which in turn makes CP1 more attractive. This positive-feedback process can lead to a corner solution, a situation in which all consumers subscribe to the CP with the first priority. To ensure an interior equilibrium in which the CP with the first priority does not corner the market ( $\sigma_i > 0$  for  $i = 1, 2$ ), we make the following assumption.

*Assumption 1.*  $t > \frac{\lambda}{(\mu - \lambda)} 2$ .

The assumption says that the content of the two CPs are sufficiently differentiated to prevent a corner solution.<sup>15</sup>

The following lemma specifies a sufficient condition under which an interior market-sharing equilibrium is stable and the market share of the content provider with the first priority decreases as the ISP's capacity increases.

*Lemma 1.* Assuming that  $\mu > \frac{3\lambda}{2}$ , we have a stable interior equilibrium in a discriminatory network with  $\tilde{x} \in (1/2, 1)$ . The market share of the CP with content delivery priority decreases as the ISP's capacity increases, that is,  $\frac{d\tilde{x}}{d\mu} < 0$ .

The main intuition for this result is that an increased capacity of an ISP makes congestion less important and reduces the relative quality differential (i.e., difference in waiting costs) across the two CPs. In the rest of the article, we assume that  $\mu > \frac{3\lambda}{2}$  to focus on the stable equilibrium. Our analysis thus proceeds with  $\frac{d\tilde{x}}{d\mu} < 0$ .

In the discriminatory network, the ISP's profit is given by

$$\max_{\tilde{a}} \tilde{\pi}_m = \tilde{a} + f \quad \text{s.t.} \quad v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x} - \tilde{a} \geq 0, \tag{9}$$

where  $f$  denotes the ISP's revenue from the provision of first priority to CP1. We do not specify a particular trading mechanism that determines  $f$ . Instead, we take a more general approach that can encompass various trading protocols. When both CPs compete to acquire the priority right, the winner is typically determined by the maximum willingness to pay. As each content provider knows that its market share will be  $\tilde{x}$  if it acquires the priority right but  $(1 - \tilde{x})$  if the other CP acquires the priority, its maximum willingness to pay for the priority is given by  $m_i(2\tilde{x} - 1)\lambda$ . For instance, if the priority right is sold through a first-price ascending auction, CP1 will receive the priority with  $f = m_2(2\tilde{x} - 1)\lambda$ , which is CP2's maximum willingness to pay for the right.<sup>16</sup> Alternatively, we can envision a situation in which the ISP makes sequential take-it-or-leave-it offers: the ISP makes the first offer to CP1, and if it is not accepted by CP1, it makes another offer to CP2. In such a scenario, the ISP can extract all surplus from CP1 by charging  $f = m_1(2\tilde{x} - 1)\lambda$ .

We adopt a framework that can encompass both scenarios above and the full range between them to represent different surplus divisions between the ISP and the CP that acquires the priority. To encompass the full range of bargaining protocols including the above two extremes,

<sup>14</sup> The explicit formula for  $\tilde{x}$  is given by  $\tilde{x} = \frac{1}{4\lambda}(2\mu + \lambda - \sqrt{(2\mu - \lambda)^2 - \frac{8\lambda^2}{t(\mu - \lambda)}})$ .

<sup>15</sup> See the proof of Lemma 1 in Appendix A to see that this condition ensures an interior solution.

<sup>16</sup> Economides (2008) discusses several consequences of the departure from net neutrality regulation based on the auction of prioritization through which only one group of content providers is entitled to the right to the fast lane.

let  $\theta(0 \leq \theta \leq 1)$  denote the ISP's bargaining power that measures the proportion of rent extraction from CP1.<sup>17</sup> Then, the price of the first priority will be given as

$$\begin{aligned}
 f &= \theta m_1(2\tilde{x} - 1)\lambda + (1 - \theta)m_2(2\tilde{x} - 1)\lambda \\
 &= (m_2 + \theta \Delta_m)(2\tilde{x} - 1)\lambda,
 \end{aligned}
 \tag{10}$$

where  $\Delta_m = m_1 - m_2(\geq 0)$  and  $\theta \in [0, 1]$ . As expected, the more bargaining power the ISP has, the higher the priority price will be, that is,  $\frac{\partial f}{\partial \theta} = \Delta_m(2\tilde{x} - 1)\lambda \geq 0$ .

Therefore, we can express the ISP's profit in the discriminatory network as

$$\tilde{\pi}_m^* = \left( v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x} \right) + (m_2 + \theta \Delta_m)(2\tilde{x} - 1)\lambda.
 \tag{11}$$

When the ISP assigns the priority to CP1 at the price in (10), each content provider's profit will be respectively given by

$$\tilde{\pi}_1^* = m_1\tilde{x}\lambda - (m_2 + \theta \Delta_m)(2\tilde{x} - 1)\lambda; \quad \tilde{\pi}_2^* = m_2(1 - \tilde{x})\lambda.
 \tag{12}$$

If the two margins are equal, then the extent of ISP bargaining power is irrelevant. In the discriminatory regime both CPs—the one with and the one without priority—make the same profit.

□ **The effects of net neutrality on ISPs' profits.** Here we analyze the effects of net neutrality regulation on the ISP's profits by comparing (6) and (11). We find the following potential tradeoff: without net neutrality the ISP earns less profit from consumers due to the decreased network access fee ( $a$ ), but gains from trading the priority to one content provider ( $f$ ).

*Lemma 2.* The network access fee in the discriminatory network is lower than that in the neutral network, that is,  $\tilde{a} < a^*$ .

Lemma 2 reflects Property 1 ( $w_2 > w > w_1$ ) and the associated result  $\tilde{x} \geq x^*$ . In the discriminatory network, the network access fee the ISP can charge to end users is reduced. In the absence of net neutrality regulation, the ISP will choose to introduce a two-tiered service if its revenue from the priority trade outweighs the loss from the reduced network access fee.<sup>18</sup>

*Proposition 1.* The ISP's incentives to introduce the discriminatory network can be summarized as follows.

- (i) If  $m_2 > \Lambda$ ,  $\tilde{\pi}_m^* > \pi_m^*$  for all  $\theta \in [0, 1]$ .
- (ii) If  $m_1 < \Lambda$ ,  $\tilde{\pi}_m^* < \pi_m^*$  for all  $\theta \in [0, 1]$ .
- (iii) If  $m_2 \leq \Lambda \leq m_1$ , there exists a critical level of  $\theta^* \in [0, 1]$  such that  $\tilde{\pi}_m^* > \pi_m^*$  iff  $\theta > \theta^*$ , where  $\Lambda = \frac{t}{2\lambda} + \frac{1}{(2\tilde{x}-1)\lambda} \left( \frac{1}{\mu-\tilde{x}\lambda} - \frac{1}{\mu-\lambda} \right)$ .

Proposition 1 identifies the beneficiaries and losers of net neutrality regulation. Part (i) states that the ISP's profit is always higher in the discriminatory network if the advertising revenue margins ( $m_i$ ) are sufficiently high for both CPs. By contrast, if the advertising revenue margins are relatively low for both CPs, the ISP would prefer a neutral network. Finally, if the advertising margin is high for one CP but low for the other, the relative merits of the discriminatory network vis-à-vis the neutral network depend on the ISP's bargaining power against the CPs.

The parameter  $m_i$  represents the importance of market share for each CP. As  $m_i$ 's are increased, CPs compete more aggressively to obtain the first priority in the discriminatory network. The ISP thus receives a higher price for the premium service, which can outweigh any potential loss in access fees from end users. This also implies that if both  $m_i$ s are sufficiently low or only

<sup>17</sup> We do not pin down detailed microfoundations for the bargaining process, because such an issue is not the focus of our article.

<sup>18</sup> Even if we consider the possibility that the ISP charges CPs in the net neutrality regime and the losing CP in the discriminatory regime, our qualitative results do not change.

$m_2$  is sufficiently low but the ISP's bargaining power is low, the ISP will endogenously choose the equal treatment of both content providers even if net neutrality is *not* required.

Cheng et al. (2009) also consider the effects of a discriminatory network on the ISP's profits, and it is useful to compare our results to theirs. They adopt a different pricing scheme for the sale of the priority. More specifically, Cheng et al. assume that the ISP deals with the two competing CPs in a nonexclusive way and charges a price for priority regardless of what the rival CP does. As a result, it is possible that both CPs pay but with the end result that both CPs are on equal footing without any CP having an advantage over the other. Our model, by contrast, assumes that the ISP deals with CPs exclusively. However, a direct comparison of their results to ours is somewhat difficult. The reason is that our model considers a more general trading mechanism that can encompass a variety of bargaining environments, whereas Cheng et al. consider a mechanism in which the ISP makes a nonexclusive but take-it-or-leave-it offer. Thus, in our notation, they consider only the case where  $\theta = 1$ . In Cheng et al., the ISP is always better off with a nondiscriminatory regime because it has the option of charging a small amount for "priority" to induce both CPs to pay, without affecting the access fee that the ISP can charge to consumers. This outcome, however, depends crucially on the assumption that the priority right is sold only once. Imagine a more realistic case where content providers can quickly respond by purchasing their own priority right when the delivery of their content is disadvantaged, which is possible when the priority right is sold in a nonexclusive way. In such a case, a more reasonable outcome would be a "delayed-purchase" equilibrium in which no content provider purchases the priority right, as in Fudenberg and Tirole's (1987) model of technology adoption. Then, offering priority in an exclusive way would be the optimal policy for the ISP. In Appendix B, we compare the relative profitability of exclusive and nonexclusive sale of priority rights for the ISP, assuming that the ISP can commit to a one-time sale of the priority right if selling nonexclusively.

□ **The effects of net neutrality on CPs' profits and consumer welfare.** We now turn our attention to the effects of a discriminatory network on CPs and consumers.

*Proposition 2.* The effects of a discriminatory network on CPs and consumers are as follows.

- (i) CP1:  $\tilde{\pi}_1^* > \pi_1^*$  if  $\frac{m_1}{m_2} > \frac{2(1-\theta)}{1-2\theta}$  and  $\theta \in [0, 1/2)$ ; otherwise,  $\tilde{\pi}_1^* < \pi_1^*$ .
- (ii) CP2:  $\tilde{\pi}_2^* \leq \pi_2^* \forall m_i, \theta, \lambda$ .
- (iii) Users: aggregate consumer welfare increases.

By comparing (6) and (12), we find that the high-margin CP has a higher profit in the discriminatory network if  $\theta \in [0, 1/2)$  and  $\frac{m_1}{m_2} > \frac{2(1-\theta)}{1-2\theta}$ . The best-case scenario for the high-margin CP with priority is to capture the whole market and double the profit it would receive under net neutrality. If  $\theta > 1/2$  and thus more than half of its profit is extracted, it cannot be better off under the discriminatory regime. When  $\theta < 1/2$ , the high-margin content provider who obtains the first priority can have a higher payoff in the discriminatory regime if the margin ratio between the two content providers is sufficiently large. Note that the threshold  $\frac{2(1-\theta)}{1-2\theta}$  is an increasing function of  $\theta \in [0, 1/2)$ , which implies that the margin ratio needs to be greater for the high-margin CP to earn more profit in the discriminatory network as the ISP extracts more rent from it. Because  $\frac{2(1-\theta)}{1-2\theta}$  reaches its minimum value of 2 for  $\theta = 0$ , we can conclude that the necessary condition for the high-margin CP to prefer the discriminatory network is  $\theta < 1/2$  and  $\frac{m_1}{m_2} > 2$ .

By contrast, the low-margin CP is always worse off from the introduction of priority classes. These results show the possibility that both content providers may engage in a Prisoners' Dilemma type of game to receive the first priority but end up with lower payoffs, whereas the ISP prefers the discriminatory network. Both CPs can also be worse off in Cheng et al. (2009). Such an outcome, however, takes place only when both CPs purchase the priority right, which is excluded in our model. If only one CP pays for the priority delivery, the paying CP is indifferent between the two regimes in their model.

Finally, the result in (iii) states that the end users as a group enjoy a higher surplus in the discriminatory network. This clean result stems from our Hotelling specification where an individual consumer's surplus increases linearly with the distance of her location from the marginal consumer who is indifferent between the two CPs. Hence, the aggregate consumer surplus in the discriminatory network is greater than that with net neutrality in which the two CPs share the market evenly, thereby minimizing the total consumer surplus.

□ **The effects of net neutrality on social welfare.** With the Hotelling model for the end users, social welfare analysis of two-tiered services is fairly straightforward: there is no demand effect with pricing, as long as the market is covered. However, there are three types of benefits/costs we need to compare to analyze the effects of two-tiered pricing on social welfare: (i) total margins for CPs, (ii) total transportation costs, and (iii) total delay costs. In our welfare analysis, we assume that CPs' revenue  $r_i$  from advertisers reflects the social benefits of advertising.<sup>19</sup> The following lemma examines the effects of these factors on the short-run social welfare one by one.

*Lemma 3.* (i) The total margins for CPs are larger under the discriminatory regime relative to the neutral one. (ii) The total transportation costs are higher under the discriminatory regime relative to the neutral one. (iii) The total expected waiting costs are the same in both regimes.

The series of the results in Lemma 3 is very intuitive. First, the discriminatory regime allows the high-margin content provider to expand its market share through speedier delivery of its content. As a result, efficiency in terms of margin maximization favors the discriminatory network. Second, recalling that the total transportation costs are minimized when the marginal consumer is located at the midpoint, the two-tiered pricing with  $\tilde{x} > 1/2$  is inefficient in terms of transportation cost minimization. Finally, as far as the total delay cost is concerned, we find the invariance result as shown in Appendix A. The simple reason is that the change in regimes only affects the order of services but not the total amount of services in the network. Given a fixed network capacity, the amount required to serve all users should be the same, which implies identical average waiting times across the two regimes. As a result, the overall waiting costs are irrelevant in the *static* welfare comparison. This conclusion, however, depends crucially on the assumption that competing contents have the same waiting costs. If the waiting costs differ across content, the overall waiting costs differ across the regimes (see Section 7 for more discussion).

Considering all three channels through which net neutrality can have an influence upon short-run total welfare, we can conclude that *static* welfare implications of net neutrality regulations depend on the tradeoff between transportation cost saving and inefficient production. If the margin difference is significantly large relative to the degree of product differentiation, the discriminatory network would be preferred from the viewpoint of social welfare.

*Proposition 3.* The comparison of social welfare in the short run with and without net neutrality regulation crucially depends on the relative magnitudes of the margin asymmetry across CPs ( $\Delta_m$ ) and the transportation cost parameter. If the margin difference is significantly large relative to the degree of product differentiation, the social welfare is higher in the discriminatory network, precisely, iff  $\Delta_m > \bar{t} \equiv (\tilde{x} - \frac{1}{2})\frac{t}{\lambda}$ .

The proposition implies that if the two CPs are symmetric in their margins ( $\Delta_m = 0$ ), the short-run social welfare is higher under net neutrality regulation.

## 5. Long-run analysis with investment incentives

■ The net neutrality debate centers on future investment and innovations.<sup>20</sup> In particular, one of the main issues in the debate is how the broadband operator's incentive to expand capacity in

<sup>19</sup> See Anderson and Coate (2005) for a microfoundation of such informative advertising.

<sup>20</sup> Wu (2003), for instance, states that "[t]he argument for network neutrality must be understood as a concrete expression of a system of belief about innovation."

infrastructure would be affected by allowing preferential transmission of content. ISPs such as Verizon, Comcast, and AT&T oppose network neutrality regulation and claim that such regulation would discourage their investment incentives in broadband networks. The intuition behind their claims is simple: they face an obvious free-rider problem, unless content providers who support bandwidth-intensive Internet traffic pay a premium.

We examine the validity of this claim by investigating the marginal change in the ISP’s profit with respect to the capacity parameter  $\mu$  for the two network regimes. Denote  $\Phi(\mu)$  to be the cost associated with the capacity level of  $\mu$  with  $\Phi' \geq 0$  and  $\Phi'' \geq 0$ . Then, the ISP’s choice of optimal investment will be determined at the point where the marginal benefit and the marginal cost with respect to  $\mu$  are equal, that is,  $d\pi_m/d\mu = \Phi'(\mu)$  in the neutral network and  $d\tilde{\pi}_m/d\mu = \Phi'(\mu)$  in the discriminatory network. Note that the marginal benefits of capacity expansion can be written as follows by using the results above:

$$\frac{d\pi_m}{d\mu} = \frac{da}{d\mu} = \frac{1}{(\mu - \lambda)^2} \tag{13}$$

and

$$\frac{d\tilde{\pi}_m}{d\mu} = \frac{d\tilde{a}}{d\mu} + \frac{df}{d\mu} = \left[ \frac{1}{(\mu - \tilde{x}\lambda)^2} \left( 1 - \lambda \frac{d\tilde{x}}{d\mu} \right) - t \frac{d\tilde{x}}{d\mu} \right] + 2(m_2 + \theta \Delta_m)\lambda \frac{d\tilde{x}}{d\mu}. \tag{14}$$

Therefore,

$$\begin{aligned} \frac{d\tilde{\pi}_m}{d\mu} - \frac{d\pi_m}{d\mu} &= \left( \frac{d\tilde{a}}{d\mu} - \frac{da}{d\mu} \right) + \frac{df}{d\mu} \\ &= \underbrace{\left[ \frac{1}{(\mu - \tilde{x}\lambda)^2} \left( 1 - \lambda \frac{d\tilde{x}}{d\mu} \right) - t \frac{d\tilde{x}}{d\mu} - \frac{1}{(\mu - \lambda)^2} \right]}_{\text{changes in the effect of capacity expansion}} + \underbrace{2(m_2 + \theta \Delta_m)\lambda \frac{d\tilde{x}}{d\mu}}_{\text{the effect of capacity expansion}} \end{aligned} \tag{15}$$

changes in the effect of capacity expansion on end user access fee due to discrimination (?)      the effect of capacity expansion on the sale price of priority right (-).

As can be seen from equation (15), there are two effects to consider when evaluating the relative incentives to invest in capacity across the two regimes.

First, capacity expansion affects the network access fee the ISP can charge end users, which is the willingness to pay by the marginal end users. This *network access fee effect* is represented by the expressions in the square brackets in equation (15). More specifically, in the network with net neutrality, the location of the marginal end user does not change and remains fixed at the midpoint with a change in capacity. Capacity expansion speeds up the delivery of content *uniformly*, which enables the ISP to charge more for access. This effect is captured by the last term in the square brackets,  $\frac{1}{(\mu - \lambda)^2}$ . By contrast, in the discriminatory network, capacity expansion affects the delivery speed of content *asymmetrically* across content providers, and thus also changes the location of the marginal consumer type who is indifferent between the two content providers. Such an effect of capacity expansion in the discriminatory network is captured by the first two terms in the square brackets. The first term,  $\frac{1}{(\mu - \tilde{x}\lambda)^2} (1 - \lambda \frac{d\tilde{x}}{d\mu})$ , measures the effect of capacity expansion on the consumer’s network access fee through the improved delivery speed of content. This effect can be further decomposed into two forces. The first part,  $\frac{1}{(\mu - \tilde{x}\lambda)^2}$ , measures the increase in the marginal consumer’s willingness to pay for network access when he subscribes to the CP with priority, with the demand configuration between the two CPs fixed. Note that the benefit from a larger capacity when he subscribes to the CP with priority,  $\frac{1}{(\mu - \tilde{x}\lambda)^2}$ , is less than  $\frac{1}{(\mu - \lambda)^2}$ , which is the benefit from a capacity expansion when no CP has priority. The reason is that when a CP has priority and its content is already delivered quickly, the beneficial effect of capacity expansion on delivery speed is relatively small. However, there is a secondary effect from capacity expansion that goes in the opposite direction. Whereas the demand configuration is fixed under net neutrality, capacity expansion under the discriminatory regime favors the CP without priority. (Recall Lemma 1, which shows  $\frac{d\tilde{x}}{d\mu} < 0$ .) Thus, capacity expansion induces demand reduction for the CP with priority and thus further diminishes potential congestion for the content with the

first priority. The increment in the marginal user’s willingness to pay due to this demand effect is captured by the second part,  $-\lambda \frac{1}{(\mu - \tilde{x}\lambda)^2} \frac{d\tilde{x}}{d\mu} (> 0)$ . In addition, the capacity expansion decreases the transportation cost of the marginal consumer who subscribes to the CP with priority in the discriminatory regime (once again, recall  $\frac{d\tilde{x}}{d\mu} < 0$ ). Such savings in transportation costs will also increase the marginal consumer’s willingness to pay for network access, which is captured by the second term of the square brackets,  $-t \frac{d\tilde{x}}{d\mu} (> 0)$ . Because  $\frac{1}{(\mu - \tilde{x}\lambda)^2} < \frac{1}{(\mu - \lambda)^2}$ , the sign of the square bracketed term in (15) is indeterminate, so we cannot tell unambiguously the relative size of this network access fee effect under the neutral regime and under the discriminatory regime.

Second, capacity expansion also affects the sale price of the priority right under the discriminatory regime. This *rent extraction effect*, represented by the last term in equation (15), weakens the ISP’s incentive to invest in capacity under a discriminatory network because the relative merit from first priority and thus its value is relatively small for a higher capacity level. In other words, because the congestion problem becomes less severe for higher capacity levels, the ISP’s rent from the allocation of priority classes also decreases, which in turn leads to a weaker investment incentive under a discriminatory regime.

In general, the ISP’s investment incentive hinges upon the relative magnitudes of these two potentially opposing effects. It is *a priori* ambiguous whether the ISP has greater incentive to invest in capacity in a neutral network or a discriminatory one. Nonetheless, the analysis in this article unveils what forces can make ISPs’ innovation incentives strong or weak in each regime. Although our model allows for the possibility that the ISP can have stronger investment incentives without net neutrality regulation, we cannot dismiss the possibility of the opposite. Contrary to the ISPs’ claim that net neutrality regulations would have a chilling effect on their incentive to invest, we find that net neutrality regulations could rather boost the incentive for ISP capacity expansion because it alleviates the need to acquire the priority right and hence adversely affects the ability to extract rent from content providers.

*Proposition 4.* The ISP’s relative incentive to invest in capacity in a discriminatory network vis-à-vis a neutral network depends on two effects: the rent extraction effect and the network access fee effect. The overall effect is ambiguous.

To understand the conditions under which the ISP may have more incentives to invest under net neutrality, we can rearrange the terms in (15) as

$$\begin{aligned} \frac{d\tilde{\pi}_m}{d\mu} - \frac{d\pi_m}{d\mu} = & \underbrace{\left[ \frac{1}{(\mu - \tilde{x}\lambda)^2} - \frac{1}{(\mu - \lambda)^2} \right]}_{\text{direct effects of capacity expansion with market shares fixed (-)}} \\ & + \underbrace{\left[ 2(m_2 + \theta \Delta_m) - \frac{1}{(\mu - \tilde{x}\lambda)^2} - \frac{t}{\lambda} \right]}_{\equiv \Sigma} \lambda \underbrace{\frac{d\tilde{x}}{d\mu}}_{(-)} \end{aligned} \tag{16}$$

indirect effects of capacity expansion through changes in market shares.

The terms in the first square brackets represent the direct effects of capacity expansion on the difference in the ISP’s profits between the two regimes with the CPs’ market shares *fixed*, which is always negative. The remaining terms represent the effects of capacity expansion through the *induced changes* in the CPs’ market shares. Thus, a sufficient condition for the ISP to have higher incentives to invest in capacity is that  $\Sigma \geq 0$ . Thus, as CPs’ margins are higher and the ISP’s bargaining power becomes stronger, it becomes more likely that the ISP will have more incentives to invest in a neutral network. The reason is that the rent extraction motives are stronger under

such a situation. Another scenario under which the ISP may have higher incentives to invest in capacity occurs when  $\frac{d\tilde{x}}{d\mu}$  is close to zero. This scenario takes place if the product differentiation parameter  $t$  is sufficiently high so that the indirect effects through changes in market shares are negligible. In this scenario, the direct effect dominates and the ISP will have higher incentives to invest under net neutrality.

## 6. Net neutrality and CPs' investment incentives

■ So far, our analysis has dealt only with investment incentives of ISPs. As pointed out in Von Hippel (2005), proponents of net neutrality regulation maintain that so-called killer applications have been developed at the “edges” of the network by users, not by the “core” of network operators. Thus, another important element in the net neutrality debate is investment incentives for content providers.

A typical concern about the so-called holdup problem is that part of the return from one party's relationship-specific investments is *ex post* expropriable by his trading partner. Such concerns arise when we consider the content service providers' investments: the monopolistic ISP could expropriate any investments made by content providers. The *ex post* optimal policy for the ISP to discriminate may not be optimal from an *ex ante* investment incentive viewpoint. Thus, an interesting question to ask is whether regulation is required as a mechanism to bind the ISP to net neutrality in order to maintain the content providers' incentives to invest.<sup>21</sup>

In order to examine the effect of the discriminatory network on the content providers' R&D incentives, let us assume that a higher margin is achieved with a higher investment cost.<sup>22</sup> An irreversible investment in margin-improving R&D is characterized by a twice-differentiable function  $\Psi(\Delta_i)$  with  $\Psi' > 0$ ,  $\Psi'' > 0$ , where  $\Delta_i$  denotes the magnitude of the margin enhancement from investing, that is,  $\Delta_i = m_i - \bar{m}_i$ . We can think of  $\bar{m}_i$  as the current margin with the best technology that is freely available to content provider  $i$ , and  $m_i$  as the postinvestment margin level for  $i = 1, 2$ .

In a neutral network, each unit of margin improvement by a content provider increases its profit by  $\lambda/2$ , which is readily seen from (6). This is because there is no demand effect of margin-improving investment in the neutral network. Thus, each content provider's optimal investment in margin-improving R&D is determined by the marginal benefit-cost comparison,

$$\Psi'(\Delta_i^*) = \frac{\lambda}{2} \quad \text{for } i = 1, 2. \quad (17)$$

Similarly, in a discriminatory network, where CPs decide their investment levels followed by the competition for priority, each content provider chooses its optimal investment at the point where the marginal revenue from margin improvement is equalized to the marginal cost. One complication in a discriminatory network, however, is that the marginal benefit from investment for a CP depends on whether or not it receives the priority, which is endogenously determined by the comparison of CPs' *ex post* margins after investment. If the initial margin difference between CP1 and CP2 is sufficiently large, the unique equilibrium entails that CP1 always has a higher *ex post* margin and thus receives the priority. In this case, the high-margin content provider earns a profit of  $\tilde{\pi}_1^* = m_1 \tilde{x}_1 - f$  where  $f$  was defined in (10), and the low-margin content provider is not affected by the ISP's rent extraction. Thus, content providers' optimal investments are given by

<sup>21</sup> DeGraba (1990) presents a model to study how price discrimination in a market for a variable input affects downstream producers' long-run choices of a production technology. He shows that a monopoly supplier of a variable input will charge the low-cost downstream producer a higher price than the high-cost producer under price discrimination, and thus the downstream producers will end up choosing technology with a higher marginal cost with price discrimination than under uniform pricing, which results in a lower welfare in the long run under discriminatory pricing. Using similar reasoning, the literature on the most favored nations (MFN) clause in international trade also suggests that discriminatory or preferential tariffs rather than uniform tariffs would have a more adverse effect on investment incentives of foreign producers (Choi, 1995).

<sup>22</sup> The investment can be either revenue enhancing or cost reducing.

$\tilde{\Delta}_1^* = \bar{\Delta}$  and  $\tilde{\Delta}_2^* = \underline{\Delta}$ , where  $\bar{\Delta} \geq \underline{\Delta}$  (equality holding when  $\theta = 1$ ) with  $\bar{\Delta}$  and  $\underline{\Delta}$  being defined by

$$\Psi'(\bar{\Delta}) = (\tilde{x} - \theta(2\tilde{x} - 1))\lambda \quad \text{and} \quad \Psi'(\underline{\Delta}) = (1 - \tilde{x})\lambda. \tag{18}$$

If the initial margin difference is small, we cannot rule out the possibility that the *ex post* margin ranking can be reversed in equilibrium. However, the qualitative results do not change, as demonstrated below.

*Lemma 4.* Under the discriminatory network, one CP invests at the level of  $\bar{\Delta}$  whereas the other CP invests at the level of  $\underline{\Delta}$  in any pure strategy equilibrium.

Thus, in any pure strategy equilibrium, one CP invests  $\bar{\Delta}$  whereas the other invests  $\underline{\Delta}$ .<sup>23</sup> In other words, the identity of the CP that invests more and receives the priority may change, but the overall equilibrium investment levels do not change in a discriminatory network. Thus, we will focus on the pure strategy equilibrium where the high-margin firm invests more and retains the priority *ex post*. By comparing optimal investments under a neutral network with those under a discriminatory one, we derive the following results.

*Proposition 5.* The high-margin content provider will choose a technology with a lower margin under the discriminatory network than it will under the neutral network, that is,  $\tilde{\Delta}_1^* < \Delta_1^*$ , if and only if the ISP’s expropriation is high enough to the extent of  $\theta > 1/2$ . Otherwise (if  $0 \leq \theta \leq 1/2$ ), we have  $\tilde{\Delta}_1^* \geq \Delta_1^*$ . The low-margin content provider always chooses a technology with a lower margin under the discriminatory network, that is,  $\tilde{\Delta}_2^* < \Delta_2^*$ .

As expected, the optimal investment level for the high-margin content provider is inversely related to the ISP’s ability to extract rent from use of the fast lane. Suppose that the right to the premium service is traded through the first-price bid auction, that is,  $\theta = 0$ . Then, the high-margin CP’s profit is constrained only by the low-margin CP’s willingness to pay for the priority service. Because the high-margin CP’s margin improvement applies to larger market coverage in the discriminatory network relative to in the neutral network, the high-margin CP will have a stronger investment incentive in a discriminatory regime. Therefore, the high-margin CP chooses a technology with a higher margin under a discriminatory regime than under a neutral regime. Such merit, however, gradually decreases as  $\theta$  increases. Eventually, for a sufficiently large rent extraction (for  $\theta > 1/2$ ), the high-margin content provider’s investment incentive becomes weaker under the discriminatory regime due to rent extraction from the ISP.<sup>24</sup>

On the other hand, the low-margin content provider will always choose a technology with a lower margin under a discriminatory regime for any  $\theta \in [0, 1]$ . This is because the low-margin content provider always has a smaller market share in the discriminatory network than in the neutral network. This implies that the ISP may have a dynamic inconsistency problem when CPs have opportunities to invest either in revenue-enhancing or cost-reducing R&D that improves their margins. For instance, we can imagine a situation in which  $\bar{m}_1 < \Lambda$ , and thus the ISP *ex ante* prefers a neutral network (see Proposition 1), but once the innovation takes place the ISP prefers to switch to a discriminatory network, that is,  $\bar{m}_2 + \Delta_2^* > \Lambda$ . This situation would apply to an emerging technology for which the initial margins are low but the potential for technology improvement is huge. Of course, if such *ex post* opportunistic behavior by the ISP is anticipated by

<sup>23</sup> There can also be a mixed-strategy equilibrium at the investment stage where the two firms randomize between  $\bar{\Delta}$  and  $\underline{\Delta}$ .

<sup>24</sup> Our analysis has assumed that the CP’s investment raises its margin. Alternatively, we can consider a scenario in which CP investment raises the consumers’ gross utility from content service  $v$ , which we assumed to be constant and symmetric across CPs. This alternative specification complicates the analysis because we need to account for the channel through which investment in quality helps the CP gain market share. Nonetheless, we can derive the same qualitative results as in Proposition 5.

the CPs, their investment will be adversely affected for fear of rent extraction.<sup>25</sup> In such a scenario, net neutrality regulation may be needed to restore CPs’ investment incentives. This result accords well with the fact that content providers in general are the proponents of net neutrality regulation and the crux of their main argument concerns their innovation incentives.

The discussion above also has some implications for the optimal degree of rent extraction in the discriminatory network from the ISP’s long-run perspective. The ISP has the following intertemporal tradeoffs. First, the ISP prefers a larger rent extraction (higher  $\theta$ ) in the short-run because of a higher surplus from trading the priority. Had we considered this short-run *direct* effect only, the most desirable situation for the ISP is total rent extraction, that is,  $\theta = 1$  with  $\frac{\partial f}{\partial \theta} \geq 0$ .

From the long-run perspective, however, such total extraction may not be the best option. This is because an increase in its rent extraction can generate the adverse dynamic effect of lowering the high-margin content provider’s investment incentive for a higher  $\theta$ , which in turn can decrease the ISP’s long-run revenue from trading the priority. Therefore, the ISP’s optimal level of rent extraction will be determined by these intertemporal trade-offs. To put it mathematically, the overall effect of  $\theta$  on the ISP’s long-run profit is evaluated as

$$\frac{d\tilde{\pi}_m^*}{d\theta} = \frac{\partial \tilde{\pi}_m^*}{\partial \theta} + \frac{\partial \tilde{\pi}_m^*}{\partial \tilde{\Delta}_1^*} \frac{\partial \tilde{\Delta}_1^*}{\partial \theta}, \tag{19}$$

(+)    (+)    (-)

where the first term captures the direct rent extraction effect and the second term represents the indirect effect through CPs’ investment incentives. Needless to say, the ISP will choose  $\theta$  by  $\frac{d\tilde{\pi}_m^*}{d\theta} = 0$ . For an explicit solution, if we consider a quadratic function  $\Psi(\Delta_i) = \Delta_i^2/2k$ , where  $k$  is a cost efficiency parameter in the investment, then the optimal level of  $\theta$ , denoted by  $\tilde{\theta}$ , is derived in the following proposition.

*Proposition 6.* The ISP’s long-run profit is maximized at  $\tilde{\theta} = \frac{\Delta_m}{(2\tilde{\alpha}-1)k\lambda}$ . The ISP does not prefer full rent extraction if  $\tilde{\theta} = \frac{\Delta_m}{(2\tilde{\alpha}-1)k\lambda} \in [0, 1)$ . The ISP’s optimal level of rent extraction is decreasing in the CPs’ R&D efficiency, but increasing in the margin differential, that is,  $\frac{\partial \tilde{\theta}}{\partial k} < 0$  and  $\frac{\partial \tilde{\theta}}{\partial \Delta_m} > 0$ .

As the content provider’s R&D process is more efficient (or as parameter  $k$  increases), the adverse effect of the ISP’s rent extraction on the high-margin content provider’s innovation incentive becomes greater, with all other things being equal. Thus, the ISP’s preferred level of rent extraction becomes relatively small. In addition, if the cost differential between the two content providers increases, the ISP will have a stronger incentive to extract more rent from content providers due to the short-run direct effect, *ceteris paribus*.

## 7. Discussion and extensions

□ **Heterogeneity in delay costs.** In the basic model, we assumed that the waiting costs due to congestion are identical across content. However, content and applications differ in their sensitivity with respect to delay in delivery. In general, data applications such as email can be relatively insensitive towards moderate delivery delays from the users’ viewpoint. By contrast, streaming video/audio or voice over Internet protocol (VoIP) applications can be very sensitive to delay, leading to jittery delivery of content. With such heterogeneity concerning delay costs, one may argue that network neutrality treating all packets equally regardless of content is not an efficient way to utilize the network in the presence of capacity constraints. It also has been claimed by opponents of net neutrality regulation that the imposition of net neutrality requirements may impede the development of time-sensitive applications such as remote medical supervision.

<sup>25</sup> One question that can be asked is why the ISP and the CPs cannot write a contract to solve this holdup problem. We can easily imagine a situation in which the magnitude of margin improvements is observable by the ISP, but not verifiable in court. Then it cannot be included in the contract and the holdup problem cannot be mitigated through an appropriate contract.

To investigate these issues, the model needs to be modified to allow the possibility of different waiting costs across applications. More specifically, let us assume  $\tau$  to be the waiting cost per unit time for the high-margin content service that would be provided through the fast lane, whereas that for the low-margin content service is still set to one per unit time for consistency with the analysis thus far. Because we are particularly interested in the case where the content with higher waiting costs is given priority and delivered first, we focus our attention on the case of  $\tau \geq 1$ .<sup>26</sup>

The marginal consumer who is indifferent between the two content services under the neutrality regime, denoted by  $x^{**}$ , is given by

$$v - \frac{1}{\mu - \lambda} \tau - tx^{**} - a = v - \frac{1}{\mu - \lambda} - t(1 - x^{**}) - a. \tag{20}$$

Thus, we have

$$x^{**} = \frac{1}{2} - \frac{\tau - 1}{2t(\mu - \lambda)} \leq x^* = 1/2, \tag{21}$$

which means that under net neutrality the demand for the content with higher waiting costs decreases compared to the case of identical waiting costs. Similarly, under the discriminatory regime the location of the marginal consumer will be given by<sup>27</sup>

$$\tilde{x} = \frac{1}{2} - \frac{\tau(\mu - \lambda) - \mu}{2t(\mu - \lambda)(\mu - \tilde{x}\lambda)}. \tag{22}$$

By comparing  $\tilde{x}$  and  $x^{**}$ , we find that the high-margin content provider always faces a higher demand for its content service with the first priority relative to in the neutral network, that is,  $\tilde{x} > x^{**}$  for any  $\tau \geq 1$ . This finding can be readily derived by the fact that the difference between  $\tilde{x}$  and  $x^{**}$ ,  $\tilde{x} - x^{**}$ , increases in  $\tau$  and that  $(\tilde{x} - x^{**})|_{\tau=1} = \tilde{x} - x^* > 0$ . Therefore, the qualitative results derived with identical waiting costs are quite robust to the relaxation of this assumption *except* with respect to the comparison of social welfare in the short run with and without net neutrality.

Now that there is the asymmetry in waiting costs across content services, Lemma 3 (iii) does not hold anymore. Assigning priority to content with high waiting costs is certainly beneficial in reducing total waiting costs. However, we cannot conclude that we have lower total waiting costs under the discriminatory regime relative to those under the neutral regime. That would be true if the market shares between the two CPs were the same across the regimes. However, giving priority to CP1's content in a discriminatory regime leads to a higher market share of content with higher waiting costs ( $\tilde{x} > x^{**}$ ). This indirect market demand effect can offset the direct effect of allocating priority service to content with higher waiting costs. In fact, our simulation exercises indicate that in most cases the induced demand effect dominates the direct effect, and thus the total waiting costs in fact increase with priority service.

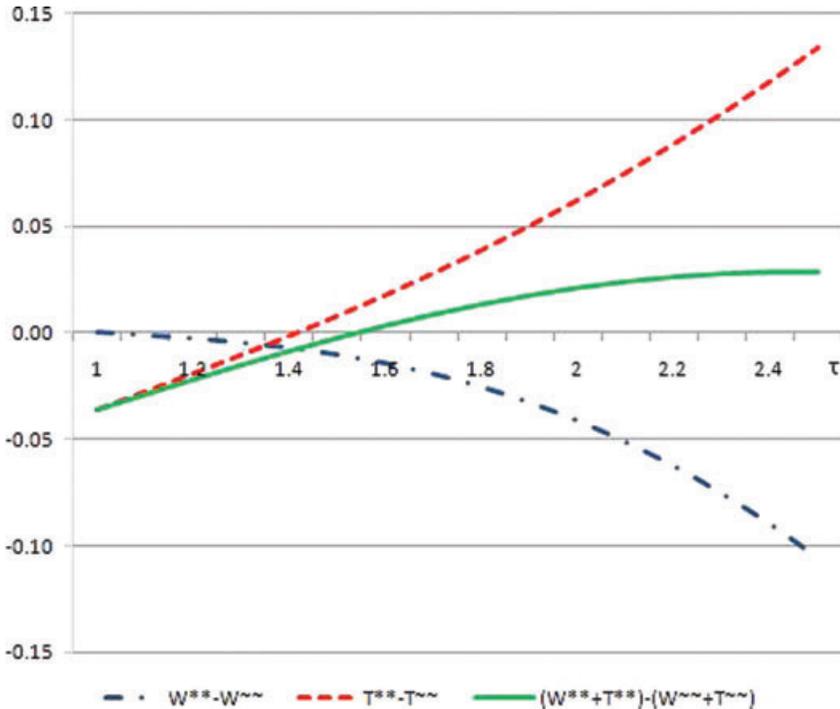
However, this does not imply that priority service in the discriminatory network reduces social welfare. As the asymmetry in waiting costs ( $\tau$ ) increases, the total transportation costs favor the discriminatory network compared to the neutral network. To see this, note that as  $\tau$  is increased from 1, the total transportation costs start to increase in the neutral regime ( $x^{**}$  departs further from 1/2) whereas they decrease in a discriminatory regime ( $\tilde{x}$  converges to 1/2 as  $\tau$  reaches  $\frac{\mu}{\mu - \lambda}$ ). Beyond the point of  $\tau = \frac{\mu}{\mu - \lambda}$ , the total transportation costs increase in both regimes, but still the transportation costs increase at a faster rate in the neutral network. As a result, the short-run welfare comparison may move toward favoring the introduction of two-tiered services in the presence of heterogeneity in delay costs across content. Thus, one may argue that

<sup>26</sup> The high-margin CP is willing to pay more for priority, but this does not necessarily mean that the CP with the greater delay cost will purchase the priority. If  $\tau < 1$ , it is possible that the content with the smaller delay cost is delivered with first priority.

<sup>27</sup> The explicit formula for  $\tilde{x}$  is given by  $\tilde{x} = \frac{1}{4t}((\lambda + 2\mu) - \sqrt{(2\mu - \lambda)^2 + \frac{8\lambda(\tau(\mu - \lambda) - \mu)}{t(\mu - \lambda)}})$ .

FIGURE 1

THE EFFECTS OF WAITING-COST ASYMMETRY ON SOCIAL WELFARE



network neutrality, which treats all packets equally regardless of content, is not an efficient way to utilize the network in the presence of heterogeneity in delay costs.

To illustrate our discussion above, we provide a simulation result. By Proposition 3, we know that social welfare is higher in the neutral network with the assumption that both CPs have the same margin ( $m_1 = m_2$ ) for their click-throughs. Our simulation result shows that the social welfare can be higher in a discriminatory regime if  $\tau$  is sufficiently large even with  $m_1 = m_2$ , which confirms our intuition.<sup>28</sup>

Figure 1, based on simulation results in Table 1, illustrates that a discriminatory network may perform better (i.e., incur lower total costs) relative to a neutral network for a sufficiently large  $\tau$  due to the effects discussed above.

□ **Possibility of quality degradation.** One interesting implication of our analysis is that degrading the nonpriority packet may be necessary to extract rent more effectively and thus restore the ISP's incentives to invest in the discriminatory network. So far, to our best understanding, the opponents of net neutrality have claimed that they have no incentive for degradation even under the discriminatory regime.<sup>29</sup> Let us discuss this inquiry in our framework of whether the ISP has incentives to degrade the quality of nonpriority packets—deliberately slow down their delivery speed—for the purpose of extracting rent more effectively and restoring incentives to invest in the discriminatory network. Quality degradation for basic service can be easily modelled by allowing the ISP to choose a waiting time higher than  $w_2$  in (3) for nonpriority packets.

<sup>28</sup> If content with higher waiting costs provides a higher margin, this will further favor a discriminatory network.

<sup>29</sup> This may be due to the ISP's fear of public backlash that would provide impetus for net neutrality regulation once such an intention is revealed. For incentives to degrade the quality of a subset of products, see Deneckere and McAfee (1996) and Hahn (2006). See also Cremer, Rey, and Tirole (2000) for incentives to selectively degrade the quality of interconnection between Internet backbone providers.

**TABLE 1** A Numerical Simulation the Effects of Waiting-Cost Asymmetry

$\tau$	Demands			Waiting Costs		Transport Costs		Difference in Total Costs $(W^{**} + T^{**}) - (\tilde{W} + \tilde{T})$
	$\tilde{X}$	$X^{**}$	$\tilde{X} - X^{**}$	$\tilde{W}$	$W^{**}$	$\tilde{T}$	$T^{**}$	
1.0	0.691	0.500	0.191	0.500	0.500	0.286	0.250	-0.036
1.2	0.648	0.450	0.198	0.548	0.545	0.272	0.253	-0.022
1.4	0.608	0.400	0.208	0.587	0.580	0.262	0.260	-0.009
1.6	0.570	0.350	0.220	0.620	0.605	0.255	0.273	0.003
1.8	0.534	0.300	0.234	0.646	0.620	0.251	0.290	0.013
2.0	0.500	0.250	0.250	0.667	0.625	0.250	0.313	0.021
2.2	0.467	0.200	0.267	0.683	0.620	0.251	0.340	0.026
2.4	0.436	0.150	0.286	0.695	0.605	0.254	0.373	0.028

Note: All parameters satisfy stable market-sharing equilibrium conditions ( $\mu = 4, \lambda = 2, t = 1$ ).

We find that the ISP can have incentive to do quality degradation in the discriminatory network, but not in the neutral network. This is because in the neutral network the ISP’s quality degradation only decreases the network access fee without yielding a higher rent extraction. In addition, as is seen from (11), the high-margin content provider will have a larger market share with such quality degradation than without it. The enlarged asymmetry in the demands for content can make the ISP earn more from the trade of the first priority, but reduces the ISP’s revenue from the network access fee. As long as the former effect outweighs the latter, the possibility of quality degradation would make the discriminatory network more profitable for ISPs.

Once again, a question of interest is how the possibility of quality degradation affects the investment incentives of the ISP. With the possibility of quality degradation, the ISP need not be concerned anymore about the rent extraction effect that adversely affects its investment incentives to capacity expansion. Because the ISP is now free of the problem that the relative quality difference between the two CPs decreases as capacity expands, the possibility of quality degradation would increase ISPs’ incentives to expand capacity.

□ **Integration/strategic alliance of ISPS and CPs.** Another important issue in the debate on net neutrality is the impact of integration of ISPs and content providers on market competition and innovation incentives. One concern expressed by net neutrality proponents is the possibility that the integrated ISPs may confer unfair advantages to their own content over content provided by competitors. Consider, for instance, a recent merger of AT&T with SBC which has a partnership with Yahoo. The question is whether AT&T would have an incentive to give its partner Yahoo site preferential treatment over competing sites such as Google in the absence of net neutrality regulations. To address this question, we need to analyze whether the ISP may have incentives to offer the first priority to the affiliated content provider over the nonaffiliated one.

In our simple model, it turns out that under net neutrality, vertical integration has no impact on allocation of resources either in the short run or in the long run. Therefore, there is no antitrust concern about vertical mergers between the ISP and CP; if there is a vertical merger, it is driven by efficiency reasons. Even without net neutrality, it can be shown that the allocation of the first priority is the same across different vertical structures in that the high-margin CP always receives the first priority. Therefore, the concern that the ISP may give its own sister division preferential treatment over competing sites is unfounded, at least in the short run.

However, a vertical integration in the discriminatory regime can have impacts on the ISP’s capacity investment. To see this, let us consider a vertical merger between the ISP and the high-margin CP, and denote the merged firm’s profit as  $\tilde{\Pi} = \tilde{a} + m_1 \tilde{x} \lambda$ , where  $\tilde{a} = v - \frac{1}{\mu - \tilde{x} \lambda} - t \tilde{x}$ . Then, the merged firm’s investment incentives can be expressed as

$$\frac{d\tilde{\Pi}}{d\mu} = \frac{d\tilde{a}}{d\mu} + m_1 \lambda \frac{d\tilde{x}}{d\mu}. \tag{23}$$

Notice that the merged firm's investment incentives in capacity ( $\mu$ ) do not depend on  $\theta$ , because the sale of the first priority is internal to the organization.<sup>30</sup> By comparing (23) and (14), the comparison of investment incentives with vertical integration and without vertical integration depends on the relative magnitude of  $2(m_2 + \theta \Delta_m)$  and  $m_1$ . Noting that  $\frac{d\tilde{x}}{d\mu} < 0$ , the ISP's investment incentives with vertical integration are higher than those under no vertical integration if  $\theta$  is sufficiently high, precisely,  $\theta > \theta_I \equiv \frac{1}{2} - \frac{m_2}{2\Delta_m}$ . The reason is that with vertical integration the ISP does not need to deliberately limit its capacity in an effort to command a higher sale price for the first priority. However, if  $\theta$  is sufficiently small, the result can be reversed. More specifically, if  $\theta < \theta_I$ , an independent ISP has greater incentive to invest than a vertically merged one. Note that this condition is identical to the one ensuring that CP1 benefits from a discriminatory regime. This condition holds when the independent ISP's ability to extract rent from the sale of the first priority is limited, and thus the ISP does not fully internalize the negative impact of capacity investment on the relative value of first priority. Once integrated, it fully internalizes its impact on CP1's profit and thus limits its investment to confer advantage to its own CP division.

Vertical integration can also alleviate the holdup problem under the discriminatory regime. It thus could be an alternative way to solve the holdup problem if the ISP cannot commit to net neutrality. When  $\theta$  is high, vertical integration increases both the ISP's and vertically integrated CP's investments.

## 8. Concluding remarks

■ This article provides an economic analysis of net neutrality regulation. In particular, our analysis focuses on the effects of net neutrality regulation on the investment incentives of Internet service providers and content providers as well as on social welfare. To address these questions, we use a simple model based on the queuing theory to capture the congestion in the network. We have shown that the ISP's incentives to invest in a multitiered network vis-à-vis in a nondiscriminatory network under net neutrality regulation depends on a potential tradeoff between the two sides of the market: the network access fee from end users and the revenue from content providers through the potential trade of the first priority in delivery. We also compare the CPs' incentives to invest in cost reduction/quality enhancement as well as social welfare across different regulatory regimes. We find that the relationship between the net neutrality regulation and investment incentives is subtle. Even though we cannot draw general unambiguous conclusions, we identified key effects that are expected to play important roles in the assessment of net neutrality regulations.

We conclude by mentioning some limitations of our simple model and discussing potential avenues for future research. First, we note that the model in the previous sections made many simplifying assumptions with regard to pricing strategies of several players. For instance, we assumed away the ability of content providers to charge end users directly. Consideration of these possibilities considerably complicates the analysis. In this regard, the burgeoning literature on two-sided markets may be useful in further analyzing these issues (see Armstrong, 2006 and Rochet and Tirole, 2006 for details). In the framework of two-sided markets, ISPs will play the role of platforms that provide a link between content providers and end users. Caillaud and Jullien (2003), for instance, show that the equilibrium in two-sided markets depends crucially on the pricing scheme used. Thus, it would be important to analyze the implications of allowing a more sophisticated pricing scheme in this model. In particular, it would be an important extension to allow competition between content providers when micropayments between content providers and consumers are possible.

Second, one may consider introducing diversity in the types of investments that can be made by content providers. More specifically, we can imagine two types of investments: firm-specific

<sup>30</sup> If the merger took place between the ISP and the high-cost CP, the incentive to invest will depend on  $\theta$ .

investments, whose effects are limited to the investing content providers, and investments that have spillover effects. For the first type of investment, we can think of investments that enhance the value of content or reduce the cost of content provision. For the second type, we can consider an investment in compression technology, which not only reduces the delivery speed of the investor's content but relieves congestion in the network and helps delivery speed of other content providers. Net neutrality regulation may have a differential effect across different types of investments and impact the choice of investment.

Finally, our basic framework assumes that the ISP market is characterized by monopoly power. This is a reasonable approximation in many geographical markets. However, it is not the only market condition prevailing. One important extension of the model would be to introduce competition in the ISP market and analyze how the effects of net regulation can play out. Most concerns expressed by net neutrality proponents are rooted in monopoly power and concentration in the ISP market. One important policy question would be whether the presence of competition in the ISP market can mitigate any problems associated with discrimination and make net neutrality regulation irrelevant.

### Appendix A

□ **Proofs for lemmas and propositions**

*Proof of Lemma 1.* As more consumers subscribe to the CP with the first priority, the waiting costs for both types of CPs increase, but the marginal effect on the waiting cost for the nonpriority CP is greater. As a result, we may end up with a situation in which all consumers subscribe to the CP with the first priority. Ironically, in this outcome no one has priority because everyone is treated equally within the priority class. To prevent this outcome from prevailing, we require the two CPs to be sufficiently differentiated. An interior solution with market sharing requires  $\tilde{x} < 1$ , that is,

$$v - a - \Gamma_1(\tilde{x} = 1) < v - a - \Gamma_2(\tilde{x} = 1), \tag{A1}$$

where  $\Gamma_1(\tilde{x}) = \frac{1}{\mu - \tilde{x}\lambda} + t\tilde{x}$  and  $\Gamma_2(\tilde{x}) = \frac{\mu}{\mu - \lambda} \frac{1}{\mu - \tilde{x}\lambda} + t(1 - \tilde{x})$ . This condition is satisfied if  $t > \frac{\lambda}{(\mu - \lambda)^2}$ .

Moreover, for the adjustment process to yield a stable (interior) equilibrium, we need to assume

$$\Gamma'_1(\tilde{x}) > \Gamma'_2(\tilde{x}) \text{ for all } \tilde{x} \in [1/2, 1]. \tag{A2}$$

By taking the derivatives of  $\Gamma_1(\tilde{x})$  and  $\Gamma_2(\tilde{x})$ , we can explicitly write condition (A2) as

$$\frac{\lambda^2}{(\mu - \lambda)(\mu - \tilde{x}\lambda)^2} < 2t. \tag{A3}$$

Because the left-hand side of (A3) can be alternatively expressed as

$$\frac{\lambda^2}{(\mu - \lambda)(\mu - \tilde{x}\lambda)^2} = \frac{\lambda}{(\mu - \tilde{x}\lambda)} \frac{\lambda}{(\mu - \lambda)(\mu - \tilde{x}\lambda)} = \frac{\lambda}{(\mu - \tilde{x}\lambda)} (2\tilde{x} - 1)t, \tag{A4}$$

where the last equality comes from equation (7) which defines  $\tilde{x}$ , the stability condition holds if the inequality  $\frac{\lambda}{(\mu - \tilde{x}\lambda)} (2\tilde{x} - 1) < 2$  is met. Because  $\frac{\lambda}{(\mu - \tilde{x}\lambda)} (2\tilde{x} - 1)$  is increasing in  $\tilde{x}$  of which the maximum value is one, it can easily be seen that the above inequality is satisfied even for  $\tilde{x} = 1$ , if  $\mu > \frac{3}{2}\lambda$ . On the other hand, by totally differentiating (7), we find the following relationship of

$$\text{sign}\left(\frac{d\tilde{x}}{d\mu}\right) = \text{sign}\left(\frac{\lambda^2}{(\mu - \lambda)(\mu - \tilde{x}\lambda)^2} - 2t\right).$$

Hence, with the assumption of  $\mu > \frac{3\lambda}{2}$ , we have  $\frac{d\tilde{x}}{d\mu} < 0$  from (A3).

*Proof of Lemma 2.* Note that  $a^* = v - \frac{1}{\mu - \lambda} - \frac{1}{2}t$  and  $\tilde{a}^* = v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x}$ . The difference between network access fees is given by

$$a^* - \tilde{a} = \frac{1}{\mu - \tilde{x}\lambda} - \frac{1}{\mu - \lambda} + t\left(\tilde{x} - \frac{1}{2}\right). \tag{A5}$$

Because  $\tilde{x}$  defined by (7) satisfies the equality of

$$t(2\tilde{x} - 1) = \frac{1}{\mu - \tilde{x}\lambda} \frac{\mu}{\mu - \lambda}, \tag{A6}$$

we can prove the given result by dividing (A6) by 2 and substituting (A6) into (A5) as

$$a^* - \tilde{a} = \frac{1}{\mu - \tilde{x}\lambda} - \frac{1}{\mu - \lambda} + \frac{1}{2} \frac{1}{\mu - \tilde{x}\lambda} \frac{\mu}{\mu - \lambda} = \frac{(2\tilde{x} - 1)\lambda}{2(\mu - \tilde{x}\lambda)(\mu - \lambda)} > 0 \because \tilde{x} > 1/2.$$

*Proof of Lemma 3.* (i) Let  $\tilde{M}$  and  $M$  denote the total margin in a discriminatory and in the neutral network, respectively. Then, the difference of these two is

$$\begin{aligned} \Delta M &\equiv \tilde{M} - M = (\tilde{x}m_1 + (1 - \tilde{x})m_2)\lambda - \frac{m_1 + m_2}{2}\lambda \\ &= \left(\tilde{x} - \frac{1}{2}\right)\Delta_m\lambda \geq 0 \quad \because \tilde{x} > 1/2 \text{ and } \Delta_m \geq 0. \end{aligned}$$

(ii) Let  $\tilde{T}$  and  $T$  denote transaction cost in a discriminatory and in the neutral network, respectively. Then, it is easily seen that

$$\Delta T \equiv \tilde{T} - T = \left(\int_0^{\tilde{x}} tx \, dx + \int_{\tilde{x}}^1 t(1-x) \, dx\right) - \frac{t}{4} = \left(\tilde{x} - \frac{1}{2}\right)^2 t \geq 0.$$

(iii) Let  $\tilde{W}$  and  $W$  denote the total waiting cost in a discriminatory and in the neutral network, respectively. Recall that the expected waiting cost for each end user in the neutral network is given by  $w = \frac{1}{\mu - \lambda}$ . With the total mass of end users normalized to one, we have that  $W = w = \frac{1}{\mu - \lambda}$ . By contrast,  $\tilde{W}$  is calculated as the weighted average of  $w_1 = \frac{1}{\mu - \lambda_1}$  and  $w_2 = \frac{\mu}{\mu - \lambda} w_1$ , with weights  $\tilde{x}$  and  $(1 - \tilde{x})$ , respectively. Using  $\lambda_1 = \tilde{x}\lambda$ , we find that

$$\begin{aligned} \tilde{W} &= \tilde{x}w_1 + (1 - \tilde{x})w_2 = \frac{\lambda_1}{\lambda} \frac{1}{\mu - \lambda_1} + \left(1 - \frac{\lambda_1}{\lambda}\right) \frac{\mu}{\mu - \lambda} \frac{1}{\mu - \lambda_1} \\ &= \frac{\lambda(\mu - \lambda_1)}{\lambda(\mu - \lambda_1)(\mu - \lambda)} = \frac{1}{\mu - \lambda} = W. \end{aligned}$$

*Proof of Lemma 4.* Consider an equilibrium in which CP1 has a higher *ex post* margin (that is,  $\bar{m}_1 + \Delta_1 > \bar{m}_2 + \Delta_2$ ) and receives the priority. In such a case, CP1 solves the following problem:

$$\max_{\Delta_1} (\bar{m}_1 + \Delta_1)\tilde{x}\lambda - [(\bar{m}_2 + \Delta_2) + \theta((\bar{m}_1 + \Delta_1) - (\bar{m}_2 + \Delta_2))](2\tilde{x} - 1)\lambda - \Psi(\Delta_1).$$

Then, it is clear that CP1's optimal investment level for this case is given by  $\bar{\Delta}$  by the first-order condition. Similarly, CP2's optimal investment problem is given by

$$\max_{\Delta_2} (\bar{m}_2 + \Delta_2)(1 - \tilde{x})\lambda - \Psi(\Delta_2).$$

Thus, CP2's optimal investment level is given by  $\underline{\Delta}$ . We have derived these equilibrium investment levels assuming that CP1 receives the priority.

For the investment levels  $(\bar{\Delta}, \underline{\Delta})$  to be sustainable as an equilibrium, we need to check whether CP2 will have an incentive to invest up to the level where it will be the firm that receives the priority with a higher *ex post* margin, given that CP1 invests at  $\bar{\Delta}$ . This deviation requires an investment level of at least  $\bar{\Delta} + (\bar{m}_1 - \bar{m}_2)$  by CP2. Because we know that  $\bar{\Delta}$  is the optimal investment level for the firm that acquires the priority right and the objective function is concave, the optimal deviation investment level is given by  $\bar{\Delta} + (\bar{m}_1 - \bar{m}_2)$ . The optimal deviation payoff for CP2 is thus given by

$$\tilde{\pi}_2^d = (\bar{m}_1 + \bar{\Delta})\tilde{x}\lambda - (\bar{m}_1 + \bar{\Delta})(2\tilde{x} - 1)\lambda - \Psi(\bar{\Delta} + (\bar{m}_1 - \bar{m}_2)).$$

However, we have

$$\begin{aligned} \tilde{\pi}_2^d &= (\bar{m}_1 + \bar{\Delta})(1 - \tilde{x})\lambda - \Psi(\bar{\Delta} + (\bar{m}_1 - \bar{m}_2)) \\ &= [\bar{m}_2 + (\bar{\Delta} + (\bar{m}_1 - \bar{m}_2))](1 - \tilde{x})\lambda - \Psi(\bar{\Delta} + (\bar{m}_1 - \bar{m}_2)) \\ &= \max_{\Delta_2} (\bar{m}_2 + \Delta_2)(1 - \tilde{x})\lambda - \Psi(\Delta_2) \\ &= (\bar{m}_2 + \underline{\Delta})(1 - \tilde{x})\lambda - \Psi(\underline{\Delta}). \end{aligned}$$

This implies that CP2 has no incentive to deviate from investing  $\underline{\Delta}$ .

Finally, we check whether CP1 will have an incentive to deviate by reducing its investment to the level where it ends up at a lower *ex post* margin given that the CP2 invests at  $\underline{\Delta}$ , which is possible if  $\bar{m}_1 < \bar{m}_2 + \underline{\Delta}$ . By following the logic above, the optimal deviation investment level is given by  $\underline{\Delta} - (\bar{m}_1 - \bar{m}_2)$  because  $\underline{\Delta}$  is the optimal investment level for the firm that fails to receive the priority right and the objective function is concave. The optimal deviation payoff for CP1 is given by

$$\tilde{\pi}_1^d = [\bar{m}_1 + (\underline{\Delta} - (\bar{m}_1 - \bar{m}_2))](1 - \tilde{x})\lambda - \Psi(\underline{\Delta} - (\bar{m}_1 - \bar{m}_2)).$$

Let  $\bar{\Delta}(\theta)$  and  $\tilde{\pi}_1^*(\theta)$  denote CP1's investment level and its corresponding payoff in the putative equilibrium. Note that  $\bar{\Delta}(\theta)$  is decreasing in  $\theta$  and  $\bar{\Delta}(\theta = 1) = \underline{\Delta}$ . By the envelope theorem, we also know that  $\tilde{\pi}_1^*(\theta)$  is a decreasing function

of  $\theta$ . Thus, for all  $\theta \in [0, 1]$ , we have

$$\begin{aligned} \tilde{\pi}_1^*(\theta) &\geq \tilde{\pi}_1^*(\theta = 1) \\ &= [\bar{m}_1 + \bar{\Delta}(\theta = 1)]\tilde{x}\lambda - [(\bar{m}_2 + \underline{\Delta}) + 1 \cdot ((\bar{m}_1 + \bar{\Delta}(\theta = 1)) - (\bar{m}_2 + \underline{\Delta}))](2\tilde{x} - 1)\lambda - \Psi(\bar{\Delta}(\theta = 1)) \\ &= [\bar{m}_1 + \bar{\Delta}(\theta = 1)](1 - \tilde{x})\lambda - \Psi(\bar{\Delta}(\theta = 1)) \\ &= (\bar{m}_1 + \underline{\Delta})(1 - \tilde{x})\lambda - \Psi(\underline{\Delta}) \\ &> [\bar{m}_1 + (\underline{\Delta} - (\bar{m}_1 - \bar{m}_2))](1 - \tilde{x})\lambda - \Psi(\underline{\Delta} - (\bar{m}_1 - \bar{m}_2)) = \tilde{\pi}_1^d. \end{aligned}$$

The last inequality above comes from the definition of  $\underline{\Delta}$  and the preceding equality comes from the fact that  $\bar{\Delta}(\theta = 1) = \underline{\Delta}$ . Thus, CPI does not have an incentive to deviate, either. Taken together, we can conclude that  $(\bar{\Delta}, \underline{\Delta})$  is an equilibrium. This is the unique equilibrium if the margin difference  $(\bar{m}_1 - \bar{m}_2)$  is sufficiently large. Otherwise, we can have another pure strategy equilibrium in which the roles are reversed. We can proceed in a similar way to show that  $(\underline{\Delta}, \bar{\Delta})$  can also be an equilibrium if  $(\bar{m}_1 - \bar{m}_2)$  is sufficiently small.

*Proof of Proposition 1.* Recall that we have derived the ISP's profits under a neutral regime and discriminatory regime as  $\pi_m^* = a^* = v - \frac{1}{\mu - \lambda} - \frac{t}{2}$  and  $\tilde{\pi}_m^* = (v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x}) + (m_2 + \theta\Delta_m)(2\tilde{x} - 1)\lambda$ , respectively. The difference between these two is given by

$$\begin{aligned} \tilde{\pi}_m^* - \pi_m^* &= \left( v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x} \right) + (m_2 + \theta\Delta_m)(2\tilde{x} - 1)\lambda - \left( v - \frac{1}{\mu - \lambda} - \frac{t}{2} \right) \\ &= \left( \frac{1}{\mu - \lambda} - \frac{1}{\mu - \tilde{x}\lambda} \right) + (2\tilde{x} - 1) \left\{ (m_2 + \theta\Delta_m)\lambda - \frac{t}{2} \right\} > 0 \\ \iff (m_2 + \theta\Delta_m) &> \Lambda \equiv \frac{t}{2\lambda} + \frac{1}{(2\tilde{x} - 1)\lambda} \left( \frac{1}{\mu - \tilde{x}\lambda} - \frac{1}{\mu - \lambda} \right). \end{aligned}$$

Hence, if  $m_2 > \Lambda$ ,  $\tilde{\pi}_m^* > \pi_m^*$  for all  $\theta \in [0, 1]$ , which proves the statement (i). Conversely, because the term  $(m_2 + \theta\Delta_m)$  is increasing in  $\theta$ , we get  $\tilde{\pi}_m^* < \pi_m^*$  for all  $\theta \in [0, 1]$  when  $m_2 < \Lambda$ . Finally, if  $m_2 \leq \Lambda \leq m_1$ , there exists a critical level of  $\theta^* \in [0, 1]$  such that  $\tilde{\pi}_m^* > \pi_m^*$  iff  $\theta > \theta^*$ .

*Proof of Proposition 2.* The statements in (i) and (ii) can be proved in a straightforward manner by comparing the expressions for the CPs' profits across the regimes. Concerning the statement in (iii), let us denote the aggregate consumer welfare in the neutral network and the discriminatory network by  $CS$  and  $\tilde{CS}$ , respectively. Notice that the marginal consumers in the neutral network and the discriminatory network are located at  $x^* = 1/2$  and  $\tilde{x} (> 1/2)$ , respectively, and they receive zero payoffs. This implies that  $CS = 2 \int_0^{1/2} tx \, dx$  and  $\tilde{CS} = \int_0^{\tilde{x}} tx \, dx + \int_0^{1-\tilde{x}} tx \, dx$ . Aggregate consumer welfare increases in the discriminatory network, which is simply shown as

$$\tilde{CS} - CS = \int_0^{\tilde{x}} tx \, dx + \int_0^{1-\tilde{x}} tx \, dx - 2 \int_0^{1/2} tx \, dx = \int_{1/2}^{\tilde{x}} tx \, dx - \int_{1-\tilde{x}}^{1/2} tx \, dx > 0.$$

*Proof of Proposition 3.* The comparison of social welfare across the two different regimes can be readily seen by the sign of  $\Delta M - \Delta T$ .

$$\begin{aligned} \Delta M - \Delta T &\equiv \tilde{M} - M - (\tilde{T} - T) = \left( \tilde{x} - \frac{1}{2} \right) (m_1 - m_2)\lambda - \left( \tilde{x} - \frac{1}{2} \right)^2 t \\ &= \left( \tilde{x} - \frac{1}{2} \right) \left\{ \Delta_m \lambda - \left( \tilde{x} - \frac{1}{2} \right) t \right\} > 0 \\ \iff \Delta_m \lambda - \left( \tilde{x} - \frac{1}{2} \right) t &> 0 \iff \Delta_m > \left( \tilde{x} - \frac{1}{2} \right) \frac{t}{\lambda}. \end{aligned}$$

Thus, if the margin difference is significantly large compared to the degree of product differentiation to the extent of  $\Delta_m > (\tilde{x} - \frac{1}{2}) \frac{t}{\lambda}$ , the social welfare is higher even in the discriminatory network, precisely.

*Proof of Proposition 5.* First, let us recall that the marginal changes in the CPs' profits are given by  $\Psi'(\Delta_i^*) = \frac{\lambda}{2}$  and  $\Psi'(\tilde{\Delta}_i^*) = \{\tilde{x} - \theta(2\tilde{x} - 1)\}\lambda$ . With the assumption of  $\Psi'' > 0$ , we get  $\Delta_i^* > \tilde{\Delta}_i^*$  if and only if  $\frac{\lambda}{2} > \tilde{x} - \theta(2\tilde{x} - 1)$ . From the following relationship,

$$\begin{aligned} \frac{1}{2} - \{\tilde{x} - \theta(2\tilde{x} - 1)\} &= (2\tilde{x} - 1) \left( \theta - \frac{1}{2} \right) > 0 \\ \iff \theta &> \frac{1}{2} \quad (\because \tilde{x} > 1/2), \end{aligned}$$

we derive the result (i) that  $\theta > \frac{1}{2}$  is the necessary and sufficient condition for  $\Delta_i^* > \tilde{\Delta}_1^*$ . For  $\theta \in [0, 1/2]$ , we have the opposite case of  $\tilde{\Delta}_1^* \geq \Delta_1^*$ . Similarly, the comparison between  $\Psi'(\Delta_2^*) = \frac{\lambda}{2}$  and  $\Psi'(\tilde{\Delta}_2^*) = (1 - \tilde{x})\lambda$  yields the result of  $\tilde{\Delta}_2^* < \Delta_2^*$  for  $\forall \theta \in [0, 1]$ .

*Proof of Proposition 7.* Recall that  $\tilde{\theta}$  must be set such that  $\frac{d\tilde{\pi}_m^*}{d\theta} = \frac{\partial \tilde{\pi}_m^*}{\partial \theta} + \frac{\partial \tilde{\pi}_m^*}{\partial \Delta_1^*} \frac{\partial \Delta_1^*}{\partial \theta} = 0$ . Because we can easily obtain both  $\frac{\partial \tilde{\pi}_m^*}{\partial \theta} = \Delta_m(2\tilde{x} - 1)\lambda$  and  $\frac{\partial \tilde{\pi}_m^*}{\partial \Delta_1^*} = \theta(2\tilde{x} - 1)\lambda$  from (11) as well as  $\frac{\partial \Delta_1^*}{\partial \theta} = -(2\tilde{x} - 1)k\lambda$  from  $\Psi'(\tilde{\Delta}_1^*) = \{\tilde{x} - \theta(2\tilde{x} - 1)\}\lambda$  with  $\Psi(\Delta_i) = \Delta_i^2/2k$ , the optimal level of rent extraction is derived as

$$\begin{aligned} \frac{d\tilde{\pi}_m^*}{d\theta} &= \Delta_m(2\tilde{x} - 1)\lambda - \theta(2\tilde{x} - 1)^2k\lambda^2 \\ &= (2\tilde{x} - 1)\lambda \{\Delta_m - \theta(2\tilde{x} - 1)k\lambda\} = 0 \text{ at } \theta = \tilde{\theta}. \end{aligned}$$

Hence, we get  $\tilde{\theta} = \frac{\Delta_m}{(2\tilde{x} - 1)k\lambda}$ . The results of  $\frac{\partial \tilde{\theta}}{\partial k} < 0$  and  $\frac{\partial \tilde{\theta}}{\partial \Delta_m} > 0$  are immediate from the comparative statics for  $\tilde{\theta}$  with respect to  $k$  and  $\Delta_m$ , respectively.

### Appendix B

□ **Exclusive versus nonexclusive priority.** CP<sub>*i*</sub>'s willingness to pay for the exclusive right to the fast lane will depend upon whether CP<sub>*j*</sub>, for  $i, j \in \{1, 2\}$  and  $j \neq i$ , will be granted that right if CP<sub>*i*</sub> does not buy it. In this appendix, we derive conditions under which selling an exclusive right is preferred to the selling of a nonexclusive right. For this purpose, we assume that the ISP has all the bargaining power and has the ability to make take-it-or-leave-it offers to CPs, as in Armstrong (1999). Let us denote by  $b_i$  and  $l_i$  the benefits and losses associated with the exclusive right, respectively. The net gain to accepting the exclusive right is the sum of  $b_i$  and  $l_i$ : CP<sub>*i*</sub> will pay up to

$$b_i + l_i = \left( \tilde{x}\lambda m_i - \frac{\lambda}{2} m_i \right) + \left\{ \frac{\lambda}{2} m_i - (1 - \tilde{x})\lambda m_i \right\} = (2\tilde{x} - 1)\lambda m_i \tag{B1}$$

for the exclusive right. Therefore, the highest rent that the ISP can obtain by selling the exclusive right to first priority is to sell to the firm with the competitive advantage in margin (CP1), in which case its rent is

$$R_{ex} = (2\tilde{x} - 1)\lambda m_1. \tag{B2}$$

In the discriminatory regime, the ISP can obtain a total profit of<sup>31</sup>

$$\tilde{a} + R_{ex} = \left( v - \frac{1}{\mu - \tilde{x}\lambda} - t\tilde{x} \right) + (2\tilde{x} - 1)\lambda m_1. \tag{B3}$$

Now let us explore whether it is optimal for the ISP to sell the right to first priority exclusively to only one CP. Suppose that the ISP offers to sell the right to CP<sub>*i*</sub> for a charge  $R_i$  and that both CPs accept. Then in order for  $i$  to agree to pay this charge (given that  $j$  has also agreed), we must have  $R_i$  no greater than the profit loss from not having the right when firm  $j$  does, which is just  $l_i$ . Therefore, the most the ISP can get from selling the nonexclusive right is just  $l_1 + l_2$ .<sup>32</sup> In this case, the ISP will obtain a profit

$$\begin{aligned} a^* + R_{non} &= \left( v - \frac{1}{\mu - \lambda} - \frac{t}{2} \right) + \sum_{i=1}^2 \left( \frac{\lambda}{2} m_i - (1 - \tilde{x})\lambda m_i \right) \\ &= \left( v - \frac{1}{\mu - \lambda} - \frac{t}{2} \right) + (2\tilde{x} - 1)\lambda \frac{(m_1 + m_2)}{2}. \end{aligned} \tag{B4}$$

Therefore, the ISP prefers selling the right exclusively to one CP if and only if

$$(\tilde{a} + R_{ex}) - (a^* + R_{non}) > 0 \iff R_{ex} - R_{non} > a^* - \tilde{a}.$$

With simple algebra, we know that  $R_{ex} - R_{non} = (2\tilde{x} - 1)\lambda \frac{\Delta_m}{2}$  and  $a^* - \tilde{a} = \frac{(2\tilde{x} - 1)\lambda}{2(\mu - \tilde{x}\lambda)(\mu - \lambda)}$ . A simple manipulation yields that the ISP will use the exclusive scheme if the margin differential is sufficiently large to the extent of

$$\Delta_m > \frac{1}{(\mu - \tilde{x}\lambda)(\mu - \lambda)}. \tag{B5}$$

<sup>31</sup> It is important that the ISP may be able to commit to granting the exclusive right to CP2 in the event that CP1 rejects the ISP's offer, if it is in the interests of the ISP to do so. A simple scheme that appears to avoid some of these commitment issues is to auction off the exclusive right to the highest bidder. In this case, the bidding will stop when CP2 drops out at the price  $b_2 + l_2$ .

<sup>32</sup> This assumes that the ISP can make discriminatory offers to CPs. Note that Cheng et al. (2009), by contrast, assume that the ISP makes nondiscriminatory offers. With nondiscriminatory offers, an exclusive contract would be more attractive to the ISP.

## References

- ANDERSON, S.P. AND COATE, S. "Market Provision of Broadcasting: A Welfare Analysis." *Review of Economic Studies*, Vol. 72 (2005), pp. 947–972.
- ARMSTRONG, M. "Competition in the Pay-TV Market." *Journal of the Japanese and International Economies*, Vol. 13 (1999), pp. 257–280.
- . "Competition in Two-Sided Markets." *RAND Journal of Economics*, Vol. 37 (2006), pp. 668–691.
- BALACHANDRAN, K.R. "Purchasing Priorities in Queues." *Management Science*, Vol. 18 (1972), pp. 319–326.
- BUSINESS WEEK ONLINE. "Rewired and Ready for Combat." November 7, 2005. Available at [www.businessweek.com/magazine/content/05\\_45/b3958089.htm](http://www.businessweek.com/magazine/content/05_45/b3958089.htm) (accessed on June 29, 2010).
- CAILLAUD, B. AND JULLIEN, B. "Chickens and Egg: Competition among Intermediation Service Providers." *RAND Journal of Economics*, Vol. 34 (2003), pp. 309–328.
- CERF, V.G. "Prepared Statement in the U.S. Senate Committee on Commerce, Science, and Transportation Hearing on 'Net Neutrality.'" February 7, 2006.
- CHENG, H.K., BANDYOPADHYAY, S., AND GUO, H. "The Debate on Net Neutrality: A Policy Perspective." *Information Systems Research*, (2009).
- CHIRICO, F., VAN DE HAAR, I., AND LAROUCHE, P. "Network Neutrality in the EU." Discussion Paper no. DP 2007-030, TILEC, 2007.
- CHOI, J.P. "Optimal Tariffs and the Choice of Technology: Discriminatory Tariffs vs. the 'Most Favored Nation' Clause." *Journal of International Economics*, Vol. 58 (2010), pp. 560–579.
- . "Tying in Two-Sided Markets with Multi-Homing." *Journal of Industrial Economics*.
- CREMER, J., REY, P., AND TIROLE, J. "Connectivity in the Commercial Internet." *Journal of Industrial Economics*, Vol. 48 (2000), pp. 433–472.
- DeGraba, P. "Input Market Price Discrimination and the Choice of Technology." *American Economic Review*, Vol. 80 (1990), pp. 1246–1253.
- DENECKERE, R.J. AND MCAFEE, P. "Damaged Goods." *Journal of Economics & Management Strategy*, Vol. 5 (1996), pp. 149–174.
- ECONOMIDES, N. "Net Neutrality, Non-Discrimination and Digital Distribution of Content through the Internet." *A Journal of Law and Policy for the Information Society*, Vol. 4 (2008), pp. 209–233.
- AND TÅG, J. "Net Neutrality on the Internet: A Two-Sided Market Analysis." Working Paper no. 07-14, NET Institute, New York University, 2007.
- EDELSON, N.M. AND HILDERBRAND, D.K. "Congestion Tolls for Poisson Queuing Processes." *Econometrica*, Vol. 43 (1975), pp. 81–92.
- FUDENBERG, D. AND TIROLE, J. "Understanding Rent Dissipation: On the Use of Game Theory in Industrial Organization." *American Economic Review: Papers and Proceedings*, Vol. 77 (1987), pp. 176–183.
- GROSS, D. AND HARRIS, C.M. *Fundamentals of Queueing Theory*, 3rd ed. New York: Wiley, 1998.
- HAHN, J.H. "Damaged Durable Goods." *RAND Journal of Economics*, Vol. 37 (2006), pp. 121–133.
- HAHN, R.W. AND WALLSTEN, S. "The Economics of Net Neutrality." Working Paper no. RP06-13, AEI-Brookings Joint Center, 2006. Available at [ssrn.com/abstract=943757](http://ssrn.com/abstract=943757) (accessed June 29, 2010).
- HERMALIN, B.E. AND KATZ, M.L. "The Economics of Product-Line Restrictions with an Application to the Network Neutrality Debate." *Information Economics and Policy*, Vol. 19 (2007), pp. 215–248.
- KOCIS, V. AND DE BIL, P. "Network Neutrality and the Nature of Competition between Network Operators." *International Economics and Economic Policy*, Vol. 4 (2007), Special Issue on the Digital Economy and Regulatory Issues, pp. 159–184.
- MENDELSON, H. "Pricing Computer Services: Queueing Effects." *Communications of the ACM*, Vol. 28 (1985), pp. 312–321.
- AND WHANG, S. "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue." *Operations Research*, Vol. 38 (1990), pp. 870–883.
- NAOR, P. "The Regulation of Queue Size by Levying Tolls." *Econometrica*, Vol. 37 (1969), pp. 15–24.
- ROCHET, J.C. AND TIROLE, J. "Two-Sided Markets: A Progress Report." *RAND Journal of Economics*, Vol. 37 (2006), pp. 645–667.
- VALLETTI, T. AND CAMBINI, C. "Investments and Network Competition." *RAND Journal of Economics*, Vol. 36 (2005), pp. 446–468.
- VAN SCHEWICK, B. "Towards an Economic Framework for Network Neutrality Regulation." *Journal on Telecommunications and High Technology Law*, Vol. 5 (2007), pp. 329–392.
- VON HIPPEL, E. *Democratizing Innovation*. Cambridge, Mass.: MIT Press, 2005.
- WU, T. "Network Neutrality, Broadband Discrimination." *Journal on Telecommunications and High Technology Law*, Vol. 2 (2003), pp. 141–178.
- YOO, C.S. "Network Neutrality and the Economics of Congestion." *Georgetown Law Journal*, Vol. 94 (2006), pp. 1847–1908.

## DAMAGED GOODS

RAYMOND J. DENECKERE

*Department of Economics  
University of Wisconsin—Madison  
Madison, WI 53706*

R. PRESTON MCAFEE

*University of Texas at Austin  
Austin, TX 78712*

*Manufacturers may intentionally damage a portion of their goods in order to price discriminate. Many instances of this phenomenon are observed. It may result in a Pareto improvement.*

### 1. INTRODUCTION

The 486SX processor of Intel Corporation was initially produced in a curious way. Intel began with a fully functioning 486DX processor, then *disabled the math coprocessor*, to produce a chip that is strictly inferior to the 486DX but more expensive to produce. Nevertheless, in 1991, the 486DX sold for \$588, and the 486SX for \$333, a little over half the price of the chip that is less expensive to produce (Frenkel, 1991).

We will argue in this paper that this is not an isolated incident, and that many manufacturers intentionally damage a portion of their production.<sup>1</sup> The obvious reason for doing so is to permit price discrimination. By producing an inferior substitute, the manufacturer can sell to customers who do not value the superior product so much, without decreasing demand for the superior product very much.<sup>2</sup> The

We thank Bruce Smith, Hal Varian, and seminar participants at Cornell, Harvard, Michigan State, Montreal, NYU, Northwestern, Princeton, Rice, SMU, Texas, Washington, and Yale for helpful discussions.

1. The phenomenon is sufficiently well known among marketing professionals that it has a name: *crimping the product*.

2. Most authors agree that it is possible to price discriminate with differentiated products, by charging distinct markups on the goods. For example, Jean Tirole (1988, p. 134) argues that "It should not be inferred that price discrimination does not occur when differentiated products are sold to different consumers," and specifically cites

novelty of this paper is not in noting that damaging a high quality good may be a less expensive way to produce a low quality good than directly manufacturing the low quality good. Indeed, we presume that damaging the superior product is the least expensive way to produce the inferior product. Our insight is that this may be a strict Pareto improvement: the manufacturer and *all* types of consumers strictly benefit from the price discrimination.

The simplest way to see the welfare effects is in the case of two types of consumers. Suppose that a manufacturer, selling only the high quality good, chooses to sell only to the high demand types. When the manufacturer price discriminates and sells a low quality good to the low demand types, the low demand types obviously benefit. Why may the high demand types benefit as well? Let  $M_H$ ,  $M_L$  represent the monopoly prices for the high demand consumers purchasing the high quality good, and the low demand consumers purchasing the low quality good, respectively. Further suppose that at these prices, the high demand consumers would prefer the low quality good, but if the prices were slightly closer together, the high demand consumers prefer the high quality good. In order to introduce the low quality good, then, the manufacturer must reduce the gap between the two monopoly prices.

Since there is a zero first-order effect from reducing  $M_H$  on profits, the profit maximizing way to narrow the gap is to reduce  $M_H$  and increase  $M_L$  slightly. Note that low demand consumers still benefit from introducing the damaged good, because the increase in  $M_L$  is compared to not being served at all. Finally, as the manufacturer only introduces the low quality good when profits increase, the very existence of the good tells us that profits go up.

This paper adds to a standard result—that if price discrimination expands output, then a welfare improvement tends to occur—in two ways. First, the standard result assumes no resale. In contrast, we permit free resale, limiting the manufacturer to the more common

---

the example of quality differentiated services. In our context, the case for price discrimination is compelling. At lower cost, the manufacturer could have sold the high quality good, but instead chose to damage a portion of production, in order to be able to charge a lower price on the product. Thus, it is as if the manufacturer were selling the high quality product at two distinct prices—one without damage, and one with. This would not be price discrimination if the manufacturer's costs on the damaged good were lower, but since the manufacturer has incurred costs to damage the good, the manufacturer must be discriminating.

case of second degree (incentive compatible) price discrimination.<sup>3</sup> Second, incentive compatible price discrimination strengthens the case for a Pareto improvement, in some cases making price discrimination a strict Pareto improvement for all types.<sup>4</sup>

The paper is organized as follows. In Section 2, we will present a number of examples of manufacturers intentionally damaging a portion of their production. These examples divide naturally into two categories: those where there are two distinct uses for the product, and those where there is only a single use for the product. With two distinct uses for the product, the natural model is one where there are two types of consumers; for example, the educational and business markets correspond to distinct groups of software users. A model of the dual use environment is analyzed in Section 3. When there is only one use for the product, such as the 486 chip, the natural model involves a continuum of consumers with different reservation prices. We analyze this case in Section 4. We end the paper with some concluding remarks.

## 2. DAMAGED GOODS

How common is it for manufacturers to intentionally damage a portion of their production? In this section, we will argue that, throughout history and across a broad variety of different industries, manufacturers damage some of their production solely for the purpose of enhancing their discriminatory abilities. We document four examples, and provide a brief summary of a variety of other examples.

3. The most general results known to date (Varian, 1985), while phrased in terms of third degree price discrimination, do allow for demand interdependencies, and hence could be applied to demand structures derived from self-selection models. However, our paper states conditions on the demand primitives (rather than conditions on endogenous variables) and focuses on Pareto improvements (rather than mere welfare increases).

4. Varian (1985) constructs an example where third degree price discrimination results in a Pareto improvement. However, because of the absence of demand interdependencies, the high demand market obtains the same price and utility as in the absence of price discrimination; the Pareto improvement arises because a new market is served. In contrast, we consider the case of second degree price discrimination, where the seller does not condition on observable characteristics of the buyers, but instead offers an incentive compatible price menu, so that buyers self-select into categories by their choice of good to purchase. While it is known that net welfare gains can be obtained in the case of second degree price discrimination, to our knowledge, no one has shown that Pareto improvements may arise, nor investigated the circumstances that tend to lead to Pareto improvements. The main theoretical contribution of our paper is to prove that incurring costs in order to damage production may make everyone better off, and that this is a quite plausible outcome in the dual use case, while a less plausible outcome in the single use case.

## 2.1 THE 486SX

With the introduction of the 486 microprocessor, Intel provided a significant improvement in performance over its predecessor, the 386. The 486 microprocessor improved on the 386 in a number of ways. First, it makes more efficient use of internal clock cycles, thereby performing some operations in one step that would have taken several steps on a 386. Second, it contains an 8 kilobyte internal cache memory, allowing the 486 faster access to instructions than if it had to fetch them from slower external DRAM. The final advantage of the 486 is that it contains a 387 compatible math coprocessor, which handles floating point numerical computations. Installing the coprocessor in the same microprocessor eliminates time consuming communication between the processor and the coprocessor, as occurred with the 386 series. The combined effects of these improvements is that a 486 outperforms a 386-387 combination, even when running at lower clock speeds (operations per second).

In response to fast 386 based microprocessors produced by competitor Advanced Micro Devices, Intel decided to introduce a low cost, high performance alternative to the 386: the 486SX. Intel renamed the original 486 the 486DX. Unlike the 386SX processor<sup>5</sup> after which it was named,

the 486SX is an exact duplicate of the 486DX, with one important difference—its internal math coprocessor is disabled. (Frenkel, 1991)

Although it is more costly for Intel to produce the 486SX, it sold in 1991 for substantially less: \$333 as opposed to \$588 for the 486DX. As with the 386, it is possible to improve numerical calculations on a 486SX by purchasing the 487SX math “coprocessor.” Unlike the 387SX, however, the 487SX is not a real coprocessor.

The 487SX math coprocessor is really a 486SX with the floating-point unit (FPU) enabled. Keep in mind that the 486SX is actually a 486DX with the FPU disabled. So in reality, the 487SX is really a 486DX. . . . In fact, the 487SX doesn’t coprocess at all. It simply disables the 486SX processor and performs like the 486DX that it really is. (Frenkel, 1991)

5. The 386SX was a 386 with a smaller internal data bus, which allowed PC manufacturers to use the 386SX as a “drop-in” replacement for the previous standard 286, without having to redesign the computer for the 386, yet increasing power and compatibility with the 386 generation of chips.

To obtain the full capabilities of a 486DX based machine, an owner of a 486SX based personal computer must therefore purchase the equivalent of *two* 486DX microprocessors: one with its internal coprocessor purposely disabled, and one that, while labeled a coprocessor, is actually a fully functioning 486DX that disables the operation of the 486SX.

It is not possible to purchase a 487SX and use it as a standalone processor, because Intel designed the 487SX so that it needs the presence of the 486SX to operate, although none of the 486SX's processing ability is actually used. Moreover, it would not be economic to purchase a 487SX alone even if it did work without the 486SX, for Intel sells the 487SX at \$799, significantly more than the 486DX.

So why wouldn't the owner of a 486SX computer who finds a strong need for the math coprocessor not just scrap the 486SX and upgrade to a 486DX? Intel has reconfigured the pins on the 486SX, so that the SX socket won't accept the 486DX.

Intel ceased to manufacture 486SXs with disabled math coprocessors in 1991, and began removing the coprocessor. Intel could make the microprocessor small enough to be "surface mounted," that is, mounted without a socket, thereby freeing up space that is at a premium in the fast-growing segment of notebook computers (Seymour, 1991).

## 2.2 IBM LASERPRINTER E

In May 1990, IBM announced the introduction of the LaserPrinter E, a lower cost alternative to its popular LaserPrinter. The LaserPrinter E was virtually identical to the original LaserPrinter, except that the E model printed text at 5 pages per minute (ppm), as opposed to 10 ppm for the LaserPrinter. According to Jones (1990), the LaserPrinter E uses the same "engine" and virtually identical parts, with one exception:

The controllers in our evaluation unit differed only by virtue of four socketed firmware chips and one surface mounted chip. PC Labs' testing of numerous evaluation units indicated that the LaserPrinter E firmware in effect inserts wait states to slow print speed. . . . IBM has gone to some expense to slow the LaserPrinter in firmware so that it can market it at a lower price.

That is, IBM has added chips to the LaserPrinter E that serve as counters or idlers, chips that perform no function other than to make the machine pause and hence print more slowly. Moreover, this is

the *only* difference in the two machines. In particular, the idling only applies to text printing, so that graphics comes out at the same speed.

It is interesting that *PC Magazine* (Jones, 1990) gave a good review of the LaserPrinter E, calling it "the obvious choice" over the Hewlett-Packard IIP. For an additional \$1099, one can upgrade the LaserPrinter E to identical performance with the LaserPrinter, bringing the total cost of the upgraded LaserPrinter E to \$200 more than the original LaserPrinter.

### **2.3 SONY MINIDISCS**

Sony recently introduced a new digital recording-playback format intended to replace the analog audio cassette, but offering greater convenience and durability. To achieve the small form factor deemed necessary for success (the audio cassette's popularity derives from its small format, not its sound quality or durability) and still be able to provide 74 minutes of music, Sony's engineers devised a data compression algorithm that permits squeezing the content of an entire audio Compact Disc onto a disc which is only 2.5 in. in diameter. The MiniDisc is not only smaller than a regular CD but is also immune to the interruption of music caused by shock or vibration in portable applications. Sony accomplished this by inserting a memory buffer between the laser pickup and the digital decoding circuitry, a feature now being introduced on standard Compact Disc players.

MiniDiscs are similar in appearance to 3.5 in. computer diskettes, and come in two varieties: prerecorded and recordable. The prerecorded variety is essentially a miniature CD housed in a plastic shell: like its bigger brother, it uses a laser beam to read the information encoded on the surface of the disc; its principle of operation is therefore entirely optical. The recordable variety looks externally the same as the prerecorded discs, but uses a technology originally developed for computer data storage: magneto-optical recording.<sup>6</sup> Sony produces complementary hardware that either includes just a playback mechanism able to read both types of discs (intended primarily for portable or in-car use), or both a recording and a playback mechanism (intended primarily for home use).

Prerecorded MDs are priced in the same range, but slightly below CDs. Some 400 titles are currently available, mostly from Sony's own label. Blank MDs come in two varieties: 60-minute discs and 74-

6. For an in-depth discussion of the technology underlying MiniDiscs, see Harley (1992, 1994).

minute discs. The list prices for these discs are currently \$13.99 and \$16.99. Despite the difference in price and recording length, the two formats are physically identical:

The 60- and 74-minute discs are identical in manufacture. A code in the table of contents identifies a 60-minute disc and prevents recording beyond this length, even though there's room on the media. (Harley, 1994)

One might think that a clever user could circumvent this scheme by constructing a device that alters the table of contents. However, Sony has made this nearly impossible:

Blank MDs are polycarbonate substrates coated with very thin layers of magnetic material. A ring of polycarbonate at the inner radius is left uncoated. This area, called the "lead in," has pits impressed in it just as on a CD. The MD recorder reads the information in this un-erasable area, which includes the optimum laser power for recording and the disc playback time. . . . The only difference between a 74-minute disc and blanks of shorter playing time is the information encoded in the lead-in area: it tells the player how much recording time is available. (Harley, 1992)

Sony already has plans to make the technology available for computer storage. An MD data disk will have a maximum capacity of 128 Mb.

## **2.4 TONTINES**

While the above examples are all of recent vintage, crimping the product is by no means a new phenomenon. In fact, the British and French governments widely used the practice during the seventeenth and eighteenth centuries. The frequent wars during this period produced large government deficits that were financed with a variety of different debt instruments. Interestingly, what differentiated the liabilities was not so much their maturity structure as their risk structure. In addition to relatively riskless government bonds yielding a normal rate of return, the governments also issued life annuities carrying much higher returns. The purchaser of such an annuity would name a nominee, and receive interest as long as the nominee remained alive. Unlike modern annuities, anyone could be named as the nominee. This effectively provided the annuitant with a random return uncorrelated upon his own life. Meanwhile, by selling the annuities to a large sub-

scriber base, the government faced little or no risk.<sup>7</sup> It might be argued that these annuities satisfied a desire for gambling, i.e., were effectively government sponsored lotteries. However, the fact that these securities paid *higher* interest rates and were only available in large denominations—approximately the average annual income of the time—indicates that they were inferior commodities. Interestingly, Smith and Villamil (1993) argue that tontines were created in an attempt to price discriminate between individuals who differed in their private investment opportunities.<sup>8</sup> Such random payout bonds disappeared as investment and saving opportunities increased, although Kurdistan is apparently offering them today.<sup>9</sup>

## 2.5 OTHER EXAMPLES

It is probably not surprising that manufacturers conceal damaged goods, since something seems wasteful about damaging a good in order to extract more revenue. Three out of the four examples above concern electronics or computer hardware. This is no accident: crimping the product is extremely pervasive in these industries, and easier to document. This subsection presents some brief summaries of other examples of damaged production.

**2.5.1 IBM 2319 Disk Drive:**<sup>10</sup> A disk drive is composed of two major components. The *spindle* includes the disk platter, a motor to turn it, read/write heads, and the actual spindle that the platter turns on. The second major component is the *controller*, which connects the disk drive to the computer and controls the actions of the spindle. In the late 1960s, IBM began to face strong competition in the disk drive market, which provided memory for the IBM 360 mainframe. Five companies, notably Telex and Memorex, offered spindles superior to IBM's, which could be used with IBM controllers. In addition, Memorex offered a separate controller, which meant customers could pur-

7. In the case of a lottery bond, a bond with a random payout but without the annuity feature of tontines, offered by England, the bond was bundled with a lottery, and offered significantly higher average rates of return than bonds without lotteries, corresponding to a lower price. A tontine divides a fixed amount of money among the annuitants with surviving nominees, thereby being risk-free from the government's perspective, but risky from the individual's perspective.

8. Section 6 of Smith and Villamil (1993), upon which the above discussion is based, contains a detailed discussion of the various types of debt instruments in use during this period, as well as an extensive bibliography.

9. Source: conversations with Bruce Smith of Cornell University.

10. This material is derived from DeLamarter (1986, Chapter 12), who worked as an economist for the DOJ in the famous IBM case. For a contrasting view, see Fisher, McGowan, and Greenwood (1983).

chase Memorex disk drives and just plug them into the IBM mainframe. Worse still for IBM, it had introduced the IBM 370 mainframe, more powerful than the old 360s, but its new disk drive system, dubbed Merlin, would not be ready for introduction for several years, and thus IBM faced loss of disk drive sales for new customers as well as old customers.

To protect the market for 370 disk drives, IBM introduced a scheme internally called "Apricot," a name later changed to "Mallard." According to DeLamarter (1986) this scheme worked as follows: IBM renamed its existing 2314 disk drive the 2319, and integrated the controller, which was previously an outboard device (separate unit), into the single unit, thereby limiting the number of additional spindles that could be plugged into the unit, and undercutting the market for rivals' spindles. This, of course, didn't eliminate Memorex, which sold controllers as well. To undercut Memorex, IBM changed the controller interface as well, thus forcing any rival who wished to offer both a controller and a spindle to decode the controller communication language.

The extent of the price discrimination is summarized by DeLamarter (1986):

IBM would favor 370 customers at the expense of 360 users, offering each group essentially the same product but at widely different prices. Where IBM charged \$256,000 for eight unbundled 2314 spindles and a controller for use on the 360, a similar number of 2319 spindles along with the file adaptor [changed controller] on a 370/135 processor went for as little as \$145,415.

IBM vice president P. W. Knaplund described the 2319 relabeling as a "gimicky tactic" to "buy time."

**2.5.2** Consumer Electronics: Crimping the product is a popular tactic in consumer electronics. We have heard numerous accounts of how lower priced models of consumer electronics (such as pocket calculators, video equipment, VCRs, and multitestors) differ from their higher priced alternatives only by having some of the features disabled.<sup>11</sup> Unfortunately, the use of this strategy in consumer electronics products has proven harder to document. Nevertheless, Nagle (1987, p. 186) reports:

11. For example, in an April 1993 internet message posted to the rec.video discussion group, Terry Jeffery (UK) reported discovering undocumented features on his Cannon video-camera.

A leading manufacturer of pocket calculators . . . sold a card programmable version of one calculator for much more than the nonprogrammable version. The only practical difference between the two was a slot in the plastic case of the programmable version where the cards could be inserted.

We have been told (but were unable to verify) that a consumer electronics magazine printed an article explaining how to convert the nonprogrammable version into a programmable version. Another leading calculator manufacturer, Sharp Electronics, produces a calculator that differs from the higher priced scientific version only in that its buttons do not have the alternate functions imprinted upon them.<sup>12</sup>

Robert Harley, one of *Stereophile's* technical editors, likens Sony's MD strategy to that of the hand-held multimeter industry:

This is analogous to a trick of the hand-held multimeter industry. The same electronics are in every meter throughout the product line; the less expensive models merely have some of their features disabled. (Harley, 1992)

While deeply entrenched as a strategy in electronics and computers, crimping the product occurs in a broad range of other industries, as the next few examples demonstrate.

**2.5.3 Educational Software:** It appears that the normal way of producing "student versions" or educational versions of software is to put limiting factors into the full-featured versions, thus destroying some of their capability. We know of two examples.

Wolfram Research, Inc.<sup>13</sup> sells a student version of its popular mathematics program, Mathematica, for \$180, less than a quarter of the normal price. The student version implements the complete Mathematica program with one exception: it does not use a math coprocessor, even if one is present on the student's computer. This disabling of the math coprocessor makes some kinds of numerical calculations significantly slower. Mathematica requires a fairly powerful microcomputer to operate; most student users are therefore likely to already have a coprocessor.<sup>14</sup>

12. It could, however, be argued that some people may prefer not having access to the scientific functions: square root buttons only serve to confuse them.

13. The source for this material is conversations with Hal Varian, April 22, 1993. Varian has edited a book entitled *Economic and Financial Modelling with Mathematica* and is familiar with the marketing practices of Wolfram Research.

14. Removing the calls to the coprocessor is a simple task, which directly incurs an insignificant additional cost. However, some additional cost would arise from marketing and supporting two versions of the program.

Data Desk is an exploratory data analysis and statistics package for the MacIntosh computer. The full retail version sells for \$95.<sup>15</sup> The student version, which has "reduced data handling capabilities," sells for \$69.95.

**2.5.4 Buying Clubs:** The proliferation of discount "buying club" stores such as Sam's, Costco, and The Price Club has segmented the market for many consumer items according to the quantity purchased. Buying clubs specialize in large quantities. Purchases in these outlets, of course, undercut the normal grocery store market, and manufacturers have responded in two ways: by bundling a number of units together, to produce a minimum purchase larger than would normally be demanded by even a large family, and by producing larger sizes specifically for this market.

Creating multipacks, or bundling, incurs additional cost directly. Manufacturers turn to contract packagers to create multipacks "because they don't want the expense of designing and building dedicated in-house lines for filling or multipacking larger-size packages" (Larson, 1993). Other manufacturers, including Chinet (disposable tableware) and Mrs. Paul's (fish sticks), design and manufacture separate production runs for the warehouse market.

Grocers are quite concerned about the growth of the warehouse club market siphoning off demand from retail grocery stores, and have responded by obtaining larger sizes from manufacturers. Traditionally, Chinet sold its products in packages of fifteen. Having introduced packages of 125 units for warehouse clubs, it introduced a 40-count package for "economy aisles" in grocery stores. Nonetheless, manufacturers are segmenting the market:

With larger sizes, manufacturers are creating stumbling blocks for wholesalers [who sell to retail grocers]; making it difficult for us to buy those items with a low price per ounce.<sup>16</sup>

Interestingly, packagers note the inefficiency of these large bundles associated with consumer sales:

15. Data Desk is produced by Data Description of Ithaca, NY. Prices are as of December 1991, as reported in *The Higher Education Product Companion*, Vol. 1, No. 1, p. 18.

16. Marty Rodgers, quoted in *Progressive Grocer*, May 1992, p. 86. Much of this issue of the magazine is devoted to grocers grousing about competition from warehouse clubs. Another adds "We have seen items in competitors' stores that we were not shown. When we tracked it to the manufacturer, we were told that those items were not made for our class of trade.... I might not need a three-pack of toothpaste banded together, but I want to know it's available" (p. 92).

Many warehouse shoppers have no intention of using, say, a bundle of 24 rolls of paper towels themselves. They just don't have room, or their family isn't big enough. So they "team-buy" with a friend. Then they split up the multi-pack.<sup>17</sup>

This evidence is suggestive that manufacturers create larger sizes than are efficient to assist in segmenting the market.

**2.5.5 Chemicals:** Distinct uses for a given chemical will generally offer an opportunity for price discrimination, provided the manufacturer can reduce arbitrage. One method of reducing arbitrage is to add an adulterant to the chemical sold for a low value use, which seriously compromises the high value use. For example, Brazil added gasoline to ethanol sold as automobile fuel to prevent people from drinking the ethanol. The newspapers often contain stories of medicines with vast price differences between human and veterinary use. Even within veterinary use, medicines used for distinct animals may have different values, and the medicine aimed for low value use may come bundled with vitamins, to deter the high value users (whose animals don't need the vitamins) from switching.<sup>18</sup>

In the following two examples, there is no evidence available to us that the manufacturer actually damaged the product. Instead, there is evidence that the manufacturer *contemplated* damaging the product, and in one case, expended significant resources to compute the best means of damaging the product.

Methyl methacrylate (MM) is a plastic with a variety of industrial uses. It is also used to make dentures. According to Stocking and Watkins (1947), the two manufacturers of MM, du Pont and Rohm & Haas, followed a uniform price policy and acted as a cartel. Their pricing policy certainly corroborates this claim: they sold the powdered version of MM (polymer) for industrial uses at 85 cents per pound, and a prepared mixture consisting of powder and liquid (monomer) MM for \$22 per pound to licensed dental laboratories.

The price difference was evidently too great, and attracted bootleggers who found they could crack the powder *back to liquid*, and sell the polymer and monomer together at a profit to the dental trade. (Stocking and Watkins, 1947, p. 403; italics added)

17. John Berkeley, quoted in Larson (1993).

18. An imperfect example is cooking wine: this is ordinary wine with sufficient salt added to make it undrinkable. Cooking wine is generally sold to avoid paying taxes rather than to screen out high value users. Originally, cooking wine was developed to solve the moral hazard problem associated with cooks drinking the wine.

Rohm & Haas considered adulterating the powdered version so that it would be unsuitable for use in dentures and would be prohibited by the Food and Drug Administration for that use. A licensee of Rohm & Haas suggested

A millionth of one percent of arsenic or lead might cause them [the FDA] to confiscate every bootleg unit in the country. There ought to be a trace of something that would make them rear up. (Quoted by Stocking and Watkins, 1947, p. 403)

There is no evidence that Rohm & Haas put this policy into effect, although they called it "a very fine method of controlling the bootleg situation." However, Rohm & Haas did resort to the less effective strategy of planting a rumor that they had adulterated their powdered MM (Nagle, 1987).

According to Stocking and Watkins (1947), two chemical companies, du Pont and General Aniline, possessed exclusive U.S. rights to market "Monastral" colors, used for both paints and textile dyeing. The use in paints required low prices, while the use in textiles permitted quite high prices. Both companies held conferences and ran experiments to determine the feasibility of adding contaminants to the colors that would render them suitable for paint but not for textiles. Three distinct strategies for contamination were considered. Ground glass would damage painting rolls used in textiles, but have an insignificant effect on paints. Compounds that would cause cotton to deteriorate rapidly, but not affect paint, were also considered. Finally, compounds that would irritate skin and cause dermatitis could be added, again to prohibit use in textile dyeing.

As we have seen, manufacturers in many circumstances disable features, degrade performance, or otherwise damage products to create a lower quality good, which they may sell at a lower price without significantly reducing demand for the high quality good. The examples naturally divide into two categories: those where there is a secondary, low value use for the product (such as educational use in software, or paint dyeing for pigments), which we call the dual use case, and those where there is a single use for the product, such as laser printers and microprocessors. We now turn to the welfare implications of damaged goods in the dual use case.

### 3. THE DUAL USE CASE

There are two types of consumers, denoted  $X$  and  $Y$ , and a monopoly producer of two qualities,  $L$  and  $H$ , for low and high. Consumers buy

one good or the other, but not both, and base their purchase decision on which good yields the highest net surplus. Let  $x_L$  (respectively  $x_H$ ) denote consumer  $X$ 's demand for the low (respectively high) quality good, when this is the only version available for purchase. Similarly let  $y_L$  and  $y_H$  represent consumer  $Y$ 's respective demands. We denote the monopoly prices on these demand curves by  $M_i^z$ , for  $i \in \{L, H\}$  and  $z \in \{x, y\}$ . For example,  $M_L^x$  is the monopoly price associated with demand  $x_L$ . The monopolist's marginal cost of production of good  $L$  and good  $H$  are constant, and denoted by  $c_L$  and  $c_H$ .

Our first set of assumptions ensures that the model fits the applications discussed in Section 2:

$$0 \leq c_H \leq c_L, \quad (1)$$

$$y_H(p) \geq y_L(p) \quad \text{and} \quad x_H(p) \geq x_L(p), \quad (2)$$

and

$$\begin{aligned} (\forall p_L, p_H) \quad 0 < \int_{p_L}^{\infty} x_L(p) dp &\leq \int_{p_H}^{\infty} x_H(p) dp \\ &\Rightarrow \int_{p_L}^{\infty} y_L(p) dp < \int_{p_H}^{\infty} y_H(p) dp. \end{aligned} \quad (3)$$

Inequality (1) just formalizes the notion that  $L$  represents an altered, and hence more costly, version of  $H$ . Inequality (2) guarantees that  $H$  is indeed the high quality good, with increased demand by both types of consumers. Jointly, conditions (1) and (2) imply that if the monopolist were required to sell only one quality, he would offer high quality.<sup>19</sup>

$$(\forall p \geq c_H) \quad (p - c_L)[x_L(p) + y_L(p)] \leq (p - c_H)[x_H(p) + y_H(p)].$$

Recall that the consumer surplus associated with demand  $q$  and price  $p$  is  $CS = \int_p^{\infty} q(z) dz$ . Thus, assumption (3) says that whenever consumer  $X$  weakly prefers purchasing  $H$  to purchasing  $L$ , consumer  $Y$  strictly prefers purchasing  $H$ .<sup>20</sup> This ensures that if the low quality

19. This is the only place in our argument where we use assumptions (1) and (2). The conclusions of Theorem 1 remain valid if we impose the above condition directly, increasing the range of applicability beyond situations where (1) and (2) are satisfied. However, while the theory applies to cases where  $c_L < c_H$ , these are economically not very interesting, for it is then no longer obvious that price discrimination is occurring, and less surprising that introducing  $L$  leads to a Pareto improvement.

20. There are various more primitive assumptions that can be imposed to imply condition (3). In particular, if either  $x_L = x_H$  or  $y_L = x_L$  and  $y_H > x_H$  whenever  $x_H > 0$ , then (3) holds. Also, if  $\inf\{p \mid x_L(p) = 0\} = \inf\{p \mid x_H(p) = 0\}$  and if for all  $p_L, p_H$  such that  $x_L(p_L) > 0$  we have  $x_H(p_H)/x_L(p_L) < y_H(p_H)/y_L(p_L)$ , then (3) is satisfied. None of these appear to improve on (3) directly, which [in conjunction with assumption (4) below] is interpretable as stating that  $Y$  is more  $H$  loving than  $X$ .

good is introduced, it will be targeted towards the  $X$  segment of the market.

Our next set of assumptions serves to guarantee that the introduction of good  $L$  is profitable, and produces a Pareto improvement. Let  $M_H^{xy} = \arg \max_p (p - c_H)[x_H(p) + y_H(p)]$ . We assume that the  $X$  market is not served if the firm sells only one quality. That is,

$$x_H(M_H^{xy}) = 0. \tag{4}$$

If (4) fails, then typically we have the monopoly price for both markets falling between the monopoly prices for the markets individually, and thus introducing the  $L$  good will tend to increase the price charged for  $H$ . This effect need not dominate the effects of the incentive constraints, but such a consideration does not appear to lead to an economically meaningful characterization.

It is possible that, at any price for which the  $X$  consumer is willing to purchase a positive quantity of the  $L$  good, the  $Y$  consumer prefers the  $L$  good over the monopoly price for the  $H$  good: This will tend to make introducing the low quality good unprofitable. To rule this out, we assume

$$\int_{M_H^{xy}}^{\infty} y_H(p) dp \geq \int_{\bar{p}_L}^{\infty} y_L(p) dp. \tag{5}$$

The inequality (5) ensures that the  $Y$  consumer would prefer purchasing  $H$  at the monopoly price to purchasing  $L$  at the price  $\bar{p}_L$ , the lowest price for which the  $X$  demand for  $L$  is zero.

Under these assumptions, the firm will always choose to introduce the low quality good  $L$ , and its introduction is a Pareto improvement.

**THEOREM 1** *Suppose that (1)–(5) hold, and suppose that  $x_L(c_L) > 0$ . Then the introduction of the good  $L$  is a Pareto improvement. If, in addition,*

$$\int_{M_H^{xy}}^{\infty} y_H(p) dp < \int_{M_L^x}^{\infty} y_L(p) dp, \tag{6}$$

*then the improvement is strict: all three agents strictly benefit.*

All proofs are contained in the Appendix.

Theorem 1 captures the intuition provided in the introduction. If the low demand  $L$  type is not served when only the high quality good is introduced, then the introduction of the low quality good benefits the low demand type and the firm. If, in addition, the incentive constraints on the high demand type bind at the monopoly prices, then the high demand type benefits as well, because his price is decreased to deter him from buying the inferior good.

The assumptions (1–3) and (5–6) are invariant to rescaling both  $X$  demand curves or both  $Y$  demand curves. As a result, provided  $M_H^{xy}$  stays sufficiently high that the  $X$  types remain excluded from the  $H$  market, the existing theory applies to the case where there are many distinct agents of the  $X$  type and many agents of the  $Y$  type.

Theorem 1 seems relevant for several of the examples discussed in the previous section. The key question is whether the low demand type would be served in the absence of the lower quality good. Thus, chemical products, pharmaceuticals, and software used for both business and education would seem to fit this model. In contrast, the 486 microprocessor and the IBM LaserPrinter E seem best modeled by a continuum of consumer types, rather than two distinct markets.

#### 4. THE SINGLE USE CASE

Index the consumers by the value  $v$  of the high quality good. We assume  $v$  has cumulative distribution function  $F$ , with continuous density  $f$ , and that  $F$  has support  $[a, b]$ . The value of the low quality good to a type  $v$  consumer is  $\lambda(v)$ . We assume

$$\lambda(a) \leq a \quad \text{and} \quad (\forall v) \quad 0 \leq \lambda'(v) < 1. \quad (7)$$

The monopolist has constant marginal cost  $c_H$  for the high quality good and constant marginal cost  $c_L$  for the low quality good. In keeping with the applications previously discussed, we also assume<sup>21</sup>

$$a \leq c_H \leq c_L < b. \quad (8)$$

Note that the demand for the high quality good at price  $p$  is given by  $1 - F(p)$ , and that the demand for the low quality good at price  $p$  is given by  $1 - F(\lambda^{-1}(p))$ . Consequently, as in Section 3, assumptions (7) and (8) imply that if the firm produces only one quality, it produces high quality:

$$(\forall p \geq c_H) \quad (p - c_H)[1 - F(p)] \geq (p - c_L)[1 - F(\lambda^{-1}(p))].$$

When offering only high quality, the profit maximizing price  $p_1$  must satisfy

$$0 = p_1 - c_H - \frac{1 - F(p_1)}{f(p_1)}. \quad (9)$$

21. Forcing  $c_H \geq a$  is convenient to ensure an interior solution to the firm's maximization problem, but not necessary. In particular, it is possible to place assumptions directly on the inverse hazard rates used below.

To ensure uniqueness of a solution to eq. (9), we employ the usual hazard rate assumption, familiar in all adverse selection models:

$$(\forall x \in (a, b)) \quad x - \frac{1 - F(x)}{f(x)} \text{ is increasing.} \quad (10)$$

We also assume that  $\lim_{x \rightarrow b} [1 - F(x)]/f(x) = 0$ . This is satisfied if  $F$  is analytic or if  $f(b) > 0$ . Conditions (8) and (10) then imply that a solution to eq. (9) exists and satisfies  $a < p_1 < b$ .

The condition for the low quality good, analogous to (10), will also prove useful. Selling only the low quality good, the firm earns

$$\pi_L(p) = (p - c_L)[1 - F(\lambda^{-1}(p))]. \quad (11)$$

Thus,

$$0 = \frac{\partial \pi_L}{\partial p} = -\frac{f(\lambda^{-1}(p))}{\lambda'(\lambda^{-1}(p))} \left( p - c_L - \lambda'(\lambda^{-1}(p)) \frac{1 - F(\lambda^{-1}(p))}{f(\lambda^{-1}(p))} \right). \quad (12)$$

To guarantee that the profit function  $\pi_L$  has a unique maximum for every  $c_L \in (a, \lambda(b))$ , we assume

$$\lambda(v) - \lambda'(v) \frac{1 - F(v)}{f(v)} \text{ is increasing.} \quad (13)$$

We now turn to the case of two qualities. Note that when the firm offers only high quality, some segment of the market remains unserved, since  $p_1 > a$ . Introducing  $L$  can draw some of these consumers into the market, and hence is potentially profitable. However, by introducing  $L$  the seller necessarily cannibalizes some of his high quality market. Whether or not introducing  $L$  is profitable depends upon the strength of these two opposing forces. We will now set up the seller's optimization problem, and provide necessary and sufficient conditions under which he is willing to introduce  $L$ .

By (7), the premium a consumer is willing to pay for the increase in quality of  $H$  over  $L$ ,  $v - \lambda(v)$ , is increasing in  $v$ . This ensures that if both goods are offered for sale, high quality will be targeted towards high valuation consumers, and low quality towards low valuation consumers. More precisely, let  $v_H$  be the consumer type who is indifferent between purchasing either good,<sup>22</sup>

$$v_H - p_H = \lambda(v_H) - p_L. \quad (14)$$

22. The critical value  $v_H$  may lie outside the range  $[a, b]$ , but profit maximization ensures that this will never happen in equilibrium.

and let  $v_L$  be the type who is indifferent between purchasing  $L$  and not purchasing at all,<sup>23</sup>

$$p_L = \lambda(v_L). \quad (15)$$

Then consumers in the interval  $[v_L, v_H]$  purchase  $L$ , and consumers in the interval  $[v_H, b]$  purchase  $H$ .<sup>24</sup>

It is useful to express the monopolist's profits in terms of the types of consumers making purchases rather than the prices directly:

$$\begin{aligned} \pi &= (p_H - c_H)[1 - F(v_H)] + (p_L - c_L)[F(v_H) - F(v_L)] \\ &= [v_H - \lambda(v_H) + \lambda(v_L) - c_H][1 - F(v_H)] + [\lambda(v_L) - c_L] \\ &\quad \times [F(v_H) - F(v_L)] \\ &= [v_H - \lambda(v_H) + c_L - c_H][1 - F(v_H)] + [\lambda(v_L) - c_L][1 - F(v_L)]. \end{aligned} \quad (16)$$

That is, we can view the firm's maximization problem as maximizing  $\pi$  subject to  $a \leq v_L \leq v_H \leq b$ . If  $v_L = v_H$ , then  $\pi$  gives the one quality outcome.

If  $v_L < v_H$ , the first order conditions for maximizing  $\pi$  are

$$0 = \lambda(v_L) - c_L - \lambda'(v_L) \frac{1 - F(v_L)}{f(v_L)}, \quad (17)$$

$$v_H - c_H - \frac{1 - F(v_H)}{f(v_H)} = \lambda(v_H) - c_L - \lambda'(v_H) \frac{1 - F(v_H)}{f(v_H)}. \quad (18)$$

If  $v_L = v_H$ , then the right-hand side of (17) must be nonpositive, and the left-hand side of (18) is zero, in accordance with (9). In order to ensure a unique solution for  $v_H$ , we need the following regularity condition, which also ensures that a solution to the first order conditions (17) and (18) yields a global profit maximum.<sup>25</sup>

$$v - \lambda(v) - [1 - \lambda'(v)] \frac{1 - F(v)}{f(v)} \text{ is increasing.} \quad (19)$$

23. Without loss of generality, we can assume that  $a \leq v_L \leq v_H \leq b$ , since values outside this range produce zero quantities.

24. Note that  $[v_L, v_H]$  is a nontrivial interval if and only if  $p_L < \lambda(p_H)$ .

25. When  $c_L < a$ , a variety of other cases emerge. The possible solutions to these cases are as follows: (i)  $a = v_L = p_1 = v_H$ , (ii)  $a = v_L = p_1 < v_H$ , and (iii)  $a = v_L < p_1 \leq v_H$ . In case (i), the low quality good is not introduced. In case (ii),  $p_H > p_1$ , and high  $v$  types are worse off when the low quality good is introduced. Case (iii) requires further assumptions to make a Pareto comparison.

26. Provided  $f$  is differentiable, the assumption (19) is equivalent to the remarkably weak condition  $(\partial/\partial v) [1 - \lambda'(v)][1 - F(v)]^2/f(v) < 0$ . By (10), then, (19) holds if  $\lambda$  is convex, or not too concave.

Under these conditions, we have:

**LEMMA 1** *Suppose eqs. (7)–(19) hold. Then introducing  $L$  is profitable if and only if*

$$\lambda(p_1) - c_L - \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)} > 0. \tag{20}$$

Henceforth, we will therefore assume that (20) holds. We now turn to the conditions under which introducing  $L$  produces a Pareto improvement. The next lemma provides some immediate insights into the welfare consequences of introducing good  $L$ .

**LEMMA 2** *Suppose (7)–(20) hold. Then  $v_L < p_1 < v_H < b$ .*

Lemma 2 shows that, selling two qualities, the monopolist sells to more consumers but sells fewer units of the high quality good, relative to selling only the high quality. Note that introducing the low quality good makes consumers with valuations in  $(v_L, p_1)$  always strictly better off. Obviously, by (20), the monopolist benefits as well. A Pareto improvement therefore occurs if and only if  $p_H \leq p_1$ , for then high valuation customers are made better off as well.

The next result provides conditions sufficient to ensure that  $p_H < p_1$ , so that all market participants other than consumers with valuations below  $v_L$  (who do not get to purchase under either scenario) are strictly better off.

**THEOREM 2** *Suppose that (7)–(20) hold, that  $[1 - F(v)]/f(v)$  is nonincreasing, and that  $\lambda'(v) [1 - F(v)]/f(v)$  is nondecreasing. Then  $p_H < p_1$ , that is, introducing the low quality good is a Pareto improvement.*

The hypotheses of Theorem 2, in conjunction with (7)–(21), are not vacuous, as the following example demonstrates.

*Example 1:* Let  $F(v) = 1 - e^{-(v-a)/\alpha}$  for  $v > a$ , with  $b = \infty$ , and  $\lambda(v) = \beta v + e^{-\beta v} + (1 - \beta)a - e^{-\beta a}$ . The parameters are assumed to satisfy  $a < c_H + \alpha$ ,  $(1 + \alpha\beta)e^{-\beta a} < 1$ , and

$$c_H \leq c_L < \beta[c_H + (1 + \alpha\beta)e^{-\beta(c_H + \alpha)}] + (1 - \beta)a - e^{-\beta a}.$$

These are satisfiable if  $\beta$  is near zero and  $\alpha > 1$ . All of the assumptions are strictly satisfied, and thus are robust to perturbations in a smooth  $C^1$  metric.

Nevertheless, the conditions guaranteeing a Pareto improvement for the single use case are much more stringent than for the dual use case. Indeed, assumption (20) fails for many specifications of the environment, as the next result shows.

**LEMMA 3** *Suppose  $\lambda(v)/v$  is nondecreasing. Then (20) fails.*

A sufficient condition for (20) to fail is that  $\lambda(0) = 0$  and  $\lambda$  is convex, for that implies the hypothesis of Lemma 3. Recall that a convex  $\lambda$  is a sufficient condition for (19) to hold (see footnote 26). This makes (20) seem somewhat unnatural. However, when (20) fails, we won't ever observe a good damaged to produce a lower quality good. In spite of the large number of examples of the use of this strategy, clearly most goods sold are not intentionally damaged by the manufacturer. Thus, it may be that (20) only holds for a small percentage of all goods sold, but still on a large number of goods.

It might be countered that the conditions of Theorem 2 are sufficient rather than necessary, and hence that the above conclusion is premature. That this is not the case is shown by our next result. For "large"  $c_L$ , we have an exact characterization of the conditions under which a Pareto improvement occurs. Suppose  $f$  is continuously differentiable, and let  $\bar{c}_L$  just make (20) fail, that is,

$$\bar{c}_L = \lambda(p_1) + \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)}.$$

Then we have

**THEOREM 3** *For  $c_L$  close to  $\bar{c}_L$ , one has  $p_H < p_1$  if and only if  $\lambda$  is convex.*

Thus, convexity of  $\lambda$  is necessary for a general result. However, as shown in Lemma 3, convexity of  $\lambda$  tends to make introducing  $L$  unprofitable. The reader may wonder why the conditions for a Pareto improvement are so much more stringent in the single use case than in the dual use case. The intuition is as follows. In the dual use case, assumption (4) implies that lowering the price of good  $H$  below its monopoly price will not cannibalize profits from the  $L$  market. In the single use case, whenever the  $L$  good has positive market share, lowering the price of good  $H$  necessarily results in cannibalization.<sup>27</sup> This has two consequences. First, introducing the  $L$  good is now no longer necessarily profitable. Technically, the first order effect of a reduction in the price of good  $H$  is no longer zero (it is negative). Secondly, because of the cannibalization, the optimal response to a lower price of good  $L$  is now more likely to be a price increase, making a Pareto improvement harder to achieve.

27. The single use model also differs from the dual use model in that individual demand is inelastic (up to the reservation price). Introducing unit demand into the dual use model produces qualitatively similar results to the general dual use case; thus the distinction between the two cases appears to have more to do with discrete types versus a continuum of types than with downward-sloping demand.

## 5. CONCLUSION

For many products, it appears that the cost effective way to segment the market and price discriminate is to damage an existing product to produce a lower quality product, rather than improve the quality of an inferior product or produce the product separately. Although many of our examples are in chemicals and electronics, we think the practice is much more widespread. For example, outlet malls are often located much farther away from major cities than land prices would seem to dictate, as the rent gradient would appear to bottom out once low value use, such as farming, commences. It is often impossible to find the defects in apparel products labeled as "seconds" and sold in discount stores, as we recall from our graduate student days.<sup>28</sup>

We have argued that the phenomenon of manufacturers damaging goods naturally divides into two distinct cases based on customer characteristics. In one case, there are two uses, a high value use and a low value use, and this case seems best modeled as two distinct markets (Sec. 3). Most of the chemicals and pharmaceuticals fit this category, along with business versus educational uses of software. In this case, the conditions for price discrimination to be a Pareto improvement are not severe, and boil down to the assumption that, absent price discrimination, the high value market is more profitable than the low value market.

In the second case, there are not two separate markets for the products, but rather a group of consumers with distinct use values for the two products. Most of the electronics examples would seem to fit this category. This case seems more naturally modeled with a continuum of use values. The restrictions necessary for price discrimination to produce a Pareto improvement seem more severe and unnatural in this case.

In modeling the phenomenon, we have not endogenized the quality of the inferior good, mainly because of the resulting complexity of the mathematical description of preferences, which must be defined over all possible qualities, and firms' costs, which are now a function not only of quantity but also of quality. It is clear from some of the examples that quality is endogenous. For example, the slowdown of

28. The obliteration of manufacturers' tags for high fashion clothes by discount houses might seem like an example of damaging goods, but has an alternative, compelling explanation. Many upscale retailers, such as Neimann-Marcus, accept returns of high fashion clothes without a receipt. If the same item could be purchased unblemished at a discounter, it would not be possible for the upscale retailers to accept returns without proof of purchase, to the annoyance of their customers. Consequently, manufacturers rip labels or mark through them, as a signal that the item was sold by a discounter.

the IBM LaserPrinter was chosen by IBM. Characterizing the conditions for a Pareto improvement when quality is endogenous would appear to be an interesting, if daunting, research goal.

At least two of the damaged goods, the Intel 486SX and the IBM LaserPrinter E, appear to have been introduced in response to competition by another producer. This is difficult to explain, particularly in the LaserPrinter case. In this case, IBM's regular LaserPrinter was significantly faster than the Hewlett-Packard IIP, and thus the products were significantly differentiated. By introducing a product comparable to the IIP, IBM makes the market for the slower printers more competitive, thereby reducing prices for the slower printers. This should have the effect of further undercutting the market for the faster printer, as Hewlett-Packard responds to IBM's LaserPrinter E with a price cut.

It is possible as a theoretical matter that IBM's ability to punish Hewlett-Packard for competitive pricing is enhanced by the introduction of the similar product. Not only does IBM gain the ability to take a large percentage of HP's sales by aggressive pricing, but it limits the cost of such punishment by permitting higher prices on the faster machine. Thus, the set of equilibria to the repeated pricing game may grow, admitting some equilibria with higher profits than existed absent the introduction of the LaserPrinter E. Nevertheless, this seems an unlikely explanation for IBM's behavior.

An alternative explanation involves the perceived need of manufacturers to offer a full line of products. Offering a full line clearly makes consumers feel more comfortable, perhaps because the firm is less likely to exit, more likely to support the products with technical help and product updates, and more likely for the products themselves to be of high quality because of experience with the industry. Further research into price discrimination by imperfectly competitive firms seems warranted.

#### APPENDIX: PROOFS

*Proof of Theorem 1.* From (3), if the firm sells both goods, it sells  $H$  to  $Y$  and  $L$  to  $X$ . Let  $p_H$  be the price of  $H$  and  $p_L$  be the price of  $L$ . The firm solves the following maximization problem:

$$\max_{p_H, p_L} (p_L - c_L)x_L(p_L) + (p_H - c_H)y_H(p_H) \quad (P)$$

subject to

$$\int_{p_H}^{\infty} y_H(p) dp \geq \int_{p_L}^{\infty} y_L(p) dp, \tag{IC_y}$$

$$\int_{p_H}^{\infty} x_H(p) dp \leq \int_{p_L}^{\infty} x_L(p) dp. \tag{IC_x}$$

Let  $(p_H^*, p_L^*)$  solve (P). First, note that  $x_L(p_L^*) > 0$ . Suppose by way of contradiction that  $x_L(p_L^*) = 0$ . Then  $p_H^* = M_H^x = M_H^{xy}$ , and by (5) we may assume that  $p_L^* = \bar{p}_L$ . Consider the following deviation:  $p_L = p_L^* + \Delta p_L$ ,  $p_H = p_H^* + \Delta p_H$ . Choose  $\Delta p_L < 0$ , and let  $\Delta p_H$  be determined as follows. If  $IC_x$  does not bind, let  $\Delta p_H = 0$ . If  $IC_x$  binds, choose  $\Delta p_H$  so that  $IC_x$  holds with equality. In either case, by (3) and (5),  $IC_y$  is satisfied. This deviation increases profits when  $\Delta p_L$  is sufficiently small, for

$$\begin{aligned} \left. \frac{\partial \pi}{\partial p_L} \right|_{p_L = \bar{p}_L, p_H = M_H^y} &= [(M_H^y - c_H)y_H'(M_H^y) + y_H(M_H^y)] \frac{dp_H}{dp_L} \\ &\quad + (\bar{p}_L - c_L)x_L'(\bar{p}_L) + x_L(\bar{p}_L) \\ &= (\bar{p}_L - c_L)x_L'(\bar{p}_L) < 0. \end{aligned}$$

This contradicts the hypothesis that  $(p_H^*, \bar{p}_L)$  solved (P).

That  $x_L(p_L^*) > 0$  implies  $IC_x$  does not bind. For suppose it does. Then (3) implies  $IC_y$  does not bind, implying that  $p_H^* = M_H^y$ , which implies that  $IC_x$  does not bind by (4). There are two remaining possibilities: either  $IC_y$  does not bind, or it does [which occurs when (6) holds].

If  $IC_y$  does not bind, then  $p_H^* = M_H^y$ , so both  $X$  and the firm are better off, and  $Y$  obtains the same (monopoly) price as when the firm only offers one quality.

Now suppose  $IC_y$  binds, that is, (6) holds. It must therefore be the case that  $p_H^* < M_H^y$ , for  $p_H^* \geq M_H^y$  and (6) imply  $p_L^* > M_L^x$ , and lowering both prices increases profits. Hence if (6) holds, a strict Pareto improvement occurs and all three agents benefit.  $\square$

*Proof of Lemma 1.* Suppose that

$$\lambda(p_1) - c_L - \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)} \leq 0,$$

and that  $v_H > v_L$ . Then by (17) and (13) we have  $v_L \geq p_1$ , so  $v_H > p_1$ . Now

$$\frac{\partial \pi}{\partial v_H} = f(v_H) \left[ - \left( v_H - c_H - \frac{1 - F(v_H)}{f(v_H)} \right) + \left( \lambda(v_H) - c_L - \lambda'(v_H) \frac{1 - F(v_H)}{f(v_H)} \right) \right],$$

and by (19)

$$\frac{1}{f(v_H)} \frac{\partial \pi}{\partial v_H} (v_H) < \frac{1}{f(p_1)} \frac{\partial \pi}{\partial v_H} (p_1).$$

The hypothesis and (9) then imply that

$$\frac{\partial \pi}{\partial v_H} (p_1) \leq 0,$$

contradicting that  $v_H$  is chosen optimally.

Conversely, suppose that

$$\lambda(p_1) - c_L - \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)} > 0,$$

and that  $v_H = v_L$ . Then  $v_L = v_H = p_1$ , and so

$$\frac{\partial \pi}{\partial v_L} (v_L) = -f(v_L) \left( \lambda(v_L) - c_L - \lambda'(v_L) \frac{1 - F(v_L)}{f(v_L)} \right) < 0,$$

contradicting that  $v_L$  is chosen optimally. □

*Proof of Lemma 2.* If  $v_L = v_H$ , then  $v_H = p_1$ , and by (20)

$$\frac{\partial \pi}{\partial v_L} = -f(p_1) \left( \lambda(p_1) - c_L - \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)} \right) < 0.$$

Thus  $v_L < v_H$ . By (13) and (17),  $v_L < p_1$ . Therefore, by (13) and (18),

$$\begin{aligned} v_H - c_H - \frac{1 - F(v_H)}{f(v_H)} &= \lambda(v_H) - c_L - \lambda'(v_H) \frac{1 - F(v_H)}{f(v_H)} \\ &> \lambda(v_L) - c_L - \lambda'(v_L) \frac{1 - F(v_L)}{f(v_L)} = 0. \end{aligned}$$

By (9) and (10),  $v_H > p_1$ . Finally, if  $v_H = b$  and if  $f(b) > 0$ , then by (7) and (8)

$$\frac{\partial \pi}{\partial v_H} (b) = f(b) [-b + c_H + \lambda(b) - c_L] < 0.$$

If  $v_H = b$  and  $f(b) = 0$ , then  $\lim_{x \rightarrow b} [1 - F(x)]/f(x) = 0$  implies that

$$\frac{\partial \pi}{\partial v_H}(v_H) < 0$$

for some neighborhood around  $b$ . In either case, this contradicts the optimality of  $v_H$ .  $\square$

*Proof of Theorem 2.* By (9), (15), (17), (18),

$$\begin{aligned} p_H - p_1 &= v_H - \lambda(v_H) + \lambda(v_L) - p_1 \\ &= \frac{1 - F(v_H)}{f(v_H)} [1 - \lambda'(v_H)] + \lambda'(v_L) \frac{1 - F(v_L)}{f(v_L)} - \frac{1 - F(p_1)}{f(p_1)} \\ &= \left( \frac{1 - F(v_H)}{f(v_H)} - \frac{1 - F(p_1)}{f(p_1)} \right) \\ &\quad - \left( \lambda'(v_H) \frac{1 - F(v_H)}{f(v_H)} - \lambda'(v_L) \frac{1 - F(v_L)}{f(v_L)} \right) < 0, \end{aligned}$$

by the monotonicity assumptions and Lemma 3.  $\square$

*Proof of Theorem 3.* First note that (17) implies

$$\frac{\partial v_L}{\partial c_L} = \frac{1}{\lambda'(v_L) \left[ 2 + \frac{1 - F(v_L)}{f(v_L)^2} f'(v_L) - \frac{\lambda''(v_L)}{\lambda'(v_L)} \frac{1 - F(v_L)}{f(v_L)} \right]}$$

and

$$\frac{\partial v_H}{\partial c_L} = \frac{-1}{[1 - \lambda'(v_H)] \left[ 2 + \frac{1 - F(v_H)}{f(v_H)^2} f'(v_H) + \frac{\lambda''(v_H)}{1 - \lambda'(v_H)} \frac{1 - F(v_H)}{f(v_H)} \right]}$$

Thus, evaluating at  $c_L = \bar{c}_L$ , we have  $v_H = v_L = p_1$  and

$$\begin{aligned} \left. \frac{\partial p_H}{\partial c_L} \right|_{c_L = \bar{c}_L} &= [1 - \lambda'(p_1)] \frac{\partial v_H}{\partial c_L} + \lambda'(p_1) \frac{\partial v_L}{\partial c_L} \\ &= \frac{-1}{2 + \frac{1 - F(p_1)}{f(p_1)^2} f'(p_1) + \frac{\lambda''(p_1)}{1 - \lambda'(p_1)} \frac{1 - F(p_1)}{f(p_1)}} \\ &\quad + \frac{1}{2 + \frac{1 - F(p_1)}{f(p_1)^2} f'(p_1) - \frac{\lambda''(p_1)}{\lambda'(p_1)} \frac{1 - F(p_1)}{f(p_1)}} \end{aligned}$$

By (13) and (19), this is positive if and only if  $\lambda''(p_1) > 0$ . That is, near  $\bar{c}_L$ ,  $p_H$  falls as  $c_L$  is decreased if and only if  $\lambda$  is convex.  $\square$

*Proof of Lemma 3.*  $\lambda(v)/v$  is nondecreasing if and only if  $v\lambda'(v) \geq \lambda(v)$ . Thus,

$$\begin{aligned} \lambda(p_1) - c_L - \lambda'(p_1) \frac{1 - F(p_1)}{f(p_1)} &\leq \lambda'(p_1) \left( p_1 - \frac{1 - F(p_1)}{f(p_1)} \right) - c_L \\ &= \lambda'(p_1)c_H - c_L < 0. \end{aligned} \quad \square$$

## REFERENCES

- DeLamarter, T., 1986, *BIG BLUE: IBM's Use and Abuse of Power*, New York: Dodd, Mead & Company, 161–175.
- Fisher, F.M., J. McGowan, and J.E. Greenwood, 1983, *Folded, Spindled and Mutilated: Economic Analysis and U.S. vs. IBM*, Cambridge, MA: The MIT Press.
- Frenkel, G., 1991, "For Intel, a 486 Chip by Any Other Name is Still the Same," *PC Week Supplement*, supplement to *PC Week*, 8(5), S37–S38.
- Harley, R., 1992, "Industry Update: Japan," *Stereophile*, 15(11), 50–63.
- Harley, R., 1994, "Sony MDS-501 Home Minidisc Recorder/Player & MZ-E2 Portable Minidisc Player," *Stereophile*, 17(9), 115–129.
- Jones, M., 1990, "Low-Cost IBM LaserPrinter E Beats HP LaserJet IIP on Performance and Features," *PC Magazine*, 8, No. 10, May 29, 33–36.
- Larson, M., 1993, "Packaging for Success in Warehouse Clubs," *Packaging*, 38(7), 26–28.
- Nagle, T., 1987, *The Strategy and Tactics of Pricing: A Guide to Profitable Decision-Making* (p. 168). Englewood Cliffs, NJ: Prentice-Hall.
- Seymour, J., 1991, "Intel: Is it Still 'The Computer Inside'?", *PC Week*, November 11, 73.
- Smith, B. and A. Villamil, 1993, "Efficiency, Randomization and Commitment in Government Borrowing," Mimeo, Cornell University, June.
- Stocking, G. and M. Watkins, 1947, *Cartels In Action: Case Studies in International Business Diplomacy* (pp. 402–405). New York: The Twentieth Century Fund.
- Tirole, J., 1988, *The Theory of Industrial Organization*, Cambridge, MA: The MIT Press.
- Varian, H., 1985, "Price Discrimination and Social Welfare," *American Economic Review* 75(4), 870–875.

## **Subsidizing Creativity through Network Design: Zero-Pricing and Net Neutrality**

Robin S. Lee and Tim Wu

**T**he “net neutrality” debate, as it has emerged over the last five years, is a social, political and economic debate over the public information network known as the Internet and the duties of its private carriers, which include telephone and cable companies and other Internet service providers (ISPs). In the early 2000s, questions surrounding the rights of Internet carriers to block certain network attachments and control access to emergent applications or content providers led to a call to protect “network neutrality” (Wu, 2003).<sup>1</sup> The debate raises familiar questions for students of travel or communications infrastructure: for hundreds of years, courts and governments have struggled over the duties that carriers like ferries, railroads, or telephone companies owe to the public by virtue of their necessity to economic and social welfare. While the net neutrality debate has many aspects, in this paper we focus on one crucial issue: the de facto rule

<sup>1</sup> In March 2005, the Federal Communications Commission fined an Internet service provider, Madison River, for blocking an Internet telephony service. In December 2006, the FCC imposed net neutrality rules on AT&T as a condition of its merger with Bell South, forcing it to commit itself “not to provide or to sell to Internet content, application, or service providers, including those affiliated with AT&T/BellSouth, any service that privileges, degrades or prioritizes any packet transmitted over AT&T/BellSouth’s wireline broadband Internet access service based on its source, ownership or destination” (Reardon, 2006). In July 2008, the FCC criticized cable firm Comcast for deliberately interfering with transmissions occurring through BitTorrent peer-to-peer software, which allows large and/or popular files to be downloaded simultaneously by many users, and required that such interference cease.

■ *Robin S. Lee is Assistant Professor of Economics, Stern School of Business, New York University, New York City, New York. At the time this paper was written, he was Research Scientist, Yahoo! Research (Microeconomics), New York City, New York. Tim Wu is Professor of Law, Columbia University Law School, New York City, New York, and a Fellow at the New America Foundation, Washington, D.C. Their e-mail addresses are <rslee@stern.nyu.edu> and <wu@pobox.com>.*

prohibiting consumers' ISPs from charging fees to content providers for access to their customer base.

The Internet can be understood as an information network that intermediates between different groups of agents comprising *users* and *content providers*, with the critical understanding that users can also act as content providers (for the purposes of discussion, we will use the term "content" loosely, referring to all types of media, applications, retailers, and services available online). Since the Internet's spread through academia in the 1980s and mass popularity in the 1990s, it has maintained a pricing structure that is unique among information networks: users and content providers typically pay ISPs *access fees*—fixed fees to get on the Internet at all—and *usage fees*—variable fees paid based on time or bandwidth usage; however, there have not generally been any additional charges for one user of the network to reach another user or content provider. For example, content providers such as Google and Wikipedia, while paying for their own Internet access and usage, do not directly pay the ISPs of users they reach.

This began to change in the mid-2000s as certain Internet service providers began to discuss their desire to charge certain users of the Internet—notably, large application or content providers—additional fees to reach their subscribers. For example, AT&T, an ISP, might demand that content providers such as Google and Wikipedia pay AT&T to access AT&T's customers; failure to comply would result in AT&T blocking traffic from those sites to its customers (and also preventing its own customers from reaching those sites). Following usage in the telephone system, we refer to these fees as *termination fees*: AT&T would charge content providers a fee to deliver their packets much like it charges other telephone networks a fee to "terminate" their calls.

Today, through historical practice, there exists a *de facto* ban on termination fees—also referred to as a "zero-price" rule (Hemphill, 2008)—which forbids an Internet service provider from charging an additional fee to a content provider who wishes to reach that ISP's customers. The question is whether this zero-pricing structure should be preserved, or whether carriers should be allowed to charge termination fees and engage in other practices that have the effect of requiring payment to reach users. This paper begins with a defense of the *de facto* zero-price rule currently in existence. We point out that the Internet, as an intermediary between users and content providers, exhibits pricing dynamics similar to other intermediaries in "two-sided markets." In particular, we posit that the Internet's absence of payments from content creators to users' ISPs facilitates the entry of content creators. In that respect, the rule provides an alternative implementation of the policy goals provided by the intellectual property system and achieves functions similar to copyright and patent law. The rule also helps avoid the problems of Internet fragmentation, in which content providers who do not reach agreements with ISPs cannot access all customers, and consumers on a single ISP are foreclosed from accessing their content.

We then consider some of the main arguments against restricting termination fees. Although we concede that limited instances may justify certain deviations from this norm, ultimately we argue that a zero-price rule has helped make the Internet distinctive from other networks in terms of its level of creativity and social usefulness.

## **Zero-Pricing and Net Neutrality**

The general practices specified by the phrase “network neutrality” emerged not as a closely considered policy decision, but as a consequence of how the Internet was designed and how it spread. We claim, however, that an economic case for the pricing rules inherent in a neutral network can be found in the theory of two-sided markets. This theory suggests the de facto ban on termination fees may be interpreted as a policy that provides a subsidy to content creation and provision. For a number of reasons, this subsidy appears to have been one of the forces generating the enormous wave of innovation in services and information in the last few years as well as spawning new forms of creative activity.

### **Internet Pricing Contrasted with Telephone and Cable Television Networks**

Currently, Internet users pay access and usage fees to their service provider and then can reach any other user who is similarly connected to the Internet. The overall network does not, by its own design, distinguish between content providers and users. Consequently, content providers—who may also be users—are also able to reach an audience consisting of every single Internet user. These norms and expectations, which have created a de facto ban on termination fees, stands in sharp contrast to what is standard practice on other important information networks, like the telephone and cable networks.

One reason for the differences between networks is rooted in history. The Internet was conceived by various visionaries, particularly the Department of Defense researchers J. C. R. Licklider and Robert W. Taylor, as a “network of networks” or an “intergalactic network” that would make it possible for users of any single computer network to reach users on any other network (Licklider and Taylor, 1968). In its original, noncommercial setting, fees to access the network were paid by universities, government, and research departments. There was no motive to charge termination fees to content providers. Government regulations designating the Internet a noncommercial network also discouraged any such fees. On the technological side, Internet protocols were designed to create a network that could be universally used by different parties with very little centralized knowledge as to who was reaching whom; as a result there was little power to track or bill for reaching certain entities. In the early years, the complexity of trying to incorporate billing capabilities might have doomed the project.

In contrast, telephone and cable television networks were designed from the outset as commercial networks, where payments were the prerequisite to connectivity at all. The Bell telephone system was, from its origins, extremely careful about who would be allowed to connect to Bell customers. In the two historically “competitive” periods in wire-line telephony—from the 1890s to the 1910s and from the 1980s through the early 2000s—calling a Bell customer generally meant paying a termination fee to the owner of the local switch. Since the 1970s, termination fees on the telephone system have been regulated based on fears that the Bell companies would use their “termination monopoly”—their exclusive access to

customers—to charge exorbitant rates and bankrupt any would-be competitor. Cable television networks are priced in a similar, but in some ways opposite, manner. From the beginning, cable companies were faced with demands for fees to access the content they needed from broadcasters, copyright owners, or other sources of content. As opposed to charging to reach their users, cable companies currently pay content providers for the right to carry content on their cable networks.

### **The Consequences of Pricing Decisions in Two-Sided Markets**

The historical reasons why the Internet has developed without termination fees does not address the issue of whether the principle is economically justified. One potentially powerful justification emerges from the economic theory of two-sided markets.

Generally, the Internet, as a network (or network of networks), can be seen as a market or intermediary that facilitates the interaction of two main groups: users and content providers. Other well-studied examples of two-sided markets include payment systems, such as credit cards or online services; hardware-software markets like videogames or operating systems; retail marketplaces such as bazaars, shopping malls, or auction houses; matching markets, such as nightclubs or job sites; and advertising exchanges, such as online advertising platforms as well as commercial telephone books.

One prevalent aspect of two-sided markets is the presence of “network effects” (Farrell and Saloner, 1985; Katz and Shapiro, 1986; Liebowitz and Margolis, 1994), where the value of a service generally increases in the number of users. Most obviously, the network is more useful to users the more people there are to e-mail, videoconference, or instant message. There are also indirect network effects in that users benefit from more content providers, and these providers also benefit from having more users to reach. “Negative” network effects are also possible: for example, in networks with congestion, the value of using a network may decrease with additional users.

The recent literature on two-sided markets has studied the pricing decisions of a network provider or platform intermediary (Rochet and Tirole, 2003, 2006; Caillaud and Julien, 2003; Armstrong, 2006; Hagiu, 2006) in the presence of such network effects and has shown that in settings with transactions costs or limits on side-payments that are allowable between agents, the division of prices charged between sides of the market matters greatly. When there are such frictions, charging fees to one side as opposed to the other—even if the sum of prices across sides is the same—can affect who uses the network, overall transaction volume, and ultimately the efficiency of the market.

Theory predicts that the optimal pricing decision for a social planner as well as for a monopolist platform provider might involve “subsidizing” one side of the market over the other(s)—that is, possibly charging one side below the marginal cost of providing service. A canonical example of this phenomena are credit card companies. Credit cards companies do not charge both sides of the market—merchants and consumers—equally. Rather, they charge a transaction fee to

merchants and typically reward consumers for using their cards with frequent flier miles or other benefits. Because using a credit card is often not only “free” for a consumer, but in fact subsidized, the effect is to encourage consumers to use credit cards more than they would if fees were charged to both sides of the market. The division of pricing matters in this case partially because merchants are often prohibited from charging consumers different prices depending on whether the method of payment is by cash or credit; in other words, since merchants are unable to “unwind” these fees, pricing has a “nonneutral” effect.

If one side is subsidized at the optimum, theory predicts that the side receiving the subsidy will be the side that either has a higher elasticity of demand with respect to price, or the side that exhibits stronger cross-side network effects—that is, an additional agent on one side of the market increases utility on the other side by more than an agent on the other side would for the original side. Another way of interpreting this result is that since demand and demand elasticities for one side of the market change when the number of agents on the other side increases, the optimal prices charged by a platform may in fact be lower than those predicted if the platform ignored these network effects and assumed demand was fixed. All of these factors lead to more complex pricing dynamics than in traditional one-sided markets, which if ignored can lead to misguided regulatory or antitrust policies (Evans, 2003; Wright, 2004a).

In contrast, if there were no transaction costs or limitations on side-payments, the division of pricing between two sides of a market would not influence transaction volume; instead, any division would merely be a transfer between different sides of the market. In these standard settings, any distinction between how the two sides of the market are charged is unnecessary. Such is the case with a value-added tax, as whether firms pay a value-added tax during the stages of production or consumers pay a sales tax at the point of purchase does not affect transaction volume. Indeed, Rochet and Tirole (2006) use the nonneutrality of price structure—that is, for a given price level, the allocation of prices across sides influences transaction volume—as their definition of what makes a market two-sided.

These insights provided by the literature on two-sided markets naturally apply to the Internet.<sup>2</sup> First, as discussed, network effects are prevalent: the decision by users to subscribe to Internet access depends on which content providers are online; similarly, the decision of a content provider to invest in the creation of new content and applications depends on the number of users that can be reached. Consequently, the willingness to pay (and elasticity of demand) of an agent varies according to usage by others.

Secondly, there are practical limits on the side-payments that can be easily conducted between different sides of the Internet. Due to informational asymmetries, access or subscription charges to content significantly reduce usage as con-

<sup>2</sup> Others have explored this connection: for example, Economides and Tåg (2007) offer an application of the two-sided market framework to modeling the Internet, while Hurwitz (2006) and Hemphill (2008) also make the connection in their policy papers.

sumers find it difficult to determine the quality of on-line content before they have used it and thus often choose not to use it at all rather than to pay an up-front fee. Moreover, payments for Internet services in general have a variety of frictions: limited consumer access to credit cards or electronic payment systems; issues with security and fraud; problems of pricing micro-transactions for very limited access; and costs of monitoring whether purchased electronic content is being used only for agreed-upon purposes. Widespread adoption of usage-based pricing by content providers is in many cases thus either impractical or inefficient; it is not surprising, then, that many content providers have chosen to rely on advertising revenues and to provide content at no charge to the user. The presence of these myriad challenges preventing content providers from passing potential termination fees on to consumers thus implies that the division of pricing will be nonneutral. This tempers the “indirect extraction” argument (raised for example by Hemphill, 2008) that a ban on termination fees would be unraveled via side-payments.

### **Subsidizing Content Creation and Invention**

The literature on two-sided markets combined with the economic realities of the Internet suggests that the ban on termination fees serves important economic and potentially social functions.<sup>3</sup> First, the rule provides a direct subsidy for the production of content and inventions. Second, it also cheapens market entry by making it easier for users to switch and become content providers themselves. Vinton Cerf (2006), a co-designer of the TCP/IP protocol, captures both these points: “Because the network is neutral, the creators of new Internet content and services need not seek permission from carriers or pay special fees to be seen online. As a result, we have seen an array of unpredictable new offerings . . . [E]ntrepreneurs need not worry about getting permission for their inventions will [sic] reach the end users . . . This is a direct contrast to closed networks like the cable video system, where network owners control what the consumer can see or do.”

Providing subsidies for the production of creative works and innovation is a typical goal of many government programs, including the copyright and patent laws and institutions such as the National Institute of Health or National Endowment of the Arts. Such subsidies are most often justified in economic terms as a remedy for market failure in the production of creative works or new inventions (Landes and Posner, 1989). Since both creative works and inventions have characteristics of a public good, such intervention may be necessary to avoid underproduction.

The pricing structure of the Internet can be seen as an alternative means of subsidizing creativity and innovation (as discussed in Benkler, 1999). As economic analysis suggests, setting a preferable price or ruling out certain types of fees for content providers may encourage creation of content or new inventions that would not otherwise occur. As Lessig and McChesney (2006) note, “more than 60 percent of Web content is created by regular people, not corporations,” and over 100 million

<sup>3</sup> Other effects of banning payments from content providers to Internet intermediaries (that is, using a “bill-and-keep” system) can be found in DeGraba (2000), Hemphill (2008), and Wright (2004b).

blogs have so far been documented (*Technorati*, 2008). The Internet, as a platform, has spawned thousands of new firms and millions of sites, from mass content projects such as Wikipedia to search indexers and content aggregators such as Google and Yahoo!.

In addition to the traditional justifications for subsidizing creativity and invention, there are special reasons that subsidizing in this context might be useful. The Internet content and applications market can be understood as a “hit-driven industry,” where hit products like those from Google, Yahoo!, Amazon, eBay, and others create massive spillover value for users. But extreme uncertainty plagues the creation of distinctly new content, and for every hit product there have been numerous failures, with names like Go.com, Pets.com, and Webvan.com. Given that the returns to content production are skewed and the expected value of a new online venture is low, sufficiently low costs of entry may have been and may continue to be crucial.

For similar reasons, the lack of termination fees has also been critical to the “long-tail” model of revenue generation: many Internet businesses rely on an extremely diverse product range that caters to individual niche markets (Anderson, 2006). A diverse collection of websites that yield small value individually but high value when considered as a group might not exist if faced with higher operating costs. Even a more targeted or asymmetric scheme of only levying fees on “successful” or large content providers still has the effect of depressing content creation as it reduces the potential gains to innovation for small entrepreneurs with the dream of making it big.

Of course, for a given price level, subsidizing content comes at the expense of *not* subsidizing users, and subsidizing users could also lead to greater consumer adoption of broadband. It is an open question whether, in subsidizing content, the welfare gains from the invention of the next “killer app” or the addition of new content offset the price reductions consumers might otherwise enjoy or the benefit of expanding service to new users. It may prove useful for economic theory to further illuminate and explore this tradeoff. However, given the possibility of users acting as content providers, a more accurate description of a ban on payments from content providers to Internet intermediaries is this: it is a subsidy to the creative and entrepreneurial at the expense of the passive and consumptive.

### **The Danger of Fragmentation**

Setting aside concerns over incentives for creativity, there is another important concern about allowing Internet service providers to charge fees to content producers: it would almost certainly result in service providers “competing” for content, as seen in other platform industries, by charging different fees and bargaining on exclusive arrangements with content providers. In turn, such bilateral agreements would inevitably lead to fragmentation—where certain content would only be available on certain service providers—and hence multiple “Internets.” For example, cable television is a fragmented network in this sense: not only do users

of cable television face a menu of prices for different numbers of channels, but in addition, certain channels are available only in some geographic areas.

Despite arising in equilibrium, these arrangements need not be efficient in networked industries with externalities and an incomplete contracting space (for example, Segal, 1999; Farrell and Weiser, 2003; Whinston, 2006). Potential welfare losses could also be significant, as consumers would find themselves foreclosed from accessing content available only on rival service providers, and content providers would find themselves unable to reach certain segments of the population captive to service providers with whom no agreement had been reached.<sup>4</sup>

Such arrangements would also be anathema to the principle of universality subscribed to by the designers of the Internet. One visionary of the Internet, Tim Berners-Lee (2008), put it this way: “It is of the utmost importance that, if I connect to the Internet, and you connect to the Internet, that we can then run any Internet application we want, without discrimination as to who we are or what we are doing. We pay for connection to the Net as though it were a cloud which magically delivers our packets. We may pay for a higher or a lower quality of service. We may pay for a service which has the characteristics of being good for video, or quality audio. But we each pay to connect to the Net, but no one can pay for exclusive access to me.” Economists might rephrase this principle of universality in a language of positive network externalities and avoiding fragmentation, but the fundamental social value remains the same.

### **The Transactions Cost Argument for a Zero-Price**

While the two-sided markets analysis might seem to justify setting different prices between agents, a crucial question remains: Why a price of zero?

There is a crucial difference between a low price and a zero-price: the ban on termination fees eliminates an entire class of transaction costs. Similar to the way in which developers can write an application for Microsoft Windows without bargaining with Microsoft at all, content providers can reach all consumers without having to negotiate individually or to pay separate fees to every service provider. For example, assuming Internet access, content providers do not have to negotiate with any service provider to get their initial production started. Transaction costs, of course, can be overcome, but their presence or absence matters.

The most obvious beneficiary of the absence of termination fees and related

<sup>4</sup> An argument is sometimes made that allowing exclusive arrangements might help new intermediaries to enter network markets (Lee, 2009)—in this case, the argument would be if new Internet service providers could offer exclusive content, it would be easier for them to differentiate themselves from existing providers and gain market share. However, we argue that this argument should not be given great weight in the context of Internet service providers. First, this literature also has noted that a standardization of network platforms often improves welfare (Farrell and Saloner, 1985; David and Greenstein, 1990); in this case, the benefit of standardization arises from maintaining the Internet as the sole platform and ISPs as simply conduits to the greater network. This argument is strengthened given there is no sole ISP, and hence no monopoly rents extracted upon standardization. In addition, just because net neutrality may prevent ISPs from competing on content, they may still compete in other ways: for example, they can differentiate themselves on quality of service.

transaction costs are media forms like blogs; there are millions of content providers for blogs, which are of highly variable quality. If each content provider somehow had to negotiate to gain access to users, the transaction costs alone might endanger their existence in the first place (Heller, 2008). Furthermore, social media sites such as MySpace and Facebook may not have been able to launch before the viability of their sites had been established. For new entrepreneurs or content providers, it has been unnecessary to reach agreements with every carrier to maximize the number of users and contributors accessible—and hence the value of their content—from the start. If content providers had to run a gauntlet of fees before being widely available, many business models would not have been feasible and many content providers may not have entered.

We note the lack of fees for providing Internet content lies in sharp contrast to other networks such as cable television, which involve intensive negotiation over prices for a channel's inclusion in a cable package. In these contexts, such frictions pose significant difficulties for new content in acquiring distribution and obtaining an audience. In a sense, the transaction costs induced by this bargaining often favors the established, well-financed, or overtly commercial at the expense of niche, specialized, and unproven; as we have argued, content exhibiting the latter characteristics may very well have defined much of the Internet's value and worth.

### **What about Content Providers Charging Service Providers?**

We have not yet discussed the possibility of content providers levying fees on Internet service providers. Should these types of reverse charges be allowed? After all, if subsidization is appropriate, why stop at a zero-price and why not have ISPs pay for content? Such a regime would be similar to what is done in the cable industry, where content providers like CNN and ESPN charge cable operators a per-subscriber fee for the right to carry their content. In the Internet context, some content providers have experimented with such fees; for example, ESPN charges service providers for the right to access its ESPN360.com broadband site, as opposed to charging individual users directly (*Wall Street Journal*, 2006).

For reasons similar to those outlined above, we do not think such arrangements are desirable. The use of discriminatory termination fees, even if negative for some content providers, may still lead to positive fees levied on others. But even if asymmetric regulation would be possible—banning one direction of fees—it would still be the case that allowing for any nonzero-pricing would introduce a new class of fees to the Internet and substantially increase transaction costs, favoring some types of content providers over others. Furthermore, such fees also exacerbate the problem of fragmentation and consumer foreclosure: if some Internet service providers did not wish to pay for certain content, it would be difficult to force them to do so.

At the same time, direct regulation for this particular direction of fee payments may not be necessary, for several reasons. First, any content provider that makes a decision to charge fees will necessarily internalize the effect of having fewer consumers that can access its site; as will be discussed later, internalization is not the

case with Internet service providers levying termination fees since a collective action problem is introduced. Second, unless consumers were willing either to switch service providers or to terminate Internet access altogether in response to a particular content provider's availability, a content provider would not be able to extract rents from an ISP; it is difficult to imagine very many (if any) content providers that fit this criteria. Under these conditions, it is unlikely that a significant number of content providers (if any at all) would find it desirable to charge service providers for access. Even ESPN has shown signs of weakening its policy: it now allows anyone with a .edu or .mil domain to access its broadband site, where previously only 20 million subscribers of the ISPs that had signed agreements with ESPN had access (*New York Times*, 2008).

## Other Questions and Concerns

Mandating net neutrality via government regulation or even maintaining the current de facto prohibition on termination fees has been controversial. Here, we address a number of questions and concerns that have been raised.

### Won't Internet Service Providers Set Appropriate Fees on Their Own?

If setting prices to subsidize content providers produces socially desirable outcomes, wouldn't service providers simply do so? Wouldn't an Internet service provider internalize the externalities across groups and subsidize the "right" side on its own?

One possible answer is "yes" and that this is already happening. Since the early 2000s, despite some early stated interest in charging termination fees to content providers, no Internet service provider has actually tried to do so. Although there have been a few attempts by service providers to limit access to certain types of content, there are relatively few examples. It is unclear, however, whether that behavior is motivated by conscious behavior and internalization of externalities, or by the existing threat of regulation which functions as a form of enforcement.

At the same time, it seems implausible that Internet service providers have appropriate incentives to price according to the social optimum. First, two-sided market theory models predict even a monopolist provider does not subsidize the "right" sides as much as a social planner would, as profit-maximizing prices are higher than those imposed in the social optimum (Armstrong, 2006). Second, and perhaps more interestingly, the fact that a customer will often have a different ISP than a content provider means there will be strong incentives to charge fees even if zero-prices were socially optimal. Furthermore, because the value of content is shared not only by consumers subscribed to a particular ISP but across all consumers, ISPs do not completely internalize the impact of charging termination fees to content providers.

To see this point more clearly, consider the following: assume there are three Internet service providers for users—A, B, and C—and consider A's decision to

charge fees to all content providers wishing to reach its own customers. Service provider A would gain revenue and account for the possibility that such fees would lead to potentially less content (via depressed investment and innovation) for its own users. However, A does not completely internalize how its action would negatively influence content production for the entire Internet, as it does not account for the effect of a reduction in content on users of B and C. Similarly, B and C's incentives to increase termination fees would also be misaligned with the social optimum, since they also fail to completely internalize the effects of their own pricing decisions on users of other ISPs.

In effect, service providers face a prisoner's dilemma: it might be individually optimal for one provider to defect and charge positive fees to content providers, although if all content providers charged such fees, the outcome would be worse than had all providers refrained from doing so. In this sense, the existing *de facto* practice of zero-pricing for content providers on the Internet can be understood as a solution to this collective action problem. Given the temptation to defect, regulation in support of net neutrality—or the threat of such regulation—can play a useful role in maintaining a cooperative solution.<sup>5</sup>

### **Could Charging Content Providers Help Pay for Network Development or Reduce Web Congestion?**

Allowing for termination fees could generate more revenue for service providers; hence, some argue that relaxing the zero-price rule will provide carriers with greater incentives to invest in existing infrastructure. Hemphill (2008) writes of an implicit tradeoff in which “not only content innovation but also infrastructure innovation must be taken into account, and that subsidizing content development necessarily must be at the expense of network development.” If so, perhaps the Internet has now reached a stage of maturity in which, even if innovation in content provision is still desirable, other objectives have become more important?

Edward Whitacre, former CEO of AT&T, made a similar claim (*Business Week*, 2005): “Now what they [content providers] would like to do is use my pipes free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it. So there's going to have to be some mechanism for these people who use these pipes to pay for the portion they're using. Why should they be allowed to use my pipes? The Internet can't be free in that sense, because we and the cable companies have made an investment and for a Google or Yahoo! or Vonage or anybody to expect to use these pipes [for] free is nuts!”

There are two flawed assumptions in this line of argument. First, it is unclear

<sup>5</sup> Although allowing Internet service providers to form a cooperative may ameliorate the coordination problem, there still is the misalignment of monopolist incentives from the social optimum; furthermore, and perhaps more importantly, the existence of cooperatives may be unpalatable for collusive possibilities that may be encouraged (for example, the credit card industry, which utilizes a cooperative structure, has been the subject of antitrust scrutiny and litigation, including *United States v. Visa U.S.A., Inc. et al.* (344 F.3d 229 [2d Cir. 2003], *cert. denied*, 543 U.S. 811 [2004]).

that the ability to charge termination fees would, in fact, increase incentives to upgrade existing infrastructure. Termination fees may provide a way to increase profits of Internet service providers regardless of whether they upgrade their pipes; the impact on the *marginal* incentive to invest is indeterminate.<sup>6</sup> Instead of investing in faster or more reliable service, firms could also pay out a (greater) dividend, undertake other projects, or even invest in increasing its returns on existing content by making it scarce and exclusive. As Wu argued (Wu and Yoo, 2007): “If you can generate revenue by charging content providers to reach customers, as opposed to charging for bandwidth, something happens. The incentives become mixed, as the provider gains an incentive to maintain a level of scarcity, and thereby maximize gatekeeper revenue. So I don’t agree . . . that [termination] fees will necessarily spark more last-mile innovation.”<sup>7</sup>

The second problem, which speaks directly to Whitacre’s claim, is that the absence of termination fees does not imply Internet service providers are not properly compensated for the use of their “pipes.” Indeed, ISPs, including AT&T, are actually paid twice. First, any network that transmits a packet through AT&T’s network provides payment to AT&T determined by negotiated peering or transit agreements (and that network was also compensated when it received the original packet from another network, content provider, or end-user). Second, AT&T already charges consumers for access, and nothing in the current net neutrality regime prevents charging higher prices to consumers who utilize more bandwidth or demand faster service. In a sense, claims that content providers should pay for use of a consumer’s ISP when consumers already pay is comparable to the Postal Service demanding a recipient pay to receive a package for which a sender has already paid postage.

Similar reasoning can be used to counter arguments which claim that termination fees might serve to reduce “web congestion.” Content providers, the argument goes, design their applications without taking into account the marginal cost of higher bandwidth usage, and hence “overuse” bandwidth in their designs; that is, these content providers design applications that “spend” bandwidth with wild abandon. However, as discussed, content providers are already forced to take into account the costs of bandwidth usage: those which utilize an Internet service provider must pay the access and usage fees to make the content available in the first place, and those which are connected directly to other network providers pay fees for egress traffic based on existing peering and transit arrangements. The less

<sup>6</sup> In other words, in a hypothetical market already served by at least one Internet service provider where termination fees are allowed, upgrading pipes would not necessarily lead to an increase in profits. For example, in markets with at least two ISPs, the gains to investment may be competed away. However, termination fees may encourage network expansion to consumers not served by any ISP. In these instances, there may be a tradeoff between expanding Internet access to new markets on the one hand and incentivizing content creation and avoiding Internet fragmentation through a zero-price rule on the other. At the same time, by reducing content creation and availability, termination fees could also lower consumer willingness-to-pay.

<sup>7</sup> The “last-mile” refers to the technologies and processes (for example, coaxial, wireless, fiber) that connect an ISP to its customers.

bandwidth a content provider uses, the less it pays. Furthermore, Internet service providers can, and in certain circumstances do, charge end-users for the amount of bandwidth used regardless of what content is consumed; thus, content providers that utilize massive amounts of bandwidth will also face and internalize lower demand by consumers.

### **Would Favoring Some Content Improve Internet Service?**

In this section, we consider the possibility that allowing networks to favor some content over others could improve service overall. If hypothetically a network could recognize and prioritize packets more sensitive to delay, like video packets, over packets that are insensitive, like email, the network would in theory function better. Similarly, if packets could be transmitted over shorter distances, less congestion would occur. This reasoning has been the basis for many proposals for “quality of service” enhancements to the Internet since at least the 1990s, and has motivated the rise of content distribution networks and caching services, such as Akamai, which “mirror” content across servers located around the world and thus allow providers to pay for improved delivery of content. Indeed, network management and quality of service inherently requires some form of packet discrimination or content co-location, and are practices with which we do not necessarily take issue (Wu, 2003); in our view, they may be palatable as long as payment is not demanded from content providers by Internet service providers as a requirement for service.<sup>8</sup>

Many global schemes for prioritizing some packets of information over others have so far failed because of a collective action problem inherent in their design. The Internet is comprised of hundreds of Internet service providers and millions of content providers worldwide. So far, agreement on standards to prioritize traffic on the shared network has been impossible to reach, as has any agreement to honor any standards for prioritization. There is an obvious incentive to label every packet as a “high priority” packet on the assumption that everyone else will do the same.

However, it is an open question whether using prices could somehow overcome this problem. One proposal that has been raised to address network congestion in the last mile would be to create a *tiered structure* for consumer ISP traffic: allow *all* content to travel *freely*, but at the same time allow individual Internet service providers to create a “preferred” service for traffic, or a “fast-lane,” for a fee that does not depend on the identity of the content provider. In our view, this approach has the advantage over termination fees by allowing anyone access to faster service—incentivizing content providers to only label traffic “priority” if the cost was deemed worthwhile—while not foreclosing those who still opt for the “free” Internet. As a result, we do not feel as

<sup>8</sup> We believe caching agreements with Internet service providers and services provided by distribution networks do not impinge on content provision because these services are available to all content providers and content providers that do not use these services are still accessible by consumers. Although these services do improve the performance of certain content providers vis-à-vis others, the services are only worthwhile for content providers with significant traffic and bandwidth demands; for new entrants with low bandwidth requirements, such services provide little benefit and are a nonissue.

though a zero-pricing rule should prohibit this particular implementation, as here content providers are not *forced* to pay a termination fee to access users.

Yet although this particular solution may have desirable benefits, we raise a warning: unless sufficient bandwidth and quality of service can be guaranteed for the “free” Internet, there is a risk that such tiering will serve to sidestep the *de facto* prohibition on termination fees. For example, a priced-priority system could simply become itself a *de facto* fee charged for all content providers if the “free” Internet was of sufficiently poor quality and consumers shifted their usage behavior accordingly. In other words, even if ISPs were prevented from charging discriminatory fees to content providers (thereby reducing bargaining frictions and the potential for fragmentation), tiering still could result in transfers from content providers to Internet service providers. As argued previously, this might dampen the introduction of new content and services and eliminate the subsidy for content innovation currently provided by net neutrality.

### **Should the Net Neutrality Argument Be Applied to Other Networks?**

We have argued the Internet’s history may have created a built-in subsidy for competitive entry and creativity. There is, however, an open question of whether the principle should be extended to other networks. While a full treatment of this question is well beyond the scope of this paper, we do wish to be careful about generalizing our arguments to other industries: in particular, several institutional details may differentiate the Internet from attempts to implement a zero-pricing rule in other networks. For example, one important consequence of the Internet’s universal design is that the bandwidth used by any one content provider is dynamic and proportional to its popularity: only content that is visited or popular consumes common resources, whereas sites that are never accessed utilize zero network bandwidth. Consequently, there is effectively no opportunity cost of subsidizing new content and lowering the barrier to entry, since other content is not precluded from existing or reaching users. In contrast, in media networks such as radio or cable television, each station uses a fixed amount of bandwidth or spectrum regardless of its popularity; similarly, new products sold in stores consume physical space and inventory even if no one purchases them. Thus, even if subsidizing content may be desirable, the scarcity of airtime, spectrum, or shelf space may very well render zero-pricing unappealing and undesirable in other industries.

### **Concluding Remarks**

At its broadest, the net neutrality debate in the United States and around the world is a reincarnation of an age-old debate about the duties of firms that supply infrastructure services essential to the economy, or—in the old common law phrase—firms “affected with the public interest.” In the nineteenth century, trains and canals were the focus of this debate; in the twentieth century, it was the telephone and the electric systems; and in the twenty-first century, the Internet has seized center stage.

This paper has highlighted a potential benefit of the zero-pricing aspect of net neutrality, which prevents Internet service providers from levying termination fees on content providers. The theory of two-sided markets provides an underlying rationalization for how this practice can subsidize the creation of new content and spur innovation while avoiding fragmentation of the Internet. Several open questions remain, including how close the optimal subsidy for content creation is to a zero-price rule, and to what extent welfare gains from increased content production due to a zero-price may be offset by potentially higher access or usage fees charged to consumers.

At the same time, a more fundamental question that underlies this paper is what, if anything, sets the Internet apart from other networks, past and present? This question suggests a much broader agenda for research: namely, understanding in a more parsimonious manner how different pricing rules and other features of information networks affect their influence. We have mentioned two differences—a *de facto* ban on termination fees, and a rough proportionality between content popularity and bandwidth usage—that set the Internet apart from the other networks of our time, including telephone, cable TV, and broadcasting networks. Yet while this much may be clear, we do not have anything close to a full vocabulary for understanding the different choices implicit in the designs of different networks. And we have an even weaker understanding of what the larger effects of such choices will be. Although in this paper we have isolated one interesting effect—namely, that a ban on termination fees can be used to encourage market entry by creators and innovators—this point is far from a full understanding of networks and their larger effects on society and the world.

■ *We thank the editors and Scott Hemphill for their helpful comments.*

## References

- Anderson, Chris.** 2006. *The Long Tail*. New York: Hyperion.
- Armstrong, Mark.** 2006. "Competition in Two-Sided Markets." *RAND Journal of Economics*, 37(3): 668–91.
- Benkler, Yochai.** 1999. "Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain." *New York University Law Review*, 74(2): 354–446.
- Berners-Lee, Tim.** 2008. "Neutrality of the Net." <http://dig.csail.mit.edu/breadcrumbs/node/132> (accessed on November 11, 2008).
- Business Week.** 2005. "Online Extra: At SBC, It's All About 'Scale and Scope.'" November 7. [http://www.businessweek.com/magazine/content/05\\_45/b3958092.htm](http://www.businessweek.com/magazine/content/05_45/b3958092.htm).
- ▶ **Caillaud, Bernard, and Bruno Jullien.** 2003. "Chicken & Egg: Competition among Intermediation Service Providers." *RAND Journal of Economics*, 34(2): 309–28.
- Cerf, Vinton G.** 2006. Prepared Statement, U.S. Senate Committee on Commerce, Science, and Transportation Hearing on "Network Neutrality." <http://commerce.senate.gov/pdf/cerf-020706.pdf>.
- ▶ **David, Paul, and Shane Greenstein.** 1990.

- "The Economics of Compatibility Standards: An Introduction to Recent Research." *Economics of Innovation and New Technology*, 1(1–2): 3–41.
- DeGraba, Patrick.** 2000. "Bill and Keep at the Central Office as the Efficient Interconnection Regime." OPP Working Paper No. 33, Office of Plans and Policy, Federal Communications Commission, Washington DC.
- Economides, Nicholas, and Joacim Tåg.** 2007. "Net Neutrality on the Internet: A Two-sided Market Analysis." NET Institute Working Paper No. 07-45.
- Evans, David S.** 2003. "Some Empirical Aspects of Multi-sided Platform Industries." *Review of Network Economics*, 2(3):191–209.
- ▶ **Farrell, Joseph, and Garth Saloner.** 1985. "Standardization, Compatibility, and Innovation." *The RAND Journal of Economics*, 16(1): 70–83.
- Farrell, Joseph, and Philip J. Weiser.** 2003. "Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age." *Harvard Journal of Law & Technology*, 17(1): 85–134.
- Hagiu, Andrei.** 2006. "Pricing and Commitment by Two-Sided Platforms." *RAND Journal of Economics*, 37(3): 720–37.
- Heller, Michael A.** 2008. *The Gridlock Economy*. New York: Basic Books.
- Hemphill, C. Scott.** 2008. "Network Neutrality and the False Promise of Zero-Price Regulation." *Yale Journal on Regulation*, 25(2): 135–180.
- Hurwitz, Justin.** 2006. "Neighbor Billing and Network Neutrality." *Virginia Journal of Law and Technology*, 11(9): 1–33.
- ▶ **Katz, Michael, and Carl Shapiro.** 1986. "Technology Adoption in the Presence of Network Externalities." *Journal of Political Economy*, 94(4): 822–41.
- ▶ **Landes, William M., and Richard A. Posner.** 1989. "An Economic Analysis of Copyright Law." *Journal of Legal Studies*, 18(2): 325–63.
- Lee, Robin S.** 2009. "Vertical Integration and Exclusivity in Platform and Two-Sided Markets." <http://pages.stern.nyu.edu/~rslee/papers/VIExclusivity.pdf>.
- Lessig, Lawrence, and Robert W. McChesney.** 2006. "No Tolls on the Internet." *Washington Post*, June 8, A23.
- Licklider, Joseph C. R., and Robert W. Taylor.** 1968. "The Computer as a Communications Device." *Science and Technology: For the Technical Men in Management*, April, no. 76, pp. 21–31. (Reprinted in 1990 in *In Memoriam: J. C. R. Licklider: 1915–1990*, Report 61, pp. 21–41. Palo Alto, CA: Systems Research Center, Digital Equipment Corporation.)
- Liebowitz, S. J., and Stephen E. Margolis.** 1994. "Network Externalities: An Uncommon Tragedy." *Journal of Economic Perspectives*, 8(2): 133–50.
- The New York Times.** 2008. "ESPN to Offer Sports Events on the Web Free to Some." February 4. <http://www.nytimes.com/2008/02/04/business/media/04espn.html> (accessed November 10, 2008).
- Posner, Richard.** 1998. *Economic Analysis of Law*. 5<sup>th</sup> ed., pp. 36–50. New York: Aspen.
- Reardon, Marguerite.** 2006. "FCC approves AT&T–BellSouth merger." CNET News, December 29. [http://news.cnet.com/FCC-approves-ATT-BellSouth-merger/2100-1036\\_3-6146369.html](http://news.cnet.com/FCC-approves-ATT-BellSouth-merger/2100-1036_3-6146369.html).
- ▶ **Rochet, Jean-Charles, and Jean Tirole.** 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association*, 1(4): 990–1029.
- Rochet, Jean-Charles, and Jean Tirole.** 2006. "Two-Sided Markets: A Progress Report." *RAND Journal of Economics*, 37(3): 645–67.
- ▶ **Segal, Ilya.** 1999. "Contracting with Externalities." *Quarterly Journal of Economics*, 114(2): 337–88.
- Technorati.** 2008. "State of the Blogosphere." <http://technorati.com/blogging/state-of-the-blogosphere> (accessed on November 12, 2008).
- The Wall Street Journal.** 2006. "ESPN Charges Net Providers For Right to Offer Broadband Web Site." August 1. <http://online.wsj.com/article/SB115439535367922979.html>.
- Whinston, Michael D.** 2006. *Lectures on Antitrust Economics*. Cambridge: MIT Press.
- Wright, Julian.** 2004a. "One-sided Logic in Two-sided Markets." *Review of Network Economics*, 3(1): 44–64.
- ▶ **Wright, Julian.** 2004b. "Pricing Access to Internet Service Providers." *Information Economics and Policy*, 16(3): 459–73.
- Wu, Tim.** 2003. "Network Neutrality, Broadband Discrimination." *Journal of Telecommunications and High Technology Law*, vol. 2, pp. 141–79.
- Wu, Tim, and Christopher Yoo.** 2007. "Keeping the Internet Neutral?: Tim Wu and Christopher Yoo Debate." *Federal Communications Law Journal*, 59(3): 575–92.



## The Benefits and Risks of Mandating Network Neutrality, and the Quest for a Balanced Policy<sup>\*</sup>

JON M. PEHA<sup>1</sup>

Carnegie Mellon University

A fundamental issue in the network neutrality debate is the extent to which network operators should be allowed to discriminate among Internet packet streams to selectively block, adjust quality of service, or adjust prices. This paper first reviews technology now available for traffic discrimination. It then shows how network operators can use this technology in ways that would make the Internet less valuable to Internet users, and why a network operator would have financial incentive to do this if, and only if, it has sufficient market power. A particular concern is that network operators could use discrimination to extract oligopoly rents from upstream markets that are highly competitive. This paper also shows how network operators can use the very same technology to discriminate in ways that benefit Internet users as well as the network operator. Thus, network neutrality supporters are right to fear unlimited discrimination in some cases, while network neutrality opponents are right to fear a policy that imposes strict limits on discrimination. From this, we argue that the network neutrality debate should be refocused on the search for a *balanced policy*, which is a policy that limits the more harmful discriminatory practices in markets where there is insufficient competition, with little interference to beneficial discrimination or innovation. We apply this balanced policy in a few controversial scenarios as examples. There has been too little attention on the possibility of a nuanced balanced policy, in part because the network neutrality debate is focusing on the wrong issues. This paper argues that the debate should shift toward the complex details of differentiating harmful discrimination from beneficial discrimination, and away from high-level secondary questions like whether discrimination is inherently just, who ought to pay for certain Internet services, how important general design principles are, what abstract rights and freedoms consumers and carriers deserve, or whether network operators can give their affiliates special

---

Jon M. Peha: [peha@cmu.edu](mailto:peha@cmu.edu)

Date submitted: 2007-06-07

<sup>\*</sup> An earlier version of this paper was presented at the Federal Communications Commission (FCC) in February 2006 and at the Telecommunications Policy Research Conference (TPRC) in September 2006.

<sup>1</sup> Jon M. Peha, Carnegie Mellon University, Associate Director of the Center for Wireless and Broadband Networking, Professor of Electrical Engineering and Public Policy, [peha@cmu.edu](mailto:peha@cmu.edu), [www.ece.cmu.edu/~peha](http://www.ece.cmu.edu/~peha).

Copyright © 2007 (Jon M. Peha). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at <http://ijoc.org>.

treatment. Reality is more complex than these questions would imply, and none of them will serve as a basis for a sufficiently specific and effective policy.

## Section 1: Introduction

As the Internet approaches its 40<sup>th</sup> birthday, “network neutrality” has suddenly become its most controversial issue. Why now? One reason is that the technology itself has been changing, giving networks extensive abilities to treat some classes of traffic differently from others. As we will show, some forms of this discrimination could harm Internet users, and this has many network neutrality advocates concerned. On the other hand, we will also show that some forms of discrimination enabled by the same technology would benefit users. There is therefore a danger that imposing a broadly defined network neutrality policy could prohibit carriers from adopting these valuable practices.

The other reason why this controversy is occurring now is that competition for consumer access to the Internet has been declining. After all, if there were rigorous competition, network operators who use discrimination to harm consumers or fail to use discrimination to benefit consumers would lose customers to their rivals. Dial-up access was naturally competitive, but consumers have been switching to broadband, and most consumers currently have one or perhaps two last-mile broadband providers from which to choose. At the same time, attempts to encourage competition over the same physical connection have largely subsided in the U.S. Without competition, if there are discriminatory practices that increase carrier profits but harm consumers, it may take regulation or the threat of regulation to deter these practices. At this point, few people are seriously advocating complete common carrier regulation of these monopoly and duopoly markets as this could limit innovation and discourage the entry of new competitors. However, under the banner of network neutrality, policymakers could attempt to limit some discriminatory practices as long as they believe the regulation will do less damage than the discrimination would.

Thus, policymakers face the following fundamental challenge.

**Can we limit how network operators can discriminate in a manner that**

- **prevents them from fully exploiting market power in ways that seriously harm users, and**
- **does not prevent them from using discrimination in ways that greatly benefit users?**

We refer to a policy that effectively balances these two competing objectives as a *balanced policy*. More specifically, we will argue that the type of discrimination that deserves closest scrutiny in a balanced policy is *discrimination that allows a provider of last-mile broadband Internet access to extract oligopoly rents from upstream competitive markets*.

To address the fundamental question above, we must understand the types of discrimination that are technically possible, their impact on users, the economic incentives carriers may have to use these techniques, and the implications for policymakers. Thus, Section 2 describes what is now technically possible with respect to discrimination. Section 3 shows how these capabilities can be used to benefit

Internet users, while Section 4 shows how a network operator with sufficient market power could use the same capabilities to the detriment of users.

Of course, most advocates seem to disagree that policy should revolve around the two-part question above. The network neutrality debate has repeatedly been framed in ways that obscure this question. Instead, we hear about the inherent evils of discrimination, violations of revered Internet traditions, the basic freedoms of consumers and divine rights of carriers, whether content providers or network operators are carrying an unfair burden, how all forms of regulation are always wrong, vertical integration and unfair alliances, and more. While some of these perspectives are useful, none of them make it sufficiently clear how network operators should be allowed to use emerging technology. Moreover, all of them have distracted policymakers from more important concerns. Section 5 summarizes and critiques some of the common ways that the network neutrality debate has been framed and misframed in light of the basic challenge described above and observations from Sections 3 and 4. Section 6 discusses what an effective balanced policy might allow or prohibit. This paper is concluded in Section 7.

## **Section 2: The Technical Basis of Discrimination**

Unfortunately, engineers, economists, and lawyers have different definitions for discrimination. In this paper, discrimination occurs whenever a network treats some network traffic or some network users differently from others.

In a packet-switched network such as the Internet, information is sent through the network one packet at a time, where a packet consists of some information to be carried across the network and some "header" information used by the network devices to make the transfer. For example, the header might indicate the sender and the recipient of the packet. A single email message or web page may yield many packets that are sent separately and reassembled at the destination. Moreover, networks are "layered" such that a higher-layer packet is stuffed inside a lower-layer packet, like a letter inserted into an envelope, which is placed in a box and then mailed. The postal system uses information written on the box, but not "application layer" information inside the envelope. Traditionally, Internet packets were sent with equal priority and "best effort," i.e., with no guarantee of delivery. This is not discriminatory by the above definition.

Times have changed. There are a variety of techniques through which networks can now favor some packets or packet streams over others. We first discuss criteria that networks can now consider when deciding who should get better service. We then discuss methods they can use to give the favored group better service.

Some criteria are easier to use for discrimination than others. Among the easiest are fields in the header of an Internet protocol (IP) packet, because every IP packet contains this information, and it is easy to find within the packet. For example, this information includes the identity (more specifically, the IP address) of the sender and recipient. Figure 1 shows some of the header fields that reveal useful information for discrimination. If the network places a device where it can monitor traffic entering the network, the device also knows about the physical location of the source, and it knows information in the

link-layer header which could reveal who manufactured the device attached to the network. However, it is difficult to infer much about a packet stream from a single packet, and larger messages have historically only been reassembled at the destination from a series of packets, so until recently more sophisticated forms of discrimination were not practical.

Protocol	Data Field(s)	Reveals something about:
Link layer protocol, such as Ethernet (802.3), WiFi (802.11), Bluetooth (802.15), DOCSIS (cable), many more	MAC address of source and destination	Manufacturer of device that is attached to network. (In some but not all cases, MAC addresses are fixed when a device is manufactured, and it is possible to identify the manufacturer from this address.)
IP	IP address of source and destination	Identity of sender, identity of recipient, location of sender, location of recipient. (e.g., was the IP address allocated through an ISP in the U.S.?)
IP	transport protocol (e.g., TCP, or UDP)	Type of application. (Some applications typically use TCP, and some use UDP.)
IP	differentiated service code point in IP version 4 / traffic class in IP version 6	Type of application, priority desired by sender. (Rarely used today. This may change when IP v6 becomes common.)
IP	packet length	Type of application. (Some applications generate larger packets than others.)
TCP or UDP	source port, destination port	Type of application (e.g., port 21 for file transfer, 23 for telnet, 25 for email, 80 for web traffic, although some applications choose unpredictable port numbers and evade port filters.)

Figure 1: Examples of header data that can easily be used as a basis for discrimination.

New technology has emerged that makes it practical for networks to collect much more information about a packet stream. One is *flow classification*, which is available today (e.g., [1]). By examining the sizes of packets in a stream, the amount of time between consecutive packets, and the amount of time since the packet stream began, one can make reasonable determinations about the nature of the packet stream. For example, a steady 30 kb/s stream of packets that lasts for ten minutes could be voice over IP (VoIP). Note that the network operator learns nothing about the content of the conversation, only the nature of the application. Indeed, this technique works equally well when the voice information is encrypted.

Another approach is *deep packet inspection*, which is also available today (e.g., from Cisco [2], Allot [3], P-cube [4], Packeteer [5]). Deep packet inspection is *stateful*, which means it maintains information about every packet stream going through it. It can categorize traffic based on the content of many consecutive packets in combination, rather than only what it can learn from the packet it is currently handling. A device using deep packet inspection is also aware of the information at the *application layer*, which means instead of looking only at the information needed to get the packet to its destination, as illustrated in Figure 1, the device seeks to understand the data that an application software running at the destination would use. That application could be a web browser, a VoIP client, a video display, or an email user agent. As a result, it is possible to tell whether a packet stream is VoIP, email, web browsing, instant messaging, video streaming, file transfer, or peer-to-peer file sharing. It is possible to examine in detail the content of the email, or web page, or downloaded file. It is also possible to distinguish music files from text from pictures and to search for keywords within any text.

All of this requires a great deal of processing, which is why cost-effective products were not available until recently, but processors are much faster and cheaper than they used to be. While it still may be challenging to do complex processing at speeds needed in the backbone links with greatest capacity, providers of last-mile broadband service can always use these techniques closer to the edge of the network where links have lower capacity. This requires more devices whose cost must be justified by increased profit, but the technical challenge becomes much easier. For example, my laboratory at Carnegie Mellon University has successfully used deep packet inspection at 150 Mb/s to determine which network applications each computer on campus is running and which remote servers they are accessing as part of an effort to determine what puts a computer at greater risk from dangerous malware. (In our work, we take many precautions to conceal the identity of the users and otherwise protect their privacy, but these precautions make the processing more complex rather than less.)

In a stateful system, every packet may cause the monitoring device to look into a database. It is not difficult to include information in these databases that is not traffic-related, such as billing information or demographic information. For example, the recipient (destination IP address) of the packet may be mapped to something that indicates that this is a premium customer who gets special treatment, or that this is a competitor to the network operator who does not.

Between deep packet inspection and flow classification, it is cost-effective for a network operator to gain unprecedented knowledge about what is happening on the network and to selectively improve or degrade service for some. Now let us consider what advantages the network might bestow on traffic it wants to favor.

One old and simple way to favor some users is through preferred *interconnection*, i.e., to allow them to connect to the network with a higher-capacity link, or to pay less for the same capacity. This is still an option to discriminate among users, although it alone does not allow the network to discriminate among traffic from a given source.

Finer-grain discrimination is possible if it is embedded in the *traffic control* algorithms, i.e., the algorithms that control the flow of packets through the network. These algorithms can greatly influence the *quality of service* (QOS) of a packet stream. QOS typically involves the amount of time it takes a packet to traverse the network, the rate at which packets can be sent, and the fraction of packets lost along the way. Consider a congested communications link. Many packets sit in a buffer, waiting to be transmitted on that link. The *scheduling algorithm* determines when each waiting packet is actually transmitted, and how often packets from a given stream are transmitted. When the number of waiting packets becomes too large, a *dropping algorithm* will select some to be discarded. A *traffic shaping* algorithm may spread packets out so they do not arrive in a single large burst. An *admission control* algorithm may block entire packet streams temporarily on the grounds that it would not be possible to meet QOS requirements for the current streams and the new one if this new stream were admitted. If these algorithms discriminate, they can give favored streams smaller queuing delays, lower loss probabilities, higher data rates and/or lower blocking probabilities.

Discrimination can also be built into the *routing algorithm*, which decides where a packet should be forwarded next. Some packets might be sent over the quickest and most reliable path, while others may be sent the slow way. A particularly undesirable packet may experience "black-hole routing," which has the same effect as dropping the packet entirely. In cases where there are multiple possible destinations – for example, load balancing across multiple servers -- favored packets may go to the server with the shorter line. There are even cases where packets are sent to a destination quite different from the destination specified by the sender. This is *redirection*. For example, if a user attempts to connect to a server that no longer exists, the network might redirect the packets to a different server.

Some network neutrality policies have focused on prioritization, and it is clear how prioritization is at work in the preceding traffic control algorithms. However, there can be discrimination without obvious prioritization. One can simply provide *separate channels* for different classes of traffic. Favored traffic may be sent over a lightly used wavelength in a fiberoptic cable, while other traffic goes over a heavily used wavelength. The channel separation can also be logical instead of physical. Favored traffic may be sent over a separate virtual local area network (VLAN), or a separate *service flow* in a cable system operating under the Data Over Cable Service Interface Specification (DOCSIS) standard [6]. Traffic flows over the same physical channel, but one logical channel has higher priority when competing for limited resources than another logical channel.

Of course, users care about more than QOS. They also care about *price*. Once a network operator can determine in detail what a user is doing, the operator can charge for that. Thus, a user may pay more, depending on which applications are being used, with whom the user is communicating, or even whether the user remembers to include text praising his ISP in every email. This is typically known as "content billing" or "content charging," and it too is already available in today's network products. In many ways, implementing content billing is easier than implementing discriminatory traffic control. For traffic control, one must decide almost immediately which packets to favor. For billing, one merely has to decide by the end of the month, so traffic analysis can be done offline.

Finally, a network operator might discriminate by providing unequal access to various services. Favored packet streams might be carried over an efficient multicast mechanism, so the sender does not have to send a separate copy of the content to every recipient [7]. This is particularly useful for those who broadcast video, music, or other content simultaneously to multiple users over the Internet. Also, some users may have better access to information caches, so needed content can be retrieved locally rather than from a remote part of the network. Others may not be allowed to use caches associated with the network operator, or may be charged more for interconnecting their own caches.

In summary, network operators have powerful means to differentiate among network traffic, including examination of packet headers, deep packet inspection, and flow classification. Once they have used these techniques to choose what to favor, they can improve quality of service or price for the favored class through some combination of preferred interconnection, discriminatory traffic control algorithms (including scheduling, dropping, traffic shaping, admission control and routing), separate channels, content billing, and access to services like caching and multicast.

### Section 3: The Benefits of Discrimination

In this section, we discuss why discrimination is valuable for both users and carriers.

One obvious use of discrimination is security. A network operator may use deep packet inspection to determine whether a packet stream is carrying a virus or a dangerous piece of spyware. A broader examination of traffic patterns may reveal that a given source is participating in a denial of service attack on another user. A network neutrality policy that prohibits networks from dropping dangerous traffic of this kind would damage network security.

Redirection in combination with deep packet inspection can further improve security. My laboratory at Carnegie Mellon University is developing tools that use deep packet inspection to identify spyware. Once detected, it is possible through redirection to send users to a website with anti-spyware and anti-virus tools that can eliminate the threat. Redirection is also commonly used to provide useful instructions to those who try to connect with servers that are down or to enable users to pay for wifi hotspots before they begin normal operation.

Another useful role for blocking is to deny service from an unauthorized device. By ensuring that only authorized devices are attached to the network, the network can prevent customers from using equipment that will operate in "promiscuous mode" to observe their neighbors' traffic, or that consumes more of the shared resources than is allowed, or that accesses adult-only material contrary to the customer's stated wishes. (The latter might occur, for example, when a child of the customer seeks content that the customer has restricted.)

Instead of blocking packet streams, the network might discriminate with respect to quality of service, price, or both to ensure that resources will be shared fairly and no one will "starve." The ready availability of high-capacity always-on connections to the network has made it possible for a small number of users to generate the vast majority of network traffic on many commercial broadband networks, while filling some communications links to capacity. Today, peer-to-peer file transfers are the primary cause, but other applications may have a similar impact in the future. Moreover, some of these applications are not "TCP-friendly," which means when congestion occurs on these bottleneck links, these applications do not reduce their rate of transmission to allow the congestion to subside. An application like this will send out data as fast as it can, while the TCP-friendly applications deliberately send fewer and fewer packets. Therefore, one Gb of traffic that is not TCP-friendly degrades performance for its neighbors more than one that is TCP-friendly. Network operators may wish to give traffic from these applications lower scheduling and dropping priorities, or limit the amount of traffic they can send per day, or charge them more for consuming more network resources. This discrimination benefits the applications that might otherwise be starved of network resources.

Discrimination with respect to QOS is also important because different applications have different QOS needs. In a VoIP application, the recipient may play out packets 50 ms after they are first sent across the network. Thus, most packets must be received within 50 ms because any arriving after 50 ms are useless. Best effort delivery could lead to completely unacceptable QOS for a VoIP application if there is congestion. On the other hand, for a large file transfer, there is no specific maximum allowable delay, but a low average delay is helpful, whereas for email, delay is of little importance. If sophisticated traffic control algorithms take these QOS requirements into consideration, it is possible to give packets high priority when and only when they need high priority to meet QOS requirements, thereby meeting QOS requirements for many more users on a given network. Alternatively, it is possible to serve the same number of customers at the same QOS with less network capacity, making the network less costly. This benefits Internet users and network operators.

Perhaps as a compromise, some network neutrality proposals would allow discrimination with respect to QOS as long as there is no discrimination with respect to price [8]. Although the policy's goals are laudable, this is not effective as users would have no incentive to accept anything less than the highest priority. Discriminatory pricing gives users incentive to provide accurate information about their real QOS needs, to avoid wasting resources, and to refrain from transmitting when the network is congested by shifting usage to off peak hours. Indeed, by adjusting prices dynamically based on congestion levels, thereby convincing some users to delay their transmissions, pricing actually becomes a form of congestion control that has quantifiable advantages over more traditional technical approaches [9]. Limited resources are allocated most efficiently when price to users is a function of "cost" to network

operators. In this case, cost is the opportunity cost of carrying a given traffic stream, since allocating resources to carry one stream means those resources cannot go to another stream. These costs can be quantified [10], and the cost per bit of a stream with strict QOS requirements is greater than the cost per bit when QOS requirements are lax. All else being equal, the cost per bit of carrying traffic that arrives sporadically in large bursts is greater than the cost of carrying traffic that arrives in a steady stream, and the cost of carrying traffic that is TCP-friendly is less than the cost of carrying traffic that is not. Since the QOS requirements, burstiness, and back-off behavior of traffic are highly dependent on the application type, the public may be well-served by networks that charge different prices per bit for different applications.

Unfortunately, these efficient pricing mechanisms may lead to higher prices and potentially greater profit when the network is congested versus when it is uncongested. Thus, although such prices may give users incentives for efficiency, they may give network operators reason to prefer congestion, i.e., to profit from providing inadequate capacity. (More on this in Sections 4 and 6.)

Note that the incentive to discriminate with respect to QOS and price is based on the assumption that there are limited resources. In fact, a network has a choice on that. Networks can deploy far more communications capacity than is usually needed, so congestion is simply not a problem. Their reward is simple traffic control that can be run on cheaper processors, simple billing systems, and pricing that can be easily explained to customers. Alternatively, they can put money into sophisticated traffic control and billing instead of communications capacity. The best strategy depends on whether processing or communications gets cheaper at a faster rate. Throughout the 1990s, as progress in fiberoptics decreased the cost of communications at an astounding rate, this kind of discrimination made little sense and flat-rate pricing was the dominant model. Some believe this trend will continue [11], but others disagree. Thus, there are risks in embedding this conjecture into our laws and regulations.

#### **Section 4: The Damage from Discrimination**

The previous section showed that the technologies for discrimination in Section 2 can be beneficial to users. In this section, we show how a network operator has incentive to use the same technologies to the detriment of users, if and only if it has sufficient market power. (Market power generally comes from lack of competition although there may be cases where a network operator with competition has this power because it has monopoly control over the termination of a call [12].)

Note that in some cases, Internet users can take countermeasures in response to attempts by the network operator to discriminate and these may actions may prompt reactions by the network operator. These countermeasures range from use of virtual private networks to conceal information from the network operator to shifting usage from home to work where there may be more competition among broadband providers. Some forms of discrimination are relatively easy to circumvent. Others are nearly impossible. The resulting arms race could affect outcomes. This topic is largely outside the scope of this paper, but is discussed in a companion paper [12].

*Protecting Legacy Services from Competition*

The dominant broadband providers are cable companies and telephone companies that have incentive to protect their traditional offerings from video and voice over IP. In a competitive market this would be the standard “innovator’s dilemma” [13], and in time, either the market leader or an upstart rival would bring the novel IP-based product to market. However, in the absence of competition, the market leaders may prefer to stifle innovation indefinitely. Network operators can simply prohibit these rival services in their user agreements and then block the traffic. Alternatively, it is relatively simple to degrade quality of service of VoIP to the point that it cannot seriously compete with traditional telephony. The same approach is also possible with streaming video, although it is not as effective because video streaming applications can be designed to tolerate QOS that would be unacceptable for VoIP [12]. A third practical approach is simply to detect the voice or video traffic and charge extra for it, so the IP-based services are no longer a competitive threat. Vendors are already building and marketing products to network operators with the stated purpose of determining when customers use “revenue bypass” [14] applications like VoIP, and adding extra charges accordingly for this behavior.

*Charging Oligopoly Rents in the Broadband Market*

Obviously, a company that dominates the broadband market can exact oligopoly rents from the broadband market itself, by which I mean the market for transport of bits. Profit is maximized through perfect price discrimination, i.e., where each user is charged precisely what that user is willing to pay. Users here include consumers, businesses, and content and service providers. This implies that the *benefit of the Internet to each user is zero.*

To approach perfect price discrimination, the network operator can divide users into categories, and estimate willingness to pay within each category. That the operator has extraordinary information about what each user does over the Internet, along with external information about credit, housing, and much more, should make this task considerably easier. It is as if a grocery store could adjust the price of any item based on all the food you have ever purchased, when, where, and at what price, as well as your credit history and the value of your house.

Further improvements in discrimination are possible by offering multiple services, such that those willing to pay even more for better service will choose to do so, and those who are more sensitive to price will choose the cheaper services [15]. This can be achieved by intentionally degrading the quality of service for those paying less. Equipment is already being deployed to degrade QOS for this purpose. As one vendor [16] put it, “Service providers can sell the same thing to customers with different willingness to pay and therefore catch the consumer surplus.” And “to maximize revenues for value added services there must be a clear, perceived difference in the performance . . . . Bottlenecks are the foundation of this differentiation . . . . Note though that bottlenecks may be actual resource bottlenecks, or managed gates in the network.” Adding managed gates in a network specifically to degrade QOS would push away many customers if there were competitors who did not do this, but can be quite profitable for a network operator with sufficient market power.

*Charging Oligopoly Rents in Competitive Upstream Markets*

The types of discrimination described in Section 2 are particularly dangerous because a network operator can extract oligopoly rents not just in the broadband market, but also in any upstream market, i.e., any market that depends on Internet access for operation. This includes electronic commerce for any and all products, communications services like VoIP and videoconferencing, information distribution markets like video streaming and MP3 music sales, on-line advertising, and network equipment that attaches to the Internet. This strategy works even if the upstream market is highly competitive. For example, there are many online bookstores. A network operator could charge extra for each book sold online by any vendor, effectively pushing total book prices to where they would be if there were only one online bookstore. (This extra charge could be paid by assigned to either consumer or content/service provider.) In the absence of competition or regulation, a network operator's ability to identify distinct upstream markets for this purpose is limited only by what the technology can reveal about the content of network traffic. As we have seen, network operators can consider the sender and recipient of the traffic, the application, the content, the time of day, and much more. Thus, not only can a network operator charge different amounts for 4-MB journal articles and 4-MB MP3 music files, it can charge more for an MP3 song that is among the ten most popular in the country than for a song that it is not. (And through its monitoring, the network operator may know more about which music is popular than anyone in the music industry does.)

Note that a network operator can effectively extract oligopoly rents from upstream markets without ever entering those markets. For example, it can charge for each iTunes downloaded without affiliating with Apple, and despite Apple's strenuous objections. However, in many cases, it might be convenient for the network operator to either enter the market for given content or service, or to partner with an affiliate who is doing so. If the network operator does have an affiliated partner, then the operator can do more than block rivals; the operator can redirect requests that were intended for one of these rivals to its partner. For example, the customer types the name of her favorite e-commerce site, but is instead shown the site of a competitor affiliated with the network operator.

In practice, network operators would probably focus their attention on a few upstream markets with big companies that are generating significant margins. For example, the "Cisco Service Control Solution" is advertised as enabling three steps that allow the extraction of rents from upstream markets [17]. First, analyze network traffic to identify markets to enter as either a competitor or partner to existing players. Second, adjust the QOS of the relevant traffic. This can provide incentives for the current content or service providers to partner with the network operator, even if they might not have done so otherwise. Alternatively, adjusting QOS for current providers could yield a competitive advantage if the network operator decides to compete with the current providers. Third, use content billing to charge for use of the relevant services.

Once network operators have identified the upstream markets from which they can extract greater profits, they can also attempt to match price to willingness to pay in these upstream markets, just as described previously for the broadband market. If perfect discrimination were possible, *network*

*operators could then drive consumer surplus to zero in every upstream market -- a terrible blow to Internet users.*

Again, network operators can exploit all of the information available regarding a user's online behavior and they have far more information than upstream content and service providers do. For example, a network operator knows more about the location of sender and receiver, and can add a surcharge to every VoIP call that is based on what telephone companies would charge for the same call, or on the credit rating of the parties involved. Even a monopoly VoIP provider would not be able to charge the user this much. Moreover, unlike firms in the upstream markets, network operators have information about multiple markets. Thus, if there is a relationship between a user's interest in streaming video on demand and in peer-to-peer file sharing, the network operator might increase its profits by charging additional fees to those high-volume customers who do both. Even a monopolist in the streaming video market could not use such a strategy to increase profits at the expense of Internet users.

As in the broadband market, network operators can also deliberately degrade service where it is helpful in capturing profits in upstream markets. As one equipment vendor put it, the ability to adjust QOS for each upstream market "enables revenue sharing schemes or value-based pricing rather than only 'bit retailing.'" [16]. An alternative to intentionally degrading QOS is selectively limiting applications. For example, Cisco [17] suggests offering a basic service in which all traffic other than email and web browsing is blocked. Users who want peer-to-peer file sharing would pay a surcharge, and those wanting VOIP would pay an even larger surcharge. Thus, people who want additional services would be required to pay more, even when these services do not place greater demands on the network. Cisco further suggests that surcharges would be waived for content and services that come from providers affiliated with the network operator. Content billing makes all of this easy.

Note that Internet users include both consumers and content or service providers. Many network operators are considering pricing schemes through which both sides of an exchange would pay in some way for the last-mile connection to the consumer. This makes it easier to conceal how the network is differentiating among upstream markets. For example, if there is a greater difference between the monopoly price and competitive price in online book sales than in online CD sales, the network might impose greater charges on book merchants than on CD merchants. This may raise fewer objections than charging consumers who buy books differently from consumers who buy CDs.

This would also allow network operators to separately charge oligopoly prices to both sides. Viewers of an online newspaper could be charged based on the value of this specific news content, while advertisers are simultaneously charged based on the value of disseminating the ads to this particular set of readers. Moreover, these advertisers might have considered many online outlets and that competition could drive down advertising rates. However, if all of these media outlets go over the same network to reach the viewer, then the network operator can charge a monopoly price where one would not otherwise have been possible.

*Further Exploiting Upstream Market Failures*

Market failures in upstream markets can provide additional opportunities for network operators. This is certainly the case when products in upstream markets are “sticky,” i.e., because of switching costs, once a customer has chosen that product, she will be reluctant to switch to a competitor. An important example is email. There may be little reason to choose one email provider over another, but once a customer has notified people of her email address, it can be unwieldy and laborious to switch. Network operators can exploit this by offering an e-mail service that is available only to customers, then using blocking, QOS manipulation, or pricing to make rival email services inaccessible or unattractive. In a duopoly, network operators could use this technique to reduce effective competition.

There are also opportunities in an upstream market where benefits per user increase substantially with the number of users [18], perhaps because of a positive externality or a strong economy of scale. For example, the benefits per user of instant messaging increase as more people join the network. In this case, a network operator may choose to turn an upstream market into a monopoly by blocking or degrading service for rivals. As the winner becomes dominant, benefits of this system grow, and so does the extra revenue that the network operator can extract from this service. The network operator may extract this benefit from users by partnering with the dominant company, but it can extract the benefit through content billing without partnerships.

Network operators can also have incentive to block or discourage online activities that benefit the users involved, but decrease profit of someone else, i.e., for which there is a significant externality. For example, operators may block any anti-spyware software that removes certain kinds of adware, in return for payment from the adware company and its advertisers. Similarly, network operators may block applications that legally or illegally use or disseminate certain intellectual property, in return for payments from the owner of that intellectual property.

*Stifling Free Speech for Fun and Profit*

A network operator with sufficient market power clearly has the ability to stifle speech and sometimes it will have the incentive. This may be particularly important in political spheres given the Internet’s growing role in raising campaign donations, disseminating candidate information, and mobilizing volunteers. Network operators could simply limit access to websites that are of use to candidates they oppose. This would cost far less than what these companies already spend on lobbying and campaign contributions, and it would probably have more impact.

Such limitations on political speech may seem alarmist, but there is certainly precedence. For example, in 2003, Cumulus Broadcasting and Cox Radio banned the radio play of music from the Dixie Chicks after one member criticized President George W. Bush and the war in Iraq, even though the multi-Grammy-winning artists had the most popular country song in the U.S. at the time and none of their antiwar sentiments were reflected in their songs [19]. Radio stations have the right to play only what they wish. After all, there are many radio stations, so if listeners are unhappy with the offerings of one station, they can try another. However, users of broadband Internet do not have so many options.

Members of the Telecommunications Union in Canada were reminded of this during a labor dispute in 2005, when the ISP Telus blocked access to a website that was trying to disseminate the union's views [20].

### Section 5: Misleading Characterizations of the Network Neutrality Issue

There is no consensus on exactly what network neutrality means in practice or why the issue might be important. Indeed, the most specific proposals tend to come from those who want network neutrality to sound foolish so they can discredit it. This section reviews a number of prominent characterizations of the issue. We argue that none of these should be used as the primary basis for specific regulation or legislation.

*Network neutrality should not be about banning all discrimination.*

As was discussed in Section 3, discrimination can be used in ways that benefit users, potentially improving security, improving quality of service, decreasing infrastructure costs, and allocating resources to those who benefit the most from them. Moreover, if discrimination were inherently bad, then it should be banned even in a highly competitive market, but there is no obvious reason for regulatory intervention if such a market existed.

*Network neutrality should not be about prohibiting vertical integration or affiliate relationships.*

Some discriminatory practices that harm consumers may involve vertical integration as network operators favor their own businesses in upstream markets. However, as shown in Section 4, broadband operators could achieve similar results without vertical integration, and even without affiliating with another business. For example, a network operator can charge consumers ten cents per minute for each VoIP phone call, or even just for each Vonage VoIP phone call, without permission from Vonage. Thus, simply prohibiting network operators from providing better service to itself and affiliates accomplishes little. Moreover, banning vertical integration can do harm as there are forms of vertical integration that may yield significant cost savings or other benefits [21, 22].

*Network neutrality should not be about protecting the rights or "freedoms" of consumers.*

The Federal Communications Commission endorsed four freedoms for consumers [23, 24]. Under these principles, consumers should have the ability to access the legal content of their choice, run the applications of their choice, attach the devices of their choice, and receive meaningful information about their service plans. The latter was later changed to a right to competition among network providers, application and service providers, and content providers. This important step in the network neutrality debate gave us useful policy objectives to consider, and variations have been enshrined in a number of proposals for regulation and legislation. However, it is not entirely clear from these freedoms alone how to achieve the stated objectives. What does it mean to have access to content? If it is possible

to download a file but at a painfully slow rate and for an extremely high price, is that acceptable access? If not, on what basis would a regulator decide whether the price is too high or the QOS too poor?

These stated freedoms also do not help the regulator when objectives clash. For instance, I could exercise my right to choose any application by deliberately launching a denial of service attack on my neighbor, depriving him of his freedoms. Or perhaps denial of service is not my intent, but that is still the effect of my resource-intensive application. On what basis can the FCC decide whether to protect my freedom or to protect my neighbor?

Worse yet, these "freedoms" must really be met by the industry as a whole rather than a specific company. If content becomes inaccessible because two companies cannot agree on the terms of interconnection, how can the FCC decide which company has violated its customers' freedoms by making unreasonable demands? If there is no competition, who should be held responsible? Moreover, in a highly competitive market, these objectives can be met if some network operators support consumer freedoms, so how can the FCC determine who among the competing firms has acted unfairly?

These statements of rights or principles clearly have their place, but if we are to develop (or at least evaluate and discard) regulatory constraints, regulations must be based on the acceptable or unacceptable behavior of network operators rather than the inherent rights of consumers.

*Network neutrality should not be about "who pays" for Internet service or infrastructure.*

This issue is of great short-term interest to a few prominent stakeholders, but its broader significance is limited. Today, both consumers and content providers pay the network operator that provides last-mile service directly. If a stream passes through one commercial network, that network is paid by both parties. Otherwise, the consumer pays one network and the content provider pays the other. Some network operators have tried to argue that content providers get a "free ride" because they pay directly for one last-mile connection but not both. Of course, this is no different from cellular telephone calls in the U.S., where both sender and receiver pay for "air time," and we do not hear similar cries about inherent injustice. Some network neutrality advocates would like to permanently enshrine this existing business model for the Internet.

On the other hand, there are network operators who would like content or service providers to pay fees for the last-mile connection to the consumer, in addition to their own last-mile connection. For example, a consumer might pay a monthly fee for her connection to the Internet, while Amazon might pay for each purchase made by the consumer over that connection. Otherwise, the network will block or degrade service for traffic from Amazon. Some self-proclaimed defenders of Internet users call this "double charging," but there are many business models where costs are shared by multiple parties who benefit. There are also communications services where one side pays disproportionately. Callers generally pay the full cost of long-distance telephone calls, and in some countries (other than the U.S.), the same is true for cellular. Again, we do not hear claims that these models are inherently unjust.

Each model has its pros and cons for Internet users, as well as the network operator, and these are largely beyond the scope of this paper. One case in which both consumers and content providers may benefit if the latter pays more of the last-mile costs is the distribution of free, advertiser-supported content. This business model makes it easier for the content providers, and ultimately the advertisers, to pay the communications costs. This saves consumers money and potentially allows advertisers to reach more people. On the other hand, if a consumer *purchases* the content, it should not matter to the consumer whether she pays the network directly or she pays the content-provider who then pays the network (except where transaction costs are different).

Thus, a shift in who pays is not always bad for Internet users, but in some cases it could be. As demonstrated in Section 4, a network operator with market power may be able to adopt discriminatory pricing models that are more harmful to consumers if that operator has the flexibility to charge both sides whatever the market will bear on a discriminatory basis. For instance, a provider of VoIP services might be charged more than a provider of videoconference services, even though the latter clearly requires more network resources. Thus, it is the exertion of market power through discrimination that we must watch for.

Proposals to treat consumers differently from service or content providers create another risk. They assume that consumers cannot also provide content or services, which may actually sanction network operators to reduce the choices available to consumers. Can't a proud parent run a server that gives the world access to baby pictures?

*Network neutrality should not be about whether network operators can differentiate their services.*

Differentiation is not a big issue in regions with only one broadband provider, but if rigorous competition were ever to emerge, some fear that a network neutrality policy would prevent a network operator from offering a unique set of services, and this would turn broadband access into a commodity [25]. One partially avoids this problem by adopting a policy that imposes no constraints if significant competition emerges. Even with only two competitors, if a network neutrality policy only limits discrimination that exploits market power in the last mile, there are still ways for carriers to differentiate themselves. Offering proprietary content as AOL did in the dial-up market would be allowed, provided that the network does not discriminate in favor of this proprietary content.

*Network neutrality should not be about preserving the traditional "end-to-end design principle."*

Under the end-to-end design principle [26], the network provides relatively simple services, while much of the complexity of providing sophisticated services is born by the devices at the edge of the network. This principle has served the Internet well. Among other things, it has facilitated innovation at the edges of the network [27, 28]. However, as discussed in Sections 2 and 3, there has already been a shift away from this principle for sound technical reasons. For example, networks use virus detection mechanisms that improve network security, and caching mechanisms that improve performance. So the shift is not inherently bad. It should become a concern if network operators use this shift to limit the use of new kinds of devices at the edge. Usually, network operators would encourage any innovation that

makes broadband services more valuable, but not when they are trying to extract oligopoly rents, as discussed in Section 4. It is therefore the latter that we should watch for.

### Section 6: Defining a Balanced Net Neutrality Policy

So what should a network neutrality policy be about? We have argued that it should balance two objectives. Based on the results of Section 4, the policy should limit discriminatory practices that allow network operators to exploit their market power to significantly harm Internet users. Impact on upstream markets is especially important, because it is harder to prevent network operators from extracting oligopoly rents in the broadband market itself without onerous regulation, and because the potential consumer surplus that could be extracted in all of the upstream markets combined is probably far greater than that of the broadband market alone. Network operators may extract rents in upstream markets by entering these markets, but this is not essential. Based on the results of Section 3, the policy should try not to interfere with the network operators' ability to use discrimination that benefits users.

It remains to be seen exactly how these objectives can be balanced. It may be impossible for a policy to prohibit *all* forms of harmful discrimination and allow *all* forms of beneficial discrimination, but perfection need not be the goal. We can start by preventing the most harmful cases. A reasonable heuristic may be possible from the following observations. To extract oligopoly rents in upstream markets, a network operator will exploit differences in willingness to pay from one upstream market to another, which means the differences in network prices across these upstream markets will not reflect the costs of providing the service alone. Thus, we might allow discrimination, but seek evidence of prices that are out of line with underlying costs as a possible sign of more harmful forms of discrimination. While it is difficult to quantify the "cost" of carrying a given stream, it is much easier to determine which of two streams would cost more, and regulators can make use of such comparisons

*At a high level, the regulator should be concerned if a network operator with market power is discriminating among traffic streams based on content, application, sender, receiver, or device, in a way that is not justified*

- *by differences in cost (or opportunity cost) of carry the traffic, or*
- *by reasonable security precautions.*

This principle leads us to the following properties, which deserve serious consideration as part of a balanced policy.

A policy designed to protect beneficial uses of discrimination might allow the following:

- Network operators could provide different quality of service to different classes of traffic using explicit prioritization or other techniques. These techniques can be used to favor traffic with stricter quality of service requirements, and/or traffic sent using a higher-priced service.

- Network operators could charge a different price for different classes of traffic. The higher price would be justified because the traffic requires superior quality of service, consumes more of a limited resource, has a greater adverse effect on other traffic, or is otherwise linked to cost (or opportunity cost).
- Network operators could block traffic that poses a threat to security, or that a reasonable network engineer might believe poses a threat to security.
- Network operators could charge the senders of information, recipients, or both.
- Network operators could offer proprietary content or unique services to their customers (without using their dominant control over the last-mile connection to favor their content or service).
- Network operators could block traffic originating from an attached device that one might reasonably believe is harmful to the network or its users, such as one that does not follow prescribed protocols and algorithms.
- Network operators could use *any* form of discrimination they wish, if the broadband market becomes truly competitive.

A policy designed to limit harmful uses of discrimination would not allow the following, if and only if, the broadband market is not highly competitive.

- A network operator could not charge more for stream A than for stream B if stream B requires at least as many scarce resources as stream A. One cannot charge more for a steady 50 kb/s VoIP stream than for a steady 50 kb/s gaming application where the QOS requirements are the same. (Such discrimination has occurred when banning virtual private networks from lower-priced services, for example [28].)
- A network operator could not charge one user more than another for a comparable information transfer or monthly service unless the disparity can be justified by a difference in cost (or opportunity cost). This applies whether the user is the sender or receiver, and whether the user is a consumer, content provider, or service provider.
- A network operator could not block traffic based on content or application alone, unless one can reasonably believe that the traffic poses a security threat.
- A network operator could not degrade quality of service for traffic based on content alone.
- A network operator could not block traffic from a properly functioning device, while carrying traffic from devices known to be technically equivalent.

- A network operator could not offer lower quality of service or higher price for traffic that competes with a legacy circuit-switched service than it offers for comparable traffic that does not compete with a legacy service.
- A network operator could not offer content or services directly or through an affiliate at a data rate or quality of service that is not available to competitors at a comparable price. It similarly could not make network-level services like multicast available to itself or affiliates and not to competitors.

Some believe we cannot develop rules about what is and is not allowed without basing them on the unfathomable intent of the network operator, but none of the rules above depend on intent.

Note that the above restrictions go beyond the traditional role of the Department of Justice's (DoJ) antitrust division. Today, the DoJ would presumably act if a network operator used its market power to limit competition in an upstream market, but probably would not act if a network operator used its market power to extract monopoly rents in an upstream market while allowing competition. For example, a monopoly network operator may be prevented from adding excessive fees to all MP3 downloads that compete with its own service, but not from adding an excessive fee on all MP3 downloads (without a fee on other downloads of comparable size). Either of these policies could have the effect of forcing consumers to pay monopoly prices in the upstream market for music downloads, while the network operator pockets monopoly rents. Of course, DoJ policies can be changed if the DoJ is selected as an enforcement agent for network neutrality, or that responsibility could instead be given to the FCC which has a broader "public interest" mandate.

Perhaps the greatest danger from an overly broad network neutrality proposal is that it could undermine security. Many staunch network neutrality advocates have agreed that discrimination for network security should not be prohibited, but further refinement is still needed. For example, one bill [29] would allow discrimination to improve security, provided that it is not based on application, service, or content. However, it is entirely possible that application, service, and content, allow the operator to conclude that a stream contains a dangerous virus or worm. Other proposals [30, 31] would allow the operator to drop packets for security if and only if a user opts in to this service. However, it is much more effective to keep a dangerous worm out of the network entirely, rather than let it in and merely try to protect some of the users. No matter how the security carve-out is defined, it should protect network operators when they block traffic that they reasonably believe is a security threat, even if they are wrong. There will be false positives and false negatives. If a network operator drops all packets that it believes with 95% certainty are dangerous, should that operator be subject to fines or lawsuits 5% of the time? On the other hand, there must be limits to this flexibility. A network operator should not be allowed to block all encrypted traffic on the grounds that it could conceivably be a security threat.

In some cases, the balancing act is more difficult. Section 4 shows how network operators with market power have incentive to intentionally degrade QOS for some traffic, even when there is excess capacity to provide excellent QOS. If one thinks of the network capacity as fixed, this practice is clearly bad for the user whose QOS is unnecessarily poor. On the other hand, if network operators were

prohibited from this practice, they might have incentive not to increase the capacity of the network, which could harm consumers in the long run.

There will also be more subtle tradeoffs. If a network operator charges more for packet stream A than for stream B when the streams are identical in every way except that one is VoIP, then this is clearly a violation of network neutrality. However, if a network operator can present a reasonable technical explanation as to why it should charge more for the VoIP stream, but the VoIP service provider alleges that it is charging too much more, then the matter is more complicated. The question of how a network neutrality policy could resolve issues like this requires much closer scrutiny. It may even be impossible to resolve that kind of dispute without plunging into detailed price regulation. Nevertheless, even if a network neutrality policy can prohibit only the more obvious abuses of market power, that policy may still have significant benefit.

Network neutrality policies also differ in the extent to which regulatory decisions are made in advance or only after complaints about the alleged misdeed. The above list implies that some decisions should be made through an ex post complaint process. If it is important to allow network operators to use discrimination against traffic that they reasonably believe is a security threat, but not against anything they claim is a security threat, then someone must decide what is reasonable. This probably occurs after a complaint about a network's security policy. Nevertheless, we should strive toward producing and continually updating a set of unambiguous a priori principles that describe what is and what is not allowed, so the complaint process yields few surprises. Companies need regulatory certainty before they can make significant investments. This applies to providers of cable modem and DSL services, potential broadband wireless or broadband-over-powerline competitors, content providers, service providers, and e-commerce merchants.

In fairness, we must note two potential counterarguments to the "balanced policy" suggested above. First, some may question the objective of not harming Internet users. Others might instead try to maximize social welfare, which would include the profits of network operators as well as the benefits to users. All else being equal, it is certainly good to increase these profits, but we assume that transfers from consumers to monopolists would not be considered to be in the public interest.

Even among Internet users, there are winners and losers, and policymakers could consider this. For example, if video streaming over the Internet becomes popular, a policy that allows a network operator to charge much more for this application will harm companies that distribute video and consumers who enjoy their content, but it may allow network operators to provide less expensive service to consumers who want nothing but email access. One can even define scenarios where one group of consumers wins, one loses, and overall consumer surplus increases [15]. Further research is required to determine whether such scenarios are likely to occur often in practice. However, as a general trend, the more a network operator can discriminate on characteristics that are somehow correlated to a user's willingness to pay, the more that operator can increase profit at the expense of consumer surplus.

Others may object to this balanced policy because their goal is to encourage network operators to extend their broadband networks to more of the nation, which is also a worthy goal. Imagine that all

consumers are placed in one of three categories: 1) those in regions that will have broadband regardless of whether there is a network neutrality policy; 2) those in regions that will not have broadband regardless of whether there is a network neutrality policy; and 3) those in regions that will have broadband only if there is no network neutrality policy. Consumers in the first category could be better off with an effective and balanced network neutrality policy, if one can be crafted. Consumers in the second category are unaffected by network neutrality. Consumers in the third category are harmed by network neutrality. In effect, network operators will serve these latter customers only if the operators can extract oligopoly rents from upstream markets. This reduces the value of broadband Internet to users, but at least they have it. Network neutrality then could help consumers in the first category and hurt those in the last, at least in the short run. Given that broadband is spreading, it may be more accurate to say that the consumers in the third category get broadband service earlier if there is no network neutrality protection, but once broadband arrives, it will always be less valuable as a result. This could be a high price to pay in the long run.

## Section 7: Conclusion

Technology has emerged that will give network operators unprecedented ability to discriminate among network traffic based on sender, recipient, content, application, attached device, demographics, and many other characteristics. Network operators can use this information to selectively block traffic, degrade quality of service, and increase prices. This technology is not hypothetical or futuristic; it is here today, and equipment is being marketed explicitly for these purposes.

People following the network neutrality debate know that content and service providers like Google and Vonage may have to pay more if policymakers do not limit discriminatory practices, but even network neutrality advocates are not discussing some important broader dangers. While it is obvious that an unregulated monopoly in last-mile broadband Internet access can bring monopoly prices to the broadband market, it is not obvious that an unregulated monopoly could have the ability and incentive to bring monopoly prices to every upstream market, including electronic commerce for any and all products, communications services like VoIP and videoconferencing, information distribution markets like video streaming and MP3 music downloads, on-line advertising, and network equipment, even when these markets are actually competitive. If perfect discrimination could be achieved, then the network operator could drive consumer surplus to zero in the broadband market and all upstream markets, meaning that all Internet users including consumers, content providers, and service providers would derive no value from the Internet. Network operators may even limit political discourse, at least as it pertains to their business. Luckily, perfect discrimination is not achievable, the equipment to support discrimination is not free, and duopoly competition in the larger markets will inhibit some of these practices, as should the fear of future actions from policy-makers. Nevertheless, there are real dangers that have been somewhat overlooked in the debate, including dangers that are not addressed under existing antitrust policy.

At the same time, we should not underestimate the dangers of imposing a network neutrality policy, especially one that is broad. Network neutrality policies could limit or even prohibit discrimination, and many forms of discrimination are beneficial to Internet users. Discrimination can be used to improve

security, to increase quality of service, to allocate resources to those who need them the most, to prevent starvation, and to decrease total infrastructure costs. If a network neutrality policy were to prohibit such practices, as many current proposals do, there would be collateral damage that deserves serious consideration. We must be sure that we do not adopt a cure that is worse than the disease.

We should try to devise a *balanced policy*, which does not limit the more useful forms of discrimination or constructive innovation, but that prevents a network operator with great market power from using the forms of discrimination that are especially harmful to users. It might be useful to create the concept of "harmful discrimination" which is more limited than "discrimination," much the same way that "harmful interference" is more limited than "interference" in spectrum management. Policymakers should pay particular attention to any attempts to protect legacy services (telephony, video distribution) or to extract oligopoly rents from upstream markets.

Unfortunately, the network neutrality debate has repeatedly been framed in ways that obscure this central issue. Attempts to describe discrimination as inherently wrong are dangerously unproductive, both because discrimination can be beneficial, and because discrimination is not a problem in the absence of market power. Attempts to clarify the rights and freedoms of consumers and of network operators are useful when describing policy objectives, but these rights cannot serve as a useful basis for enforceable regulation, as it is often unclear who is at fault when someone's rights are violated, or what to do when rights come into conflict. The questions about who should pay for services, vertical integration, differentiation among network operators, and the end-to-end design principle are all noteworthy, but they are secondary issues that have distracted policymakers from the more central concerns of a balanced policy.

Misframing the issue inevitably leads to problematic policy proposals. Because the critical role of market power has sometimes been absent in the debate, some network neutrality proposals might apply to any broadband service, which according to the FCC is any service of 200 kb/s or more. Conceivably, data services in a 3G cellular market could some day be subject to severe limits on discrimination even if that market proves to be highly competitive. Also, because some stakeholders stress their concerns about competition from network operators and their affiliates, some network neutrality proposals would only limit discrimination that favors network operators or their affiliates. Because network operators have ways of increasing their profits at the expense of users without these affiliations, such policies would not achieve their intended goals, and these policies may limit some beneficial practices. Finally, because network operators and content and service providers are so focused on whether the latter will have to pay more to the former for "access" to consumers, both sides often forget to debate whether those extra payments can be discriminatory, which is what makes them most dangerous.

This paper has indicated what an effective balanced policy might allow or prohibit in a few cases if such a policy can be defined, and the results differ greatly from most current policy proposals. However, many cases still have not yet been addressed in detail, here or elsewhere. It may ultimately be difficult to both prohibit harmful applications of discrimination and allow beneficial applications. This will disappoint both those who want to prohibit every theoretically possible form of harmful discrimination and those who want to protect any unlikely but conceivable form of welfare-enhancing discrimination. There

may still be plenty of room for reasonable compromise. We will not know what is possible until more detailed proposals are considered by the broader community.

Those members of Congress who have placed network neutrality onto the legislative agenda have forced the community to address an important issue, and warned network operators that some forms of discrimination may lead to sanctions. This is a great service. The same can be said for the FCC Commissioners who supported the consumer freedoms [23, 24]. However, much work remains before an effective and enforceable policy is defined. Success depends on moving the debate from vague principles to specific details about what practical forms of discrimination should and should not be allowed, and where one can prohibit the harmful without prohibiting the beneficial.

## References

- [1] Caspian, *Caspian Media Controller QoS Operation*, April 2006, [www.caspiannetworks.com/PDF/QoS\\_Overview.pdf](http://www.caspiannetworks.com/PDF/QoS_Overview.pdf)
- [2] Cisco Systems, *Deploying Premium Services Using Cisco Service Control Technology*, 2005, [www.cisco.com/application/pdf/en/us/guest/products/ps6150/c1031/cdccont\\_0900aecd8025258e.pdf](http://www.cisco.com/application/pdf/en/us/guest/products/ps6150/c1031/cdccont_0900aecd8025258e.pdf)
- [3] Allot Communications, *NetEnforcer Overview*, [www.allot.com/index.php?option=com\\_content&task=view&id=45&Itemid=44](http://www.allot.com/index.php?option=com_content&task=view&id=45&Itemid=44)
- [4] P-Cube, *Service Control White Paper: The Next Step in Networking for Cable Operators*, 2003, [www.p-cube.com/doc\\_root/products/Engage/WP\\_SC\\_Cable\\_MSOs\\_110202.pdf](http://www.p-cube.com/doc_root/products/Engage/WP_SC_Cable_MSOs_110202.pdf)
- [5] Packeteer, *Gaining Visibility into Application and Network Behavior*, 2006, [www.packeteer.com/resources/prod-sol/VisibilityDrillDown.pdf](http://www.packeteer.com/resources/prod-sol/VisibilityDrillDown.pdf)
- [6] Cable Television Laboratories, *Network Data Management - Usage for IP-Based Services, Service Specification - DOCSIS 1.1 Service Flow Metering*, Sept. 2002, [www.ipdr.org/service\\_specs/DOCSIS\(TM\)/DOCSIS\(TM\)1.1-3.1-A.0.pdf](http://www.ipdr.org/service_specs/DOCSIS(TM)/DOCSIS(TM)1.1-3.1-A.0.pdf)
- [7] A. S. Uluagac, J. M. Peha, "IP Multicast Over Cable TV Networks," *Lecture Notes in Computer Science*, B. Stiller et al. (Eds.), LCNS 2816, Springer-Verlag, pp. 168-180, 2003.
- [8] Rep Edward Markey, *Network Neutrality Act of 2006*, H.R. 5273, May 2, 2006.
- [9] J. M. Peha, "Dynamic Pricing and Congestion Control for Best-Effort ATM Services," *Computer Networks*, Vol. 32, 2000, pp. 333-45, [www.ece.cmu.edu/~peha/pricing.html](http://www.ece.cmu.edu/~peha/pricing.html)

- [10] Q. Wang, J. M. Peha, and M. A. Sirbu, "Optimal Pricing for Integrated-Services Networks," in *Internet Economics*, Joseph Bailey and Lee McKnight editors, MIT Press, 1997, pp. 353-76, [www.ece.cmu.edu/~peha/pricing.html](http://www.ece.cmu.edu/~peha/pricing.html)
- [11] G. R. Bachula, Testimony Before the Senate Commerce Committee, Feb. 7, 2006, <http://commerce.senate.gov/pdf/bachula-020706.pdf>
- [12] W. Lehr, J. M. Peha, M. A. Sirbu, S. Gillett, "Scenarios for the Network Neutrality Arms Race," *Proc. 34th Telecommunications Policy Research Conference (TPRC)*, Sept. 2006.
- [13] C. M. Christensen, *The Innovator's Dilemma*, Harper Collins Publisher, New York, 2000.
- [14] Sandvine, *Sandvine Network Demographic Management*, [www.sandvine.com/general/getfile.asp?FILEID=15](http://www.sandvine.com/general/getfile.asp?FILEID=15)
- [15] H. R. Varian, "Versioning Information Goods," 1997, [www.sims.berkeley.edu/~hal/Papers/version.pdf](http://www.sims.berkeley.edu/~hal/Papers/version.pdf)
- [16] Operax, *Efficient Network Resource Control - A Source of Competitive Advantage*, Sept. 2005, [www.operax.com/docs/efficient\\_network\\_resource\\_control\\_claes\\_sept2005\\_final.pdf](http://www.operax.com/docs/efficient_network_resource_control_claes_sept2005_final.pdf)
- [17] Cisco Systems, *Cisco Service Control: A Guide to Sustained Broadband Profitability*, [www.democraticmedia.org/PDFs/CiscoBroadbandProfit.pdf](http://www.democraticmedia.org/PDFs/CiscoBroadbandProfit.pdf)
- [18] B. Van Schewick, "Towards an Economic Framework for Network Neutrality Regulation," *Proc. Telecommunications Policy Research Conference*, Sept. 2005, <http://web.si.umich.edu/tprc/papers/2005/483/van%20Schewick%20Network%20Neutrality%20TPRC%202005.pdf>
- [19] S. Renshaw, Testimony, Senate Commerce Committee Hearing on Media Ownership, July 8, 2003, [http://commerce.senate.gov/hearings/testimony.cfm?id=831&wit\\_id=2340](http://commerce.senate.gov/hearings/testimony.cfm?id=831&wit_id=2340)
- [20] J. Windhausen, *Good Fences make Bad Broadband*, Attachment N, Public Knowledge White Paper, [www.publicknowledge.org/pdf/pk-net-neutrality-attach-20060206.pdf](http://www.publicknowledge.org/pdf/pk-net-neutrality-attach-20060206.pdf)
- [21] J. Farrell, P. J. Weiser, "Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age," *Harvard Journal of Law and Technology*, Vol. 17, No. 1, Fall 2003.
- [22] C. S. Yoo, "Beyond Network Neutrality," *Harvard Journal of Law and Technology*, Vol. 19, No. 1, Fall 2005.

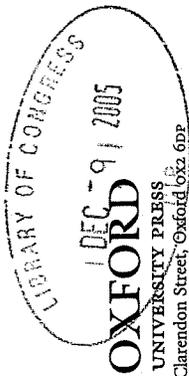
- [23] M. J. Powell, *Preserving Internet Freedom: Guiding Principles for the Industry*, Feb. 8, 2004, [http://hraunfoss.fcc.gov/edocs\\_public/attachmatch/DOC-243556A1.pdf](http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf)
- [24] Federal Communications Commission, *New Principles Preserve and Promote the Open and Interconnected Nature of Public Internet*, August 5, 2005, [http://hraunfoss.fcc.gov/edocs\\_public/attachmatch/DOC-260435A1.pdf](http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-260435A1.pdf)
- [25] J. G. Sidak, Testimony, Senate Commerce Committee Hearing on Network Neutrality, Feb. 7, 2006, <http://commerce.senate.gov/pdf/sidak-020706.pdf>
- [26] D. D. Clark, M. S. Blumenthal, "Rethinking the Design of the Internet: The End to End Arguments vs. The Brave New World," *Proc. Telecommunications Policy Research Conference*, August 2000, [www.tprc.org/abstracts00/rethinking.pdf](http://www.tprc.org/abstracts00/rethinking.pdf)
- [27] L. Lessig, Testimony, Senate Commerce Committee Hearing on Network Neutrality, Feb. 7, 2006, <http://commerce.senate.gov/pdf/lessig-020706.pdf>
- [28] T. Wu, "Network Neutrality, Broadband Discrimination," *Journal of Telecommunications and High Technology Law*, Vol. 2, pp. 141-78, 2005
- [29] Rep. Jim Sensenbrenner, *Internet Freedom and Non-Discrimination Act of 2006*, H.R. 5417, May 18, 2006.
- [30] Sen. Ron Wyden, *Internet Non-Discrimination Act of 2006*, S. 2360, March 2, 2006.
- [31] Sen. Olympia Snowe, *Internet Freedom Preservation Act*, S. 2917, May 19, 2006.

# Economic Transformations

*General Purpose Technologies and Long-Term  
Economic Growth*

RICHARD G. LIPSEY  
KENNETH I. CARLAW  
CLIFFORD T. BEKAR

OXFORD  
UNIVERSITY PRESS



Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide in

Oxford New York  
Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto

With offices in  
Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan Poland Portugal Singapore  
South Korea Switzerland Thailand Turkey Ukraine Vietnam  
Oxford is a registered trademark of Oxford University Press  
in the UK and in certain other countries

Published in the United States  
by Oxford University Press Inc., New York  
© Richard G. Lipsey, Kenneth I. Carlaw, and Clifford T. Bekar  
The moral rights of the authors have been asserted  
Database right Oxford University Press (maker)

First published 2005

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above

You must not circulate this book in any other binding or cover and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloguing in Publication Data

Data available

Typeset by SPI Publisher Services, Pondicherry, India  
Printed in Great Britain on acid-free paper by

Clays Ltd., St. Ives plc.

0-19-928564-0(hbk) 9780199285648

0-19-929089-x 9780199290895

1

HC 79  
T4257  
2005  
COPY 1

2005/19638

# Contents

- *List of Figures*
- *List of Tables*
- *List of Boxes*
- *Foreword*
- *Preface*
- *Acknowledgements*

## Part I GROWTH, TECHNOLOGICAL CHANGE, AND GENERAL PURPOSE TECHNOLOGIES

- 1. Technology as Revolution
  - Pervasive Economic Change
  - Some Comments on Procedure and Method
  - A Preview
- 2. Two Views of Economic Processes
  - Comparisons and Contrasts
  - Functioning of the Market
  - Appendix: Technological Choices and Trajectories
- 3. A Structuralist-Evolutionary Decomposition
  - The S-E Decomposition
  - Incentives and Behaviour
  - Innovating and Non-Innovating Behaviour
  - Path Dependence and the Arrow of Time
  - Appendix: The Evolutionary Hand and Factor Substitution
- 4. Technology and Technological Change
  - Microeconomics of Technological Change
  - General Purpose Technologies
  - Is There a Testable GPT Theory?
  - How Do We Know a New GPT When We See One?
  - Appendix: Measuring Technological Change
    - An Overview of Total Factor Productivity
    - TFP and Costly Technological Change
    - Counterfactual Measures
- 5. A Survey of GPTs in Western History (Part A): 10,000 BC to AD 1450
  - Initial Conditions
  - An Evolving Hunter-Gatherer Society

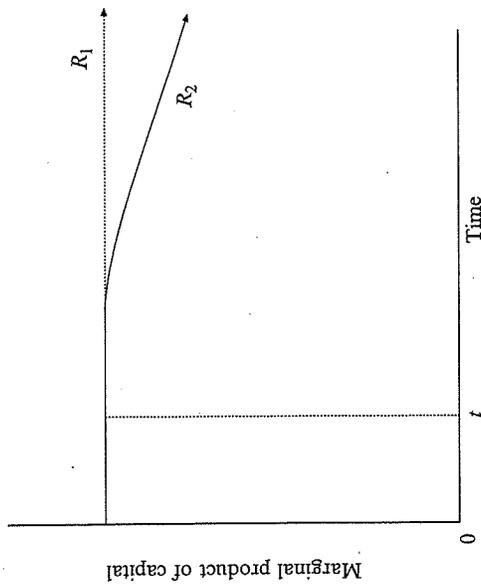


Figure 4A.1. Alternative time paths for the marginal product of capital

investment in existing technologies, as in case 2 in an earlier section. Changes in TFP will thus be zero (which is why TFP is emphatically not a measure of technological change). The second curve,  $R_2$ , falls on the assumption that no new GPTs are invented after time  $t$  so that returns eventually fall as innovation possibilities get used up.

So whether or not there are externalities in the form of technology transfers for which the recipients would have paid more to obtain than they actually did, and whether or not there is a discrepancy between private and social rates of return, the technological complementarities that arise from radically new technologies have been a major (we would say *the* major) source of growth over the last three centuries.

When presenting these ideas at various seminars we have encountered very strong feelings to the effect that if we cannot show how to measure our counterfactual concept of the effect of technological change on output, we should not criticize the methods that are used, specifically TFP. We reject this argument on three grounds. First, if TFP does not measure the output effects of technological change, confusion may result from pretending that it does. Second, if concepts cannot be published and discussed before they can be measured, very few really innovative concepts would ever be developed. Any such concept, or other idea, requires much investigation and developing to refine it before methods of measuring it can be worked out. To ask that a new concept be eventually measurable is not an unreasonable demand; to ask that it be not discussed until it is shown to be measurable is unreasonable. Third, one of us is currently developing measures of the impact of technological change on output, by developing independent measures of technological diffusion, application, and investment, then relating these patterns to actual patterns of TFP change. In another line of measurement development, we are generating data from the model provided in Chapter 14, and comparing what is known from the model where the growth of technology is directly measurable with TFP made under various measurement methodologies.

## A Survey of GPTs in Western History (Part A): 10,000 BC to AD 1450

We now start a tour through the West's technological history, stopping briefly to consider what we see as the most important transforming GPTs and using our S-categories (inputs, technology, facilitating structure, policy, policy structure, an performance) to organize the wealth of historical facts. Since books have been written on each of the GPTs covered in this and the next chapter, what we say about them should be taken as illustrative rather than exhaustive. We concentrate more on the broad consequences of new technologies than is typical in the growth literature, illustrating how technological change has the power to transform society's economic, political, and social institutions. We do not, however, enter into some of the many controversies over the causes of these changes.

This tour provides a base of factual material to guide our subsequent theorizing. It also provides many cautionary tales concerning the wealth of important detail that is omitted in any formal modelling of GPT-based growth. (See Chapter 11 for an overview of such models.) Since it is important for our subsequent theorizing to have at least a basic knowledge of the most important GPTs in history, we put all the material in Chapters 5 and 6 on record. Readers are encouraged, however, to pick and choose those GPTs that most interest them. Any material that is skipped can be referred to later when references to specific GPTs become important.

Starting from the neolithic agricultural revolution, we identify twenty-four transforming GPTs listed in Table 5.1. (We discuss in footnote 1 those classifications that have seemed questionable to some readers.)<sup>1</sup>

Although others might expand or contract our list by a few items, it illustrates several important points. First, the current ICT revolution is not unique; there have

<sup>1</sup> The idea of replanting seeds and selecting those that produced the most desired crops is a process technology with a wide variety of applications of product outputs. So is the idea of selecting animals for breeding to create a wide variety of docile and useful animals. Writing is a process technology that produces a wide variety of written materials—although it uses goods as inputs such as clay, a stylus or pen ink, and paper. The first factory was an organizational innovation in the production of textiles, but later the system spread to the production of a wide variety of other manufactured goods. The same can be said about mass and lean production, which can be regarded as separate GPTs in their own right or as developments of a single GPT, the factory system. Biotechnology is a process technology that is built on three key innovations: how to segment, recombine, and reduplicate sections of DNA. Nanotechnology is likewise a process technology that is built on several innovations concerning how to manipulate individual atoms and molecules, and combine them in almost any way that does not violate the laws of physics.

Table 5.1. Transforming GPTs

No.	GPT	Date <sup>2</sup>	Classification
1	Domestication of plants	9000–8000 BC	Pr
2	Domestication of animals	8500–7500 BC <sup>3</sup>	Pr
3	Smelting of ore	8000–7000 BC	Pr
4	Wheel	4000–3000 BC <sup>4</sup>	P
5	Writing	3400–3200 BC	Pr
6	Bronze	2800 BC	P
7	Iron	1200 BC	P
8	Waterwheel	Early medieval period	P
9	Three-masted sailing ship	15th century	P
10	Printing	16th century	Pr
11	Steam engine	Late 18th to early 19th century	P
12	Factory system	Late 18th to early 19th century	O
13	Railway	Mid 19th century	P
14	Iron steamship	Mid 19th century	P
15	Internal combustion engine	Late 19th century	P
16	Electricity	Late 19th century	P
17	Motor vehicle	20th century	P
18	Airplane	20th century	P
19	Mass production, continuous process, factory <sup>5</sup>	20th century	O
20	Computer	20th century	P
21	Lean production	20th century	O
22	Internet	20th century	P
23	Biotechnology	20th century	Pr
24	Nanotechnology <sup>6</sup>	Sometime in the 21st century	Pr

Note: P, product; Pr, process; O, organizational.

been (GPT-driven) 'new economies' in the past. Second, GPTs have not been common in human experience, averaging between two and three per millennium over the last 10,000 years. Third, the rate of innovation of GPTs had been accelerating over the whole period. We start with millennia between GPTs, then centuries.

<sup>2</sup> Many of these dates are approximate and are based on rough estimates of when their use in the West became widespread enough for the technology to be identified as a GPT from contemporary experience although many were first innovated centuries and even millennia ago.

<sup>3</sup> We include items 1 and 2 but not more modern agricultural developments because the domestication of plants and animals were truly generic developments with many uses that go far beyond food to such things as clothing, containers, shelter, transport, and power (many of which are still being worked out), while later innovations had a much narrower range of mainly agricultural uses.

<sup>4</sup> There is little evidence regarding the origins of the wheel but it was certainly not in use before the agricultural revolution and was in common use by about 3000 BC.

<sup>5</sup> Although continuous process techniques began to evolve with the rationalization that followed the electrification of factories in the late 19th century, we date the emergence of mass production as a GPT at Henry Ford's innovations in the first decade of the 20th century.

<sup>6</sup> Nanotechnology has yet to make its presence felt as a GPT but its potential is so obvious and developing so quickly that we are willing to accept that it is on its way to being one of the most pervasive GPTs of the 21st century.

In the eighteenth century there are two important GPTs, four in the nineteenth century, and seven in the twentieth. The time from first discovery to a fully developed GPT has also accelerated (although not smoothly). Several millennia pass between the discovery of iron and the onset of the Iron Age; hundreds of years passed between the introduction of the waterwheel to Europe in Roman times and its widespread, multipurpose use in the late medieval period; just over a century between Papin's first steam engine and the innovation of the high-pressure steam engine that turned a useful technology into a GPT; while from the nineteenth century onwards, the gestation period between first introduction and emergence as a full GPT has typically been measured in a few decades.

These technologies fall into six main classes, with some overlap. Notice that at any one time there may be several GPTs in existence and even more than one in particular class (e.g. the dynamo and the internal combustion engine).

1. *Materials technologies*: domesticated plants<sup>7</sup>; domesticated animals<sup>8</sup>; bronze; iron; biotechnology.
2. *Power*: domesticated animals; waterwheel; steam engine; internal combustion engine; dynamo.
3. *Information and communications technologies*: writing; printing; computer; Internet.
4. *Tools*<sup>9</sup>: wheel.
5. *Transportation*: domesticated animals; wheel; three-masted sailing ship; railway; iron steamship.
6. *Organization*: factory system; mass production; lean production.

To check that each of these belongs on our list we need to relate each to our fourfold definition of GPTs given in Chapter 4. Stated briefly, a GPT is a technology that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects. The class of technologies, on our list whose inclusion might be thought to be questionable are transportation technologies, and this only with respect to the criterion of multiple uses. For example from a very broad perspective ships are for one purpose, to transport people and goods—but this is no different than power technologies, which in the broadest

<sup>7</sup> Although their first use was mainly for food, plant products provide many varied materials such as cotton, flax, materials for making baskets, clothing, and small boats, and many other vegetable-based products such as herbal medicines.

<sup>8</sup> Although domestic animals were very probably first used for food, they became a major power source and stayed so for millennia, as well as providing many materials such as leather, feathers, furs, and fertilizer.

<sup>9</sup> Of course there have been many other important tools in the history of technology but most are too specialized to have achieved the status of a transforming GPT. The laser is a tool-GPT but, as mentioned in the text, it fitted too well into the existing facilitating structure to be classed as a transforming GPT. As we have argued, the idea of mechanization that is embodied in most factory machinery is a GPT, not a GPT.