VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Edited by David Evans

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Market Definition: Use and Abuse

Dennis W. Carlton

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Market Definition: Use and Abuse

Dennis W. Carlton

Market definition is a crude though sometimes useful tool for identifying market power. The ambiguity in what analysts mean by market power (price above marginal cost, or excess profits) cannot be resolved by market share. When used to analyze a merger or U.S. Sherman Act Section 2 case, it is not just the level of market shares, but also the changes in market shares that are relevant to calculate whether any increase in market power occurs. Despite this, in Section 2 cases courts often use market definition to figure out whether market power exists, a question that can be especially problematic to answer by using market definition. In Section 2 cases, the full antitrust analysis is difficult because any increase in market power typically has to be weighed against any benefits of the alleged bad act. The procedure for defining a market in a merger case or Section 2 case can be rigorously described, but the information required to implement the procedure is typically unavailable. Few analysts (or courts) follow the rigorous procedure in either merger or Section 2 cases. Instead, most markets are defined with some guidance from theory and some qualitative knowledge. Econometric studies using market definition may be helpful both in testing various definitions and in understanding the economic consequences of either the merger or the bad act.

My view is that the definition of a market and the use of market shares and changes in market shares are at best crude first steps to begin an analysis. I would use them to eliminate frivolous antitrust cases when shares are low, but would use them cautiously for anything else. Their usefulness in Section 2 cases is especially weak. Despite their limitations, when they can be used to eliminate frivolous antitrust cases, that use can contribute enormous value to society.

The author is Professor of Economics, University of Chicago, Research Associate at the National Bureau for Economic Research in Cambridge, MA, and Deputy Assistant Attorney General at the U.S. Department of Justice. He thanks Thomas Barnett, David Evans, Kenneth Heyer, James O'Connell, Richard Schmalensee, Hill Wellford, and Gregory Werden for helpful discussions. The views expressed in this paper do not necessarily represent those of the Department of Justice.

I. Introduction

Market definition and the market shares based on it continue to be a central focus of many antitrust cases. This is so despite the well understood limitations of such a methodology in providing an accurate guide to the competitiveness of an industry. The simplicity of the methodology is both its strength and weakness. Its strength is that it is easy to understand and seems intuitively correct—high market shares indicate that competition is weak, while low ones indicate the reverse. The weakness of the methodology is its failure to identify when high market shares may in fact not convey accurate information about an industry's competitiveness, or conversely when low market shares can mask a lack of competition. Although some may call for the elimination of the methodology as an analytic tool because of its limitations, its great strength is that it may prevent decisionmakers from making egregious errors. I think its best use is to provide safe harbors so that firms in relatively competitive industries are not harassed with senseless antitrust suits and, if they are, such suits can be dispensed with at summary judgment.

A "market" can be rigorously and precisely defined quantitatively, but the information to do so is typically not available. Instead, markets are often defined based on qualitative information, leading to the possibility of errors. I make some practical suggestions to mitigate such errors. When markets are correctly defined, it is the change in market shares that is central to the antitrust analysis, though this is not how courts typically use market definition and shares to analyze cases that are brought under Section 2 of the U.S. Sherman Act (Section 2 cases). Unfortunately, there is only a weak link between change in market share and change in competitive performance, and that is why market definition and the use of market shares are very crude tools of analysis. That is why their best use is as safe harbors to quickly screen out frivolous cases from those where the economic forces governing industry behavior need to be carefully studied. But, I explain why even this use of market definition and market shares can be problematic in Section 2 cases.

Although market definition, together with the calculation of market shares, is a crude methodology, if it is to be used, there are certain logical principles that one should follow. Otherwise, this methodology will become even cruder or, worse yet, misleading. Once one has defined a market, one must understand why market shares are a very imprecise way of characterizing competition and are, at most, the beginning point for an analysis, not the endpoint. The government agencies responsible for antitrust, the U.S. Federal Trade Commission (FTC) and U.S. Department of Justice (DOJ), recognize this limitation—it is explicit in the *Horizontal Merger Guidelines*, for example—but courts often have less experience in antitrust matters and that can create problems with the use of market shares.[1]

1 *See* U.S. Dep't of Justice & Federal Trade Comm'n, Horizontal Merger Guidelines (1992, revised 1997) *available at* http://www.ftc.gov/bc/docs/horizmer.htm.

This paper is organized as follows. Section II explains the purpose of market definition, namely the identification of "market power", a term whose meaning is often ambiguous. The section explains that it is the change, not the level, of market power that is relevant in most antitrust cases. Despite this, most single-firm conduct (hereinafter Section 2) cases focus on the level of market power, a calculation for which market definition surprisingly turns out to be particularly problematic.[2] Section III explains how economic theory combined with applicable assumptions tells us precisely what we want to know about the economic effect of mergers, cartels and various types of Section 2 behavior. Using Section III as a framework, Section IV explains the economic principles underlying market definition and market share analysis, emphasizing the sometimes extreme information requirements one must have to define markets, or lacking that information the arbitrariness of market definition. This analysis naturally leads to a discussion of the limitations of market definition and market shares as tools to use to arrive at the correct answer. It pays special attention to feasibility of implementation, and discusses merger and Section 2 cases separately. Section V explains how market definition can be a useful research tool, while Section VI discusses some common mistakes made in applying market definition. Section VII describes how one would apply market definition in two complicated settings: one where research and development (R&D) is central and the other where goods are interrelated as complements, such as in two-sided markets where different market participants exert strong effects on each other. Section VIII concludes with a discussion of how the best use of market definition and market shares is as a safe harbor.

II. What Is the Purpose of Market Definition?

This section makes four points. First, it answers what the goal of market definition is, namely to measure market power. Second, it explains an ambiguity in the definition of market power. Third, it explains why it is the change in market power, not the level of market power, that is relevant to most antitrust analyses. Finally, it explains the limitations of using predicted changes in market shares to estimate the change in market power.

Markets are defined so that when one calculates the share that a firm (or group of firms) comprises, one can assess whether that firm has significant market power. Roughly speaking, "market power" means that the industry's behavior deviates from perfect competition. One standard definition of market power is the ability to set price profitably above the competitive level, which is usually taken to mean marginal cost. For this definition to make sense there must be a possibility that competition could establish the competitive level. Let's suppose that is so—for example, consider an industry where there are constant returns to

2 Some of what I label single-firm conduct cases (e.g., tying, vertical restraints) are covered by Section 1 of the Sherman Act. I mean to include those cases when I refer to Section 2 cases.

scale (it costs C to produce each unit) and many firms. We can contrast price in that industry to an industry with only one (or a few firms) and ask whether the price in the latter case is above the competitive price, C. If it is, we can then ask whether the deviation is big enough to be considered a significant enough deviation from the competitive level to justify an antitrust concern that could trigger an antitrust intervention as, for example, when the market power is created by merger or some other action. Of course, any such intervention carries the risk that the decision will be in error and will do more harm than good.

As far as I know, there are no judicial standards to determine how large a deviation of price from C constitutes significant. The consequence of declaring a specific deviation level as "significant" is that antitrust decisions based on market shares will be made and therefore a decision theoretic framework in which one trades off the expected costs of type I and type II errors is the only one capable of answering the question of what constitutes a significant level. I have never seen any quantitative attempt to use such a framework to answer the question of how large a deviation of price from C should be considered significant. Furthermore, there is a time dimension that must also be analyzed. For how long should a price elevated above marginal cost persist before we attach the label significant? Answers to these questions can be specified based not on any such quantitative assessment but based on what seems reasonable. So, for example, Areeda and Turner suggest using a 5 percent threshold in a discussion about what might constitute a significant price increase.[3]

Before readily accepting this 5 percent threshold, I note that numerous attempts to measure the gap between price and marginal cost estimate gaps in excess of 5 percent for industries that many would consider to be relatively competitive in that there is free entry and several firms. Roughly speaking, a monopolist facing a demand elasticity of 20 would price at about 5 percent above constant marginal cost, but many (most?) firms face much lower elasticities. Perhaps, in light of this, 5 percent may be okay to use to determine whether the change in market power is significant but a higher number may be appropriate to determine whether the level of market power is significant.[4]

3 Philip E. Areeda & Donald F. Turner, 2 Antitrust Law 347 (vol. 2, 1978). Notice that if one uses a 5 percent price deviation (or any specified percent) as a criterion for significant deviation, then there can be a logical problem. Consider the following. Firm A and Firm B merge in New York City causing prices to rise there from $100 to $105, or a five percent increase. The product is also shipped for $100 to Chicago and therefore, the Chicago price rises from $200 to $205, a two and a half percent price increase. Is it sensible to say that a New York City consumer has suffered a significant loss, but not the Chicago one, if each consumes one unit of the product? The problem arises because a percent criterion does not measure the deadweight loss to society, nor does it measure the harm to consumers.

4 Marginal cost can be difficult to estimate. If one approximates it as average variable cost, then one may erroneously measure that there is a gap between price and marginal cost when there is none as, for example, when price equals marginal and average cost and the marginal cost is upward sloping. In this situation, average variable cost underestimates marginal cost. Similarly, economic profit, which requires the calculation of a competitive rate of return, can be difficult to estimate.

Suppose that unlike the previous example in which a competitive price could be defined, the industry is one in which there cannot be an equilibrium where price equals marginal cost. A good example is an industry in which there is a fixed cost of entry and then Cournot competition. Suppose further that there is free entry. The free entry condition guarantees that (economic) profits are zero (i.e., a competitive rate of return is earned on capital), but price will exceed C, marginal cost. There is often confusion between pricing at marginal cost and earning zero profits. In most industries, there is a deviation from perfect competition in that price exceeds marginal cost, yet free entry can still guarantee zero (expected) economic profit. Suppose profits are zero yet price exceeds marginal cost. Should we attach the label "market power" to describe this circumstance, or should we reserve that label for the case in which price exceeds marginal cost and profits are positive? Alternatively, as my textbook suggests, should we label the first situation as "market power" and the second as "monopoly power"?[5] Courts and analysts often fail to specify what definition they are using.

The fact that typically it is difficult to calculate either marginal cost or economic profits foreshadows that the direct determination of the level of market power is going to be hard no matter what definition is used. That is one reason why analysts use market share as a proxy for market power, but, as we will soon see, it may be no easier to define markets to calculate market share than it is to measure market power directly.

THE FACT THAT TYPICALLY IT IS DIFFICULT TO CALCULATE EITHER MARGINAL COST OR ECONOMIC PROFITS FORESHADOWS THAT THE DIRECT DETERMINATION OF THE LEVEL OF MARKET POWER IS GOING TO BE HARD NO MATTER WHAT DEFINITION IS USED.

Although we have been discussing the level of market power, it is the change in market power (which includes any changes in future market power or, alternatively stated, in the durability of market power) that is (or should be) the focal point of most antitrust analysis. (This is not quite right. It is the change in welfare that should be the ultimate focus. But changes in market power can be informative about changes in welfare.) In a merger setting[6], it is a comparison between the market power in two different industry structures that one must analyze in order to predict whether price will rise post-merger. For example, all else equal, is a market where there are five firms with shares 15, 15, 20, 25, 25 significantly less competitive than a market in which the first two firms merge so that there are only four firms with shares 30, 20, 25 and 25? This strikes me as a well-posed question. Notice that the pre-merger level of market power is irrelevant for answering the question. It is only the change in market power that matters. One can answer a question about the change even though

5 DENNIS W. CARLTON & JEFFREY M. PERLOFF, MODERN INDUSTRIAL ORGANIZATION 93 (4th ed. 2005).

6 Cartels and mergers involve similar considerations. For simplicity, I focus on mergers throughout the paper.

one does not know the initial level. Indeed, one can see why a market power definition based on price (P) in excess of marginal cost is particularly convenient to use here. Let P_2 be the post-merger price and P_1 be the pre-merger price. The change in market power equals (P_2 – C) minus (P_1 – C) or $P_2 - P_1$. As long as C is unchanged as a result of the merger, the change in market power is measured as the change in prices. Notice how this approach focuses on the change in price (in the absence of other changes). To the extent that the merger creates efficiencies, so that the marginal cost of the merging parties will fall, this will make an analysis that focuses only on price in a hypothetical where costs do not change a conservative one in the sense that if a merger does not significantly raise price under the assumption of unchanged costs, one would reach the same conclusion if one took further account of any cost efficiencies.[7]

Consider now a Section 2 case in which the issue is whether some alleged bad act (e.g., exclusive dealing) harmed competition. How should one measure whether there is significant market power? Should one measure it before or after the alleged bad act? Following the same logic as in the merger case, one should focus on the change in market power as a result of the alleged bad act and ask how much market power exists absent the alleged bad act and compare it to the market power that exists with the alleged bad act, keeping all else constant. The conceptual difficulty is that the alleged bad act may have some efficiency justification, but price must typically rise in order to create the incentives to generate the efficiency. Indeed, an increase in market power may be desirable if it enables the firm to provide a higher quality product.[8]

For example, exclusive territories can provide incentives for firms to engage in the provision of services by giving them the ability to raise price as a result of the elimination of competition. Therefore, the product characteristics (including service) are not being held constant when one compares the price with and without the alleged bad act. This means that even if the alleged bad act is desirable in that it creates incentives for the provision of valued services to at least some consumers, and even if there are perfect substitutes to the product both with and without services, the analyst who looks at only price will mistakenly conclude that market power is created even though none is. The analyst concludes this

7 Suppose price rose but quality improved. Although the next section shows how to handle this case precisely, for purposes here one should focus on the quality-adjusted price. Suppose price falls, but not as much as marginal cost. Consumers and society gain, so there should be no antitrust concern even though market power has increased. Suppose price rose, but some costs (e.g., fixed costs) fell. Then one would have to do a more complicated analysis to determine whether total welfare rose if one believes that total welfare, not just consumer surplus, should be the proper objective of antitrust. These examples illustrate that it is the change in welfare, not market power, which is the ultimate focus of analysis. *See* Dennis W. Carlton, Does Antitrust Need to Be Modernized? (Economic Analysis Group, Discussion Paper No. 07-3, 2007) and Ken Heyer, *Welfare Standards and Merger Analysis: Why Not the Best?*, 2(2) Competition Pol'y Int'l 29 (2006).

8 I use the term "product quality" broadly to include not just the physical characteristic of the product, but also the way it is sold.

because the analyst observes a lower price in the absence of the alleged bad act and, therefore, incorrectly reasons that the bad act created additional market power. This is why Section 2 cases can be much more complicated than a typical merger case. One expects a price increase as a result of the alleged bad act if the alleged bad act harms competition, but one could also expect a price increase even when the alleged bad act does not harm competition but improves product quality. Therefore, looking only at the behavior of price before and after the alleged bad act does not answer whether the bad act really is harmful. One must dig further and examine, for example, in the case of exclusive distribution, whether some consumers are served better and whether rival manufacturers can still obtain efficient distribution. It is typically hard to trade off the benefit to some consumers from the improved service against the harm to others as a result of the elevated price. Moreover, especially when the services have been provided for many years, it would be wrong to postulate that a reduction in price from elimination of the special services associated with exclusive territories will not harm consumers. For the short term, that may be so, but eventually as the failure to educate consumers mounts over time, the long-run impact on demand could be substantial.

Despite the logic of looking at the change in market power, courts in Section 2 cases often inquire about only the level of market power. In doing so, they are trying to create a safe harbor and shortcut the need to investigate whether market power increased and harmed competition. I discuss this point more fully in Section IV.

Because it is change in market power that is (or should be) the focus of an antitrust analysis, when one is using market shares as a proxy for market power one must focus on the change in shares that results from some particular antitrust decision. But it may be hard to predict the change in share. For example, if Firm A merges with Firm B, the industry will be more concentrated as a result and the analysis measures how that concentration changes as a result of the merger. The concentration measure is based on the pre-merger market shares of the individual firms as in, for example, the Herfindahl-Hirschman Index (HHI) index of concentration, which equals the sum of the squared market shares of firms. So, if there are five firms, each with a market share of 20, and two merge so that the new firm has a share of 40, the HHI rises from 2000 to 2800. We then ask whether that increase warrants concern that price might rise.[9] Notice that I have assumed that the post-merger share of the merged firm equals the sum of the pre-merger shares. That may be so the day after the merger, but need not remain so in the new equilibrium post-merger. When it is not so, then this method will be inaccurate as a guide to predicting how price will change based on how industry

9 In answering that question, the linkage between a change in HHI and a change in price could also depend on the level of HHI.

concentration (which depends on market shares) will change.[10] And, of course, this analysis presumes that a change in concentration will cause a change in price, a relationship that may not be true. Similarly, in a Section 2 context, one should be interested in answering how the alleged bad act alters the market share of the firm engaged in the action. If there are not observations on market share both before and after the alleged bad act began, this calculation could be a source of difficulty.

III. Getting It Exactly Right

As a theoretical matter, if one knows the structure of demand for a product and all its substitutes, knows the cost curves of firms that currently produce (or could produce) the product, and knows the game that describes the competitive environment (e.g., static Cournot, static Bertrand, dynamic trigger strategies), then one can write down a model whose equilibrium reflects the outcome of all these economic forces. This is of course a tall order, but it is critical to know what one would want to measure before turning to proxies, such as market share.

Consider the case in which a merger is to occur. Suppose that Firm A is a dominant firm facing a competitive fringe with supply curve $S^*(p)$. Firm A wishes to merge with a large segment of the competitive fringe so that after merger the competitive fringe will have supply of only $S^{**}(p)$ where $S^{**} < S^*$ for all p. If industry demand is $D(p)$, then the demand pre-merger facing the dominant firm is $D(p) - S^*(p)$ and the profit maximization yields that the pre-merger price p^* is determined by:

$$\frac{p^* - mc}{p^*} = \frac{-1}{E^*}, \qquad (1)$$

where mc = marginal cost of Firm A, E^* = elasticity of demand facing Firm A which equals

$$\frac{1}{s}E^D - E^S\left(\frac{1-s}{s}\right),$$

where E^D = demand elasticity of $D(p)$, E^S = supply elasticity of $S(p)$, and s = share of sales of Firm A.

Landes and Posner use Equation (1) to develop insights about how to define markets in their seminal 1981 paper.[11] It is of course easy to see that the deviation of price from marginal cost depends not only on share s (in the way intuition sug-

10 This method can be adapted as long as one can use pre-merger shares to predict post-merger shares. I show how this can be done in the next section.

11 William M. Landes & Richard A. Posner, *Market Power in Antitrust Cases*, 94 Harv. L. Rev. 937 (1981).

gests: the firm has more market power when s is larger), but also on E^D and E^S, elasticity concepts that depend on how demand or supply changes as price changes. A share will not necessarily reflect either of these elasticities accurately.

If Firm A merges, then the exact calculation of how price changes is the difference between the pre-merger price p^* and the post-merger price p^{**} which is calculated exactly as in Equation (1) but with $S^{**}(p)$ replacing $S^*(p)$. We see that p^{**} will depend on not just how the merger affects the shares of the dominant firm but also on supply and demand elasticities. We could enhance the model and recognize that the merger could lower Firm A's marginal cost, and that could easily be reflected in the calculation of p^{**}.

We can expand the analysis to include market structures other than a homogeneous product with a dominant firm and competitive fringe. Suppose, for example, that each firm i faces demand $d_i(p_1, p_2, \ldots, p_n)$ where $i = 1, 2, \ldots, n$ is a listing of all products. If we know each firm's costs, and know the competitive game (e.g., Bertrand), we can solve for equilibrium prices pre-merger and post-merger. One does not necessarily need to know the cost curves if one is willing to specify the game. For example, if the game is Bertrand, then one can use profit maximization to derive an equation like (1), and calculate mc from p and the elasticity. This is a now standard type of merger simulation used to estimate so-called "unilateral" effects.

There is no reason to limit these simulations to cases where Bertrand is the competitive game, where the competitive game remains unchanged pre- and post-merger, where product quality is unchanged, or to static situations. If one allows for dynamic (repeated) games, one can address what the *Merger Guidelines* call "coordinated effects". All of these complications are difficult to implement, but at least theoretically, these models allow the analyst to focus on what are the underlying forces that matter in influencing how the price will change as a result of the merger. These models show exactly why in the case of merger, market shares or changes in them, however measured, cannot possibly be anything but a crude guide to market power or its change, or to the change in price resulting from a merger.

THESE MODELS SHOW EXACTLY WHY IN THE CASE OF MERGER, MARKET SHARES OR CHANGES IN THEM, HOWEVER MEASURED, CANNOT POSSIBLY BE ANYTHING BUT A CRUDE GUIDE TO MARKET POWER OR ITS CHANGE, OR TO THE CHANGE IN PRICE RESULTING FROM A MERGER.

Now consider Section 2 cases. In Section 2 cases, again the theoretically correct model can be described, though it may be difficult to implement in practice. Let a be the alleged bad acts(s) and let a^* be the act(s) that would occur if a were not allowed. Then, the analyst needs to compare $p(a)$ to $p(a^*)$ where p is the vector of all prices of the relevant products and a and a^* are actions that influence demand (e.g., selling effort) and costs. (The acts could also influence the types of competitive game.) A full analysis of the competitive

consequences of act *a* as compared to act *a** requires an analysis of not just prices, but also how the different acts affect the quality of the product to (some) consumers. For example, if *a* represents vertical restrictions designed to increase sales information to the consumer, then the demand curve for a firm will be affected by whether *a* or *a** occurs. Similarly, the supply capabilities of the firm and its rivals could depend on the firm's actions. Taking these effects into account one can then calculate, at least theoretically, whether banning *a* and replacing it with *a** leads to an increase in welfare.[12]

Let me summarize this section. Although perhaps difficult to implement empirically, theoretical models produce clear results about how to calculate the effect of mergers or alleged bad acts under Section 2 on prices and consumer plus producer welfare. I do not mean to suggest that the assumptions underlying the models are not contentious, or that these models can easily be implemented.[13] I do mean that theory tells us how price and welfare will be determined and therefore theory tells us how to calculate the effect of either mergers or Section 2 behavior.

There is no model that I am aware of where market share (or more precisely its change) is the only variable that matters in predicting the change in either price or welfare. Moreover, it is clear from most models, especially those involving differentiated products, that there is no theoretical need even to define a market to get to the correct answer. At best, market definition and market shares can be used as a shortcut to start the analysis, especially when the correct analysis is hard to do.

---

12 Even if in the context of a particular case one could show that an act caused a decline in welfare, it could still be incorrect to create antitrust liability for the act. The goal of antitrust law should be to create a decision process that leads to maximization of expected welfare. Since the legal process consumes resources and since courts (and economists) can be wrong, it is sensible to create safe harbors for certain types of conduct, even if an economist can show that it is possible that the conduct could under certain circumstances harm welfare. For example, even though it is well-known that above-cost pricing can theoretically harm competition by driving inefficient rivals out of business, there is a safe harbor for such behavior. Even though it is well-known that the choice of product variety or advertising can theoretically fail to maximize welfare, the choice of product quality or advertising typically does not subject a firm to antitrust inquiry. Such behavior falling within safe harbors is sometimes called "competition on the merits". Choosing the appropriate safe harbors is an exercise that should depend on the frequency with which a practice is used in ways that harm society compared to its frequency of use in ways that benefit society, the ability of courts to identify the two uses, the harms from incorrect identifications, and the benefits from correct identifications. As experience with the effect of an act accumulates, the safe harbors should be adjusted. The calculation described in the text of the net effect of an act on social welfare should be done only for those acts that fall outside safe harbors. (A logic similar to that for creation of safe harbors applies to the creation of per se violations.)

13 For a critique of how these models have been used, *see* Dennis W. Carlton, *The Relevance for Antitrust Policy of Theoretical and Empirical Advances in Industrial Organization*, 12 Geo. Mason L. Rev. 47 (2003) and Dennis W. Carlton, *Using Economics to Improve Antitrust Policy*, 2004(2) Colum. Bus. L. Rev. 283 (2004).

Merger cases are typically much easier to analyze than Section 2 cases. Merger cases will usually be handled by answering whether price will rise as a result of the merger. Section 2 cases will usually be handled by asking whether the price increase is offset by some beneficial product change. A focus on the level of market power (rather than its change) can allow a court to provide a safe harbor for either merger or Section 2 behavior if the level of market power after the merger or alleged bad act is low. Courts often use market share to decide that market power is low and we now turn to an examination of whether they can do that in a rigorous way.

IV. Market Definition

We have seen that the theoretically correct analysis may be difficult to implement empirically. In such cases, it is reasonable to resort to a simpler analysis as a first step and that is exactly what market definition and the use of market shares is designed to do. I will discuss merger cases separately from Section 2 cases because, as I have already explained, merger cases are logically easier to analyze.

A. MARKET DEFINITION IN MERGER CASES

1. Mergers: Theory of Market Definition

In a merger case, one uses market shares to calculate industry concentration so as to determine the level of industry concentration and the change in industry concentration as a result of the merger. The implicit assumption is that increases in industry concentration lead to increases in price. (The effect of any particular change in concentration could depend on the level of concentration.) A typical starting assumption is that the post-merger share of the merged firm equals the sum of the pre-merger shares of the merging firms. This of course may not be so as, for example, when entry is easy. In such a case, the use of pre-merger market shares in this way may be inappropriate. But let's suppose that we are in an industry where post-merger the share of the merged firm is well predicted by the sum of the pre-merger shares of the merging firms, so that the use of pre-merger market shares is sensible. There are two virtually equivalent ways to define markets.

One is to rely on demand substitution to identify products and the geographic areas where they are sold and then separately to consider as market participants all those who would supply the product at the current price plus, say 5 percent. This is roughly the approach of the *Merger Guidelines*. A second and virtually equivalent approach is to combine this procedure into one step and define the market to include all those products and areas that constrain prices of the product under analysis from either the demand or supply side. Product A is a demand substitute for Product B if a price increase in B causes consumers to substitute to A. Product A is a supply substitute for Product B, if a price increase in B causes firms that produce A to shift their capacity to the production of B.

To see the difference between the two alternative ways of defining a market, consider the following example. There are two products golf clubs for right-handed golfers (right-handed golf clubs) and golf clubs for left-handed golfers (left-handed golf clubs). Consumers do not substitute between them, so there is no substitution on the demand side. For simplicity, assume that initially firms make either right-handed or left-handed golf clubs. A monopolist of right-handed golf clubs could profitably raise price by 5 percent above current levels as a result of a merger of all current and potential producers of right-handed golf clubs. Firm A makes left-handed golf-clubs, but could and would switch to producing right-handed golf clubs if the price of right-handed golf clubs rose by 5 percent, holding constant the price of left-handed golf clubs, so there is supply substitution. Under the *Merger Guidelines*' approach, the market is right-handed golf clubs, but when calculating shares, one considers all those right-handed golf clubs that would be produced by Firm A and other firms if only right-handed golf club prices rose 5 percent.[14] Under the second approach, the market would consist of right-handed golf clubs plus left-handed golf clubs (somehow appropriately weighted, perhaps by value), and shares would be calculated accordingly. I will follow the first approach, but recognize that the second approach can also be a sensible way to proceed. Since market shares are only crude proxies for market power, these roughly equivalent approaches for calculating shares should not differ and, if they do, one should delve deeper into the underlying economics.[15]

The *Merger Guidelines* recognize the need to define a time dimension, a magnitude of increase, and a benchmark price to approach the question of whether a merger raises an antitrust concern by increasing market power. For example, one could ask whether, after the merger, prices could be profitably increased above current levels[16] by a significant amount (e.g., 5 percent) for a significant time (e.g., 2 years). The *Guidelines* define a market to be consistent with this

14 Typically, one uses the likely capacity of the firm to produce a product as a measure of its market participation. Needless to say, capacity can be hard to measure or even define. As a technical matter, this artifice of holding constant all prices of products outside the market need not be a correct description of what would happen if the price of the product under analysis rose. For example, in the example in the text, the price of left-handed golf clubs could rise as left-handed golf club producers start producing right-handed golf clubs, causing less switching to right-handed golf clubs than in the text. This strikes me as one of many details that should not matter to the analysis and if they do, the analyst must think hard about the underlying economics using the theory of the previous section.

15 Proxies obviously can lead to erroneous conclusions under certain hypotheticals. I am not saying that these two approaches always yield the same result, but if they don't one should reexamine the underlying economics to make sure it is not a peculiarity of the proxy that is generating a strange result. *See* Jonathan B. Baker, *Market Definition: An Analytical Overview* (Nov. 2006) (mimeo, Am. Univ. Wash. Coll. of Law), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=854025.

16 The *Merger Guidelines* use the expected future price (absent merger) if that can be predicted to be different from the current price. They also indicate they may use the competitive price if the current price exceeds it. The logic for the latter approach presumably is that the competitiveness of the industry is expected to increase in the future.

phrasing of the issue. A market is defined by thinking about a hypothetical monopolist. A monopolist of all of the products in a market would raise price profitably above current levels by, say 5 percent, for some time, say two years, on the assumption that the prices of all the products excluded from the market remain unchanged. In this thought experiment of using a hypothetical monopolist, there is not necessarily a unique set of products that determines the market, nor is there an unambiguous methodology of how to raise the price of each product in the market (should each go up by 5 percent or just on average rise by 5 percent?).[17] These strike me as details that again, if they matter, would cause me to pause about the usefulness of the proxy of market shares and to delve more deeply into the underlying economics as described in the previous section.

Aside from determining which products belong in the market, one must determine the geographic scope of the market. I would handle this in the same ways as product market definition is handled: by treating location as a product characteristic and asking the same type of questions as one does for inclusion of a product in the market. For example, apples in Chicago are in the same market as apples in Milwaukee, if an increase in the price of apples in Chicago would induce buyers to switch to buying apples in Milwaukee in such quantities as to defeat a price increase. Suppose no buyer would literally go to Milwaukee to buy these apples, but instead that DC Transport would pick them up and bring them to sell in Chicago. Technically, DC Transport has become a market participant in the market for apples in Chicago. Alternatively stated, there is supply substitution between apples in Milwaukee and those in Chicago. I would treat these two cases—one involving the buyer traveling, the other involving DC Transport traveling—in the same way. One could define the market to be apples in Milwaukee and Chicago, or one could define it using the other approach, in which the market is only Chicago, but DC Transport is a participant in that market. Again, this seems like a detail.[18] If it matters, one should delve more deeply into the underlying economics.

2. Mergers: Practical Implementation of Market Definition

The theory underlying market definition for mergers is logically coherent. A separate issue is whether it is able to be implemented. It is possible to describe an

17 One could add the condition, as the *Merger Guidelines* do, that one use the smallest market and when it is necessary to add products to the candidate market one adds products to the market sequentially with the closest substitute product to the candidate market being added. Regarding which price to focus on, one could focus on the price of the products of the firms involved in the transaction when asking whether price will rise and one could assume that the hypothetical monopolist sets the price of each product in the market optimally. I return to these points in the next section.

18 The *Merger Guidelines* define the geographic area based on the location of production, not consumption. Although this initially may seem odd, it really is not. Because there is an assumption of no geographic price discrimination in this part of the *Guidelines*, they come to the same result as I do above. Notice that the prices in Chicago and Milwaukee become linked in my example.

econometric procedure to define markets.[19] For any set of products ($a_1, a_2, \ldots, a_n$), estimate econometrically a demand system in which the demand for product a_i depends on its own price and that of all other products. Suppose that Product 1 is the product under analysis, such as when two producers of Product 1 want to merge, and that we have ordered the products so that Product 2 is the closest substitute for (Product 1) and so on.[20] Now, assuming costs are known, calculate the price that a monopolist of just a_1 would charge. If that price exceeds the current average price for a_1 by, say, 5 percent, stop. If not, add a_2, and calculate the optimal prices for a_1 and a_2. If (by some measure) the average price of a_1 and a_2 rises above current levels by, say, a 5 percent, stop. If not, continue. In this way, a market can be defined.

This econometric approach requires a tremendous amount of information about a demand system, information that is typically not available. Moreover, if it were available it seems odd to use it only in this way. The reason is that with such a detailed demand system available, it might well make sense to calculate directly the effect of the proposed merger. This can be done by a merger simulation, as described in the previous section, where one uses the demand system combined with various assumptions of the competitive game (e.g., Cournot or Bertrand) and perhaps cost, to predict what the new pricing will be if there is a merger.[21] This direct approach requires no market definition, but utilizes all the same information required to define a market. It is a much more refined way of making predictions on pricing than one based solely on market share. Indeed, this methodology can also account for the fact that products outside the market can affect the price under analysis and the prices of those products may themselves change in response to the merger, in contrast to the procedures for market definition under the *Merger Guidelines*.[22] Market definition, with its dichotomous "in" or "out" classification (is a product in or out of the market?), is a crude simplification; a merger simulation can be a more accurate approach that automatically takes account of demand and supply substitutability.

19 *See, e.g.*, Gregory J. Werden, Market Delineation Algorithms Based on the Hypothetical Monopolist Paradigm (Economic Analysis Group, Discussion Paper No. 02-8, 2002), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=327282.

20 It is a bit tricky to define exactly what one means by closest substitute to a_1. One could say it is the product a_2 such that the joint pricing of a_1 and a_2 allows the price of a_1 to be the highest. When the market consists of more than one product, it is less clear what a unique sensible definition is and differences in this definition can lead to differences in the products included in the market. Moreover, the procedure of adding the closest substitute does not necessarily lead to the smallest market in which a hypothetical monopolist would raise the price of a_1 by 5 percent. Again, these strike me as details that if they mattered to the analysis then one should examine more deeply the underlying economics.

21 As discussed in the previous section, one could at least theoretically assume a repeated game (and so deal with a coordinated effects analysis). Although possible in theory, such simulations are not commonly used in antitrust matters, unlike merger simulations based on static games (e.g, Bertrand).

22 The *Merger Guidelines* would look at price changes in other products, entry responses, and other supply responses, but after the market is defined.

The drawback of merger simulation is that it requires not only extensive demand estimation, but also assumptions about how firms will compete. Even if one has information on the former, many are uncomfortable about assumptions on the latter.[23]

There are really two responses to this reluctance to use merger simulation. The merger simulation, when done under different assumptions, is really a way of revealing to the analyst the constraints on pricing that the demand system imposes and makes transparent all the underlying assumptions. The different merger simulations allow the analyst to see whether these constraints hold under a variety of assumptions. Second, if instead of doing a merger simulation, one defines a market and uses pre-merger market shares to calculate the change in the HHI, one is assuming that these market shares allow one to predict the price effect of a merger. That is, the price is assumed post-merger to depend on (pre-merger) market shares in a simple way (e.g., price is assumed to depend on just the HHI). There is no such model that I am aware of that has this property. There are models in which price depends on current concentration and other things such as elasticities, but not only are those models premised on assumptions that may not be relevant to the industry under analysis, worse yet, in such models there may not be a profit incentive to merge.[24] This is all a very long way of saying that the use of changes in market shares to calculate the change in the HHI is a very crude methodology for predicting whether a merger will increase price. The use of market shares is at best viewed as a crude merger simulation, but lacks the logical consistency underlying merger simulation. Its main attractiveness is its simplicity.

But there is a further problem. I had assumed that a detailed econometric demand system together with knowledge of costs was available. When it is not, then it is not possible to delineate a market with the precision that its definition demands.[25] Instead, one attempts to use various types of evidence to do one's best to see whether the price constraining effect of one product on another will be sufficient to prevent a significant price rise. Although the clear theoretical construct of market definition can guide one, the absence of estimates of the demand (or cost) system subject this exercise to possible error and arbitrary judgments.

---

23 *See* Carlton (2003, 2004), *supra* note 12.

24 For example, in a Cournot model with constant returns to scale, one can show that $(P- C)/P = HHI/E$ where P is price, C is cost, E is the absolute value of the industry demand elasticity, and HHI is the sum of squared market shares. *See* Stephen W. Salant et al., *Losses from Horizontal Merger: the Effects of an Exogenous Chance in Industry Structure on Cournot Nash Equilibrium*, 98 Q. J. Econ. 185 (1983), and Joseph Farrell & Carl Shapiro, *Horizontal Mergers: An Equilibrium Analysis*, 80 Am. Econ. Rev. 107 (1990).

25 To define a market using the hypothetical monopolist test, one must specify marginal cost. To do a merger simulation, one could also use cost information, or alternatively infer cost from the profit-maximizing conditions that emerge from equilibrium of the assumed competitive game.

These errors can be mitigated by some of the types of econometric studies that I describe in the next section.

One alternative path to market definition in the absence of detailed econometric estimates of a demand system is simply to ask consumers to which products they would turn if price of the product under analysis rose by, say, 5 percent. Notice that this set of products does not satisfy the market definition under the *Merger Guidelines* because it may include products that attract so few switches that those products would not prevent a price increase. Therefore, although this method is simple, markets defined in this way will tend to be overbroad unless one includes only those products for which there is significant substitution (how much is significant?—well if I define it precisely then I am back to an approach like that of the *Guidelines*). However, consumer responses as to their switching possibilities can give one a rough estimate of demand price elasticities and cross elasticities, and those can assist in defining a market.[26]

I have not discussed critical loss analysis, because it is not an alternative method for defining markets. When done correctly (as Harris and Simons recognize)[27], it is simply a rephrasing of the hypothetical monopolist test. It asks what is the critical amount of demand that has to be lost in response to a price rise before the price rise is unprofitable. That is a question about how big the demand elasticity has to be to make a price increase unprofitable. Critical loss can help one describe this critical demand elasticity, but it is not a new analytic tool and has been misused.[28]

THE METHODOLOGY OF MARKET DEFINITION AND MARKET SHARES IS AN EXTREMELY CRUDE WAY OF ASSESSING A MERGER'S COMPETITIVE EFFECT, ESPECIALLY SINCE MARKET DEFINITION IS USUALLY NOT BASED ON THE EXTENSIVE QUANTITATIVE INFORMATION REQUIRED TO DEFINE IT RIGOROUSLY.

The methodology of market definition and market shares is an extremely crude way of assessing a merger's competitive effect, especially since market definition is usually not based on the extensive quantitative information required to define it rigorously. The methodology can certainly be informative in many cases, but it is only the first

26 For a more skeptical view of the value of relying on consumer responses, *see* Ken Heyer, *Predicting the Competition Effects of Mergers by Listening to Customers*, 74 ANTITRUST L.J. 1 (forthcoming 2007). Another procedure to define markets is to identify products whose prices are highly correlated. Stigler and Sherwin recommend this procedure. *See* George J. Stigler & Robert A. Sherwin, *The Extent of the Market*, 28(3) J. L. & Econ. 555-85 (1985). Although the procedure can sometimes be a useful way to start an analysis, it has quite serious drawbacks. *See* CARLTON & PERLOFF, *supra* note 54, at ch. 20 and Gregory J. Werden and Luke M. Froeb, *Correlation, Casualty, and All that Jazz: The Inherent Shortcomings of Price Tests for Antitrust Market Delineation*, 8 REV. INDUS. ORG. 329 (1993).

27 *See* Barry Harris & Joseph Simons, *Focusing Market Definition: How Much Substitution Is Necessary?*, *in* RESEARCH IN LAW & ECONOMICS 207 (Richard O. Zerbe, Jr. ed., 1989).

28 *See* Carlton (2004), *supra* note 12.

step in an analysis that must delve into the economic facts of the industry. It can be a useful guide, but only if subsequent analysis confirms its message. The methodology's 0 or 1 nature (i.e., "in" or "out") together with the arbitrariness of certain decisions (e.g., why hold the price of products outside the market constant?), emphasizes its crudeness. Still, the use of market shares (or changes in them) is simple, and it can be thought of as the first step in a merger analysis. Its best use is likely to provide a safe harbor when industry concentration and shares of merging firms are low.

B. MARKET DEFINITION IN SECTION 2 CASES

We have already discussed how the central issue in a Section 2 case is whether some alleged bad act enables additional market power to be exercised, and, if so, whether any exercise of additional market power is offset by the additional provision of valuable services made profitable as a result of the price increase. Estimating market power while adjusting for services provided can be difficult and it is even more difficult to figure out whether an increase in market power is offset by improved services—the traditional pro-competitive explanation for many alleged bad acts.

Instead of focusing on whether the alleged bad act increases market power, the courts typically focus on whether there is market power and, if so, whether the alleged bad act is justified on pro-competitive grounds. One reason, I think, for this current emphasis on the level of market power (whether it is measured before or after the bad act often seems not to be a focus of attention) is because at the summary judgment stage, a case can be thrown out if there is no market power, while it is thought to be more difficult to get the case thrown out at summary judgment if one concedes market power but defends by claiming that the action is pro-competitive. Because the courts focus on existing levels of market power, this has required markets to be defined in Section 2 cases to see whether market power exists (presumably, after the alleged bad act has occurred). My experience is that courts ask whether market power exists in the presence of the alleged bad act, a question with the potential to be answered in a misleading way if one ignores the efficiency justification for the alleged bad act, as I explained in a previous section. Moreover, such an analysis fails to consider whether the bad act creates any additional market power. Still, the procedure does have a logic because if there is no market power after the alleged bad act, then the antitrust inquiry ends.

To answer the question of whether the firm has market power, some have tried to adapt the procedure of the *Merger Guidelines* to define a market in a Section 2 context. As a logical matter, this initially seems fine with the benchmark price now no longer being the current price but rather the competitive price. So the hypothetical monopolist test to define a market is as follows: consider all those products such that a hypothetical monopolist of those products would raise price above the competitive level by, say, 5 percent. One then calculates the market

share of the firm in this market and if it is high one concludes that there is market power. But what sense does this make? Suppose the current price is $10. If one knows that the competitive price is $5, then the market definition exercise is useless! One can observe whether the current price ($10) exceeds the competitive price ($5) and the deviation is the measure of market power. There is no need to define a market and calculate market share in order to see whether the market share is so high that one can safely conclude that $10 is higher than $5. Alternatively, if one does not know the competitive price, there is no way to implement this market definition test.[29]

But a bit more analysis shows that the logic of using the competitive price as the benchmark price is not necessarily correct. In a merger case, we typically use the current price as the benchmark, not the competitive price. That is sensible because the relevant question is whether the merger will raise price from current levels. By similar logic, in a Section 2 case we should use the price that would prevail in the absence of the bad act as the benchmark price in order to define a market and calculate market shares in an effort to determine whether the firm has enough market power so that it could possibly use (or have used) the bad act to elevate price.[30] The hypothetical monopolist test for market definition in a Section 2 case should be: include all those products such that the hypothetical monopolist would raise price by 5 percent above the benchmark price, defined as the price that would prevail absent the bad act. If the firm's market share is low, the inquiry should end.[31] It may sometimes be difficult to figure out the benchmark price, though not always. For example, if the bad act has not yet taken effect, the current price can be used as the benchmark price.[32] But when, as will commonly occur, this is not the case, the analyst could have difficulty.

In this situation, one is in the uncomfortable position of realizing how arbitrary market definition can be in Section 2 cases and how this arbitrariness can lead to errors. Perhaps the best one can say is that one might look at similar firms and throw out the antitrust case if there are enough of them—but that is a cop-out

29 It is also correct to say that in the absence of cost information, one cannot define a market in a merger case using the *Guidelines* in the rigorous way I described earlier.

30 If possible, the expected post-bad act market share of the firm should be used. The well-known *Cellophane* fallacy arises when one uses the post-bad act price as the benchmark price.

31 If one concludes that there is market power, then as described previously, one should compare the price effect of the bad act to any efficiency effects associated with the bad act. The change in market share pre- and post-bad act may give insight into the likely price effect.

32 If the benchmark price is known and the price after the bad act is known, then, as already explained, there is no need to go to the effort to define a market. If the benchmark price is not known, one cannot define the correct market. If the benchmark price is known, but the price after the bad act is not known, then one may benefit from defining a market and asking whether the bad act is likely to allow the firm to achieve a sufficiently high market share that market power concerns arise. If not, the inquiry ends.

unless one can define what "similar" means. If one is able to establish a benchmark price because there is a consensus that in some areas (or time periods) there are no bad acts, one can then use econometric techniques to try to use those benchmark areas and their characteristics to calculate the benchmark price in any area. This can be a useful approach, and one I describe in the next section.

V. Is Market Definition a Useful Tool for Understanding Market Behavior?

So far I have discussed market definition only in the context of antitrust cases, but what about as a research tool to understand economic behavior? Should economists study market definition and market shares in their academic research and if so wouldn't such studies be relevant in antitrust cases? It is undeniable that most of the current interest in market definition stems from its use in antitrust cases. But, although it is no longer as popular as it once was, there was a flourishing literature in relating market performance to market structure measured by market shares. This literature has been heavily criticized,[33] because, among other reasons, a market share does not have the same economic effect across industries, which differ enormously, and because market share is an outcome of industry fundamentals, not a basic characteristic of them. Such studies are sometimes still used in academic studies and can be done properly. They are used in antitrust studies and, under appropriate circumstances, can be a powerful tool not just for checking market definition, but also for understanding the economic behavior of the industry.[34]

Consider a proposed merger between two firms. One may well be able to use the past historical relationship between price and concentration to predict the effect of the merger. One could use regression analysis to estimate this relationship, though caution is needed to deal with the determination of concentration.[35] Simply analyzing the relation between price and concentration over time may tell one nothing about the relation of competitiveness to concentration absent a theory explaining why concentration might be changing over time. However, it is sometimes possible to construct such theories and to use the estimated relationship between price and concentration as a predictor of a merger's effects. For example, in the railroad industry where tracks were laid many years ago, it seems sensible to predict the effect of a merger of two railroads that will

33 *See, e.g.*, Carlton & Perloff, *supra* note 5, at ch. 8.

34 *See, e.g.*, Dennis W. Carlton & Hal Sider, *Market Power and Vertical Restraints in Retailing: An Analysis of FTC v. Toys 'R' Us*, *in* The Role of the Academic Economist in Litigation Support (Daniel Slottje ed., 1999); Carlton (2003), *supra* note 12; and Carlton (2004), *supra* note 12.

35 The statistical issue is whether concentration should be treated as an exogenous or endogenous variable.

reduce the number of railroads serving a route from 3 to 2 by comparing pricing on routes with 3 railroads to those with 2, after adjusting for other route characteristics. In fact, a recent paper by Peters, analyzing the airline industry shows that such predictions based on the historical relationship of price to concentration are often as or more accurate than those based on merger simulation.[36]

Such econometric studies can also shed light on the appropriate market definition. For example, suppose there is a question whether Product B is in the same market with Product A. A regression reveals that there is a relation between the price of Product A and market concentration based on a market definition excluding Product B, but no relation based on a market definition including Product B. Under appropriate statistical circumstances, that can be quite informative as to the correct market definition and can indicate that Product B is not in the same market as Product A. I have often found these types of econometric analyses helpful in understanding both market definition and predicting the consequence of mergers.[37]

Similarly, in the context of Section 2 cases, one can use econometric techniques to explore the direct effect of a bad act if one is fortunate enough to have data on periods when the bad act was in use and not in use. Again, one has to make sure that one can deal with the statistical issue of exogeneity properly, but if so these studies can be valuable. One can also use the same type of studies as just described in the merger context to test which definitions of market make sense and are useful for prediction.

VI. Common Mistakes in Defining Markets

Although I have stressed the limitations of the methodology of using market definition and market share, I have also explained that the methodology still can sometimes be useful if done in a way that captures the underlying economics, especially in the context of merger cases. In this section, I list a few of what I have found to be common mistakes:

(1) **Firm 2 is producing at capacity. Hence, it cannot increase supply to offset a hypothetical price increase by Firm 1, and accordingly should be excluded as a participant from the market.** This logic correctly recognizes that Firm 2's zero supply elasticity means that increases in Firm 2's output cannot constrain Firm 1's price. But it fails to recognize that Firm 2's existing production constrains Firm 1's ability to raise price. Suppose it costs $1 to make one unit of wheat. In equilibri-

36 Craig T. Peters, *Evaluating the Performance of Merger Simulation: Evidence from the U.S. Airline Industry*, 49 J. Law & Econ. 627-49 (2006).

37 *See* Carlton (2003), *supra* note 5, and the similar views of Coleman and Scheffman *in* David T. Scheffman & Mary Coleman, *Quantitative Analyses of Potential Competitive Effects from a Merger*, 12 Geo. Mason L. Rev. 319 (2003).

um, 1,000 units are sold at $1 each. Imagine 1,000 wheat farmers including Firms 1 and 2 each of which produces one (and only one) unit. Each wheat farmer likely faces a highly elastic demand precisely because of the output of the others, and would on its own be unable to increase the price of wheat. Excluding capacity constrained wheat farmers would incorrectly indicate that Firm 1 has market power.[38]

(2) **Firm 1 produces steel. It has several long-term customers. The capacity to serve those customers should not be considered in calculating the market for steel in evaluating a merger involving other firms.** If the customers have signed long-term, fixed-price contracts with Firm 1, but the steel can be resold, then the capacity to produce that steel should be in the market, but should not be attributed to Firm 1. If the steel cannot be resold, the contract will not be breached, and the output produced by these customers does not affect other customers of steel, then the steel sold to these customers should be excluded from the market. However, if the output of these customers does constrain the prices of the products of other steel customers, then the steel output to these customers should be included in the market, but should not be attributed to Firm 1. The presence of these customers constrains the price that these other steel customers can pay for steel. If there is no fixed-price contract, then the capacity is attributable to Firm 1. The price to long-term customers will be set in the marketplace where the price reflects competition amongst many other steel producers.

(3) **Used goods sell for a lower price than new goods and therefore are not part of the same market as new goods.** Used goods sell for a lower price than new goods for many reasons, including the fact that they have fewer years of service to provide. Whether they are in the same market as new goods depends on how good a substitute they are for various demanders. For example, if used goods have greater reliability problems than new goods, there may be a class of consumers willing to pay a (length-adjusted) price that reflects a premium for the reliability. That could mean that used and new goods do not tightly constrain each other's prices, but that is an empirical question.[39]

VII. Market Definition in Complicated Settings

I now discuss two somewhat complicated settings and see how useful market definition can be. Since we have already seen its limitations in even relatively sim-

38 The elasticity of the residual demand curve facing a single farmer equals E/s where E is the aggregate demand elasticity and s is the market share of our single farmer. This elasticity facing a farmer will be large for small s.

39 *See* Dennis W. Carlton & Robert Gertner, *Market Power and Mergers in Durable-Good Industries*, 32 J. L. & Econ. 203 (1989).

ple settings, we should not be surprised that its limitations are even more severe as the circumstances become more complicated. We discuss two settings. One is where R&D is important. In such settings, I ask whether it is sensible to think of an "R&D innovation market", a concept that was used by the DOJ in the 1990s. The second setting is one involving what are called "two-sided markets".[40] These are markets where multiple inputs and outputs require coordination in order to produce desirable products. One example is a mall in which the mall owner must account for the fact that some stores attract customers to the mall, yet those customers buy at other stores in the mall. Another example is an operating system for computers, where the owner of the operating system wishes to induce application programmers to write applications programs for its operating system so as to make its operating system attractive to users. In such cases, there are interactions between different sides of the market that should be internalized. So, for example, the mall owner subsidizes the rent of the bookstore, but charges a high rent to the restaurant. Or, the owner of the operating system subsidizes application programmers, but charges users a high price for the operating system. Other common examples of two-sided markets include dating clubs, game stations and games, and card payment systems. As far as I know, there has been no recognition yet by courts of market definition in two-sided markets.

A. INNOVATION MARKETS

An innovation market consists of the future innovations in some area, presumably measured by the resources devoted to R&D in the particular area.[41] Shares are calculated for each firm in the obvious way. Notice that this analysis is focused on an input (R&D) not the output (new products). It is a departure from the usual procedures of basing market definition on products. It would be a justifiable procedure if it were easy to predict which R&D will lead to which new product, but in many (most?) cases it is not possible to do this. The success of R&D is highly uncertain and predicting from where R&D breakthroughs will come is very hard. Perhaps pharmaceuticals are an exception because one can see exactly how far along a drug development is in its U.S. Food and Drug Administration trials. Yet a market for particular drugs in development really is not an R&D market, instead it is a market for a future product (uncertain as it may be). This is different from a market based on R&D for a particular general type of product. Moreover, we lack a theoretical framework for defining markets for R&D innovation markets. What is the analogue to a 5 percent price increase? What price is being measured if the product cannot be defined? Furthermore, the

40 See David Evans & Richard Schmalensee, *The Industrial Organization of Markets with Two-Sided Platforms,* 3(1) Competition Pol'y Int'l 151–179 (2007) and Jean-Charles Rochet & Jean Tirole, *Two-Sided Markets: A Progress Report*, RAND J. Econ. (Autumn 2006).

41 Richard J. Gilbert & Stephen C. Sunshine, *Incorporating Dynamic Efficiency Concerns in Merger Analysis: The Use of Innovation Markets*, 63 Antitrust L. J. 569 (1995).

link between R&D concentration and new product output is quite weak.[42] For all these reasons, I am skeptical that the already crude theoretical construct of market can be of much use in analyzing industries where R&D is key.[43] Of course, this does not mean that the effect of a merger on innovation in such industries requires no analysis. Instead, it does mean that the tool of innovation markets is likely to be of little help in the analysis.

B. TWO-SIDED MARKETS

In two-sided markets, one party (e.g., mall developer, owner of a computer operating system) internalizes the externalities across agents by effectively taxing and subsidizing different groups so that a total package is produced. There has been a literature questioning the empirical relevance of these two-sided markets (or the related concept of industries with network economies). In such markets, without the coordinating ability of a third party, markets cannot produce the efficient result. The lack of a third party could then indicate either no need for one or the existence of a market failure.[44] For purposes of this discussion, I assume that a third party is needed and does exist in order to coordinate activity among different groups. What is a sensible procedure to define a market in such a case?

To take a concrete example, suppose two shopping malls want to merge.[45] To simplify, assume that there are no surrounding competing retail stores that are not in malls. We start out by recognizing that a mall owner puts together a portfolio of stores that complement each other and whose existence he coordinates by lowering the rent of one type of store to stimulate demand (and elevate rent) at another. Suppose that the mall owner charges each store a rent based on its retail sales. Following an approach similar to the *Merger Guidelines*, we ask: Which nearby malls must a hypothetical monopolist control in order for it to be profitable for the merged firm to raise the price by, say five percent? But just as in the earlier discussion of market definition when multiple products were in the market, one must define what "price" means. Is it the rent of one particular retail store, average rent, or total rent that has to rise? Or, to complicate matters fur-

42 *Antitrust Modernization Committee Hearings* (2005) (statement of Richard J. Gilbert, Professor, Univ. of Cal. at Berkeley, "New Antitrust Laws for the 'New Economy'," Nov. 8, 2005), *available at* http://www.amc.gov/commission_hearings/pdf/Statement_Gilbert.pdf.

43 For a more detailed critique, *see* Dennis W. Carlton & Robert Gertner, *Intellectual Property, Antitrust and Strategic Behavior*, *in* Innovation Policy & The Economy (Adam Jaffee & Joshua Lerner eds., 2003).

44 For example, when cars were being developed, the car manufacturers could have perhaps benefited from subsidizing location of gas stations. The fact that no such subsidization occurred shows either the market was inefficient, or alternatively, that whatever inefficiency existed, it was too small to cause it to be corrected. *See* S.J. Liebowitz & Stephen E. Margolis, *Network Externality: An Uncommon Tragedy*, 8(2) J. Econ. Persp. 133-50, 133 (1994).

45 For an application of market definition to credit cards, *see* Eric Emch & Scott Thompson, *Market Definition and Market Power in Payment Card Networks*, 5 Rev. Network Econ. (2006).

ther, if customers pay a parking fee, or are provided with elegant surroundings, how should those be changed when this hypothetical monopolist raises "price"? In the earlier discussion of market definition when the market contained multiple products, I recognized the ambiguity in the definition of price but said that I doubted that it should matter much, though I indicated a preference to focus on the products of the merging firms, rather than all products in the market. But here, there is no one type of retail store to focus on.[46] Therefore, one should focus on an aggregate measure of rent. Moreover, we know that because of the two-sided nature of the market it is unlikely that it is optimal for the hypothetical monopolist to raise rents to all stores by 5 percent. Indeed, the whole point of having a mall is to charge different rents to different types of stores. Failure to allow the hypothetical monopolist to set rents optimally could lead one to a misleading market definition and, depending on the circumstances, to either overstate or understate the market power of a mall owner. For example, one might conclude that post-merger there is no market power (i.e., a very broad market in which the post-merger mall owner has a small share) when with optimal pricing the market is narrower and the mall owner has a large market share reflecting market power created by the merger. Conversely, if one ignores the ways in which mall owners can compete to attract customers directly or indirectly through low rents to some stores, one could find market power when in fact there is none. My sense is that this problem of using the right price will make market definition in two-sided markets more difficult than in the typical case and will therefore further limit reliability of market definition and market shares.

MY SENSE IS THAT THIS PROBLEM OF USING THE RIGHT PRICE WILL MAKE MARKET DEFINITION IN TWO-SIDED MARKETS MORE DIFFICULT THAN IN THE TYPICAL CASE AND WILL THEREFORE FURTHER LIMIT RELIABILITY OF MARKET DEFINITION AND MARKET SHARES.

VIII. Conclusion

Market definition is a crude though sometimes useful tool for identifying market power. The ambiguity in what analysts mean by market power (price above marginal cost, or excess profits) cannot be resolved by market share. When being

46 Notice that the product is malls, not individual retail stores. If one does mistakenly focus on rent to only a particular type of retail store, one must recognize the two-sided nature of the market in which feedback effects occur in other retail stores in the mall. An increase in the percent of sales charged as rent to the bookstore could lead to higher book prices and fewer customers to the bookstore and, thereby, to all other stores in the mall. The fall in mall customers leads to a decline in sales in other retail stores and a decline in rents from these stores. Failure to understand this feedback effect could lead one to overestimate the profitability to the mall owner of raising rents to the bookstore and, thereby, lead one to define markets too narrowly and overestimate market power.

Notice that this type of feedback effect can also arise in one-sided markets, when a firm sells complementary products. The price increase in one product will adversely affect sales of the other, and that effect will temper the profitability of a price increase in the initial product.

used to analyze a merger or Section 2 case, it is not just the level of market share, but also the changes in market shares that are relevant to calculate whether any increase in market power occurs. Despite this, in Section 2 cases courts often use market definition to figure out whether market power exists, a question that we have shown can be especially problematic to answer by using market definition. In Section 2 cases, the full antitrust analysis is difficult because any increase in market power typically has to be weighed against any benefits of the alleged bad act. The procedure for defining a market in a merger case or Section 2 case can be rigorously described, but the information required to implement the procedure is typically unavailable. Few analysts (or courts) follow the rigorous procedure in either merger or Section 2 cases. Instead, most markets are defined with some guidance from theory and some qualitative knowledge. Econometric studies using market definition may be helpful both in testing various definitions and in understanding the economic consequences of either the merger or the bad act.

My view is that the definition of a market and the use of market shares and changes in market shares are at best crude first steps to begin an analysis. I would use them to eliminate frivolous antitrust cases when shares are low, but would use them cautiously for anything else. Their usefulness in Section 2 cases is especially weak. Despite their limitations, when they can be used to eliminate frivolous antitrust cases, that use can contribute enormous value to society.

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Time to Rethink Merger Policy?

Jordi Gual

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1.
Competition Policy International is a free publication. To order or download additional copies, visit eSapience.org.

Time to Rethink Merger Policy?

Jordi Gual

This paper provides a critical analysis of some of the key features of merger policy as understood and practiced in leading jurisdictions such as the European Community and the United States. It focuses first on a discussion of the gradual move of merger policy towards the examination of unilateral effects. The critical appraisal of this process is based on the practical and theoretical shortcomings of the economic models that underlie the growing prominence of unilateral effects as the key anticompetitive factor arising from a proposed merger. The paper stresses that even if unilateral effects were to lead to an increase in the conventional measures of anticompetitive performance (such as markups), it is not clear that this implies less competitive behavior for many of the most relevant industries in today's advanced economies. Finally, the paper also examines the relation between competition and welfare, and argues that even if competition does indeed diminish due to a merger, it is not a straightforward conclusion that this is not good in terms of economic welfare when the incentives to innovate and the dynamic welfare gains that arise from new products and production processes are taken fully into consideration.

The author is Professor at IESE Business School and Chief Economist and Head of Research at "la Caixa".

I. Introduction

This paper provides a critical analysis of some of the key features of merger policy as understood and practiced in leading jurisdictions such as the European Community and the United States. It focuses first on a discussion of the gradual move of merger policy towards the examination of unilateral effects. The critical appraisal of this process is based on the practical and theoretical shortcomings of the economic models that underlie the growing prominence of unilateral effects as the key anticompetitive factor arising from a proposed merger. The examination of why non-cooperative behavior is judged to be anticompetitive leads naturally to a discussion of the conceptual and empirical problems associated with the assessment of the level of competition in modern industries. Finally, the paper examines the last step in merger policy, the link between changes in competition levels and economic welfare and argues that from a dynamic perspective it is not at all clear that less static competition leads to lower levels of welfare. This undermines the key relationship behind the well-known substantial lessening of competition test that has come to dominate merger policy practice.

The paper briefly reviews the trend towards the inclusion of unilateral effects analysis in merger policy, with a focus on recent changes in the European Community, in Section II. The paper proceeds to an examination of the key shortcomings of this approach (Section III) and how the difficulties are compounded when the measure and welfare implications of both the level of competition and its rate of change are assessed (Sections IV and V). Section VI concludes with a few remarks on the implications of this analysis in the design and implementation of merger policy.

II. Europe Follows the United States: The Adoption of a New Competition Standard for Mergers

As recently highlighted by Vickers in his 2003 paper, competition policy is haunted by the meaning of words.[1] The debate on the reform of the EC merger regulation and its substantive competition test provides a vivid example of this problem. Much of the controversy revolved around the scope of the dominance concept and the extent to which it includes some post-merger, oligopolistic situations where competition may be harmed despite the absence of collusive intent (what has come to be known as unilateral effects). These situations, it was argued, are embraced by the alternative concept, the substantial lessening of competition (SLC) test.

1 J. Vickers, *How to reform the EC merger test?, in* EC Merger Control: A Major Reform in Progress (G. Drauz & M. Reynolds eds., 2003) (based on a speech given at the 2002 EC/IBA merger control conference, Brussels, Nov. 2002).

For some observers, both tests had in practice led to substantially convergent outcomes, with broadly similar assessments of competitive situations at both sides of the Atlantic. Moreover, by the time of the EC merger reform, the European Court of Justice had not yet ruled on potential damages to competition arising from non-cooperative behavior, and therefore the jurisprudence did not exclude that this possibility could be part of the conventional dominance concept. Nevertheless, those in favor of an adaptation of the test were able to convince the legislator of the need to move towards a broader framework, to ensure that no important cases were left out of the regulation.

At the end of the day, however, the final wording of the revised *EC Merger Regulation*[2] is barely different from the old version, moving from preventing "a concentration which creates or strengthens a dominant position as a result of which effective competition ... would be significantly impeded" to "a concentration which would significantly impede effective competition ... in particular as a result of the creation or strengthening of a dominant position." The almost unchanged text reflects the need to provide continuity and consistency with the interpretation of the previous Regulation provided by EC court decisions, but it means that the intended extension of the concept was almost completely excluded from the articles of the Regulation, and left for a detailed explanation in the recitals and the European Commission's *Horizontal Merger Guidelines*.

Recitals 24 to 26 of the regulation explicitly argue that:

> "under certain circumstances, concentrations involving the elimination of important competitive constraints that the merging parties had exerted upon each other, as well as a reduction of competitive pressure on the remaining competitors, may, even in the absence of a likelihood of coordination between the members of the oligopoly, result in a significant impediment to effective competition."[3]

The *Merger Guidelines*, in a section entitled "Non-coordinated effects", provide a detailed list of factors "that may influence whether significant non-coordinated effects are likely to result from a merger."[4] Crucially, the list includes the

2 Council Regulation (EC) No. 139/2004 of Jan. 20, 2004 on the control of concentrations between undertakings (the EC Merger Regulation), 2004 O.J. (L 24) 1 (Jan. 29, 2004).

3 Council Notice on Guidelines on the assessment of horizontal mergers under the Council Regulation on the control of concentrations between undertakings, 2004 O.J. (C 31) 5 (Feb. 5, 2004).

4 *Id.* at paras. 24-38.

degree of substitutability between the products of the merging firms, a key parameter if we wish to assess the presence of unilateral effects for a differentiated products market; and the availability of excess capacity to non-merging firms, the corresponding crucial aspect for the presence of unilateral effects in the homogeneous product case.

The final outcome of the debate reflected the views of influential academics and policy makers. For example, John Vickers forcefully argued in favor of the SLC test, and so did John Fingleton, among others.[5]

The 2003 *Interim Report* by Marc Ivaldi, Bruno Jullien, Patrick Rey, Paul Seabright, and Jean Tirole (hereinafter IDEI report) provides an attempt to justify technically the difference between dominance and a substantial lessening of competition.[6] The authors argue that the dominance test does not encompass all the range of anticompetitive outcomes. For example, using a standard Nash equilibrium concept for homogeneous goods competition, a merger of two firms in a five-firm industry with constant and equal marginal costs would lead to an increase in the markup of 5.25 percent. Similarly, in a similar industry, with three firms, one with a 60 percent share and two equal firms with market shares of 20 percent, the merger of the smaller firms raises prices by 6.75 percent. The new short-run equilibrium involves therefore higher prices achieved through unilateral effects. However, as the authors point out, the first example must involve some gains to be achieved through the reduction of fixed costs, and the second leads to a market structure that need not be less competitive in repeated interactions, since the new rival may well end up being a more viable competitor.

THE REVISION OF THE *EC MERGER REGULATION* UNDERTAKEN RECENTLY BY THE EUROPEAN COMMISSION HAS LED TO AN ENLARGEMENT OF THE RANGE OF ANTICOMPETITIVE EFFECTS THAT MAY BE CONSIDERED IN EC MERGER CASES, ADDING EXPLICITLY THE UNILATERAL EFFECTS THAT MAY ARISE AS A RESULT OF NON-COOPERATIVE BEHAVIOR.

In conclusion, the revision of the *EC Merger Regulation* undertaken recently by the European Commission has led to an enlargement of the range of anticompetitive effects that may be considered in EC merger cases, adding explicitly the unilateral effects that may arise as a result of non-cooperative behavior. Some ana-

5 Vickers (2003), *supra* note 1 and J. Fingleton, *Does Collective Dominance Provide Suitable Housing for All Anticompetitive Oligopolistic Mergers?*, *in* INTERNATIONAL ANTITRUST LAW & POLICY: FORDHAM CORPORATE LAW, ch. 12 (B. Hawk ed., 2006).

6 M. Ivaldi et al., The Economics of Unilateral Effects, Interim Report for DG Competition, European Commission (Institut d'Economie Industrielle, Univ. Toulouse, Nov. 2003) [hereinafter IDEI Report], *available at* http://idei.fr/doc/wp/2003/economics_unilaterals.pdf.

lysts think that this new approach will lead to excessive EC intervention.[7] Others do not believe that is the case.[8] In what follows I intend to explore the issue further, by assessing the conceptual and practical robustness of unilateral effects analysis.

III. The Popularity of Unilateral Effects

The growing use of unilateral effects arguments and models in merger cases is mostly due to their increased popularity among academic economists. Unilateral effects models appear to be grounded solidly in economic theory and, to a certain degree, they offer a range of fairly consistent theoretical results. This is in sharp contrast with the alternative coordinated effects stories in which mergers are forbidden because of the potential increase in the likelihood of collusion of the remaining market players. The theory of collusion is perceived as less definitive, with a variety of possible equilibria, and the more or less informal establishment of lists of conditions that lead to the potential anticompetitive behavior. Academic economists feel more at ease with the unilateral effects theory and, arguably, this is what led to its dominance in the United States[9] This has, as in many other policy fields, been exported to the European Community.

Note, however, that Judge Posner, in the second edition of his well-known book on antitrust law, does not even mention unilateral effects.[10] Nevertheless, the distinction between unilateral and coordinated effects appears already in a 1991 paper by Robert Willig in the run-up to the revised U.S. *Horizontal Merger Guidelines*.[11] It has been fully articulated at the textbook level by Massimo Motta and inspired the new EC policy detailed in the IDEI report.[12] The fourth edition of Kwoka and White's casebook also highlights the increased role of unilateral effects.[13] However, the debate goes on. As recently as 2006, Dan Rubinfeld has

7 *See, e.g.*, D. Ridyard, The Commission's New Horizontal Merger Guidelines: An Economic Commentary (Global Competition Law Centre, Working Paper 02/05, Feb. 2005).

8 J. Vickers, *How does the prohibition of abuse of dominance fit with the rest of competition policy?*, *in* European Competition Law Annual 2003 (C.-D. Ehlermann & I. Atanasiu eds., 2006) (based on a speech given at the 8th Annual EU Competition Law and Policy Workshop at the European University Institute, Florence, Jun. 2003).

9 *See* J. Baker, *Why did the Antitrust Agencies embrace unilateral effects?*, 12 Geo. Mason L. Rev 31 (2003) and J. Baker, *The Case for Antitrust Enforcement*, 17 J. Econ. Persp. 27 (2003).

10 *See* R. Posner, Antitrust Law (2nd ed. 2001).

11 R. Willig, *Merger Analysis, Industrial Organization Theory, and Merger Guidelines*, 1991 Brookings Papers on Econ. Activity: Microeconomics 281 (1991).

12 M. Motta, Competition Policy: Theory & Practice (2004).

13 J.E. Kwoka & L.J. White, The Antitrust Revolution: Economics, Competition, & Policy (4th ed. 2003).

argued that unilateral effect theory is less conclusive than coordinated effects, theoretically debatable, and with less case experience to build on.[14]

The examination of unilateral effects theory as a conceptual and practical basis for merger analysis should start from the key observation that merger policy—as opposed to other areas of antitrust—is not about assessing behavior based on observed facts. It is about anticipating behavior, and this means that the standard of proof—the degree of confidence that is required in order to make a finding—may have to be stronger than in other areas of competition policy. This fact has implications for both the quality of the theory used and for the soundness of its empirical application. It also has consequences in terms of the design of the process by which mergers are approved (with or without conditions) or forbidden.

In the process of merger assessment, authorities may decide to minimize either type I errors (blocking efficient mergers) or type II errors (allowing anticompetitive mergers). If the key concern is to minimize type I errors, the process should be designed so that all mergers are allowed in principle, and specific deals are contested when the authorities can show with a high degree of confidence that the merger is anticompetitive.

If the goal is to minimize type II errors, the ideal approach would be to block all mergers in principle, and allow specific operations only if it could be shown convincingly (in this case by the parties involved) that these operations are procompetitive.

In principle the procedure used by the European Community correctly focuses on minimizing type I errors, with the correct presumption that in a market economy companies will try to increase their size by mergers and acquisitions, with the goal of improving their efficiency. Of course, the minimization of type I errors will crucially depend on how convincingly the potential anticompetitive effects have to be shown. Until recently, no efficiency considerations could be claimed, and this tended to increase the probability of blocking good mergers. Similarly, the standard of proof was not very high. Only recently the European Court of First Instance has made it explicit that the anticompetitive effects have to be very likely (have to occur "in all likelihood" is the phrasing used by the Court).

For our purposes, however, the key issue is whether unilateral effects theory, and its use in practice, provides a sound basis for the analysis of the presence of anticompetitive effects. That is, with unilateral effects models, do we increase the probability of correctly assessing the existence and importance of the anticompetitive effects of a merger?

14 *See Antitrust Modernization Commission Hearings* (2005 to 2006) (statement of D. Rubinfeld, Prof. L. & Econ., Univ. Cal. Berkeley, Jan. 19, 2006), *available at* http://www.amc.gov/commission_hearings/pdf/rubinfeld_statement_final.pdf.

Of course, to the extent that we conclude—as discussed in the coming sections—that unilateral effects theory misleadingly categorizes as anticompetitive economic situations involving effective competition, it is clear that unilateral effects will increase the probability of type I errors.[15]

But even if we were to accept unilateral effects theory, the key issue is whether its insufficient robustness, both conceptual and in its practical application, will diminish our chances of correctly assessing prospective mergers. Is it possible with unilateral effects theory and practice to achieve a standard of proof as high as needed in merger analysis? Is it higher than the one achieved with coordinated effects theory?

From the perspective of the theory, the economic model on which authorities base decisions should be particularly robust. That is, it should be valid under alternative circumstances, even if analytical consistency and formal rigor diminish in importance. It is unclear that this requirement is satisfied by unilateral effects theory, and it may very well be better accomplished by the old coordinated effects analysis. Even if it is hard to show formally that many mergers increase the likelihood of cooperative behavior, this is an intuitive result under a variety of well-known scenarios and there is not a lot of controversy on the set of observables that have to be present to make the case convincingly. On the contrary, in unilateral effects theory one can show with several simple comparative static exercises that unilateral effects may occur, but their generality and magnitude is uncertain, particularly when we assess competition in dynamic industries where the conventional oligopoly model is less well developed.

Indeed, it should be emphasized that the generality of the results underpinning the theory behind unilateral effects is unclear. As developed in the IDEI report and also by Werden and Froeb, the results are basically tied to the static oligopoly model, exploiting the relationships that this model yields in terms of the relation between firm-level and industry-level markups with measures of perceived elasticity of demand.[16] The literature does analyze the implications in terms of entry and a more dynamic examination of the market (relationships such as those highlighted in Section V below), but this is left as a complement, and the corresponding results are rather less conclusive than the ones that are key to the implementation of the unilateral effects approach.

The application of unilateral effects theory, based on formal oligopoly models, faces an additional hurdle. It does not fit adequately with the existing U.S. *Merger Guidelines* or its equivalent in the European Community. This is so

15 Fingleton argues, of course, exactly the opposite: that the exclusion of the unilateral effects from the dominance test increased the likelihood of type II errors, which to him should be of particular concern to authorities. *See* Fingleton, *supra* note 5.

16 *See* G. Werden & L. Froeb, *Unilateral Competitive Effects of Horizontal Mergers*, *in* Handbook of Antitrust Economics (P. Buccirossi ed., forthcoming 2007).

because these *Guidelines* are precisely the result of an indirect use of economic models as the background for merger analysis. Economic theory and its empirical application allow today a more direct assessment of the parameters of interest, circumventing the *Guidelines*. But on the other hand, the formal framework has also its shortcomings and, as argued before, it may be worthwhile to retain some of the flexibility of the proxy analysis used in the *Guidelines*.

It is certainly the case that, for example, market definition is a conceptual shortcut. It was designed years ago, as an intermediate step, so that antitrust authorities could compute measures of market structure and its changes, and use that as a proxy of the changes in the degree of competition. The attempt to match the *Guidelines* with oligopoly theory is fraught with difficulties. The *Guidelines* focus the definition of the market on the assessment of own and cross-price elasticities, and leave the assessment of competitive reactions and entry for additional stages after the market has been defined. However, oligopoly theory highlights the need to be explicit about the reaction of rivals in determining the extent to which a would-be monopolist would be able to increase prices significantly for a certain amount of time. Thus, modern (static) oligopoly theory would assess the extent of the market by explicitly considering the competitive reaction of rivals, while the *Guidelines* leave that reaction for consideration after the market has been defined, in the context of competitive reactions and (non-sunk costs) entry. In fact, modern oligopoly theory, as highlighted repeatedly by authors such as Tim Bresnahan and Judge Posner, makes the assessment of the size of the market irrelevant, to the extent that the increased possibility of raising prices after the merger is completely captured by the changes in the elasticity of the residual demand curves.[17]

From an empirical point of view, the paper by Greg Werden, Luke Froeb, and David Scheffman provides a comprehensive analysis of the conditions that unilateral effects theory should satisfy for its use in merger analysis in practice.[18] One of the key conditions is that it should be shown that in the past the theory used has been applicable to the industry under examination and that, for its specific use in a case, it fits the facts to a reasonable degree.

This requirement is related to the broader discussion of how reliable the models developed by modern economic analysis are. Following a long tradition in modern economic methodology, these models are based on deductive introspective reasoning, and not on asking the actors. Their assumptions need not be real-

17 *See* T. Bresnahan, Comments on "Reforming European Merger Review: Targeting Problem Areas in Policy Outcomes" by Kai-Uwe Kühn (Nov. 2002) (mimeo, Stanford Univ.), *available at* http://www.stanford.edu/~tbres/research/Reforming%20European%20Merger%20Review.pdf and Posner, *supra* note 10.

18 G. Werden et al., *A Daubert Discipline for Merger Simulation*, Antitrust Mag. (Summer 2004). *See also* D. Neven, *Competition Economics and Antitrust in Europe*, 48 Econ. Pol'y 741, 766 (2006).

istic, provided that they offer a parsimonious and reasonably accurate ex post explanation of the facts. Can this type of model be used reliably for the assessment of future situations? Will the estimated parameters remain stable after structural change? Is the behavior of companies predictable and as hypothesized in the model? It seems to me that given the importance of the decisions to be taken, not only should we require that the model explains well the facts of the past, but we should also request that it explains well the behavior of companies after mergers of the past. This is crucial. As George Akerlof has recently pointed out, modern economic theory has a built-in bias against alternative theories, with very low power of statistical tests and a low probability of rejecting the null when false.[19] Thus, "in almost every instance a large number of parsimonious models can be fitted statistically, making it hard—if not all but impossible—to statistically reject all variants of the model."[20] One wonders whether this is a framework that provides useful guidance for hypothetical scenario analysis such as the one needed in merger policy.

GIVEN THE IMPORTANCE OF THE DECISIONS TO BE TAKEN, NOT ONLY SHOULD WE REQUIRE THAT THE MODEL EXPLAINS WELL THE FACTS OF THE PAST, BUT WE SHOULD ALSO REQUEST THAT IT EXPLAINS WELL THE BEHAVIOR OF COMPANIES AFTER MERGERS OF THE PAST.

An example of the wealth of models available, and how easily they fit the data, is provided by the MCI Worldcom and Sprint merger, where the parties involved presented dramatically different empirical elasticity estimates using structural oligopoly models.[21] As it is well-known the merger was abandoned due to the opposition of antitrust authorities, but to a certain degree the events post-merger vindicate the arguments used by the companies in their defense of the deal. The focus of the discussion was the long distance market. The applicants argued that this was an industry in which margins were quickly collapsing and in the midst of a structural change provoked by technological breakthroughs (the Internet) and regulatory changes (unbundling of local networks, etc). The long distance market is a market where the extent of product differentiation is limited, and in which the key competitive features are the high investment costs involved; and the risk of competitive entry by new technologies (email, chat through the internet, webphones, etc), powerful companies (the "Baby Bell" operating companies which could provide jointly long distance and local calling), and new competitive providers with brand new fiber. History has shown the importance of all these features. New entrants with new fiber have failed, but the

19 G. Akerlof, *The Missing Motivation in Macroeconomics*, AM. ECON. REV. (forthcoming 2007), working paper (Oct. 2005), *available at* http://www.econ.yale.edu/~shiller/behmacro/2005-11/akerlof.pdf.

20 *Id.* at 46.

21 These authors review other cases (*Volvo/Scania*) of substantial technical discrepancies between the parties. For a description of the case, *see* KWOKA & WHITE, *supra* note 13, at ch. 4. For a detailed discussion of the technical discrepancies between the two parties involved, *see* Werden & Froeb, *supra* note 16.

new assets are there, the new services compete aggressively with traditional long distance, and the local phone companies have indeed made substantial inroads into long distance by bundling their offers together with the supply of other services such as Internet access. Despite all this, most of the disagreement focused on the alternative estimates obtained by the competing parties with regards to the extent of differentiation between the competitors involved. Whether or not MCI/Worldcom and Sprint were close substitutes or not, is certainly key to the extent to which a standard oligopoly model generates substantial price increases when simulating a merger, since the price increases depend directly on the value of the estimated cross-price elasticity. However, it is less than clear that this is the appropriate framework to assess the competitive implications of the merger at a moment of structural change in the industry. Similarly, much of the discussion focused on the implied markups resulting from the estimated elasticities, thus neglecting the fact that this is an industry where scale economies are substantial and markups have to be assessed together with the fixed costs.

Ultimately, the choice of the economic models behind merger policy should be very careful, since there is already substantial debate on the efficacy of antitrust laws and merger policy is a particularly sensitive area. It is very hard to evaluate ex post merger policies and there is considerable disagreement as to their effectiveness.[22] This means that policymakers should be especially careful in this area. Judge Posner says "it is hard enough to prove collusion; it is even harder to prove that a proposed merger will create a dangerous probability of future collusion,"[23] and we may add that it is even harder to show that after the merger prices will unilaterally increase.

IV. What Is the Appropriate Level of Competition?

The assessment of mergers involves the forecast of the competitive situation that will prevail after the merger. This is, as we have seen, a very complex exercise.

22 In 1999, the U.K. Office of Fair Trading (OFT) commissioned a study by National Economic Research Associates in which twelve cases of cleared mergers where examined. In the report, only two of the cases were found wrongly cleared and turned out to be anticompetitive. *See* NAT'L ECON. RES. ASSOCIATES, MERGER APPRAISAL IN OLIGOPOLISTIC MARKETS (prepared for the U.K. Office of Fair Trading, Research Paper 19, Nov. 1999). Apparently the authorities overestimated the power of buyers, the degree of substitutability, and the extent of technical change. *See also* Baker, *supra* note 9 and R. Crandall & C. Winston, *Activist Antitrust?*, 17 J. ECON. PERSP. 1, 15-20 (2003).

In addition, a paper by Tomaso Duso, Damien Neven, and Lars-Hendrik Roeller examined the effectiveness of EC policy so far. According to their work, in the period between 1990 and 2002 the Commission incurred in 23 percent type I errors and 28 percent type II errors. *See* T. DUSO ET AL., THE POLITICAL ECONOMY OF EUROPEAN MERGER CONTROL: EVIDENCE USING STOCK MARKET DATA (WZB CIC, Working Paper FS IV 02-34, Apr. 2003), *available at* http://skylla.wz-berlin.de/pdf/2002/iv02-34.pdf.

23 POSNER, *supra* note 10, at 119

Moreover, merger analysis also implies that this predicted level of competition (or lack of competition) should be assessed in terms of its impact on some measure of consumer or overall welfare.

It is important to emphasize, first, that when the anticompetitive effects of a merger are predicated on the existence of unilateral effects, the future post-merger scenario is one where non-coordinated or non-cooperative behavior prevails. This poses the question: To what extent should a non-cooperative equilibrium be considered uncompetitive? Already many years ago, Friedrich Hayek argued against this view.[24] Quoting Dr. Johnson, he highlighted the etymological meaning of competition: "the action of endeavouring to gain what another endeavours to gain at the same time." From a modern game-theoretical perspective, such a definition fits quite well with what we today qualify as non-cooperative, profit-maximizing behavior. As Hayek already recognized, such behavior may lead—due to the structural characteristics of the market—to a markup of prices over marginal costs, but this need not imply that the market should be qualified as uncompetitive. Hayek focused at the time on product differentiation as the source of a positive competitive markup (the modern monopolistic competition model), and more recent analysis has developed new sources of competitive markups (contestable markets for the case of fixed costs that are not sunk and sunk-costs competition otherwise)[25] which make it difficult to assess the proper level of competition by a simple reference to relations between price and marginal cost.

The exact meaning of non-cooperative behavior and its relation with the presence or absence of competition is exemplified by the discussion in the IDEI report. These authors distinguish carefully between unilateral effects, where there is "passive adaptation to market conditions," from tacit collusion where we find "anticipation of a response to one's own action," and behavior that "would not be in our own interest where it not for that anticipated reaction."[26] It is at least questionable whether passive adaptation to market conditions should be viewed as non-competitive behavior.

What this implies is that a priori there is no reason to expect that non-cooperative behavior leads to insufficient competition, and that in practice substantial effort should be devoted to the analysis of the broad characteristics of the equilibrium (or business environment) post-merger, in order to carefully characterize and measure the extent of competition. Clearly, the conventional analysis through structural measures such as market shares and concentration indices pro-

24 F. Hayek, Individualism & Economic Order, ch. V: The Meaning of Competition (1948).

25 *See* J. Sutton, *Market Structure: Theory and Evidence*, *in* Handbook of Industrial Organization (M. Armstrong & R. Porter eds., vol. III, forthcoming 2007).

26 IDEI report, *supra* note 6, at 3, 4.

vides a very poor approximation to actual competitive conditions. However, as argued above, even a direct measurement of the price to marginal cost markup would be incorrect when the conditions of competition involved fixed costs (sunk or not). This of course extends also to the more sophisticated analysis of residual demand, which tries to assess directly through econometric analysis a measure of the extent to which other companies will restrain the pricing of the companies involved in the merger. Such an analysis is theoretically sound in the very limited number of markets where firms do not incur fixed costs, and it is particularly inappropriate in markets where competition takes place through the escalation of fixed sunk costs. Indeed, in dynamic industries characterized by sunk-costs competition (or by network effects), the level of competition is incorrectly assessed by looking at price-cost markups. Firms gain competitive advantage by investing in advertising and research and development (R&D), or by developing (direct and indirect) networks, and the excess pricing over marginal costs need not reflect economic inefficiencies, but rather the complex set of complementary services sold by these firms and the need to obtain sufficient margins to pay for the fixed costs and achieve adequate profitability.[27]

In such a context, the discussion about the assessment of competition should focus on an analysis of the presence of (ex ante) excess profitability. That is, profits in excess of the competitive rate of return, controlling for the risk involved in each type of activity. Profitability analysis is a controversial field in competition policy. It is true that accounting data on profits is plagued with difficulties. However, some authors argue that the problem is no larger than with other data typically used in antitrust proceedings, and this is an area that probably should be more prominent in antitrust analysis when industries with these characteristics are involved.[28] Financial theory has developed good instruments for the measurement of excess returns. They involve the comparison of internal rates of return with the cost of capital adjusted for risk. In practice, the measurement of profitability must refer to short periods of time and requires the assessment of initial and terminal asset values, an exercise that—not surprisingly—turns out to be crucial for the proper evaluation of profitability in industries characterized by sunk-costs competition.

The assessment of what is an appropriate level of competition also has implications for market structure from a dynamic perspective. In industries with sunk costs or network effects, firms may anticipate further concentration, as companies try to sink more costs in order to improve their positioning in the marketplace or exploit internally the positive externalities of networks. That is to say, the merger may have as an objective the achievement of a market position that

27 On these issues, *see, e.g.*, J. Gual, *Market Definition in the Telecoms Industry*, in THE ECONOMICS OF ANTITRUST & REGULATION IN TELECOMMUNICATIONS (P. Buigues & P. Rey eds., 2004).

28 *See* OXERA, ASSESSING PROFITABILITY IN COMPETITION POLICY ANALYSIS (prepared for the U.K. Office of Fair Trading, Economic Discussion Paper #6, 2003).

is large enough to finance the increased sunk costs and at the same time anticipate the competitive move of rival firms. It is not at all clear that this type of deal should be considered anticompetitive.

A related argument has been made recently by Robin Mason and Helen Weeds, albeit with a very stylized model.[29] They argue that optimal merger policy should take into consideration that preventing some mergers (ex post) may lead to insufficient entry (ex ante). In their model, the entrant, given the uncertainty and the sunk costs it faces, will only enter if it can anticipate the merger as a potential way out, if profitability turns out to be insufficient. These authors consider that the effect they highlight is more general, applying to any decision by companies that is difficult to reverse, has uncertain returns, and is affected by the possibility of a merger.[30]

From my perspective, the key point is that the merger may be a way to relax static competition, but need not imply a softening of dynamic competition. That is to say, at any point in time, the company may find that previous entry or investment decisions have not been correct and, therefore, that it is insufficiently profitable. Mergers are a way to restore profitability, but not necessarily above the competitive rate of return given the risks involved in sunk costs competition.

THE FOCUS OF COMPETITION POLICY SHOULD THEREFORE MOVE AWAY FROM THE COMPARATIVE STATIC ANALYSIS OF THE EFFECTS OF CHANGES IN MARKET STRUCTURE TOWARDS A THOROUGH EXAMINATION WHICH FOCUSES ON THE DYNAMIC FEATURES OF INDUSTRIES.

The focus of competition policy should therefore move away from the comparative static analysis of the effects of changes in market structure (i.e., analysis such as how do markups change?) towards a thorough examination which focuses on the dynamic features of industries. This means an analysis of the sustainability of the new market structure, and leads to an examination of the extent to which a merger significantly increases or decreases the barriers to entry into the industry.

On the issue of barriers to entry, it is well-known that in general they should be considered anticompetitive when they are artificial, but need not be so when they are the result of legitimate innovation and internal growth of a company and do not involve the anticompetitive exclusion of rivals. These are the so-called strategic barriers to entry that form part of a dynamic competitive landscape.

Of course, a merger is not a case of internal growth and, from this perspective, the increase in strategic barriers is only legitimate to the extent that it is the

29 R. Mason & H. Weeds, Merger Policy, Entry and Entrepreneurship (Jul. 2006) (mimeo, Univ. Essex), *available at* http://privatewww.essex.ac.uk/~hfweeds/ffd19.pdf.

30 *Id.* at 3.

result of an unavoidable change in market structure due to insufficient profitability. In this sense, it may be seen as an instance of the failing firm defense that, within a proper analysis of profitability as mentioned above, could well be known as the insufficiently profitable firm defense.

V. Is Less Competition Always Bad?

The final stage in merger analysis is the examination of the consequences of changes in the level of competition for some measure of aggregate welfare. In simple static models of oligopolistic competition, the relationship between the degree of competition and welfare is well-known to be negative, but that need not be necessarily the case in dynamic settings, since the consequences for welfare are crucially dependent on the extent to which firms engage in product and/or process innovation. If the relationship between competition and the measure that assesses welfare is not linear, it cannot be taken for granted that a merger that substantially lessens competition will be detrimental to welfare.

A direct and simple link between competition and a dynamic assessment of welfare, as proxied by innovation, is well developed in the literature. For example, Damien Neven refers to the studies of Stephen Nickell and others, and concludes that competition matters for economic efficiency and in particular for productive efficiency and the incentives to innovate.[31] However, the economics profession is far from reaching a consensus in this area. Theoretical work by Xavier Vives provides arguments for a nonlinear relation between competition and measures of innovation activity, and so does the work of Aghion and Griffith, which also provides an empirical assessment of what they argue is an inverted U-shaped relationship.[32] In fact, Aghion and Griffith start their argument by quoting Nickell where he disarmingly asserts that "the general belief in the efficacy of competition exists despite the fact that it is not supported either by any strong theoretical foundations or by a large corpus of hard empirical evidence in its favour."[33]

The positive impact of competition on the pace of innovation (the upward sloping part of the inverted U curve) corresponds to what Aghion and Griffith term as the "escape the competition effect". From the perspective of the innovation race models that they use, it is a fairly general result that increasing the

31 Neven, *supra* note 18, at 744 and S. Nickell, *Competition and corporate performance*, 104 J. Pol. Econ. 724-46 (1996). For a review of empirical studies, *see, e.g.,* S. Ahn, Firm Dynamics and Productivity Growth: A Review of Micro Evidence from OECD Countries (OECD, Economics Department working paper 297, Jun. 2001).

32 X. Vives, Innovation and competitive pressure (Working Paper #634, Jun. 2006), *available at* http://www.iese.edu/research/pdfs/D1-0634-E.pdf and P. Aghion & R. Griffith, Competition & Growth: Reconciling Theory and Evidence (2005).

33 Aghion & Griffith, *id.* at 32.

number of players (a measure of heightened competitive pressure) leads to renewed innovation efforts, as each player tries to stay ahead of its competitors in order to reap the benefits of success, escaping from the competition through the innovation race.

Vives obtains a similar effect using fairly general oligopoly models that are firmly based on the game-theoretical oligopoly model tradition but nevertheless incorporate the dynamic efficiency gains that may be obtained through product and process innovation. In his work, again measuring the extent of competition by the presence of more competitors, the impact of more competition on R&D effort is positive to the extent that (in the presence of bankruptcy costs) the larger number of competitors diminishes expected profits, increases the chances of bankruptcy and, therefore, triggers a higher level of ex ante R&D effort to improve efficiency.

At the same time, however, both approaches find that beyond a certain threshold competition may in fact deter innovation and, as a consequence, thwart dynamic efficiency. In innovation race models, such as those encompassed by the general framework used by Aghion and Griffith, this is a Schumpeterian effect whereby in certain contexts increased rivalry ex post (or the prospect of such a level of rivalry) diminishes incentives to innovate because the possibility of appropriating innovation rents is diminished. Indeed, an antitrust policy that promotes competition but erroneously prevents large companies from developing legitimate commercial strategies may be quite counterproductive in terms of its effect on R&D and innovation.

The general class of oligopoly models discussed by Vives also generates a negative relation between increased competition and rates of innovation to the extent that the presence of more competitors diminishes the demand faced by each firm and the expected rewards from innovation.

THE ASSESSMENT OF THE CONSEQUENCES OF MERGERS SHOULD REMAIN AT THE LEVEL OF THE EXAMINATION OF THEIR EFFECTS IN TERMS OF THE VARIABLES THAT PROVIDE THE MORE STRAIGHTFORWARD EVALUATION OF THE ABSENCE OF COMPETITION.

Overall, it is apparent that once we consider competition in a dynamic setting, taking into account the crucial effects of product and process innovation on welfare, the link between increased rivalry (understood as static competition) and welfare becomes less clear-cut than is commonly assumed. If we cannot simply conclude that less competition is bad, and if this less competition comes about without collusion, explicit or tacit, then this implies from a policy perspective that the assessment of the consequences of mergers should remain at the level of the examination of their effects in terms of the variables that provide the more straightforward evaluation of the absence of competition. As argued before, these are the presence of excess profitability and the reinforcement of artificial barriers to entry.

VI. Concluding Remarks

This essay has presented a broad critical analysis of what constitutes the mainstream approach to merger policy in western economies. The paper argues, first, that unilateral effects—the fashionable new approach behind merger policy—need not in fact imply anticompetitive behavior and, in any case, are very difficult to measure and use reliably in practice. Moreover, the paper stresses that even if these effects were to lead to an increase in the conventional measures of anticompetitive performance (such as markups), it is not clear that this implies less competitive behavior for many of the most relevant industries in today's advanced economies. Finally, the paper also examines the relation between competition and welfare, and argues that even if competition does indeed diminish due to a merger, it is not a straightforward conclusion that this is not good in terms of economic welfare when we take fully into consideration the incentives to innovate and the dynamic welfare gains that arise from new products and production processes.

These three lines of criticisms of current merger policy do not dispute the fundamental idea that economic analysis should be the basis of proper merger examination. Rather, what they imply is that economic models are unavoidable abstractions of the real world that have to be handled with extreme care when used in important policy matters such as merger decisions. Despite the tremendous progress of industrial organization theory over recent decades and the phenomenal improvement in quantitative methods, the range of uncertainty regarding the appropriate model of competition for real-life industries is huge, and merger policy should be deployed with a broad portfolio of analytical tools.

The more so because, as recently discussed by George Akerlof,[34] modern positive economics is biased both theoretically and empirically against models of behavior that, despite being potentially relevant in practice, incorporate non-objective arguments in utility functions and pay attention not only to what decisionmakers do, but also to why they say they do it. This methodological bias excludes a non-trivial set of potentially powerful explanations of behavior, favoring abstract and parsimonious models over frameworks that may be less complete but are derived from the "knowledge of human nature and from the detailed facts of experience"[35] and may be more insightful or appropriate. As a consequence, and given the complexity of real-life merger cases, it may be advisable to design merger policy in such a way that a broad range of economic analysis and evidence is collected, and the improvements in detailed techniques developed by economic theory and econometrics, should be carefully complemented by case-specific analysis, and a careful assessment of industry trends. ◀

34 Akerlof, *supra* note 19.

35 *Id.* (Akerlof *quoting* John Maynard Keynes).

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Holding Innovation to an Antitrust Standard

Richard Gilbert

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1.
Competition Policy International is a free publication. To order or download additional copies, visit eSapience.org.

Holding Innovation to an Antitrust Standard

Richard Gilbert

Several antitrust cases have involved allegations of anticompetitive innovation or product design and some plaintiffs and antitrust scholars have argued that investment in research and development that excludes competition can have predatory effects similar to predatory pricing. This article analyzes several tests for predatory innovation, including the rule of reason based on total and consumer welfare and profit sacrifice tests. All of these tests are likely to produce false positives that chill incentives for beneficial investments in research and development. Most courts that have considered allegations of anticompetitive innovation, including the appellate court in *U.S. v. Microsoft*, have concluded that innovation is not anticompetitive if it has plausible efficiencies. This is close to a test of whether innovation is a sham. While a sham test may fail to identify innovations that harm competition, that risk is acceptable given the high cost of penalizing beneficial innovation.

The author is Professor of Economics at the University of California at Berkeley. He is grateful to Jonathan Baker, Matthew Hendrickson, A. Douglas Melamed, Janusz Ordover, James Ratliff, Howard Shelanski, Gregory Werden and Robert Willig for helpful discussions on this topic. Steven Albertson provided excellent research assistance.

I. Introduction

Innovation is the lifeblood of the economy. Firms should be encouraged to invest in research and development (R&D), as studies of the social rate of return to investment in R&D often yield estimates that substantially exceed the private cost of capital.[1] Nevertheless, innovation often disrupts markets, and several antitrust cases have alleged that innovation has harmed competition and, by inference, lowered economic welfare. This paper considers the standards that antitrust policy should apply to evaluate whether innovation contributes to unlawful monopolization. While innovation occurs in many different contexts, the focus in this article is on single firm conduct that creates new products or alters the characteristics of existing products. The conduct may affect markets for products that are substitutes or complements for the products sold by the innovating firm. An example of conduct that affects complements is an interface design that affects the compatibility of complementary components for a computer network. An example of conduct that affects substitutes is a product line extension for a patented pharmaceutical that has consequences for generic competition.

In an idealized world, market performance, including price and quality, would be mapped into an outcome measure, and conduct that lowers this measure would be anticompetitive. Economic welfare is an example of such an outcome measure. Total economic welfare is the sum of producer profits and consumer benefits that result from economic activity, while consumer welfare ignores profits. Whether antitrust policy should be concerned with total economic welfare or only consumer welfare is a subject of considerable controversy, although neither welfare measure correctly captures the objectives of antitrust policy.[2] Firms have wide discretion to choose the prices of their goods and services without running afoul of U.S. antitrust law, despite the fact that at least in the short run an increase in price unambiguously lowers consumer welfare and lowers total economic welfare when price is above marginal production cost. Similarly, if a firm fails to take advantage of an opportunity to create a better product, the result is an increase in the product's quality-adjusted price relative to a baseline in which the innovation occurs. A failure to innovate would lower consumer welfare and

1 Estimates of the average social rate of return to R&D range from 20 to 40 percent per annum and sometimes higher. *See, e.g.*, Zvi Griliches, *R&D and Productivity: Econometric Results and Measurement Issues*, *in* Handbook of Economics of Innovation and Technological Change 53-89 (P. Stoneman ed., 1995); E. Mansfield et al., *Social and Private Rates of Return from Industrial Innovations*, 91(2) Q. J. Econ. 221-40 (1977); and, Bernstein & Nadiri, *Interindustry R&D spillovers, rates of return, and production in high-tech industries,* 78 Am. Econ. Rev. 429 (1988).

2 *Compare* K. Heyer, *Welfare Standards and Merger Analysis: Why Not the Best*, 2(2) Competition Pol'y Int'l 29 (2006) (arguing for a total economic welfare standard) with Steven C. Salop, *Exclusionary conduct, effect on consumers, and the flawed profit-sacrifice standard*, 73(2) Antitrust L.J. 311, 336 (2006) (antitrust law focuses on consumer welfare). Joseph Farrell and Michael Katz conclude that there is a strong case for using total surplus along with other criteria for antitrust enforcement, but observe that a consumer welfare standard can perform better in some circumstances. Joseph Farrell & Michael Katz, *Welfare Standards in Competition Policy*, 2(2) Competition Pol'y Int'l 3, 28 (2006).

would lower total economic welfare if the cost of the innovation were less than the value of the quality improvement. Yet a failure by a firm, acting independently, to take advantage of an innovation opportunity would not violate the antitrust laws.[3]

Although economic welfare does not determine whether conduct is anticompetitive, measures of economic welfare can inform antitrust policy by providing objective estimates of the impact of the conduct on market performance. This article explores the utility of different welfare standards that imply alternative tests for antitrust liability arising from innovation by a single firm, including total economic welfare and consumer welfare, and others, such as profit sacrifice, that are only indirectly related to economic welfare. These standards have been applied with varying success to inform the analysis of predatory pricing. While some suggest an analogy between predatory pricing and predatory innovation, the consequences of innovation and the link between competitive effects and the incentives to invest in R&D are difficult to evaluate with any welfare measure.

Section II develops a simple model to illustrate how alternative antitrust standards may apply to innovation, with a focus on innovation that affects competition for substitutes. The model shows why conventional approaches may give incorrect signals for antitrust enforcement in an innovation context. Section III reviews how courts have responded to evaluations of anticompetitive innovation in industries where new products or changes in existing product characteristics have created incompatibilities with complementary products. Section IV examines the special case of innovation in the patented pharmaceuticals industry. Manufacturers of generic pharmaceuticals have alleged that branded drug manufacturers have harmed competition by patenting modifications to existing drugs. These patented modifications may extend the effective length of exclusivity for a drug and delay generic competition. Allegations of anticompetitive innovation in the pharmaceutical industry differ from most other innovation cases in that the affected products are substitutes, not complements for the products of the innovating firm, patents and agency considerations are important, and market conduct and outcomes are heavily influenced by legislation and regulation.

The risk of enforcement error is high in cases that allege predatory innovation. A welfare test may find that innovation is predatory when it has no anticompetitive effect or may fail to identify innovation that could make consumers worse off. The risk of excessive enforcement is much higher than the risk of too little intervention because most innovation is beneficial and would be chilled by attempts to police the rare cases in which innovation might harm welfare. Noting that antitrust policy is informed by measures of economic welfare, but intended to protect the competitive process, Section V analogizes innovation to

3 The focus in this article is on innovation by a single firm. Coordinated conduct related to innovation can raise additional antitrust concerns. For example, an agreement by competitors not to invest in R&D could be a source of antitrust liability.

other single firm conduct that has antitrust implications. The competitive impacts from a change in interface standards that prevents interoperability of complementary products are no more severe than the effects of a decision not to deal with the suppliers of these products. Given the skepticism expressed by the U.S. Supreme Court in *Verizon v. Trinko* regarding the obligation of a firm to deal with a rival, it is likely that a refusal to deal with no other anticompetitive conduct would escape antitrust liability in most circumstances.[4] A product innovation that has the same effect should not be subject to greater antitrust scrutiny.

I conclude that welfare and the efficient use of judicial resources would be best served by a policy that presumes that innovation is pro-competitive and condemns innovation by a single firm in only the most extraordinary circumstances. I stop short of endorsing a policy of per se legality for innovation by a single firm because innovation may involve other conduct, such as exclusive dealing, that should be subject to careful review. A monopolist should not be able to shield potentially anticompetitive conduct from antitrust scrutiny merely because the arrangement relates to a product innovation. In assessing whether innovation by a single firm, alone or with other conduct, violates antitrust law, courts could apply a rule of reason analysis or a different test that presumes that innovation is not anticompetitive when it has a valid business justification. Under either approach, innovation by a single firm would not have a safe harbor from Section 2 liability, but would be protected by a strong presumption that innovation is beneficial for the economy.

II. A Simple Model of Innovation

I begin with a simple model of innovation for substitute products that highlights the incentives to innovate and the competitive effects that are likely to result from the innovation. The purpose of this simple model is to illustrate how an antitrust analysis of innovation should differ from an analysis of conduct that affects the prices and outputs of existing products. The potentially anticompetitive conduct considered here is a form of predation. The allegation is that innovation by a single firm can harm welfare even if it generates benefits in the short run, just as excessively low prices can harm welfare if they result in exit or significantly impair the ability of rivals to compete and contribute to monopoly pricing in the long run. The point of this exercise is to show that an antitrust standard that isolates socially harmful innovation is extremely difficult to define, even more so than a standard that defines socially harmful pricing.

Consumers are identical in this simple model. Each consumer has a demand for one unit of a product. A product j has quality v_j and that is also the maximum amount that a consumer would pay for the product. The total number of con-

4 Verizon Communications Inc., Petitioner v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398 (decided Jan. 13, 2004) [hereinafter *Trinko*].

sumers is N. Before innovation occurs there is a single product with quality v_0. By spending an amount R, a firm can develop a new product with quality $v_1 > v_0$. The innovation is the product with quality v_1 and the size of the innovation is $v_1 - v_0$. To keep the example very simple, I assume that there are zero production costs for both the old product and the new product.

Ignoring spillover benefits or costs from the innovation that might affect other consumers or firms, and ignoring possible future benefits or costs, the innovation is socially desirable if $N(v_1 - v_0) > R$.[5] Suppose the old product was available at a price P_0 and the new product is available at a price P_1. Assume for now that $v_1 - P_1 > v_0 - P_0 > 0$.[6] These inequalities imply that all consumers purchased the old product before the innovation and switch to the new product when it becomes available. The profit from innovation depends on the prices before and after innovation and on whether the innovator also sold the old product. If a firm is the only seller of the old and the new product, its incentive to innovate is $N(P_1 - P_0)$. If the innovator does not sell the old product and becomes the only seller of the new product, its incentive to innovate is NP_1.

The private incentives to innovate depend on the prices and need not correspond to the social benefit from the innovation. A firm that is the only seller of the old and the new product would profit from the innovation if $N(P_1 - P_0) > R$. If $N(P_1 - P_0) > R > N(v_1 - v_0)$, the innovation would be privately profitable but socially undesirable. That cannot occur if consumers prefer the new product when both are available, if the quality of the old product remains unchanged when the new product becomes available, and if the products' private values to consumers are the same as their social values. Under these assumptions $v_1 - P_1 > v_0 - P_0$ implies that if $N(P_1 - P_0) > R$, then $N(v_1 - v_0) > R$. These are strong assumptions, however. Social values can differ substantially from private values due to large spillover effects,[7] and the quality of the old product could deteriorate if it is no longer in demand. Thus the innovation could be privately profitable but socially undesirable. The opposite would hold if $N(P_1 - P_0) < R < N(v_1 - v_0)$. In this case, innovation would be socially desirable, but not privately profitable. A firm that sells only the new product also can have the wrong signal for innovation. Innovation can be privately profitable but socially undesirable if

5 The left-hand side is the social benefit from the innovation and the right-hand side is its cost. The innovation has net social value if the left-hand side exceeds the right-hand side. The number of users, N, is fixed in this example. This understates the social and private values of an innovation that expands the use of the technology (i.e., increases N.)

6 In this example, a firm that is the only seller of the new technology would set a price equal to its value, but this would not be the case in a more general model with heterogeneous consumers.

7 Bernstein and Nadiri estimate social rates of return from R&D in different industries that range from 16 percent to more than 100 percent in 1981, compared to private rates of return of from 12 to 24 percent. The social benefits include productivity gains in industries other than the industry where the R&D investments occurred. Bernstein & Nadiri, *supra* note 1, at 432-33.

$NP_1 > R > v_1 - v_0$, and innovation can be unprofitable but socially desirable if $NP_1 < R < v_1 - v_0$.

A challenge for any standard applied to innovation is that antitrust analysis is likely to occur after the innovation, but ex post outcomes reveal little about whether the innovation was a good decision ex ante, when the decision was made. If the goal of antitrust policy is to promote socially desirable conduct and deter undesirable conduct, then the conduct should be evaluated based on the information that was available when it occurred. For innovation, this means an ex ante analysis of expected costs and benefits. An innovation investment could generate nothing of value and look unprofitable ex post even if its expected profit was high. Alternatively, a poor investment decision can turn out lucky and generate significant value. An innovation could be unprofitable, yet still generate social benefits for consumers and other firms that the investing firm cannot appropriate. An innovation also can generate private benefits as a stepping stone to other, more profitable discoveries, or because the innovation signals something of value to consumers, which the firm can appropriate in its reputation.[8]

An innovation can be privately profitable but not socially desirable, or socially desirable but not privately profitable. It can be profitable for some firms but not for others, or it can benefit some consumers but disadvantage others.[9] Although there are conditions under which the private incentive for innovation corresponds to the innovation's social value, this is not true in general. The market can offer too little or too much reward compared to an innovation's social value. Private and social incentives are better aligned for changes in price. A reduction in price usually increases consumer welfare and increases economic welfare (in the short run) provided that the price is above marginal production cost. A price below marginal cost is unprofitable in the short run and socially inefficient because the cost of an incremental unit of supply exceeds its value to consumers. Thus it is not unreasonable for antitrust policy to scrutinize pricing below marginal cost in order to exclude competition. For innovation, analogous conduct is an innovation that is unprofitable in the short run and excludes competition. A rule that identifies conduct with these properties as "predatory innovation" likely would lead to false positives and chill socially desirable innovation. Innovation typically involves a sacrifice of short-run profits. Firms have to invest to develop a new interface standard or a new medicine. Really good innovations make old technologies obsolete, and the prospect of developing a new

8 Pittman make a plausible case that IBM invested excessively in the 360/90 computer to signal technical superiority. Russell W. Pittman, *Predatory Investment: U.S. vs. IBM*, 2(4) Int'l J. Indus. Org. 341, 363 (1984).

9 An example is an industry with switching costs and network effects. An innovation can shift the market to a new technology, leaving the installed base of customers stranded. Consumers of digital audio tape were stranded after the introduction of compact disks reduced the supply of music in the digital audio tape format. *See, e.g.*, Joseph Farrell & Garth Saloner, *Installed base and compatibility: innovation, product preannouncement, and predation*, 76 Am. Econ. Rev. 940-55 (1986).

product or process that dramatically alters the competitive landscape drives the incentive to invent. The conditions associated with predatory conduct could exist for innovation, namely a sacrifice of profit in the short run followed by elimination of rivals and higher prices (or lower consumer surplus), even though the innovation has no predatory effect or intent.

THE CONDITIONS ASSOCIATED WITH PREDATORY CONDUCT COULD EXIST FOR INNOVATION, EVEN THOUGH THE INNOVATION HAS NO PREDATORY EFFECT OR INTENT.

I now turn to alternative tests or standards that could be applied to assess whether innovation is anticompetitive.

A. TOTAL ECONOMIC WELFARE STANDARD (TOTAL RULE OF REASON TEST)

A rule of reason (ROR) test based on total economic welfare asks whether innovation increases total economic surplus, equal to the sum of producer profits and consumer benefits. If it does not, it fails the test and may incur antitrust liability. Whether total economic welfare is an appropriate standard for antitrust enforcement is a controversial question. Economists often favor a total welfare standard because resources are allocated efficiently when total economic welfare is maximized, and no individual in the economy can be made better off without making another individual worse off.

If total economic welfare is an appropriate objective for antitrust policy, then it follows that a total ROR test is the correct standard to evaluate conduct, including innovation. But this is just a tautology, and the more serious issue is whether a total economic welfare standard would lead to sensible antitrust enforcement outcomes when applied to innovation by a single firm. A total ROR test would have to consider the impacts of innovation on the innovator and on other firms and consumers in the present and the future, and should also account for the impacts of antitrust enforcement on future incentives to innovate. This is an enormously complex undertaking. It requires an assessment of impacts on all economic agents in the industry where the innovation occurred and also in other industries that may be affected by the innovation. The difficulties associated with identifying and quantifying the impacts of innovation on consumers and firms are so large that a practical application of the total ROR test can lead to a conclusion that innovation fails the test when it has no anticompetitive element or passes the test when the innovation is arguably anticompetitive.

Rule of reason analysis is a complex undertaking whether applied to innovation or to other conduct, but the analysis is far more complicated for innovation because the benefits from innovation are uncertain and difficult to measure and innovation often has spillover benefits for other firms and consumers. Furthermore, in the context of innovation by a single firm, the analysis would take place after society has the benefit of the innovation and the issue would not be whether the innovation has value, but rather whether its value exceeds its

cost including any adverse impacts on competition. An antitrust policy that punished innovation in a specific situation where its benefits are less than its costs would be counterproductive if it deterred investments in the much more common situations where the benefits of innovation exceed its costs.

The simple example provides an illustration of innovation that fails the total ROR test, but is not anticompetitive. The net social value of the innovation is $W = N(v_1 - v_0) - R$. The innovation fails the total ROR test if $W < 0$, which it would for any significant value of R if v_1 is close enough to v_0. Suppose a new entrant makes the innovation and offers it for sale at a price P_1 and all consumers purchase the innovation at that price. The innovation is profitable if $NP_1 - R > 0$. Profits and social value are equal if, but only if, $P_1 = v_1 - v_0$. This might be the case if the old product stays in the market and competes aggressively with the new product. But why would a supplier of the old product stay in the market if it wouldn't get any sales? It is more likely that suppliers of the old product would exit or not invest to maintain the quality of the old product. Then the firm could charge a price higher than $v_1 - v_0$ for the new product if it is costly for a firm to re-enter with the old product or reinvest to improve its quality. In that case the private value of the innovation can exceed its social value.[10]

Innovation in this example fails the total ROR test because the new firm benefits at the expense of the old firm, although there is nothing anticompetitive about the firm entering the industry with a new product. Taking market share from an incumbent is an important stimulus for innovation. According to Steve Jobs, CEO of Apple Computer, "[W]hat's the point of focusing on making the product even better when the only company you can take business away from is yourself?"[11] Without the driving force of winning market share, the amount of innovation in the economy would be lower and consumers could be worse off, particularly after accounting for spillover benefits.

It is easy to underestimate the total social value of an innovation because benefits from new technologies are difficult to forecast and often occur in markets far removed from where the innovation occurred. A hypothetical example is a way to apply a thin film to glass beverage bottles that has application to liquid crystal displays. In the model terminology, the social value of the innovation can be much larger that the value v_1 in the market where the innovation occurs. When innovation has positive spillover benefits for consumers and firms in other industries, its true social value can be much larger than its value in any one industry. If $N(v_1 - v_0)$ only measures part of the social value of an innovation because other spillover benefits are hard to estimate, then it is not necessarily a

10 Another difficulty with a rule of reason standard is that benefits and costs that differ over time would have to be discounted in order to determine whether total net benefits exceed total net costs, however the choice of the discount rate often affects the sign as well as the total value of net benefits, and the appropriate discount rate can be controversial.

11 Interview with Steve Jobs, CEO, Apple, Business Week (Oct. 11, 2004), at 96.

waste of social resources to reward innovation with a payoff that exceeds the measured, but underestimated, social value.

A total ROR test that does not take spillovers fully into account can produce false positives and condemn socially desirable innovation. The total ROR test is also flawed because it can generate false negatives; an innovation can pass the total ROR test, yet be anticompetitive. Consider the following variation on the simple example. Before entry occurs, the incumbent sells the old product at a price $P_0 < v_0$. Consumers earn a total surplus $N(v_0 - P_0) > 0$. A firm enters with a new product for which $W = N(v_1 - v_0) - R > 0$, which passes the total ROR test. The new firm signs up distributors for its product under the condition that they deal exclusively with its product. Firms that offer the old product cannot make any sales; they exit the market or fail to make investments necessary to compete effectively and do not discipline the new firm's price. As a result the new firm charges the monopoly price $P^m = v_1$ and consumers are worse off. This conduct is arguably anticompetitive absent a business justification for the exclusive dealing. Yet it passes the total ROR test for the value of the innovation.

A total ROR test for innovation should account for spillover benefits and costs in the present and the future, is very complex to perform, and requires courts to assess the values of innovations, which they are not in a position to do. A total ROR test can lead to false positives and false negatives and undermine incentives for innovation. Although it is theoretically possible to construct a rule of reason standard for innovation that would condemn only socially harmful innovation, such a rule would not be practical. The benefits from innovation are hard to quantify, but likely to be large, and a ROR analysis could mistakenly assign a predatory label to conduct that has positive net social value.

B. CONSUMER WELFARE STANDARD (CONSUMER RULE OF REASON TEST)

A number of antitrust scholars have argued that antitrust policy is about protecting consumer welfare and therefore conduct should be evaluated using a rule of reason standard that emphasizes consumer rather than total welfare. Innovation would pass a consumer rule of reason test (consumer ROR) only if it does not lower consumer surplus, defined by total consumer benefits less total expenditures.[12] A consumer ROR test obviously can condemn innovation that increases total economic welfare because the consumer ROR test ignores the effects of an innovation on producer profits. Suppose there is a competitive industry with marginal production cost c_0, which is also equal to the market price. A new firm enters the market with a breakthrough technology that enables production at a cost c_1 so low that firms cannot compete using the old technology. The more efficient firm makes the old industry obsolete or greatly reduces its sales. The obso-

12 As in the case of a total economic welfare standard, consumer benefits that differ over time would have to be normalized by applying a discount rate.

lete or shrunken old industry exerts less pricing discipline on the new technology and as a result prices increase above c_0. Consumers can be worse off because the relaxation of pricing discipline from the old technology allows the firm with the new technology to increase price above c_0. But the innovation increases total welfare if $N(c_0 - c_1) > R$. The innovation could generate very large cost savings (and hence be socially desirable), yet fail the consumer ROR test even if the price increase is very small relative to the cost savings.

Some of the most important innovations in recent times have proceeded in steps with little or no consumer benefit at the early stages of the innovation. An example is a research tool such as the Cohen-Boyer technology for gene splicing. The Cohen-Boyer technology made possible major advances in medicine and agriculture that would have been difficult to predict when the technology was first discovered. Yet in its early stage, the Cohen-Boyer technology was just a tool for inserting genetic material into a cell and had no immediate consumer benefits. In its infancy the Cohen-Boyer technology, revolutionary as it was, would not have scored particularly well on a consumer ROR test.

The consumer ROR test has the advantage that it is aligned with antitrust goals if the objective of antitrust policy is to protect the welfare of consumers. Nevertheless, antitrust enforcement for innovation based on a consumer welfare standard would be difficult to do correctly and likely would generate false positives and false negatives. The consumer ROR test for anticompetitive innovation ignores the impacts that innovations can have on firm values, whether positive or negative. Furthermore, innovation can make some consumers worse off, but make other consumers better off, either through price discrimination or through spillover benefits in other markets. In theory, a consumer ROR test could take these impacts into account, but that is difficult to do in practice.

ANTITRUST ENFORCEMENT FOR INNOVATION BASED ON A CONSUMER WELFARE STANDARD WOULD BE DIFFICULT TO DO CORRECTLY AND LIKELY WOULD GENERATE FALSE POSITIVES AND FALSE NEGATIVES.

A particularly worrisome objection to a consumer welfare standard for innovation is that it can too easily fail to take into account the chilling effects of antitrust enforcement on decisions to invest in R&D. A consumer welfare analysis typically takes as given the economy's existing production possibilities. In this sense a consumer welfare analysis is ex post, after investments have been made. An ex post consumer welfare analysis can easily overlook that investments were made in the past with the expectation of future profits. These investments created the goods and services that benefit today's consumers.

There is an additional informational issue with a consumer or total welfare standard. Firms have limited information when they estimate the private (or social) value of an investment in R&D. An ex post antitrust analysis can draw on new information and information available from other firms. An innovation

may fail a ROR test ex post because, as a result of investments made by others and observed ex post, the incremental value of a firm's R&D falls short of its costs. However, the firm that made the investment could have no way to know what other investments were planned when it made its ex ante R&D decision.

Innovation is uncertain. Some innovations may not make consumers better off because they did not turn out as well as expected, although the expected benefits were positive when the investments were made. Furthermore, innovations typically build on other innovations. A particular incremental innovation may not improve consumer welfare, but that innovation builds on other innovations that generate benefits for consumers. In some cases, the profits from incremental innovations are necessary to justify the earlier innovations that consumers desire. Firms would not invest in the first place if they could not anticipate additional profits from subsequent innovations. Moreover, an antitrust standard that focuses only on consumer benefits discounts efficiency benefits and spillovers from innovations that show up as higher profits.

While a consumer rule of reason analysis may be aligned with the goals of antitrust policy, the practical difficulties of applying a consumer rule of reason analysis to innovation creates a risk that consumers would be harmed, not benefited, by a zealous application of such an antitrust standard to innovation by a single firm.

C. THE PROFIT SACRIFICE TEST

According to Janusz Ordover and Robert Willig (O-W), "predatory intentions are present if a practice would be unprofitable without the exit that it causes, but profitable with the exit."[13] I refer to this as the profit sacrifice test for predatory conduct.[14] O-W apply their test to identify predatory innovation as well as predatory pricing, arguing that pricing and innovation can have similar motives and effects. An improvement in the quality of a product is similar to a reduction in its price. Rivals may be unable to compete with the new and improved product and may exit the industry or fail to make investments necessary to remain as effective competitors. If the investment in the product would not have been profitable but for the exit of rivals, the Ordover and Willig test would ascribe predatory intentions to the investment.

A profit sacrifice test has inherent limitations to evaluate anticompetitive innovation. Innovation is about sacrificing short-term profits for long-term rewards. A firm incurs costs that reduce profits in the short run in order to devel-

13 Janusz Ordover & Robert Willig, *An economic definition of predation: pricing and product innovation*, 91(1) YALE L.J. 8-53, 9 (1981).

14 Anticompetitive conduct does not require a reduction in profit in the short run. Conduct such as exclusive dealing can harm competition with no reduction in profit. *See, e.g.*, Aaron S. Edlin, *Stopping Above-Cost Predatory Pricing*, 111 YALE L.J. 941 (2002).

op new products or processes that generate profits in the longer run. It is difficult to determine when the sacrifice of short-run profit by investing in R&D is excessive. A price below marginal cost is inefficient because the cost of an incremental unit of supply is less than its value (although pricing below marginal cost can have other benefits, such as overcoming switching costs or signaling to consumers that they will enjoy the product once they try it). There is no corresponding guidance for investment in innovation. The innovation may be economically inefficient if it costs more than the value it creates, but that entails evaluation of expected rather than realized costs and benefits, and requires complex measurement of the social and private values of the innovation.

> A PROFIT SACRIFICE TEST HAS INHERENT LIMITATIONS TO EVALUATE ANTICOMPETITIVE INNOVATION. INNOVATION IS ABOUT SACRIFICING SHORT-TERM PROFITS FOR LONG-TERM REWARDS.

A second prong of the O-W test is that the practice is unprofitable without the exit that it causes, but profitable with the exit. Successful innovation often disrupts markets and leads to the exit of firms that use technologies that are made obsolete by the innovation. Xerography was not a predatory innovation because it required a short-term sacrifice in profit and led to the exit of manufacturers of mimeograph machines. Just as a sacrifice of short-run profit says nothing about whether innovation has a predatory intent or effect, neither does the resulting exclusion of competition.

Whether a firm exits or becomes a less effective competitor as a consequence of innovation cannot control whether the innovation is anticompetitive. Significant and pro-competitive innovations often displace rivals. A possible alternative interpretation of the exit prong in the O-W test is whether an innovation would have been profitable assuming that firms remain in the market as actual or potential competitors with their old technologies, even if they have no sales because they are not competitive with the new and improved products or processes. This is a difficult inquiry not only because it is hard to conceptualize the effects of potential competitors that are displaced by the innovation, but also because the profit that the innovator could earn under the assumption that actual or potential competitors remain in the market depends on the intensity of the competition that would occur. The test would yield one profit level if, but-for exclusion, competition is assumed to be intense, and would yield another, higher profit level if, but-for exclusion, the innovator and rivals would have shared the market at a high price.

Returning to the example of a product innovation that increases value from v_0 to v_1, the profit sacrifice test would ask whether the innovation was profitable assuming actual or potential competition from firms with the old product. Suppose P_1 is the price for the new product. Ignoring production costs, the profit sacrifice test would require $NP_1 > R$ without exclusion of the old product. With

intense competition from the old product, the most that a firm can charge for the new product is $P_1 = v_1 - v_0$. At this price the innovation passes the profit sacrifice test if it is socially desirable, ignoring spillover benefits, under a total rule of reason standard; i.e., if $N(v_1 - v_0) > R$. In this respect the profit sacrifice test provides a screen for innovations that are socially beneficial. There are, however, many circumstances in which the price for the new product would be greater or less than $v_1 - v_0$, depending on the strength of actual or potential competition from the old product and other constraints that affect pricing, and there are many circumstances in which innovation generates large spillover benefits. When P_1 diverges from the social value of the innovation, the profit sacrifice test becomes less useful.

The profit sacrifice test requires that a court evaluate an innovator's profit under the counterfactual that the innovator does not benefit from changes in market conditions caused by the innovation. This is a complicated calculation. Even if it could be done accurately, there is no assurance that it leads to the right answer except in special circumstances. Profits earned from changes in market conditions may be essential to drive pro-competitive innovation. A better mousetrap can destroy the market for other mousetraps, but that is part of the reward that motivated the invention, and the incentive to innovate may be inadequate without that prospect.

Even if we cast these difficulties aside, application of a profit sacrifice test likely would ignore the spillover benefits from innovation for consumers and for firms, and for consumers in other markets and at future points in time. As with the consumer ROR test, a profit sacrifice test also runs the risk of performing the wrong calculation by ignoring incentives for innovation and by evaluating ex post rather than ex ante benefits and costs.

While there are problems with the profit sacrifice test as a test of predatory innovation, it could have value as a screen to identify when innovation is not anticompetitive, although there are also potential pitfalls in this application. Suppose an innovation produces a new product with a value of $100. There are other competitors with the same production cost that could supply a product worth $90. These other firms choose not to enter the market because with aggressive competition they can't expect to make any sales. Assuming the same production costs, a firm with a product that is worth $100 can beat competition from a product that is worth only $90; the better product can capture all sales at a price slightly less than its incremental value of $10. If other competitors choose not to enter because they do not anticipate any sales, then the innovator can charge the full value of $100. A correct application of the profit sacrifice test would use $10 as the social value of the innovation. This is its incremental value relative to other products. Yet if other products never enter the market because they are deterred by the innovation, it would be difficult to ascertain the innovation's true incremental value and easy, albeit incorrect, to conclude that the

social value of the innovation is its full value of $100 rather than its incremental value of $10.[15]

The profit sacrifice test is not a cure for the problems raised by the total or the consumer rule of reason tests to evaluate predatory innovation. It is complex to perform and can lead to false positives and false negatives.

D. NO ECONOMIC SENSE TEST

Under the no economic sense test, "conduct is not exclusionary or predatory unless it would make no economic sense for the defendant but for the tendency to eliminate or lessen competition."[16] While similar to the profit sacrifice test in some respects, the no economic sense test has important differences. The profit sacrifice test makes a positive statement that predatory intentions are present if a practice would be unprofitable without the exit that it causes, but profitable with the exit, although a finding of predatory intent is neither necessary nor sufficient for innovation to be anticompetitive. The no economic sense test instead implies that there is no antitrust liability for predatory conduct unless the conduct would make no economic sense but for the tendency to eliminate or lessen competition.[17]

A second difference is the focus in the profit sacrifice test on the short-run cost of a strategy. The profit sacrifice test compares a loss in short-run profit against future benefits from the exclusion of competition. There is no specific mention of a profit sacrifice in the no economic sense test. To some extent this is merely semantics. If conduct makes no economic sense, then it is likely because it entails a reduction in profit relative to another course of conduct. The difference in short-run profit with and without exclusionary conduct is a measure of the cost of that conduct.

15 Farrell and Katz show that the profit sacrifice also can produce false positives and false negatives when technologies have network effects. With network effects, the technologies that would represent actual or potential competition in the absence of exclusion depend on consumer expectations, which are not uniquely determined. *See* Joseph Farrell & Michael Katz, *Competition or Predation? Consumer Coordination, Strategic Pricing, and Price Floors in Network Markets*, 53(2) J. Indus. Econ. 203-31 (2005).

16 Gregory Werden, *Identifying exclusionary conduct under Section 2: The "No Economic Sense" Test*, 73(2) Antitrust L.J. 413 (2006). *See also* Gregory Werden, *Identifying Single-Firm Exclusionary Conduct: From Vague Concepts to Administrable Rules* ch. 22, Fordham Competition Law Institute, at 557-88; A. Douglas Melamed, *Exclusive Dealing Agreements and Other Exclusionary Conduct - Are There Unifying Principles?*, 73(2) Antitrust L. J. 375, 391 (2006) and A. Douglas Melamed, *Exclusionary Conduct Under the Antitrust Laws: Balancing, Sacrifice, and Refusal to Deal*, 20(2) Berkeley Tech. L. J. 1248 (2005).

17 According to Gregory Werden, conduct by a single firm is unlawfully exclusionary if it makes no economic sense but for its effect of eliminating competition and thus creating and maintaining market power. Furthermore, the conduct must be reasonably capable of making a significant contribution to maintaining monopoly power or give rise to a dangerous probability of creating monopoly power and not fall within any safe harbor or established exemption. *See id.* Werden, *Identifying Single-Firm Exclusionary Conduct: From Vague Concepts to Administrable Rules*, at 576.

Under some circumstances, conduct could harm competition even if it costs very little. Exclusive dealing, raising rivals' costs, and tying can exclude competitors without incurring significant costs in the short run. Gregory Werden offers an example of a firm that sets fire to its competitors' factories in a hypothetical world with no arson laws and costless matches. The profit sacrifice test might not catch this anticompetitive conduct because the arson does not require a sacrifice of short-run profit in this extreme hypothetical. The no economic sense test would properly alert an antitrust enforcer to possible anticompetitive conduct because it would make no economic sense for a firm to set fire to its competitors except to accomplish an anticompetitive end.

Conduct that has a valid business justification other than the exclusion of competition would escape liability under the no economic sense test. In this respect the test could exempt conduct such as innovation that is usually beneficial, although this turns on interpretation of the business justification for innovation. It makes economic sense for a firm to try to reduce its costs or raise the value of its product, even if the investment does not produce a positive return. Some investment in innovation, however, may be clearly unprofitable if it does not exclude competition. Nevertheless, there is a plausible case to assign a valid business justification to such investment because the benefits from innovation are difficult to assess and society could be better off from an innovation that excludes competitors. Alternatively, one might place innovation in the category of conduct that falls within a safe harbor for unlawful exclusion by a single firm.

Some might argue that a safe harbor for single firm innovation is unwarranted. Suppose a firm is the only supplier of an essential component of a system, such as access to a telecommunications local loop. Furthermore, suppose that the firm cannot charge a profit-maximizing price for access, but instead must accept a much lower price. As a result, other firms combine cheap access to the local loop with other complementary valued added services to offer systems, such as voice and Internet access, that consumers desire and sell these systems in competition with each other at low prices. Now suppose that the owner of the local loop invests in an innovation that makes the local loop incompatible with the value added services provided by other firms. The innovation could have an anticompetitive effect, but could escape liability under the no economic sense test if the owner of the loop could supply a plausible justification for the innovation other than the elimination of competition, or if the test provides a safe harbor for innovation. This may not be a bad result, given that the innovation could have significant social benefits. Moreover, it is consistent with the deference that courts give to firms in their decisions about when and how to deal with their rivals, as reflected in the Supreme Court's *Trinko* decision.

E. SHAM TEST

If innovation is construed to be an activity that always makes economic sense, then the no economic sense test provides a broad pass for innovation even if the

innovation may have anticompetitive consequences. That is an acceptable tradeoff. Antitrust policy has to strike a balance between over- and under-deterrence, and the risk of chilling innovation with too much antitrust enforcement is much greater than the risk of allowing some anticompetitive innovation to slip through the antitrust cracks.

If innovation always makes economic sense, then the no economic sense test is similar to a test of whether the innovation is a sham. Under a sham test, single-firm innovation would escape Section 2 liability if the innovation is not a sham. The problem, of course, is in the definition of a sham innovation. One might apply a sham innovation test in our simple example by requiring that $v_1 - v_0$ be above some minimum threshold value to establish that the innovation is not a fraud, but there is little to guide the choice of the minimum threshold. There are many innovations that appear to have a low incremental social value, yet consumers value them highly. Consider downloadable ring tones or computers that come in different colors.

A possible definition of a sham innovation is an innovation resulting from an investment that no firm would possibly make except for its adverse effect on competition, although this interpretation as well can lead to enforcement errors. Taking an existing technology as a given component of a firm's production possibilities, investment in an improvement to that technology may make no economic sense but for the improvement's adverse effects on competition, and hence the investment may fail either the no economic sense test or a sham test. But this conclusion may be incorrect. The profit from the improvement could be essential to justify the investments that created the technology that is the baseline for the improvement. If the firm could not improve the technology without incurring antitrust liability, the firm may not have invested in the underlying technology, and society could be worse off. An alternative definition of a sham innovation is whether the innovation makes at least some consumers better off. If it does, it is not a sham. This standard would be easier to apply than a no economic sense test or a minimum threshold for innovation and would be less likely to result in excessive deterrence of investment in R&D.

III. Strategic Innovation with Complements

Conventional approaches to evaluate predatory conduct can yield both false positives and false negatives when applied to innovation that changes the competitive landscape for substitute products. Given the potentially large benefits from innovation and the risks of judicial error, antitrust policy should restrain innovation by a single firm that affects substitute products only in exceptional circumstances, if at all. Further supporting this conclusion is that innovation does not preclude a rival from inventing around or improving on new technology that is the subject of an alleged predatory scheme.

In some circumstances, however, it can be difficult for rivals to invent around or improve on even a minor innovation. An example is an interface standard that affects the compatibility of complementary components for a computer network. In a series of cases decided in the late 1970s, plaintiffs alleged that IBM redesigned its mainframe computers to make them incompatible with products sold by independent vendors and chose prices and lease terms to advantage its own components. The product designs arguably achieved some cost savings or technical efficiencies, but also erected barriers to independent suppliers of peripheral components. While the coexistence of efficiencies and adverse effects on competition suggests cause for some rule of reason balancing, none of the courts involved in the IBM peripheral cases engaged in an express comparison of benefits and harms. Instead, they generally concluded that plausible efficiencies from product design placed the conduct in the category of monopolization (if it occurs) that is the result of a superior product or business acumen, and hence was not an offense under the Sherman Act.[18]

Courts have dismissed allegations of monopolization in other cases involving innovation by a single firm. *Berkey Photo* alleged that Kodak's introduction of a new camera and film format and its failure to disclose information about the new format to other camera manufacturers and film processors was part of an unlawful monopolization scheme.[19] The appellate court ruled that Kodak did not have a duty to disclose information about its products to its competitors and its introduction of a new camera and film was not anticompetitive. The court emphasized the special place of innovation in antitrust policy, stating that "Because [...] a monopolist is permitted, and indeed encouraged, by § 2 to compete aggressively on the merits, any success that it may achieve through 'the process of invention and innovation' is clearly tolerated by the antitrust laws."[20] In a more recent case, a district court held that a manufacturer of insulin pumps did not violate the antitrust laws when it modified its pumps to be incompatible with components sold by another firm.[21]

In other cases, courts have implicated product design and innovation as elements of a monopolization strategy. In *C.R. Bard v. M3 Sys., Inc.*, a manufacturer of biopsy guns and needles (C.R. Bard) changed the design of its biopsy gun

18 *See, e.g.,* California Computer Products, Inc. v. IBM, 613 F.2d 727 (9th Cir. 1979) and In re IBM Peripheral EDP Devices Antitrust Litigation, 481 F. Supp. 965 (N.D. Cal. 1979) ("Where there is a difference of opinion as to the advantages of two alternatives which can both be defended from an engineering standpoint, the court will not allow itself to be enmeshed 'in a technical inquiry into the justifiability of product innovations.'" ILC Peripherals Leasing Corp. v. IBM Corp., 458 F. Supp. 423, 439 (N.D. Cal. 1978)).

19 Berkey Photo, Inc. v. Eastman Kodak Co., 603 F.2d 263, 281 (1979).

20 *Id.* at 281.

21 Medtronic Minimed Inc. v. Smiths Med. MD Inc., 371 F. Supp. 2d 578 (D. Del. 2005).

in a way that made it incompatible with the needles sold by M3 Systems.[22] A district court held that Bard unlawfully leveraged its monopoly power in biopsy guns to obtain a competitive advantage in replacement needles by modifying its gun to accept only Bard needles. A divided panel of the U.S. Court of Appeals for the Federal Circuit sustained the verdict. The precedent value of this opinion is limited, however, because Bard advanced only limited arguments in its appeal.[23] Furthermore, the opinion is apparently inconsistent with a later Federal Circuit case in which the court held that a patent holder may exclude others from making, using, or selling the claimed invention free from liability under the antitrust laws.[24] Bard held patents on its biopsy gun and needles.

The question of predatory product design took center stage in the antitrust case brought by the U.S. Department of Justice (DOJ) and several states against Microsoft. The plaintiffs alleged a pattern of anticompetitive conduct in violation of Sections 1 and 2 of the Sherman Act. The district court found that Microsoft maintained a monopoly in the market for Intel-compatible PC operating systems and attempted to gain a monopoly in the market for Internet browsers in violation of § 2. The district court's findings with regard to anticompetitive product design identified three actions by Microsoft that interfered with competition from suppliers of rival Internet browsers:

(1) excluding Internet Explorer (IE) from the Add/Remove Programs utility;

(2) designing Windows so as in certain circumstances to override the user's choice of a default browser other than IE; and

(3) commingling code related to browsing and other code in the same files, so that any attempt to delete the files containing IE would, at the same time, cripple the operating system.[25]

The appellate court applied a test to evaluate the question of anticompetitive product design that included the following steps.[26]

- The plaintiff must demonstrate that the conduct harmed consumers (an anticompetitive effect);

22 C.R. Bard, Inc. v. M3 Systems, Inc., 157 F.3d 1340 (decided Sep. 30, 1998).

23 The jury instructions concerning monopolization may have been misleading, however Bard did not challenge the lower court's instructions in its appeal.

24 In re Independent Service Organizations Antitrust Litig., 203 F.3d 1322 (Fed. Cir. 2000).

25 U.S. v. Microsoft, U.S. Court of Appeals for the DC Circuit, 253 F.3d 34 (2001).

26 The Court described five principles, including the principle that the focus of the analysis is on the effect of that conduct, not on the intent behind it. I have condensed the first two principles into one principle dealing with competitive effects.

- if a plaintiff successfully demonstrates anticompetitive effect, then the monopolist may proffer a pro-competitive justification for its conduct; and
- the plaintiff can rebut the proffered pro-competitive justification or, if the justification stands unrebutted, then the plaintiff must demonstrate that the anticompetitive harm of the conduct outweighs the pro-competitive benefit.

The third step implies a rule of reason type of balancing of benefit and harm. The *Microsoft* court did not provide a manual for how to balance benefits and harms from innovation because the court never got to the third step in its analysis. The court concluded that Microsoft had not demonstrated any pro-competitive justifications for two of three contested elements: excluding IE from the Add/Remove utility and commingling code related to browsing and other code in the same files. Having satisfied the other requirements for a § 2 offense, the court concluded that these actions contributed to monopolization of the market for Intel-compatible personal computer operating systems.[27] For the third element—designing Windows to override the user's choice of a default browser other than IE—the court concluded that Microsoft offered a pro-competitive justification, which the plaintiff neither rebutted nor demonstrated was outweighed by the harm to competition.

The welfare implications of product design that affects interoperability are ambiguous. Markets for systems with complementary components can have multiple equilibria that have different consequences for consumer and total welfare.[28] Permitting the owner of an essential component to design the component so that it is does not interoperate with other firms' components may or may not lower consumer or total welfare, depending on the equilibrium that would have occurred with compatible components. A prohibition against incompatible technology designs can generate errors by prohibiting conduct that increases welfare, and the frequency of these errors would depend on the particular welfare standard that is applied.

THE WELFARE IMPLICATIONS OF PRODUCT DESIGN THAT AFFECTS INTEROPERABILITY ARE AMBIGUOUS.

27 Offering an inferior product can be part of a product differentiation strategy that has benefits for consumers as well as the seller. The Intel 386SX microprocessor was an Intel 386 device with a severed connection between the central processor and the math co-processor. This allowed Intel to offer consumers products with different functionality at different prices. See Raymond J. Deneckere & R. Preston McAfee, *Damaged Goods*, 5(2) J. ECON. & MGMT STRATEGY 149-74 (1996). Microsoft's design for Windows 98 did not appear to be part of a product differentiation strategy that could have similar effects.

28 Richard Gilbert & Michael Riordan, *Product Improvement and Technological Tying in a Winner-Take-All Market*, J. INDUS. ECON (forthcoming 2007) and Farrell & Katz, *supra* note 15.

None of the courts that considered cases involving product design, including *Microsoft*, engaged in a quantitative weighing of costs and benefits from the exclusionary effects of a product design according to either a total welfare or consumer welfare standard, nor did courts apply a profit sacrifice test. Most courts that have dealt with cases alleging anticompetitive innovation have applied a standard that more closely agrees with a no economic sense test, although not articulated as such.[29] Courts generally have refused to assign antitrust liability to innovation when there was a valid reason for a particular product design, and this threshold was met when the design produced plausible efficiencies. The *Microsoft* court purported to do a rule of reason balancing of the benefits and harms from the design of the Microsoft Windows 98 operating system and described a sequence of steps to perform the calculation. In fact, the court held that the design of the operating system was not anticompetitive when Microsoft could demonstrate plausible and unrebutted efficiencies, and held that design elements were anticompetitive only when Microsoft did not offer any efficiency justification. The *Microsoft* court never reached the point in its analytical roadmap that would require a comparison of benefits and adverse competitive effects from innovation.

IV. Product Line Extensions in the Pharmaceuticals Industry

The innovation cases discussed in the previous section involved complementary products that interoperate with each other. Allegations of anticompetitive innovation for substitute products have appeared in the pharmaceutical industry. Characteristics of the pharmaceuticals industry interact to create special circumstances for innovation and competition. Consumers have limited information about the therapeutic benefits of alternative prescription drugs and rely on their doctors to recommend a particular therapy. Price is often a secondary consideration. Patients and their physicians care about health outcomes and insurance often shields patients from the full price of a drug. As in most agency relationships, the objectives of the physician and his patient are not perfectly aligned. A patient's doctor may be relatively insensitive to cost even if the patient is not insured or faces a high co-payment.

Patent protection further limits the extent of price competition in the pharmaceuticals industry. Most patented drugs are available only from a single supplier. For example, in the class of statin drugs that are used to lower the levels of low density lipids (cholesterol) in the blood, atorvastatin calcium is available only as the branded drug Lipitor manufactured and sold by Pfizer. Until Pfizer's patent expires, price competition for atorvastatin calcium can occur only by sub-

29 Mark S. Popofsky, *Defining Exclusionary Conduct: Section 2, the Rule of Reason, and the Unifying Principle Underlying the Antitrust Laws*, 73(2) ANTITRUST L.J. 435, 445 (in practice, Microsoft and other courts have subjected product design to a no economic sense test).

stituting a different drug or therapy, not by substituting among different suppliers of the same drug. Other drugs could be as effective or nearly as effective as atorvastatin calcium in controlling blood lipid levels. These include other statin drugs, such as lovastatin (sold as the branded drug Mevacor), pravastatin (sold as Pravachol), or simvastatin (sold as Zocor), as well as drugs with a different mechanism of action, such as fenofibrate (sold under the brand name Tricor), gemfibrozil, bile acid sequestrants, or nicotinic acid.

In a typical market a consumer would comparison shop among many brands and types of products. If a consumer wants to purchase a car, she might consider sedans, station wagons, vans and SUVs and in each category compare different brands of new and possibly used vehicles. Armed with information from *Consumer Reports* and other sources, the consumer would choose the vehicle that offered the best value. The shopping experience is different for prescription drugs. If a consumer desires a better blood lipid profile, she cannot independently choose between the statins and other prescription drugs that can control lipid levels. She may only purchase what her doctor prescribes. Limited information about the benefits and costs of different therapies on the part of the patient, and in some cases her doctor as well, and insurance plans that isolate the consumer from drug prices act to moderate price competition between different drug therapies.

For drugs whose patents have expired, patients can benefit from price competition between different suppliers of the generic chemical compound.[30] Drugs with generic equivalents are called multi-source drugs. The original patented drug is alternatively called the pioneer or innovator drug or identified by a brand name rather than the name of the active ingredient. Many states allow pharmacists to dispense a therapeutically equivalent generic drug to fill a prescription for a branded product unless the doctor requires that the pharmacist dispense the brand. Price competition for generic equivalents can be intense because they are functionally identical products and a drug retailer is free to choose among multi-source generic suppliers when the law permits generic substitution.

The U.S. Food and Drug Administration (FDA) publishes *Approved Drug Products with Therapeutic Equivalence Evaluations*, also called the *Orange Book*, which lists all drug products approved by the FDA and has information on generic drug equivalents as well as active ingredients and proprietary names. Patent protection for the statin drug Zocor expired on June 20, 2006. In January 2007 the *Orange Book* listed eight suppliers of simvastatin in addition to Merck, the supplier of the Zocor branded product. The price of generic simvastatin is a fraction of the price of Zocor. In January 2007, packages of fifty 20mg pills of the

30 *See, e.g.*, A. Cook et al., How Increased Competition from Generic Drugs Has Affected Prices and Returns in the Pharmaceutical Industry (Congressional Budget Office, Jul.1998).

generic simvastatin were available for $14.37, while Zocor in the same package size and dose cost $137.45 from the same retailer.[31]

The Drug Price Competition and Patent Term Restoration Act of 1984 (often called the Hatch-Waxman Act after its sponsors) sought to balance the benefits of patent protection for drug innovation against the benefits of lower prices from generic competition. Prior to 1984, a generic manufacturer had to file a separate New Drug Application (NDA), which required proof of safety and effectiveness before the drug could be sold. The Hatch-Waxman Act introduced an Abbreviated New Drug Application (ANDA), which accelerates FDA approval by allowing a generic manufacturer to demonstrate that its drug is therapeutically equivalent to an already approved drug. Drugs are therapeutically equivalent if:

(1) there are no known or suspected bioequivalence problems, or

(2) actual or potential bioequivalence problems have been resolved with adequate evidence.

The FDA *Orange Book* designates drugs in conventional oral dosage forms in the first category as AA and those in the second category as AB.

A manufacturer of a pioneer drug can attempt to mitigate generic competition by introducing a related drug that provides new therapeutic benefits or by changing the delivery form or dosage strength of the drug. I refer to all of these tactics as product line extensions of the pioneer drug.[32] Many industries employ product line extensions (e.g., a low fat version of a yogurt brand). Product line extensions capitalize on consumer recognition of the underlying brand and are a valuable way to maintain or improve the market position of the brand.[33]

FDA rules and legislation such as the Hatch-Waxman Act contribute to the value of product line extensions for brand name drug manufacturers. Drugs that appear to be similar may not qualify as therapeutic equivalents and would not be listed as such in reference databases used by pharmacists. For example, a drug that differs from a pioneer drug in its delivery form would not be therapeutically equivalent to the pioneer drug and therefore would not be AB substitutable as a generic alternative. The same would apply to a similar new drug with a different

31 Prices available from Costco.com Pharmacy, *at* http://www.costco.com/Pharmacy/frameset.asp?trg=HCFrame.asp&hcban=Banner.asp&hctar=DrugInfo.asp&log=&rxbox=&fromscript=1&qf=&srch=zocor&Drug=ZOCOR&Article=ZOCOR, (accessed on Jan. 19, 2006).

32 Janis, Hovemkamp, and Lemley call this "product hopping". *See* Mark Janis et al., IP and Antitrust: An Analysis of Antitrust Principles Applied to Intellectual Property Law (Aspen Law & Business, 2001).

33 Product line extensions represent the majority of new product introductions in all industries by some estimates. *See, e.g.*, V. Kadiyali, N. J. Vilcassim, & P. K. Chintagunta, *Product line extensions and competitive market interactions: an empirical analysis*, 89(1-2) J. Econometrics 339-63 (1998) and Morris A. Cohen et al., *An Anatomy of a Decision-Support System for Developing and Launching Line Extensions*, 34(1) J. Marketing Res. 117-29 (1997).

chemical composition. Furthermore, under the Hatch-Waxman Act, if the branded drug or its product line extension is protected by a patent, the branded manufacturer can obtain an automatic stay that prevents generic entry for a period equal to the lesser of 30 months or the time required for the generic manufacturer to prove that the patent is not valid or would not be infringed.

In two recent cases plaintiffs have alleged that manufacturers of branded products have engaged in anticompetitive innovation through product line extensions.[34] *Walgreen v. AstraZeneca*[35] involved Prilosec and Nexium, drugs in the class of proton pump inhibitors used to block excess production of stomach acid. AstraZeneca, the manufacturer of Prilosec, introduced Nexium prior to expiration of patents on Prilosec. The active ingredient in Nexium is esomeprazole, which is an isomer of the active ingredient omeprazole in Prilosec. Isomers are different arrangements of the same molecule and have similar but not identical effects in the body. The plaintiffs alleged that Nexium was not therapeutically superior to Prilosec for treatment of ordinary persistent heartburn (although there was some indication that Nexium has benefits for treatment of esophageal and duodenal ulcers) and that by promoting Nexium over Prilosec AstraZeneca undermined the market for generic omeprazole. The plaintiffs further alleged that AstraZeneca spoiled the market for generic omeprazole by promoting an over the counter version of Prilosec; managed care organizations typically do not reimburse drugs that are available over the counter.

In *Abbott v. Teva*[36] the manufacturer of the drug Tricor reformulated the drug, changed the pill from a capsule to a tablet with lower dosage, and introduced the new tablet with a broader FDA indication, and on a second occasion offered a tablet with a new composition of the active ingredient with a further reduction in dosage that could be absorbed into the bloodstream without the requirement that it be taken with food. Both of the product changes were based on patented technologies. When the manufacturer made the changes, Abbott stopped marketing the older version of the drug and notified the National Drug Data File (NDDF), a widely used database of prescription drugs, that it was no longer selling the older drug. The active ingredient in Tricor is fenofibrate, which is used to control triglyceride and lipid levels. Generic manufacturers complained that they were foreclosed from the market for fenofibrate because pharmacists could not freely substitute the older drugs for prescriptions written

34 In at least one other case the U.S. Federal Trade Commission alleged that a branded drug manufacturer abused the Hatch-Waxman process and the special statutory thirty-month stay by listing a patent in the *Orange Book* that was not related to the actual drug and was used to delay generic entry. I do not address these types of allegations in this article.

35 Walgreen Co. et al. v. AstraZeneca Pharmaceuticals, U.S. District Court for the District of Columbia, Case No. 06-cv-02084-RWR.

36 Abbott Labs. v. Teva Pharms. USA, Inc., 432 F. Supp. 2d 408 (D. Del. 2006). The author consulted for Abbott Labs and Fournier in this case.

for the newer drugs, even though the older drugs were not significantly dissimilar from their newer versions.

In both cases, the improvements to the branded drug did not prevent generic or other drug manufacturers from competing with older versions of the drug. A generic manufacturer can sell omeprazole or older versions of fenofibrate without infringing patents held by the branded drug companies. The improvements only precluded the generic suppliers from obtaining automatic substitution of their drugs for the newer versions of the branded drugs. The new and old drugs were not AB substitutes for each other, and patents on the new drugs invoked the thirty-month stay of generic entry under the Hatch-Waxman Act.

In both cases the plaintiffs alleged that the conduct of the branded drug manufacturers frustrated the intent of the Hatch-Waxman Act, which was to facilitate generic competition. This is a misreading of the Act. The Hatch-Waxman Act offered a compromise between promoting generic competition and assuring a period of exclusivity for the branded product. The thirty-month stay provision was intended to protect the owner of a drug patent when challenged by generic entry.

A second objection was that the drug improvements were not significant and therefore should not be treated with deference under the antitrust laws as genuine product innovations. The basic premise is debatable. Nexium offers benefits compared to Prilosec for some patients. Some consumers prefer a tablet to a capsule and the move to a new formulation gave Abbott an additional opportunity to market the drug with a new FDA-approved indication. Furthermore, in both cases the changes to the drugs qualified for patent protection. In the case of fenofibrate, the improvements related to the absorption of the chemical in the bloodstream. While the patent office has been known to apply a low threshold for invention, it would be odd to conclude that an invention that wins a valid patent obtained by legal means is a sham. Furthermore, to the extent that a patent protects a minor invention, it should be possible for other firms to invent around the patent or sell other competitive products. This is true in the pharmaceutical industry as well as in other industries, although the cost of doing so is likely to be higher for drugs given the lack of consumer information, price insensitivity, and provisions in the Hatch-Waxman Act that limit generic entry.[37]

Plaintiffs in the Abbott case made a third objection that Abbott should not have removed the older versions of the drug from the NDDF and accepted returns of the older products. The effects of removing the older versions of the drug from the NDDF are unclear. A pharmacist could not substitute older versions of these drugs for prescriptions of the newer drugs even if they were available, because they are not AB rated with the newer drugs.

37 Organizations that provide a managed drug benefit have an incentive to identify drugs that offer similar therapeutic benefits at lower costs. These organizations are marketing opportunities for suppliers of low-cost older versions of drugs, provided that these older versions offer similar therapeutic benefits.

Clearly, many manufacturers discontinue their older products when they introduce newer versions. Ford does not sell its 2006 trucks after it moves to the 2007 model year. Suppliers of home electronics do not sell older models after they introduce new models. And some software vendors do not sell or support older versions of their software after they issue upgrades. There are legitimate reasons for a manufacturer to stop selling and even recall older products. It reduces consumer confusion and support costs and focuses retailers on the objective of promoting the new product, all of which can generate consumer benefits. A general rule that prohibits firms, even firms with monopoly power, from discontinuing older products would be unwise.

A determination that product line extension is anticompetitive should follow from the application of one or more tests for anticompetitive innovation, but all of the conventional tests have significant flaws. A total rule of reason test is likely to show that product line extensions for prescription drugs do not decrease total economic welfare. Generic competition transfers revenues from the branded manufacturer to consumers through lower prices. A revenue transfer has no effect on total economic welfare.[38] Furthermore, generic competition may cause output of the generic and the branded drug to fall relative to a baseline without generic competition. Branded manufacturers may reduce expenditures on promotion for drugs that face generic competition. Reduced promotion may lower sales,[39] which implies lower total economic welfare in the short run. Under these conditions a product line extension could increase output even in the short run, which would reinforce the conclusion that the product line extension is not anticompetitive under a total rule of reason test.

A DETERMINATION THAT PRODUCT LINE EXTENSION IS ANTICOMPETITIVE SHOULD FOLLOW FROM THE APPLICATION OF ONE OR MORE TESTS FOR ANTICOMPETITIVE INNOVATION, BUT ALL OF THE CONVENTIONAL TESTS HAVE SIGNIFICANT FLAWS.

A consumer rule of reason test could conclude that a product line extension is anticompetitive if it slows the erosion of market power, however this finding may be mistaken. Most innovations throughout the economy are extensions of existing products. Product line extensions may appear to be inconsequential, yet have significant value for consumers. Berndt et al. find that incremental prescription drug innovations in the form of supplementary approvals for new dosages, formulations, and indications account for a substantial share of drug utilization and

38 There would be a deadweight loss if higher prices resulted in lower output.

39 Scott Morton finds no significant relationship between brand advertising, including promotion expenditures, and generic entry or market share. Fiona Scott Morton, *Barriers to entry, brand advertising, and generic entry in the US pharmaceutical industry*, 18(7) INT'L J. INDUS. ORG. 1085-104 (2000). These results are not inconsistent with brand promotion increasing generic sales by expanding the potential for pharmacy substitution.

associated economic and medical benefits.[40] It would be incorrect to adopt a rule that would generally condemn these innovations.

In practice, a rule of reason analysis typically focuses attention on short-run benefits and tends to ignore the long-run benefit from innovative activity. The drug cases provide an instructive example. It is easy, but not generally correct, to conclude that consumers are harmed by a policy that delays generic competition. The ability to delay generic competition provides an incentive for firms to invest in the pioneer drugs that generic manufacturers copy. A proper balancing must account for the positive effects of product line extensions for investment in new drugs. After firms have invested to create a new product, consumers gain if the innovation is made available at its marginal cost, although a policy of zero-cost compulsory licensing for new inventions ultimately would harm consumers by undermining the incentive to invent. The situation is analogous for pharmaceutical product line extensions. One cannot measure economic benefit solely by considering the short-term benefit to consumers from generic competition. It is essential to account for the negative effects of generic competition on the incentive to create new drugs.

There are flaws in other tests for anticompetitive innovation. The profit sacrifice test compares the cost of the product line extension to its benefit assuming no exclusion of generic competition. This comparison is misleading because it assumes that the innovator product exists, although profits earned from the product line extension could be instrumental for investing in the innovator product in the first place. The profit sacrifice test also should take into account that the very conduct that threatens generic competition may be necessary for its viability. The supplier of the branded product could reduce expenditures on product promotion and physician detailing if generic competition greatly eroded profits from sales of the brand. Without support from the manufacturer, sales of the brand could fall. Fewer prescriptions for the brand mean fewer opportunities for pharmacists to make generic substitutions. As a result, sales of the generic could fall as well. Generic sales depend on doctors writing prescriptions for the generic molecule, which they likely would do for a popular branded drug that has recently gone off patent, such as Zocor, or for a drug that has been around for a long time, such as ampicillin. For drugs that are neither blockbuster products nor generics that have achieved common name recognition, generic competition could be its own undoing because sales of the generic from pharmacy substitutions depend on promotion of the brand.

If generic competition causes sales to fall, a profit sacrifice test could show predatory intent from a product line extension even though consumers as well as the brand manufacturer would be better off with the extension. Consider an extreme example in which generic competition eliminates prescriptions for a

40 Ernst R. Berndt et al., *The Impact of Incremental Innovation in Biopharmaceuticals*, 24(2) PHARMACOECONOMICS 69 (2006).

branded drug because the manufacturer stops promoting the brand. With generic competition, the brand would have zero sales and the manufacturer would not invest to improve the drug. With the product line extension, consumers benefit from consumption of the branded drug. Without the extension, doctors do not prescribe the drug and consumers are worse off. The brand manufacturer would be worse off without the extension and the generic manufacturers would be no better off if doctors are not prescribing the drug. Thus, in the alternative world that assumes no exclusion of generic manufacturers, it is possible that every participant in the market would be worse off (or no better off) than in the world in which generic manufacturers are excluded. Lower profits from sales of the brand without the product line extension also would contribute to lower consumer and producer surplus in the long run by eroding incentives for investment in innovator drugs. Nonetheless, the profit sacrifice test could ascribe predatory intentions to a product line extension that excluded generic competition.

According to Ordover and Willig, the profit sacrifice test could account for the dependence of generic sales on sales of the branded product and avoid this erroneous conclusion. The test should consider whether the generic manufacturer could profitably compete if it had to compensate the manufacturer of the brand for promotion expenditures and for any negative effects on other products.[41] This would bring the profit sacrifice test closer to a total rule of reason analysis, although it still would not consider the incentives to invent the pioneer drug in the first place.

The no economic sense test may escape some of the difficulties with the other tests, although that depends on its interpretation. One could argue that it makes no economic sense to spend millions on a product line extension for a drug unless the extension excludes generic competition. With this interpretation the no economic sense test essentially reduces to the profit sacrifice test, with its associated difficulties. Alternatively, one can interpret investment to improve a product as being outside the scope of activities that make no economic sense. With this interpretation the no economic sense test is similar to a test of whether the innovation is a sham. Given the difficulties in applying other tests to identify anticompetitive innovation in the pharmaceutical industry and the social cost of antitrust liability that deters investment in R&D, a rule that focuses on whether the innovation is a sham is good policy and consistent with the treatment of single firm innovation in Section 2 cases by most courts.

Innovation can delay entry of generic equivalents in part because provisions of the Hatch-Waxman Act, such as the automatic thirty-month stay when the holder of a drug patent sues a generic manufacturer for infringing the patent, protect innovator drugs from generic competition. The thirty-month stay creates an opportunity for strategic patenting by a branded manufacturer to delay generic

41 Ordover & Willig, *supra* note 13, at 45-7.

competition, which can be particularly effective if the Patent Office has a low threshold for patentability. If one were to conclude that innovation raises unique antitrust concerns in this industry, a logical remedy would be to ease generic substitution requirements or the application of the thirty-month stay, rather than to carve out special antitrust rules. The FDA could develop policies to facilitate generic substitution and limit new drug approvals to drugs that meet a threshold level of utility, and the U.S. Congress could further amend the Hatch-Waxman Act.[42] This would address unique causes of competitive effects from innovation in the pharmaceutical industry without imposing flawed antitrust rules.

IF ONE WERE TO CONCLUDE THAT INNOVATION RAISES UNIQUE ANTITRUST CONCERNS IN THIS INDUSTRY, A LOGICAL REMEDY WOULD BE TO EASE GENERIC SUBSTITUTION REQUIREMENTS OR THE APPLICATION OF THE THIRTY-MONTH STAY, RATHER THAN TO CARVE OUT SPECIAL ANTITRUST RULES.

V. Consistent Rules

Suppose a computer manufacturer changes an interface standard so that another firm's disk drive is no longer compatible. Unable to supply drives for this computer, other disk drive manufacturers may not be able to achieve economies of scale and may not be viable competitors in markets for disk drives. Suppose instead that the computer manufacturer simply refused to supply the information necessary for other firms to offer compatible drives. This refusal to deal would have the same competitive effect in markets for disk drives as the changed interface standard, but likely would have fewer efficiency benefits. In light of the skepticism expressed in the recent Supreme Court decision in *Verizon v. Trinko* concerning the obligation of a firm to assist a rival, it seems unlikely that the refusal to deal with no other anticompetitive conduct would incur antitrust liability. A pharmaceutical product line extension that excludes a generic competitor also has aspects of a unilateral refusal to deal. The generic manufacturer needs prescriptions for the branded product to take advantage of automatic generic substitution by the pharmacist. Although the branded product and the generic are substitutes, in a sense they are complements. The generic requires the brand to make automatic substitution sales. The strategy of introducing a newer drug along with retirement or failure to support the older version of the drug is similar to a refusal to supply the older version of the drug to allow generic substitution. Consistency suggests that product designs with exclusionary effects should have no greater antitrust scrutiny than a unilateral refusal to deal.

Antitrust policy applies a different standard to conduct by a firm with monopoly power that denies competitors access to necessary inputs or markets (other than access to the firm's own facilities) or imposes unnecessary costs on those who would deal with competitors. Such exclusive dealing can violate sections 1

42 For example, 2003 amendments permit only one thirty-month stay per ANDA.

and 2 of the Sherman Act if there are no offsetting efficiencies from the exclusionary conduct. Product designs could have effects that are similar to an exclusive dealing strategy. An extreme example is the introduction of a new computer reservation system by an airline that automatically penalizes travel agents for bookings on rival airlines. The DOJ and state plaintiffs in the *Microsoft* case alleged that the design of Windows 98 operating system increased the cost to computer vendors of offering computers with rival browsers.

Einer Elhauge supports a distinction in the treatment of single firm innovation depending on whether innovation furthers monopoly power though an increase in the firm's efficiency or by impairing rival efficiency, with no antitrust liability for the former.[43] His proposal has appeal for the rare innovations that are clearly intended to harm rivals or for design features that impose costs on rivals, but can be removed without significantly compromising the performance of the product. In the *Microsoft* case, the court concluded that two design elements were intended to impose costs on rivals and were not essential to the performance of the operating system. In many cases, however, the exclusionary effects from an innovation are entwined with the innovation's efficiency benefits and it is impossible to treat them separately. A new interface standard that permits faster data transfers but is incompatible with rival products creates efficiencies and can exclude rivals. An improvement to a branded drug creates benefits for consumers and can prevent automatic substitution by generic competitors. In such a situation it could be tempting to require alternative designs that have less of an exclusionary effect, but a search for less restrictive alternatives would involve courts in product design activities where they have little or no expertise, and would risk deterring beneficial innovation. If the exclusionary effects are an unavoidable consequence of an innovation that has actual benefits for product quality or cost, then the effects should be treated as part of the innovation and should not be a source of antitrust liability.

One might object that deference to innovation by a single firm is inconsistent with the treatment of innovation in other contexts. In merger analysis, competition authorities engage in a rule of reason balancing of likely pro-competitive effects of a merger against any likely competitive harm, and take into account both potential benefits for innovation and possible harm from a reduction of innovation.[44] Innovation benefits do not trump competitive effects in merger analysis, but plausible efficiencies can be sufficient for innovation to escape antitrust liability for monopolization. The different approaches to the treatment of innovation reflect the different treatment of unilateral conduct and mergers under the antitrust laws. Merger analysis is a prospective inquiry into the merg-

43 Einer Elhauge, *Defining better monopolization standards*, 56 Stanford L. Rev. 253, 316, 320 (2003).

44 Richard Gilbert & Willard Tom, *Is Innovation King at the Antitrust Agencies? The Intellectual Property Guidelines Five Years Later*, 69 Antitrust L. J. 43-86 (2001) describes how the agencies have incorporated innovation effects in merger analysis.

er's likely future effects, including its effects on innovation. In some cases, mergers can create market structures that are more or less likely to promote investment in R&D and these effects should be taken into account along with any risks that the merger would lessen price competition. In the innovation cases considered here, innovation has already occurred and an important concern is that antitrust enforcement would chill future incentives for innovation investments. Furthermore, as noted above, the conduct at issue in most of the cases examined in this article is similar in many respects to a refusal to deal, for which courts have been reluctant to impose obligations.

Deference to innovation in cases that allege predatory innovation is justified in part because the profit from successful innovation is the motivating force to invest in R&D. Clearly, this argument can be taken too far. Price-fixing creates profits that may motivate investment in R&D, but this is not a valid defense for price-fixing conspiracies. The relationship between profit and investment in R&D is too tenuous to justify an innovation defense for price-fixing and other naked restraints of trade.

VI. Conclusions

No single welfare measure provides an accurate guide for antitrust policy. Firms have wide discretion to choose the prices of their goods and services without running afoul of U.S. antitrust law, despite the fact that at least in the short run an increase in price unambiguously lowers consumer welfare and lowers total economic welfare when price is above marginal cost. Nonetheless, welfare measures can help to inform whether certain types of conduct should be prohibited under the antitrust laws by providing objective estimates of the impact of the conduct on market performance. Antitrust scholars have endorsed different measures to assess liability for predatory conduct. These include a rule of reason analysis that includes producer as well as consumer welfare, a rule of reason analysis that focuses only on consumer welfare, and profit sacrifice tests. All of these approaches are seriously flawed when applied to innovation by a single firm. Rule of reason analysis, whether based on consumer or total economic welfare, generally fails to measure the spillover effects from innovation, focuses on ex post rather than ex ante benefits and costs, does not adequately account for uncertainty, ignores the value of innovation as an input into future innovations, and, perhaps most importantly, does not account for the chilling effect of antitrust scrutiny on incentives to innovate. The profit sacrifice test is ill-suited to identify anticompetitive innovation because investment in R&D necessitates a sacrifice of short-run profit and therefore is not an indicator of predatory intent or effect. Furthermore, exclusion that results from successful innovations may be a necessary reward to induce socially desirable levels of R&D. When applied to product line extensions in the pharmaceutical industry, a profit sacrifice test can mistakenly identify innovation as anticompetitive even though consumers would be worse off and profits would be lower if the innovation did not occur.

Rule of reason and profit sacrifice approaches to the analysis of the competitive effects of innovative activity typically assume the existence of the innovation. By doing so, it is easy to forget that the profits earned from the exclusion of competitors provide an incentive to make the innovation in the first place. The problem is similar to an analysis of the consequences of patent licensing that ignores the effects of licensing terms on the incentives to innovate. It is easy to reach the erroneous conclusion that licensing innovations at very low royalties would increase output and promote welfare. That conclusion is clearly incorrect because such a policy would undermine incentives to invest in new innovations and would lower economic welfare in the long run.

The no economic sense test potentially addresses some of the shortcomings of the profit sacrifice test when applied to innovation, although it depends on its interpretation. The test does not raise concerns about predatory conduct unless the conduct would make no economic sense but for the tendency to eliminate or lessen competition. The test is similar to a profit sacrifice test if the definition of no economic sense turns on the ex ante profitability of the investment. If one instead concludes that innovation always makes some economic sense whatever its cost, then the no economic sense test provides a wide and deep safe harbor for innovation that is not a sham.

ANTITRUST POLICY SHOULD PROVIDE, IF NOT A SAFE HARBOR, AT LEAST A WIDE BERTH FOR INNOVATION BY A SINGLE FIRM BECAUSE INNOVATION NEARLY ALWAYS INCREASES ECONOMIC WELFARE AND THE ADVERSE EFFECTS OF INNOVATION THAT EXCLUDES RIVALS ARE TYPICALLY NO GREATER THAN THE EFFECTS OF A UNILATERAL REFUSAL TO DEAL.

Antitrust policy should provide, if not a safe harbor, at least a wide berth for innovation by a single firm because innovation nearly always increases economic welfare and the adverse effects of innovation that excludes rivals are typically no greater than the effects of a unilateral refusal to deal. Furthermore, antitrust courts are not well-equipped to analyze the effects of innovation on the entire economy and to evaluate the negative consequence that their enforcement decisions can have on future innovative efforts. A wide berth for single firm innovation can be accomplished with a rule of reason analysis that includes a strong presumption that innovation is not anticompetitive or with a no economic sense test that presumes that innovation makes economic sense even if it is not profitable ex post, provided that the innovation is not a sham. While these analytical approaches differ, they wind up essentially in the same place: innovation by a single firm is not anticompetitive if it has a plausible business justification and is not accompanied by other anticompetitive conduct. Indeed, this is what most courts have concluded when faced with allegations of predatory innovation. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

The Logic and Limits of Ex Ante Competition in a Standard-Setting Environment

Damien Geradin and Anne Layne-Farrar

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

The Logic and Limits of Ex Ante Competition in a Standard-Setting Environment

Damien Geradin and Anne Layne-Farrar

Some scholars have questioned the process by which cooperative standards are typically set, worrying about the potential for anticompetitive market power to come hand in hand with pro-competitive interoperability. To combat the perceived problems of ex post opportunism, the suggested solutions have focused on promoting procedures to facilitate ex ante competition. Since standards are generally desirable and competition often exists beforehand, many have argued that we need only formalize the ex ante competitive status quo to avoid any ex post market power trouble. Options proposed in the literature include ex ante auctions to be held during the standard definition phase or binding ex ante licensing commitments made before any vote on technologies occurs. We evaluate the various policy changes suggested with a particular eye to their unintended consequences and costs. Certainly the ex ante proposals would hold some appeal, if ex ante competition generally did not exist in their absence, but we find that they are problematic in important ways. We argue that not only are they not needed, they would tend to create more harm than good if implemented.

Damien Geradin is a partner at Howrey LLP and a professor of competition law and economics at TILEC, Tilburg University. Anne Layne-Farrar is a Director at LECG. The authors gratefully acknowledge Qualcomm, Inc.'s financial assistance. They also wish to thank Jorge Padilla for helpful comments.

I. Introduction

Few would question the pro-competitive effects that standards can bring. A standard can be defined as a set of technical specifications that seeks to provide a common design for a product or process.[1] For our purposes, we focus on standards that do not fully specify an end product, but rather specify key elements of that end product so as to enable various parts of such product and other products to successfully work together. A simple, but still high-technology, example would be modem protocols that allow PCs and server computers made by a wide range of firms to (more or less) seamlessly communicate with one another over a network, such as the Internet. When successful, such standards can improve the interoperability of complex technical products, enable welfare-enhancing cooperation among a host of disparate firms, increase consumer choice and convenience, reduce costs for consumers and producers alike, and broaden the size of the market (and thus profit opportunities) for participating firms.[2]

In order to achieve such benefits, complex standards, like modem protocols, require cooperative industry efforts. Firms—some of which produce complementary products and many of which compete against one another in various downstream markets—meet in a variety of forums to discuss and develop technical specifications to solve perceived industry interoperability problems. These forums are generally referred to as standard-setting organizations (SSO).[3]

Despite the clear benefits from the end results of SSO activities, some have begun to question the process by which cooperative standards are typically set. A number of firms and scholars have identified what they consider to be the potential for anticompetitive market power to come hand in hand with pro-competitive interoperability.[4] The concerns expressed above tend to rely on a number of theories that outline the risks that essential intellectual property (IP)

1 *See* H. Hovenkamp et al., IP and Antitrust: An Analysis of Antitrust Principles Applied to Intellectual Property Law § 35.1 (Supp. 2003-04). For alternative definitions, *see* D. Teece & E. Sherry, *Standard Setting and Antitrust*, 87 Minn. L. Rev. 1913 (2003).

2 *See FTC/DOJ Hearings on Competition and Intellectual Property Law and Policy* (2002) (statement of A. Marasco, Vice President and General Counsel, Am. Nat'l Standards Inst., Apr. 18, 2002), *available at* http://www.ftc.gov/opp/intellect/020418marasco.pdf:

> Standards do everything from solving issues of product compatibility to addressing consumer safety and health concerns. Standards also allow for the systemic elimination of non-value added product differences (thereby increasing a user's ability to compare competing products), provide for interoperability, improve quality, reduce costs and often simplify product development. They also are a fundamental building block for international trade.

3 *See* M. Lemley, *Intellectual Property Rights and Standard-Setting Organizations*, 90 Cal. L. Rev. 1889 (2002).

4 *See generally*, M. Lemley & C. Shapiro, Patent Hold Up & Royalty Stacking (Stanford Law and Economics, Olin Working Paper No. 324, Jul. 2006), *available at* http://ssrn.com/abstract=923468; C. Shapiro,

holders could impose excessively high royalty rates once their technologies have been embedded in a standard.

One such theory of ex post abuses is concerned with perceived hold-up problems. Before a standard is defined, firms compete on technology, offering different solutions to the proposed problems that standardization is intended to address. After a standard is defined, those firms that win technology selection votes within an SSO, it is argued, can potentially win market power as well. For example, it is alleged that if the circumstance arises where the firms that intend to implement the new standard already have made irreversible investments in plant and equipment, the firms holding patents on the technology comprising the standard could choose to hold up these implementers, asking more for licensing their patents than the patents' contribution to the standard warrants.[5] An implementer would be willing to pay this higher rate if it allows the firm to avoid the cost of switching to another technology—at least up to the point where the patent license fees equal the cost of moving to the next best alternative.

Part of the concern over ex post market power thus lies in the alleged unpredictable nature of the cost of licensing. Unless they enter into license agreements ex ante, implementers do not know the full cost of producing a standard, which may include royalty payments and other licensing fees. They are concerned that, for the reason expressed above, an essential IP holder may be in a position to impose excessive royalty rates, thereby negatively affecting the expected return of their investment.

Another theory points to the complex nature of technical standards that typically incorporates a multitude of complementary technologies. This is the classic Cournot complements point.[6] With many firms contributing technologies

---

footnote 4 cont'd
Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard-Setting, *in* Innovation Policy & the Economy, (A. Jaffe et al. eds., vol. 1, 2001); D. Lichtman, Patent Holdouts and the Standard-Setting Process (U. Chicago Law and Economics, Olin Working Paper No. 292, May 2006), *available at* http://ssrn.com/abstract=902646; R. Skitol, *Concerted Buying Power: Its Potential for Addressing the Patent Holdup Problem in Standard-Setting*, 72(2) Antitrust L.J. 727 (2005). For a rebuttal, *see* D. Geradin & M. Rato, Can Standard Setting Lead to Exploitative Abuse? A Dissonant View on Patent Hold-Up, Royalty Stacking and the Meaning of FRAND (Nov. 2006) (mimeo, Tilburg Law and Economics Center), *available at* http://ssrn.com/abstract=946792.

5 *See* Lichtman, *id.* at 2 ("In short, a patentee that comes into view only after a firm has invested in a given standard can hold hostage the firm's standard-specific investments. The result may be a royalty payment that far exceeds the inherent value of the underlying patented technology.").

6 *See* Lemley & Shapiro, *supra* note 4, at 16:

> The Cournot Complements effect arises when multiple input owners each charge more than marginal cost for their input, thereby raising the price of the downstream product and reducing sales of that product. Effectively, each input supplier imposes a negative externality on other suppliers when it raises its price, because this reduces the number of units of the downstream product that are sold.

that must work together in the final product, double (or more) marginalization can result. In the context of standards, this perceived problem is generally referred to as "royalty stacking" since each patent holder's royalty stacks up with all of the others to create one potentially excessive aggregate royalty for the firms hoping to implement the standard.[7]

To address these potential ex post problems, suggested solutions have focused on the promotion of procedures to facilitate ex ante competition. Since standards are generally desirable and competition often exists beforehand, many have argued that we need only create procedures to formalize the ex ante competitive status quo to assure that there will be no ex post market power trouble. Options proposed in the literature include ex ante auctions to be held during the standard definition phase or binding ex ante licensing commitments made before any vote on technologies occurs. Certainly these ex ante proposals would hold some appeal, if ex ante competition generally did not exist in their absence, but they also are problematic in important ways.

Following this introduction, Section II describes the main features of standard-setting processes, their significance, and the strategic battles that may affect them. Section III focuses on the fair, reasonable, and non-discriminatory (FRAND) licensing regime traditionally prevalent in SSOs. Under this regime, owners of intellectual property rights (IPR) that are essential to the standard typically commit to license such patents on fair, reasonable and non-discriminatory terms and conditions. The paper then turns to the theories of ex post problems and the proposed ex ante solutions. Section IV evaluates the logic behind the claims of ex post market power, including patent holdup, opportunistic behavior, and royalty stacking. Section V then assesses the limits of the suggested policy reforms meant to address ex post market power problems. Section VI concludes. We find that, while several of the proposals contain some attractive elements, most would either be difficult, if not impossible, to implement in practice or would entail a number of unattractive unintended consequences.

II. The Business of Standard-Setting

In today's technology-driven world, the importance of industry standardization, device interoperability, and product compatibility are critical for promoting innovation and competition in a number of industries.[8] To name just one example, standardization has been a key factor behind the significant growth in innovation and product differentiation in the information and communications technologies (ICT) sector.

7 *Id.*

8 *See* Marasco, *supra* note 2.

Of course, achieving product compatibility through standardization usually entails making choices, the effects of which will represent a cost. By design, after a standard has been defined it can constrain a variety of options and reduce competition between rival technologies.[9] When the technologies involved are covered by IPRs (usually patents), a standard may also raise issues related to access.[10] As will be seen below, holders of IPRs have the right to exclude others from their inventions. Those wishing to implement a standard should thus obtain a license from all the holders of essential IP.

Given the significant stakes frequently involved, the outcome of the debate over the most suitable technologies to be incorporated into any given standard have occasionally strained the SSO process.[11] Some tension is inevitable as each firm desires to promote its own solutions as part of the standard but also needs to work together with other SSO members to develop, establish, endorse, and promote the standard.[12]

Another factor contributing to SSO tensions relates to the fact that firms involved in standard-setting often wear different hats corresponding to the fundamentally different business models they adopt.[13] Consider a simplified categorization:

(i) Pure innovators or upstream-only firms (i.e., firms that develop technologies and earn their revenues solely by licensing them);

(ii) Pure manufacturers or downstream-only firms (i.e., firms that manufacture products based on technologies developed by others but that conduct no basic research of their of their own, limiting their activities to product development, and have no relevant IPRs);

(iii) Vertically integrated firms (i.e., firms that develop technologies and manufacture products based on those technologies and the technologies of others); and

(iv) Firms that do not create technologies or manufacture products, but buy products that are manufactured on the basis of patented technologies.

---

9 On the other hand, standardization promotes competition within a standard (i.e., between products implementing the standard). *See* Teece & Sherry, *supra* note 1, at 1915.

10 *See* C. Shapiro, *Setting Compatibility Standards: Cooperation or Collusion?*, *in* Expanding the Bounds of Intellectual Property § 3 (R. Dreyfuss et al. eds., 2001).

11 For case study examples, *see* B. DeLacey et al., Strategic Behavior in Standard-Setting Organizations (May 2006) (mimeo), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=903214.

12 *See* Shapiro, *supra* note 10, at 1-2.

13 *See* Teece & Sherry, *supra* note 1, at 1929.

These different firms operate in the downstream product market, the upstream technology market, or both. Naturally their incentives are asymmetric and their behavior in the standard-setting context diverges accordingly. While there is a certain degree of fluidity between the categories, the following structure of incentives can be identified:

- Pure innovators are entirely dependent on licensing revenues to continue their operations. Licensing revenues must be sufficient to cover the costs incurred in developing the technologies they seek or hope to license (including the costs of failed projects), as well as to give them sufficient incentives to engage in complex and risky projects.
- Pure manufacturers have converse incentives. As royalties represent a cost of production they have every incentive to reduce them. The lower the level of royalties payable to holders of IPRs essential to the standards they practice, the higher their potential level of profits.
- Vertically integrated firms that both develop technology and sell products have mixed incentives. On the one hand, they can draw revenue from their IPRs if they so choose. On the other hand, they will have to pay royalties to other firms holding IPRs essential to the standard for the products they manufacture. Since the bulk of the revenues (and profits) of these firms is usually made downstream through product sales, they are much less dependent than pure innovators on revenues generated by royalties.[14] In their licensing negotiations with other firms, they may well be more interested in protecting their downstream business from litigation than in charging royalties. They therefore tend to have stronger incentives to cross-license their own essential IPRs in exchange for essential IPRs held by other firms, instead of seeking royalty income.[15]
- The immediate incentives of buyers of products implementing standards relying on patented technologies are generally in line with manufacturers. They may consider that the royalties that manufacturers pay to IP holders will increase the price of the products they buy from such manufacturers. Generally, however, royalty payments and other direct licensing costs represent a small share of the total cost of production. Moreover, reducing royalty rates on some products might not necessarily lead to cheaper prices. As will be seen below, the extent to which royalty savings are passed on to buyers will vary depending on the state of competition in the downstream market. If that market is not competitive, royalty savings will not necessarily be passed on.

14 In 2004, for example, royalties roughly represented only 1.3 percent of Ericsson's total revenues. *See* LM Ericsson Telephone Company, Form 20-F: Annual Report 45 (2004), *available at* http://www.ericsson.com/ericsson/investors/financial_reports/2004/20f.pdf.

15 *See* A. Layne-Farrar & J. Lerner, To Join or Not to Join: Examining Patent Pool Participation and Rent Sharing Rules (Nov. 2006) (mimeo), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=945189. Note that IP licenses can include a wide array of terms, from upfront lump-sum fees to technology milestone payments. We use royalties as shorthand for all such terms.

In light of these widely acknowledged tensions, most formal SSOs have written IPR policies whose primary goal is to address two fundamental issues: disclosing and licensing of IPRs incorporated into a proposed or adopted standard.[16] Although their scope may vary significantly across SSOs, these procedures usually seek to encourage IPR owners to make their proprietary inventions known and available for standardization, and to allow their use by those wishing to implement the standard—all without imposing undue obligations on SSO members. SSO IPR policies are thus studies in balance. While keeping members' diverse interests in mind, they also strive to accommodate the interests of implementers to obtain access to the standardized technology, avoiding situations where IPR owners refuse to license their technology essential to the implementation of a standard to protect, for example, their positions in downstream markets.[17]

Most SSOs encourage IPR owners involved in standardization to disclose upfront (i.e., prior to the adoption of a standard) the IPRs that they consider may be essential for its implementation.[18] Early disclosure of patents "is likely to enhance the efficiency of the process used to finalize and approve standards."[19] It also:

> "permits notice of the patent to the standards developer [...] in a timely manner, provides participants the greatest opportunity to evaluate the propriety of standardizing on the patented technology, and allows patent holders and prospective licensees ample time to negotiate the terms and conditions of licenses...."[20]

As a rule, however, SSOs do not impose an obligation on IPR owners to conduct a search for, or guarantee the disclosure of, all IPRs they own that may be essential to a given standard. In most instances, this would prove extremely difficult. For

16 *See* Lemley, *supra* note 3, at 20-21.

17 *See, e.g.,* Eur. Telecomm. Standards Inst., ETSI Guide on Intellectual Property Rights (IPRs), art. 1 (2006) [hereinafter ETSI's Guide on IPR], *available at* http://www.etsi.org/legal/documents/ETSI_Guide_on_IPRs.pdf ("The ETSI IPR Policy seeks a balance between the needs of standardization for public use in the field of telecommunications and the rights of the owners of IPR.").

18 ETSI defines "Essential IPR" as meaning "that it is not possible on technical (but not commercial) grounds, taking into account normal technical practice and the state of the art generally available at the time of standardization, ... [to] comply with a standard without infringing that IPR." *See* Eur. Telecomm. Standards Inst., ETSI Rules of Procedure, Annex 6: ETSI Intellectual Property Rights Policy, art. 15 (2006) [hereinafter ETSI PRI Policy], *available at* http://www.etsi.org/legal/documents/ETSI_IPRPolicy.pdf.

19 *See* Am. Nat'l Standards Inst., Guidelines for Implementation of the ANSI Patent Policy 3 (1997) [hereinafter ANSI Guidelines], *available at* http://www.niso.org/committees/OpenURL/PATPOL.pdf.

20 *Id.*

firms with large patent portfolios and widely dispersed development groups such a search would be impracticable and could restrict the willingness of firms to participate with the SSO in the first place.[21] Even without a large portfolio, though, determining which patents are essential for a standard typically requires a difficult, and subjective, patent-by-patent evaluation. Indeed, this determination may not be feasible as the scope of a standard evolves through its development or, if the relevant IPR is a pending patent application, as claims are modified during prosecution at the patent office. The fact that the scope of such disclosure and the obligations imposed on IPR owners by the policies of some SSOs have in certain instances been the subject of conflicting and ambiguous interpretations has led some commentators to decry "the inadequacy of typical SSO disclosure policies."[22] As we argue below, these concerns are generally misplaced.

Once disclosure is made, or contemporaneously with disclosure, IPR owners are typically asked to provide an assurance or undertaking that, should their IPRs turn out to be essential for the final standard, they will license them on FRAND terms and conditions to other members of the SSO and, as is often the case, to outsiders.[23] SSOs do not mandate such commitments—which could be interpreted as compulsory licensing—but if the owner of potentially essential IPR seeks to have its technology included in a standard it has a strong incentive to provide the SSO with the assurance that it will license on FRAND terms and conditions. Given the fundamental importance of FRAND assurances, we turn next to a more detailed discussion of the concept of FRAND in the context of IP licensing.

III. IP Licensing under FRAND Commitments

IPRs are legitimate exclusive rights, which confer on their owners two basic prerogatives:

(1) the right to prevent any third party from applying or using the subject matter of the IPR;[24] and correlatively,

21 *See* Teece & Sherry, *supra* note 1, at 1946:

> An obligation to search for "implicated" IP can be extremely onerous. It is a major task to search a patent database and to compare it against the proposed standard. Patent searching is especially problematic when the standard evolves over time. Further, it is often difficult to know whether a patent "reads on" a proposed standard, as that may entail a major effort at claims construction and interpretation. A search requirement is especially onerous for IP owners who have substantial numbers of patents. Many firms in high-tech industries have thousands of patents, hundreds of which may be potentially relevant to a proposed standard.

22 *See* Skitol, *supra* note 4.

23 *See* Lemley, *supra* note 3, at 26.

24 *See* G. Masoudi, Intellectual Property and Competition: Four Principles for Encouraging Innovation, Remarks at Digital Americas 2006, São Paulo, Brazil (Apr. 11, 2006), at 3 ("In the world of physical

(2) the right to set the conditions of a license in consideration for use of the IPR and as a reward for the innovative contribution made.

Except for certain exceptional circumstances,[25] a patent owner may therefore decide not to grant any third party a license to practice the invention. These exclusive rights are recognized in all patent laws as well as in the TRIPS agreement.[26]

Modern standards typically include technologies protected by IPRs. In recognition of the exclusive rights, SSOs generally do not force their members with IPRs—usually patent holders—to grant a license for their patents. The European Telecommunications Standards Institute's (ETSI) IPR policy, for instance, does not contain any obligation to license essential IPR. Rather, it provides that a standard or technical specification may not be approved unless the owner of essential IPR provides an assurance of its intentions to license on FRAND terms. In particular, Section 6.1 of ETSI's IPR Policy states that when essential IPR is disclosed, ETSI will request—but not oblige—the owner of the IPR to undertake in writing that it is prepared to grant irrevocable licenses on FRAND terms and conditions, and as such to waive its right to refuse to offer a license to those seeking one. These waivers reflect a willingness by the patentee to forego its right to exclusivity in exchange for the opportunity to have its patented technology included in a standard. The FRAND undertaking is thus meant to ensure the dissemination of essential IPRs in a standard, keeping the standard available for adoption by members of the industry while at the same time making certain that the IP holders are able to be properly compensated for their innovations.[27]

THE FRAND UNDERTAKING IS MEANT TO ENSURE THE DISSEMINATION OF ESSENTIAL IPRS IN A STANDARD, KEEPING THE STANDARD AVAILABLE FOR ADOPTION BY MEMBERS OF THE INDUSTRY WHILE AT THE SAME TIME MAKING CERTAIN THAT THE IP HOLDERS ARE ABLE TO BE PROPERLY COMPENSATED FOR THEIR INNOVATIONS.

footnote 24 cont'd
property, enforceability means the right to exclude: for example, the ability to evict a person from your land. In the world of intellectual property, the fundamental right is similar: an enforceable IP right means the right to exclude others from using your intellectual property right at all.").

25 The European Court of Justice, for instance, has held that such exceptional circumstances may occur where the refusal to license cannot be objectively justified and would eliminate all competition in a downstream market for a new product for which there is customer demand not offered by the owner of the IPR. *See*, inter alia, ECJ Judgment of Oct. 5, 1988, Case 238/87, AB Volvo v. Erik Veng (UK) Ltd., 1989 4 C.M.L.R 122, at para. 8; Joined Cases C-241/91 P and C-242/91 P, RTE and ITP v. Commission of the European Communities (*Magill*), 1995 E.C.R. I-00743, at para. 50; and, ECJ Judgment of Apr. 29, 2004, Case C-418/01, IMS Health GmbH & Co. OHG v. NDC Health GmbH & Co. KG, at paras. 35 and 52.

26 WORLD TRADE ORG., MARRAKESH AGREEMENT ESTABLISHING THE WORLD TRADE ORGANIZATION, ANNEX 1C: AGREEMENT ON TRADE-RELATED ASPECTS OF INTELLECTUAL PROPERTY RIGHTS, art. 28 (signed Apr. 15, 1994).

27 The ETSI IPR Policy, for example, provides that IPR holders should be rewarded properly, explicitly recognizing that patent holders "should be adequately and fairly rewarded for the use of their IPR." *See* ETSI PRI Policy, *supra* note 18, art. 3.2.

If the owner of an essential IPR decides not to make a FRAND commitment, however, it does not necessarily follow that the relevant IPR will be excluded from the standard. Again using ETSI's IPR policy as an example, Article 8.1 provides that ETSI's General Assembly will examine whether alternate technical solutions exist. Where it concludes that this is not the case, the Director General may request the owner of the IPR to reconsider. However, the latter is not under any obligation to agree to license.[28]

Even with a FRAND assurance in place, standard implementers still need to negotiate and enter into license agreements with each essential IPR owner. In other words, a FRAND assurance is not, itself, a license. Actual licensing negotiations between IPR holders and each individual potential licensee is conducted outside SSOs. Most SSO IPR policies make clear that such discussions must not take place under the auspices of standard development activities, as SSOs view their role as directing technical rather than commercial issues.[29] Likewise, the reasonable and nondiscriminatory character of any license must be addressed in a commercial context. While some have tried to alter this demarcation, none have been successful. For instance, recent proposals made by some members of ETSI called for revising the current IPR policy in order to introduce the principles of "aggregated reasonable terms" and "numeric proportionality" into the definition of FRAND, but these efforts were rebuffed.[30] No consensus as to the need for or desirability of this proposed system of patent valuation could be achieved among ETSI members.

---

28 This was recently confirmed by a Working Committee of the International Association for the Protection of Intellectual Property (AIPPI) which stated the following with regard to the relationship between technical standards and patent rights: "The owner of a relevant patent can, in principle, not be forced to grant licences to other members of the organization or to outsiders. Only in a few exceptional cases should compulsory licences be admissible according to the conditions of Art. 31 TRIPS or the respective national laws" and "... [a] patent right whether owned by a member of the organization or a third party, which has been identified as relevant for a 'de jure' standard, may be used in the standard only with the consent of the owner." *See* Int'l Ass'n for the Protection of Intellectual Property, Question Q 157 The Relationship between Technical Standards and Patent Rights (2001), at paras. 3.2 and 4, *available at* http://www.aippi.org.

29 For example, ETSI's Guide on IPR provides that:

> specific licensing terms and negotiations are commercial issues between the companies and shall not be addressed within ETSI. Technical Bodies are not the appropriate place to discuss IPR issues. Technical Bodies do not have the competence to deal with commercial issues. Members attending ETSI Technical Bodies are often technical experts who do not have legal or business responsibilities with regard to licensing issues. Discussion on licensing issues among competitors in a standards making process can significantly complicate, delay or derail this process.

See ETSI Guide on IPR, *supra* note 17, § 4.1.

30 Pursuant to this proposal, called "Minimum Change, Optimal Impact," aggregated reasonable terms would mean that:

> in the aggregate the terms are objectively commercially reasonable taking into account the generally prevailing business conditions relevant for the standard and applicable

As noted, the terms and conditions of any license negotiated under the umbrella of a FRAND assurance are the result of a normal arms-length process of commercial negotiations between the licensor and an individual licensee. A commercial market-driven negotiation of license terms is not only what FRAND suggests but is also justified from an economic perspective, as it supports dynamic competition and provides incentives to innovate. Firms engaged in the development of innovative technologies "must not be restricted in the exploitation of intellectual property rights" in case their incentives to innovate are hindered.[31] SSOs recognize that an IPR owner must be free to seek compensation that is sufficient to maintain investment incentives.[32]

Equally important, given the voluntary nature of participating in an SSO, allowing IPR owners to seek adequate compensation is paramount to ensuring that those who own valuable proprietary technology remain involved in the standard-setting process. Note that SSOs are not the only option for standards development. Firms with sufficient name recognition or with clearly superior products, depending on the circumstances, may be able to choose to opt out of an SSO and try instead for a market-defined de facto standard.[33] In such cases, they may no longer be bound by that SSO's IPR policy. Securing the participa-

---

footnote 30 cont'd

> product, patents owned by others for the specific technology, and the estimated value of the specific technology in relation to the necessary technologies of the product.

In turn, numeric proportionality would mean that "compensation under FRAND must reflect the patent owner's [numeric] proportion of all essential patents." *See* Informa Telecoms and Media, Vendors Seek Compromise on LTE (Mar. 20, 2006), *at* http://www.informatm.com/itmgcontent/icoms/s/sectors/networks-infrastructure/20017342341.html.

31 *Id.*

32 *See e.g.* ETSI IPR Policy, *supra* note 18, art. 3.2 ("IPR holders whether members of ETSI and their AFFILIATES or third parties, should be adequately and fairly rewarded for the use of their IPRs in the implementation of STANDARDS and TECHNICAL SPECIFICATIONS." (emphasis in the original) *See also*, Teece & Sherry, *supra* note 1, at 1934:

> The complaints of those who believe that they are being compelled to 'overpay' for the use of others' IP embedded in the standard are frequently and forcefully stated. The more reasoned and quieter countervailing arguments focused on the social benefits of innovation and the need to compensate inventors for their efforts often are downed out by this din. The tension between static and dynamic views of efficiency is nothing new in the context of IP. But it suggests that policies that further burden IP and IP holders will only exacerbate the problem.

33 *See* Teece & Sherry, *supra* note 1, at 1918:

> In addition, many "standards" are not set by SSOs at all. Rather, they reflect the market success of a particular product in competition with other competing products. Such "de facto" or "market" standards are common in what economists term "network industries" in which consumers benefit by adopting products or processes adopted by others. Well-known examples include VHS VCRs (which "won" a "market standards" war with Sony's Betamax VCRs) and Microsoft's DOS and Windows operating systems.

tion of holders of valuable IPRs allows SSOs to adopt standards based on the best available technological solutions. The adoption of a standard incorporating second-best technology would have potentially damaging consequences negating the purpose of standardization itself.[34] It could thwart the standard's acceptance by industry and consumers alike and, as firms outside the SSO introduced incompatible products, it could lead to conflicting technologies, thereby reducing the efficiencies fostered by standardization. The ability to license IPR on FRAND terms and conditions is, in this respect, a flexible tool which secures the availability of essential IPR without unduly constraining licensors.

IV. Assessing the Complaints against the Ex Post FRAND Licensing Regime

While SSOs have significantly contributed to the development of, and the growing competition within, high-tech sectors, as explained at the outset of this article some commentators nonetheless believe that the current disclosure and FRAND licensing commitments are insufficient.[35] Without regard for the realities of the ex ante market interactions that typically occur, it has been said that the existing FRAND regime—or more generally the procedures and IPR policies of the SSOs—is inadequate to give standard implementers a sufficient degree of predictability over the costs of implementing a proposed standard. It is also claimed that the current regime is unable to prevent essential IP holders from behaving opportunistically. Finally, because many standards involve a large number of patents held by different firms, some claim that the present regime can lead to cumulative royalty rates of such a level that implementing the standard would no longer be attractive and thus useful innovations would no longer make it to the marketplace. This latter problem is the royalty stacking issue discussed earlier. We address each of these three claims in order below.

A. LACK OF PREDICTABILITY

There is little doubt that predictability over costs is an important issue for firms intending to invest in the design and manufacture of new products. Those firms and commentators who complain about the lack of predictability offered by FRAND commitments generally argue for the need to obtain more precise information about the costs of the various technologies being considered for integration

34 *See* J. DeVellis, *Patenting Industry Standards: Balancing the Rights of Patent Holders with the Need for Industry-Wide Standards*, 31 AIPLA Q. J. 301, 343 (2003).

35 *See, e.g.*, G. Ohana et al., *Disclosure and Negotiation of Licensing Terms Prior to Adoption of Industry Standards: Preventing Another Patent Ambush*, 24 Eur. Competition L. Rev. 644 (2003) and Skitol, *supra* note 4.

within a standard before the standard in question is adopted.[36] They thus claim that essential IP holders should disclose their licensing terms on an ex ante basis, typically in the form of maximum royalty rate and most restrictive terms to be offered.

These criticisms tend to overlook the fact that voluntary ex ante disclosure of licensing terms by IPR owners and ex ante negotiations of license agreements with IPR owners are already regular occurrences.[37] Neither the IPR policy of ETSI, for instance, nor the policies of many other major SSOs prevent IPR holders from disclosing and negotiating licensing terms before a standard is adopted. Much to the contrary, rights owners have a strong incentive to enter into such ex ante negotiations as they increase the likelihood that their technology will be incorporated in the standard.[38] In order to have their technology embodied in a forthcoming standard, these firms must find support among the members of the SSO. Consequently, they will seek to assure the superiority of their technology, and may also want to show that the royalties they will charge if their technology is selected will be reasonable. When the process works properly, the firm offering the best overall package—in terms of technology, ease of use, royalty rates, and other licensing terms and conditions[39]—will find the greatest number of supporters and its technology will be incorporated in the standard. Furthermore, nothing prevents a standard implementer from approaching an owner of essential IPR to inquire what its licensing terms will be. In other words, nothing prevents firms that wish to obtain information about the costs of proprietary technologies to request essential IP holders to provide them with information about the royalty rates and the other licensing terms that would apply should the technology be embedded in the standard under consideration.

B. EX POST OPPORTUNISM

One of the criticized pitfalls of the current FRAND regime is the alleged risk that owners of IPR essential to a standard will be able to unduly capture some of the

36 *See FTC/DOJ Hearings on Competition and Intellectual Property Law and Policy* (2002) (statement of S. Peterson, Corporate Counsel, Hewlett-Packard Company, Nov. 6, 2002), *available at* http://www.ftc.gov/opp/intellect/021106peterson.pdf.

37 *See FTC/DOJ Hearings on Competition and Intellectual Property Law and Policy* (2002) (statement of R. Holleman, Apr. 18, 2002) 2, *available at* http://www.ftc.gov/opp/intellect/020418richardjholleman.pdf.

38 ANSI Guidelines, *supra* note 19, at 3-4 ("A patent holder may have a strong incentive to provide an early assurance that the terms and conditions of the license will be reasonable and demonstrably free of unfair discrimination because of its inherent interest in avoiding any objection to the standardization of its proprietary technology.").

39 Potential licensors and licensees may focus their negotiations on factors other than royalty rates, such as for instance cross-licensing of IPR or ex post implementation costs. It would, for instance, be too simplistic to believe that, because A offers on an ex ante basis a lower royalty rate than B, A's technology will overall be cheaper than B's. Differences in implementation costs may be a legitimate reason for B to charge higher royalty rate than A.

economic value attributable not to the intrinsic value of those rights but to standardization itself. It is argued that if members of an SSO had known ex ante a standard being set the terms under which IPR owners would license their rights, they might have chosen an alternative technology (provided, of course, such alternative technology existed).[40] But once the standard has been adopted and implemented, switching to an alternative technology may have become too onerous for those practicing it. The argument continues that the bargaining power of the owner of essential IPR will have thus increased and that it may be able to extract more favorable licensing terms ex post standardization than would otherwise have been the case.[41] This phenomenon can be described as ex post opportunism.

Attractive at first blush, the theory of ex post opportunism overlooks several critical issues.

The first is that this theory is based on the premise that alternative technologies existed at the time of adoption of a particular standard and that the successful technology would have been chosen notwithstanding any licensing disparity.[42] In many instances of standard development, however, no suitable alternative technology exists. In the absence of substitute technologies, it cannot be argued that the standard-setting process gives additional market power to the IP holder: the technology had no competition either before or after the standards vote. The market power pre-exists the standard and is due to the uniqueness of the technology in question. Fundamental economics maintains that firms with a unique product or IP will be in a stronger position than those with products or IP for which alternatives exist. The fact that the IP is embedded in a standard adds no market power. Instead, what standardization might do is to increase the value of the IP by allowing its holder to collect royalties on larger volumes.

Firms holding patents relevant for a standard also face a number of important constraints. Regardless of whether the patented technology faces viable substitutes, the licensing price is constrained by the prices commanded by complemen-

40 *See* Teece & Sherry, *supra* note 1, at 1938-39:

> Whether the SSO would have in fact adopted another alternative had it known of the patent claims raises a complex counterfactual question: 'What would the SSO have done if the world had been different?' The answer is likely to be hotly debated, and depends on the particular facts of the standard at issue. The greater the advantages of the (patented) standard over the alternatives that were considered and rejected at the time the standard was originally set, the less likely it is that an alternative would, in fact, have been chosen.

41 D. Lichtman, *supra* note 4; C. Shapiro, *supra* note 4, at 19-20.

42 *See* Teece & Sherry, *supra* note 1, at 1939.

tary patents within the standard.[43] That is, patent prices are limited by their context. In addition, patent holders without any downstream operations (upstream firms) are constrained by the elasticity of demand for the product in the end market.[44] While vertically integrated firms can have incentives to raise rival downstream firms' prices through their licensing terms, they may also be open to cross-licensing agreements with other integrated companies, which can hold down royalty rates as well. And lastly, all firms face dynamic constraints through the formal standard-setting process. Because standards evolve over time, and many high-technology standards pass through multiple versions—mobile telecom is on its third generation (3G) currently, with 3.5G, 4G, and beyond 4G already under development—any unreasonable pricing or abuse of market power can be punished in future iterations of the standard.[45] Firms that act opportunistically in today's version of a standard may find their technologies excluded, avoided, or at least minimized in votes on tomorrow's version of the standard.

WHILE OWNERS OF IPR MAY BENEFIT FROM A BROADER ADOPTION OF THEIR TECHNOLOGIES, IMPLEMENTERS —AS WELL AS CONSUMERS— ALSO BENEFIT FROM THE OPPORTUNITY TO GAIN ACCESS TO AND USE INNOVATIVE SUPERIOR TECHNOLOGIES.

Finally, one last but important, overlooked issue relates to why, if standardization increased the value of a given IPR, the essential patent holder should not capture part of that value. The implicit assumption in the ex post opportunism claim is that all of the additional value created by the standardization process improperly accrues to patent licensors. But formal standardization is a costly cooperative effort that requires both innovators and implementers. There is no reason to assign all of the rents to one or the other. Thus, while owners of IPR may benefit from a broader adoption of their technologies, implementers—as well as consumers—also benefit from the opportunity to gain access to and use innovative superior technologies. This sharing of benefits helps to ensure participation incentives.

C. ROYALTY STACKING

Royalty stacking can be explained simply. A firm wishing to produce a good, especially one embodying a technical standard, typically needs to acquire rights to the intellectual property underlying the good. When that good is comprised of multiple complementary components, each of which is necessary for produc-

43 D. Geradin et al., Royalty Stacking in High Tech Industries: Separating Myth From Reality (Dec. 2006) (mimeo), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=949599 .

44 K. Schmidt, Licensing Complementary Patents and Vertical Integration (Nov. 2006) (mimeo, University of Munich), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=944169.

45 For a discussion of such dynamic and institutional constraints, *see* DeLacey, *supra* note 11.

tion and each of which is covered by patents held by separate firms, the aggregate royalty fees for licensing all of the required pieces can, it is sometimes suggested, add up to a very large amount—perhaps so large that it is no longer economical for the manufacturing firm to make the good.[46] This can allegedly happen even if each component's patent is offered on reasonable terms when considered individually because stacking up so many reasonable terms could lead to an unreasonable sum in the aggregate.

At least five factors are implicit in this royalty stacking proposition. First, innovation must be sequential and cumulative, so that the patents are overlapping and interrelated. Otherwise, the royalties could not stack up. Second, there must be many patents for a given product, such as one embodying a technical standard. Otherwise, the stack would be small and either inconsequential or relatively easy to negotiate out of. Third, the many patents must be held by numerous, distinct rights holders. Otherwise, negotiating the use of the many patents would be fairly straightforward, involving a limited number of bilateral discussions. Fourth, the given licensee or all licensees must have no patents to trade with licensors. Otherwise, cross-licensing would drastically reduce the risk of royalty stacking.[47] Finally, one additional assumption is required by the theoretical model to consistently predict royalty stacking: all patents should command identical rates.[48] That is, most discussions of royalty stacking (and the sole formal model) are based on inferences of one rate multiplied by all participants.[49] No allowance is made for heterogeneity among IP and IP holders.

While the first two assumptions, cumulative innovation and the presence of numerous patents, appear to hold in a great many high-technology industries, the remaining three assumptions are open to considerable debate. Assumption three, concerning fragmented rights holders, appears not to hold in some ICT industries. Empirical evidence is sparse, but existing papers on the software and

46 The roots of such propositions as royalty stacking and patent thickets can be traced back to Heller and Eisenberg who, in a seminal article published in 1998, suggest that the combination of pioneer and follow-on inventors could lead to "too many" patents in biomedical research, ending in a "tragedy of the anti-commons." *See* M. Heller & R. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 Sci. 698-701 (1998) (Patent policy might permit "...too many upstream patent owners to stack licenses on top of the future discoveries of downstream users."). The anti-commons claims have not gone unchallenged. *See* R. Epstein & B. Kuhlik, *Is There a Biomedical Anticommons?* Reg. 55 (2004). See also, R. Epstein, Studying the Course: Property Rights in Genetic Material (U. Chicago Law and Economics, Olin Working Paper No. 152, Mar. 2003).

47 This assumption raises the point that in most high-technology industries, most licensors are also licensees, and therefore will be able to reduce any eventual royalty stacking.

48 Lemley and Shapiro present the only formal model of the theory of which we are aware. *See* Lemley & Shapiro, *supra* note 4. In addition to assuming that all patents are of equal value, Lemley and Shapiro assume that all licensing negotiations occur simultaneously and that patent holders are unable to fully appropriate the rents generated by their inventions.

49 *Id.*

the mobile telecommunications industries suggest more concentration than the theoretical arguments suppose.[50] Regarding assumption four that cross-licensing is unavailable or inadequate, evidence that this is a widespread problem is again weak. For example, an empirical study of the semiconductor industry finds high levels of patenting and numerous distinct rights holders,[51] but also finds substantial evidence of cross-licensing.[52]

The last of the five assumptions, that all patents should command identical licensing rates, is the most restrictive. This view ignores the extensive literature on intellectual property valuation that makes clear all patents are not created equal.[53] When the crucial aspect of disparate patent value is incorporated into the royalty stacking theoretical model, however, the predictions are no longer so clear cut. While royalty stacking is still a possibility, it is not a foregone conclusion: some equilibria exhibit stacking while many others do not.[54] In other words, the royalty stacking theory is not robust. This finding is not surprising when you consider the ultimate goal for firms participating in standard-setting efforts: no one makes money if the product does not sell.

V. Proposals to Reshape the FRAND Model: Encouraging Ex Ante Competition

As the preceding section illustrates, many of the criticisms made against the present FRAND regime are based on inaccurate or incomplete premises. In particular, the alleged problems of hold-up and royalty stacking, while perhaps real problems in isolated incidents, do not withstand serious analysis when they try to move toward generally applicable theories. Despite the shaky underpinnings for broad application, however, in recent years a number of proposals have been made in a variety of settings to modify the current FRAND regime to mandate

50 *See* M. Noel & M. Schankerman, Strategic Patenting and Software Innovation (Center for Economic Policy Research, Discussion Paper No. 5701, June 2006); D. Geradin et al., *supra* note 43.

51 *See* R. Ziedonis, *Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms*, 50(6) Mgmt. Sci. 804 (2004).

52 As Shapiro observes, "The impressive rate of innovation in the semiconductor industry in the presence of a web of such cross-licenses offers direct empirical support for the view that these cross-licenses promote rather than stifle innovation." Shapiro, *supra* note 4, at 13.

53 For different licensing approaches, *see* M. Kamien, *Patent Licensing*, *in* Handbook of Game Theory with Economic Applications 331-54 (R. Aumann & S. Hart eds., vol. 1., 1992); L. Johnston & R. Rapp, *Modern Methods for the Valuation of Intellectual Property*, 532 PLI/Pat 817, 817-42 (1998); G. Smith & R. Parr, *Valuation of Intellectual Property and Intangible Assets*, *in* The New Role of Intellectual Property in Commercial Transaction (M. Simensky & L. Bryer eds., 1994). F. Denton & P. Heald, *Random Walks, Non-Cooperative Games, and the Complex Mathematics of Patent Pricing*, 55 Rutgers L. Rev. (2003).

54 *See* Geradin et al., *supra* note 43.

ex ante licensing. The proposed reforms are offered as ensuring greater predictability for standard implementers, as well as a means for preventing hold up and royalty stacking scenarios by promoting ex ante competition between technologies. Many of these proposals, however, have not been carefully thought through, and would be difficult or impossible to implement. Moreover, some of the proposals could raise oligopsony power concerns and artificially depress the royalties that should be paid to innovators.

A. THE SWANSON-BAUMOL MODEL OF EX ANTE AUCTIONS

In a recent paper, Swanson and Baumol suggest that ex ante price competition could take place under a system of auctions run by the SSO.[55] They propose the following thought experiment to illustrate their ex ante approach. During the development phase of a standard, the SSO would hold an auction between different technologies. IPR holders vying to have their technology incorporated in the standard would submit offers to license it to downstream standard implementers for a fee (the royalty) calculated per unit of output. The SSO members would then choose which technology should win the auction and be embodied in the standard. Swanson and Baumol argue that the outcome of such an auction would provide a benchmark for what is a fair and reasonable royalty, as it would fully reflect the degree of competition between IPR holders existing prior to adoption of the standard. When two technologies compete against each other, competitive pressure would result in lower royalties since profits in license revenues would be competed away. This reasonable royalty would of course be constrained by the price of the final product in the downstream market. If a proposed royalty were too high, such that it would result in downstream manufacturers producing at a loss, they would simply veto the technology during the auction.

As a thought experiment, ex ante competition through SSO-sponsored auctions is theoretically attractive and has the potential to lead to efficiency-maximizing outcomes. The model propounded by Swanson and Baumol has, however, some inherent limitations, most of which relate to its practical application. First, their model is based on a simple structure that makes the modeling tractable: one company holds one patent defining one good. Unfortunately, this does not reflect the reality of modern standards, which are usually comprised of tens of firms that hold hundreds or thousands of patents that define one complex good with multiple facets or components. In such a multidimensional setting, auction design quickly gets extremely complicated.[56] It is not merely a matter of picking the lowest cost option for a well-defined product. Instead, auction bidders would need to evaluate the options on price plus a host of other dimensions, including technical superiority, ease of implementation, and so forth.

55 *See* D. Swanson & W. Baumol, *Reasonable and Nondiscriminatory (RAND) Royalties, Standards Selection, and Control of Market Power*, 73(1) Antitrust L.J. (2005).

56 *See, e.g.*, P. Dasgupta & E. Maskin, *Efficient Auctions*, 115(2) Q. J. Econ. 341-88 (2000); F. Branco *The Design of Multidimensional Auctions*, 28(1) RAND J. Econ. 63-81 (1997).

Related to this point, the engineers active in SSOs typically make hundreds of different technology choices on the path to a given standard. Hundreds of items, major and minor, need attention before the standard can be defined. Taking this point to its logical conclusion, an SSO would need to run hundreds of auctions—one for each component—to fully specify the licensing price of a standard. Moreover, since many of the components rely on other components, the various auctions would be linked in complicated ways, and might need to be conducted in a particular sequence. Even if it were feasible to arrange, a multi-tiered auction of this sort would require a tremendous ex ante investment from SSO members.

Nor is it entirely clear what ex ante really means in practice. As mentioned earlier, standards generally evolve over time. Would an auction need to be held each time a technology component were modified? Every time a new technological option surfaced? Just before the final vote for a new version of the standard? These timing decisions would likely have a significant impact on the outcome of the auction.

The second assumption embedded in the Swanson and Baumol model is that competing technologies for every relevant portion of the standard will be available. As noted above, a standard will usually comprise two categories of technologies:

(1) those for which there were, at the time of development, one or several alternatives and

(2) those for which there was no suitable alternative.[57]

While price competition may take place between competing technologies,[58] there is no place for such competition between peerless technologies for which no adequate substitute exists. In this (common) scenario, ex ante and ex post licensing will be identical, as holders of non-substitutable technologies will have the same level of market power before and after a standard is adopted.[59] The model therefore offers few insights on instances where complements are stan-

57 For simplicity we ignore the intermediate category of imperfect substitutes. In that case, competition of a sort does exist, but the superior option will nonetheless have some degree of market power before the SSO determines the standard.

58 *See* Skitol, *supra* note 4, at 734:

> a patent owner's own perspective on RAND terms can be expected to be quite different at the ex ante stage—when it may be competing with *alternative* technology offerings for the proposed standard—than ex post (after the standard has been adopted with the owner's technology and those alternatives are no longer viable). (emphasis added)

59 Note that holdup theory requires sunk investments, not standard approval necessarily. Knowing that a particular component has only one feasible technical solution, implementers would be unlikely to make irreversible investments in advance of securing access to the necessary IPRs. Holdup, then, would be possible only when the IP holder did not disclose its patents at all.

dardized, save for the possibility of reducing royalties for portions of the standard for which substitutes exist, but which will remain complementary to other IPR incorporated into the standard.

Another drawback of the Swanson and Baumol model of ex ante auctions, or of any ex ante approach for that matter, is that it may hinder innovation in those cases in which the value of an invention is unclear at the moment of standardization. The significance, technical merit, and full value of an invention covered by IPRs may only be revealed over time, as the standard is implemented and adopted by end users. Freezing royalty levels and other terms and conditions at a moment where imperfect information is available to SSO members has the potential to lead to suboptimal technological choices if firms were to vote on price and other tangible elements of an offering without fully understanding the differences across technology. Plus, as information developed over time parties could have strong incentives to renegotiate, which would mean incurring the transaction costs of licensing negotiations at multiple points. Furthermore, firms with unknown technologies can benefit from introductory pricing, where initial fees are set low to encourage adoption while later fees are higher to recoup investments.[60] This kind of dynamic pricing would be made more difficult by an ex ante auction.

The final limitation raises more serious concerns. The ex ante auction model assumes that owners of essential IPR will seek to charge a royalty that is high enough to compensate their research and development efforts and low enough to win the auction and see their technology embedded in the standard. Some essential rights holders may, however, behave strategically. For instance, implementers within an SSO may use their collective power by holding a mandatory auction (either in the SSO itself or through the facilitation and encouragement of the SSO) that drives royalties below levels sufficient to reward innovation. Alternatively, rights holders might commit to charge very low royalties in order to exclude competitors from the standard concerned.[61] As seen above, vertically integrated IPR owners, for instance, have a distinct advantage over pure innova-

60 J. Farrell & P. Klemperer, *Coordination and Lock-in: Competition with Switching Costs and Network Effects*, *in* Handbook of Industrial Organization (M. Armstrong & R. Porter eds., vol. 3, forthcoming 2007).

61 Swanson & Baumol assume that SSO members will not manipulate voting. *See* Swanson & Baumol, *supra* note 55, at 17 ("We further assume that the operative SSO voting (or other decision-making) process would not be unduly susceptible to being skewed or biased by one or more SSO members, much as many antitrust decisions in the area have effectively required."). Further, they assume the absence of vertically integrated firms among essential patent holders. *Id.* at 19 ("We further assume that many downstream firms use the IP to produce perfect substitutes, but that patent owners do not also produce final products."). This of course changes the dynamics of the model as pure innovators will have much lower incentives to game the auction process along the lines described above.

tors when it comes to setting royalty rates.[62] Their revenues do not depend on the royalties charged given that they can take their profit downstream in the market for the products embodying the standard. By eliminating the pure innovator's technology during an auction, vertically integrated IPR owners stand to gain in at least two ways:

(1) they would weaken a firm that would be a rival in future innovation races; and

(2) they would be best positioned to manufacture products implementing the standard embedding their own technology.

If such a scenario was to occur—not a remote possibility considering the asymmetry of interests between SSO members—it would amount to transforming standard-setting processes into a mechanism which renders a judgment on comparative value, favoring one business model (vertical integration) over another (pure innovator).

B. PROPOSALS FOR COLLECTIVE NEGOTIATIONS OF ROYALTIES

Other authors suggest an ex ante regime based on joint negotiations of royalties between and among potential licensors and licensees before a standard is formally adopted.[63] The main difference with the Swanson and Baumol model discussed above lies in the fact that royalties would not be determined ex ante in an auction, but through collective action in the form of joint negotiations. It is this element of collective action which renders it particularly problematic.

While, voluntary ex ante term disclosure may enhance the ability of licensors and licensees to negotiate mutually advantageous terms, mandatory term disclosure poses numerous perils. It can lead to a one-size-fits-all solution that would not only homogenize licensing conditions in inefficient ways, but would also distort the way standards development now fosters competition between and amongst implementing standards participants. In the absence of mandatory disclosure of licensing terms, standard implementers may make different strategic choices. For instance, an implementer may decide to negotiate a license for patents even before it is certain they will become essential, as early negotiations may allow it to obtain better license terms than those available after the standard is adopted. These advantageous license terms would then give the firm a competitive advantage over a late-to-license implementer, whose costs of implementa-

62 P. Klemperer, *Auctions with Almost Common Values: The "Wallet Game" and its Applications*, 42(3-5) Eur. Econ. Rev. 757-69 (1998); P. Klemperer, M. Huang, & J. Bulow, *Toeholds and Takeovers*, 107(3) J. Pol. Econ. 427-54 (1999).

63 *See, e.g.*, Ohana et al., *supra* note 35; *See* Skitol, *supra* note 4, at 727.

tion might be higher. Compulsory disclosure of licensing terms would eliminate that competitive aspect of standardization processes.[64]

Joint ex ante negotiations of royalties before the adoption of a standard also could trigger serious antitrust concerns to the extent they require competing downstream firms to collaborate during royalty negotiations.[65] Such collaboration could involve restrictions of competition and could therefore fall foul of Article 81(1) of the EC Treaty and Section 1 of the Sherman Act in the United States, or equivalent antitrust provisions in other jurisdictions on several grounds.

JOINT EX ANTE NEGOTIATIONS OF ROYALTIES BEFORE THE ADOPTION OF A STANDARD ALSO COULD TRIGGER SERIOUS ANTITRUST CONCERNS TO THE EXTENT THEY REQUIRE COMPETING DOWNSTREAM FIRMS TO COLLABORATE DURING ROYALTY NEGOTIATIONS.

First, joint negotiations could lead to serious anticompetitive exercises of oligopsony power.[66] As in classic examples of the exercise of buyer power,[67] the negotiations would be primarily aimed at depressing the royalties (i.e., the price) which standard implementers would pay for gain-

64 *See* R. Taffet, Ex Ante Licensing in Standards Development: Myths and Reality, Remarks at the American Intellectual Property Law Association Spring Meeting, Chicago, IL (May 4, 2006), at 9-10.

65 *See* Swanson & Baumol, *supra* note 55, at 12-13:

> The standardization process typically involves consultation and agreements among firms that are often competing buyers of IP and also may be competing sellers in the downstream product markets. While joint decision making by competitors can sometimes promote the general welfare, it always entails the danger of misbehavior for anticompetitive purposes, such as the threat of behavior aimed at collusively reducing the price paid for intellectual property.

Nonetheless, as noted by Chairman Majoras of the U.S. Federal Trade Commission, "joint ex ante royalty discussions that are reasonably necessary to avoid hold up do not warrant per se condemnation. Rather, they merit the balancing undertaken in a rule of reason review." *See* D. Majoras, Recognizing the procompetitive potential of royalty discussions in standard setting, Remarks delivered at Stanford University (Sept. 23, 2005), *available at* http://www.ftc.gov/speeches/majoras/050923stanford.pdf.

66 *See* Swanson & Baumol, *supra* note 55, at 12-13; Teece & Sherry, *supra* note 1, at 1955:

> The SSO members would, in effect, say to the patent holder, 'We will collectively reject a standard that incorporates your patented technology unless you agree to license it to us at pre-specified rates that we collectively find acceptable.' In other contexts, this clearly would amount to a group boycott.

For a perfect example of this risk, *see* Skitol, *supra* note 4, at 729, who considers that potential licensees should make use of their buyer power to extract what they consider as a reasonable royalty rate from a potential licensors ("A patent owner's refusal to accept terms satisfactory to the group as a whole would cause the group to consider alternatives to the use of that owner's technology.").

67 *See* U.K. Office of Fair Trading, The Welfare Consequences of the Exercise of Buyer Power, no. 16 (1998).

ing access to essential IPR.[68] This would diminish the licensors' incentives to invest in research and development (R&D) and could therefore potentially hamper innovation. Joint ex ante negotiations could also give rise to the risk that potential licensees would threaten to opt for an alternative technology unless the potential licensor offered a royalty they considered appropriate. Such a threat could amount to a collective boycott.

Second, required ex ante negotiations generating uniform licensing terms would lead to a homogenization of the conditions of competition and could facilitate collusion in the downstream product market. This is a risk in any collective price negotiation, but within standards it is a special concern in light of the different objectives of firms according to their business model. Vertically integrated firms have an incentive to raise the prices facing their downstream competitors without any relevant IP in the standard. Integrated firms could therefore use the ex ante collective bargaining to signal high royalties to be charged to other downstream players, with the effect of either limiting the competition downstream (if royalties were high enough) or at least disadvantaging other downstream rivals.

Finally, if joint negotiations produce a one-size-fits-all approach, it would prevent efficient discrimination in licensing conditions. Because standard implementers are not all equally situated (as, for instance, some have wider patent portfolios to offer in exchange than others, or cover broader geographic areas, etc.), charging a similar level of royalties to all of them would prevent the adoption of flexible deals that take into account their meaningful differences.

The question then arises whether, even assuming that a proposed joint negotiations regime could survive summary condemnation under per se rules, does it benefit from the application of Article 81(3) of the EC Treaty or generate sufficient countervailing efficiencies under a rule of reason regime?[69] A detailed analysis of these requirements goes beyond the scope of the present paper, so we address only certain features that, in our view, temper a finding that such collec-

---

68 *See* Teece & Sherry, *supra* note 1, at 1955:

> One key issue concerning patents is whether the patent holder must announce the terms for a patent license in advance. If so, there are potential antitrust concerns. Typically, the other participants in the SSO are the most likely potential licensees for the patent. This raises the potential for collusive, oligopolistic 'price fixing' in the technology market.

For a different view, *see* Skitol, *supra* note 4, at 739.

69 In a December 2005 press release, the European Commission took note of the fact that ETSI's General Assembly had established a group with the mission to examine possible changes to ETSI's standard-setting rules, in particular on the issue of ex ante licensing. It stated that it had "indicated in its Guidelines on the application of Article 81 of the EC Treaty to technology transfer agreements (see IP/04/470) that such *ex ante* licensing can have pro-competitive benefits when subject to appropriate

footnote 69 cont'd on next page

tive negotiations could be deemed, on balance, to be in line with competition law.[70] For instance, the discussion that follows suggests that such negotiations could not be justified under either.

First, a collective ex ante negotiation system would have an adverse impact on the rewards granted to licensors, in particular those obtainable by non-vertically integrated holders of essential IPR. This is a particular threat in SSOs because non-vertically integrated IPR holders are virtually always in the minority.[71] It is therefore possible that a collective negotiation regime would not promote technical innovation or economic progress, but on the contrary negatively affect these objectives by under compensating innovators. It also is far from certain that end consumers would benefit from what would essentially amount to an exercise in rent-shifting between innovators and implementers. There is no empirical foundation to the proposition that the payment of lower royalties to innovators would automatically lead to lower selling prices of the products implementing the standard. Prices at the end-user level depend on a complex number of factors, not the least of which is the extent to which licensing terms impact incremental costs and the level of competition between standard implementers at the downstream product level.[72] Just as higher royalties could be internalized by such manufacturers, lower royalties would not necessarily be passed along to consumers. Nor is it clear that a system of joint negotiations of royalty rates is necessary (i.e., the least restrictive means available) to achieve the stated objective of the proponents of this ex ante regime (i.e., preventing perceived risks of ex post opportunism and increasing certainty as to the implementation

footnote 69 cont'd
safeguards" and that it would follow ETSI's forthcoming discussions with interest. *See* Press Release, European Commission, IP/05/1565, Commission welcomes changes in ETSI IPR rules to prevent 'patent ambush', (Dec. 12, 2005).

This statement from the Commission cannot be interpreted as meaning that it is prima facie favorable to the joint negotiations approach or to any of the other reforms proposed by firms and commentators in the framework of this ETSI group. It only suggests that the Commission will carefully review the various proposals made to ETSI to ensure their compatibility with EC competition rules. In fact, the same press release made clear that the Commission had carefully reviewed under Article 81 EC a prior amendment to the ETSI IPR rules designed to limit the risk of "patent ambush" and that it had cleared it subject to some modifications of its content.

70 *See* Swanson & Baumol, supra note 55, at 13-14 ("In the case of the typical SSO [...] the integration and efficiencies needed to justify outright collective bargaining on royalties are in short supply."). *See* Shapiro, *supra* note 10 ("While the law has typically looked for integration and risk-sharing among collaborators in order to classify cooperation as a joint venture and escape per se condemnation, [...] the essence of cooperative standard setting is not the sharing of risks associated with specific investments, or the integration of operations.").

71 Teece & Sherry, *supra* note 1.

72 J. Tirole, The Theory of Industrial Organization 66-75 (MIT Press 1997).

cost of a given standard).[73] As discussed, bilateral ex ante discussion, negotiation, and licensing often occurs today. In light of this, joint negotiations produce no pro-competitive benefits.

C. MANDATORY EX ANTE DISCLOSURE OF LICENSING TERMS

Recognizing the significant antitrust liability inherent in joint negotiations, some proposals have been made within SSOs for the adoption of a policy of mandatory ex ante disclosure of licensing terms. Under such an ex ante policy, SSO members would be required to disclose, prior to the adoption of a given standard, the upper limit of the consideration they would expect in order to license their essential IPRs, perhaps along with the most restrictive terms the licensor would seek. It should be noted that mere royalty rate disclosure is likely to be misleading. The picture it would convey would necessarily be imprecise, as the rate itself is but one of the various elements of consideration that need to be agreed on by licensor and licensee.

Although the resulting antitrust risk is markedly lower than that arising from joint negotiations, mandatory ex ante disclosure also has the potential to run afoul of competition provisions. If disclosure led to inefficiently uniform licensing terms and homogenous conditions of competition, the same complaints as for joint negotiations would hold. Moreover, term disclosure could facilitate anti-competitive cooperation designed to put pressure on the potential licensor to lower its royalties. Such a threat could create oligopsony concerns. Ex ante term disclosure could also facilitate collusion in the downstream product market, in that the announcements could be used as price signals obviating the need for any explicit coordination.

To illustrate this last claim, consider an industry where downstream manufacturers require various complementary patents to operate lawfully. Suppose further that the industry is populated by a number of firms where some are vertically integrated, some are pure innovation (upstream) companies, and some are pure downstream manufacturers (with no IP). In an industry like this, the vertically integrated firms have incentives to discriminate against their downstream competitors.[74] Each of the vertically integrated companies would like to see its downstream competitors pay a very high aggregate royalty rate. This could hap-

73 Deborah Majoras made this very point in a recent speech: "It may also be appropriate to consider whether joint ex ante royalty discussions are reasonably necessary to mitigate holdup." *See*, Majoras, *supra* note 68, at 10. *See, e.g.,* U.S. Fed. Trade Comm'n & U.S. Dep't Justice, Antitrust Guidelines for Collaborations among Competitors § 3.36(b) (Apr. 2000) (noting that "[t]he Agencies consider only those efficiencies for which the relevant agreement is reasonably necessary") and U.S. Fed. Trade Comm'n & U.S. Dep't Justice, Antitrust Guidelines for the Licensing of Intellectual Property § 4.2 (Apr. 1995) ("If it is clear that the parties could have achieved similar efficiencies by means that are significantly less restrictive, then the Agencies will not give weight to the parties' efficiency claim."). *See* Majoras, *supra* note 65, at 9-10.

74 *See, e.g.*, Schmidt, *supra* note 44.

pen if each of the vertically integrated companies sets a moderately high royalty rate for its patents, or alternatively if a subset of those IP holders set very high royalty rates for their patents. To achieve that end, each vertically integrated firm could use the obligation to disclose its maximum royalty to the SSO as a device to signal to the other vertically integrated firms what it intends to charge to the downstream competitors. Disclosure of the maximum royalty rate would thus allow the vertically integrated companies to collectively raise their downstream rivals' costs. This signaling device not only suppresses the need for explicit collusion, it would also allow the vertically integrated companies to justify their common rate as reasonable.

A more subtle, but equally troubling, possibility of mandatory ex ante licensing disclosures relates to those firms that hold patents for defensive purposes only. For instance, some firms focus on downstream operations and take patents only as bargaining devices should they find themselves, say, sued for patent infringement by another firm. These firms have no active plans to license their patents, and instead operate on an implicit cross-licensing basis for rival firms that might infringe their IPRs. If declaration of maximum terms is mandatory, however, declaring royalty-free and permissive terms and conditions would eliminate a patent portfolio's worth as a defensive mechanism for cross-licensing and lawsuit avoidance. At the same time, declaring high royalty rates and restrictive terms can lead to a firm's technology being bypassed during standard development stages. Mandatory disclosure therefore cuts out a great deal of operational flexibility, all for a group that would likely not contribute to any royalty stacking or hold-up even if it were able.

MANDATORY DISCLOSURE CUTS OUT A GREAT DEAL OF OPERATIONAL FLEXIBILITY, ALL FOR A GROUP THAT WOULD LIKELY NOT CONTRIBUTE TO ANY ROYALTY STACKING OR HOLD-UP EVEN IF IT WERE ABLE.

One SSO is already implementing an ex ante licensing term disclosure policy. The VMEbus International Trade Association (VITA) recently received a business letter review from the U.S. Department of Justice stating that it had no present intention to challenge, unless anticompetitive in practice, a proposal for their SSO arm (VSO) to execute a significant new patent policy requiring upfront disclosure of patents and patent licensing terms in connection with VMEbus standard-setting activities.[75] Under VSO's new policy, each member must, inter alia, declare the maximum royalty rate for all the patent claims that it represents, owns, or controls and that may become essential to implement the standard in question. In addition, each VITA member company must disclose

75 VSO is a non-profit organization that develops and promotes standards for VMEbus computer architecture. *See* VITA Patent Policy, at http://www.vita.com/disclosure/VITA%20Patent%20Policy%20section%2010%20draft.pdf; Letter from T. Barnett, Assistant Attorney General, U.S. Department of Justice, Antitrust Division, to Robert A. Skitol, Drinker, Biddle & Realth (Oct. 30, 2006), *available at* http://www.usdoj.gov/atr/public/busreview/219380.htm.

the most restrictive non-royalty terms that it will request. Such declarations are irrevocable, although patent holders may submit subsequent declarations with less restrictive licensing terms (including lower royalties). In other words, the disclosure is intended as a binding price cap for licensors. Any further joint discussion of terms within the SSO was prohibited in this proposal.

Nonetheless, a danger of such a policy, as mentioned above, is under-compensation of IP holders. For example, suppose that two firms have patented technology relevant for some component of a new standard. In this case, the ex ante disclosure process could easily resemble the ex ante auction along the lines of Swanson and Baumol. This raises several issues. First, if one technology option were superior to another, unless this fact was widely known by SSO members the lesser technology would drive the license pricing. The firm with the better, but perhaps less-known, technology would have the choice of pricing its IP below the actual contribution value to the standard or losing the auction. Even if this under-compensation were not an issue in practice for VITA, the SSO members will still face a complex set of comparisons, needing to evaluate one multidimensional option against another. Moreover, as we have observed in our discussion of the Swanson and Baumol model, there could be significant risks that some IP owners could game the disclosure process by disclosing low royalties for the sole purpose of eliminating upstream firms that rely on royalties to fund their innovation. Such a predatory approach could be funded by the rents generated on downstream markets. Disclosure of licensing terms that would be taken into account for technology selection may also induce collusion as can often be observed in bidding processes. We are not suggesting that these gloomy predictions would necessarily materialize (it is in fact too early to say), but VITA-type disclosure may increase the risk of anticompetitive behavior.

It remains to be seen whether VITA's new policy will work as planned, avoiding the potential anticompetitive consequences discussed above. It is possible that VITA will manage to balance the restrictive features of its new policy with alleged pro-competitive aspects, so that the overall effect will not be anticompetitive. Regardless of this one case, however, it would be dangerous to make sweeping statements about such ex ante term disclosure policies. The devil, as they say, is in the details, so that assessments will need to be conducted on a case-by-case basis.[76]

VI. Conclusions

Concerns over possible abuses of the formal standard-setting process continue to generate significant debate. Among the topics are the risks perceived by some for opportunistic licensing behaviors, patent hold up, and royalty stacking. While a potential for such behavior exists, at this point it appears unlikely that any of

76 In fact, the U.S. Department of Justice's review and decision not to oppose VITA's proposed plan illustrates the application of a rule of reason approach.

these problems is in fact widespread. Regardless of the extent of any ex post standardization problems, however, many of the proposed ex ante solutions would likely cause more difficulties and unintended consequences than they could correct—even assuming the solution could be implemented in practice.

In this paper we have assessed various proposals for addressing supposed ex post opportunism within standard-setting and have found most of them lacking. Systems of ex ante auctions and joint negotiations appear far too dangerous a road to take, with more potential to cause harm than to fix any ex post problems with market power. Ex ante licensing term disclosures emerge as the most sensible of the proposals, but such conduct already occurs, and imposing more onerous requirements than already exist could cause more harm than good. The trick will lie in a prudent execution of term disclosure programs, and then only where the SSO members as a whole deem the risks of ex post abuse great enough to warrant instituting mechanisms that go beyond the guarantees provided by existing FRAND commitments and voluntary ex ante licensing.

In the end, the extant FRAND regime typical of modern SSOs appears a remarkable compromise. It balances the danger to standard implementers that IP holders might refuse to license or offer only unreasonable terms against the danger to IP holders that standard implementers might press for unreasonably low royalty rates that prevent an adequate return on R&D investments. Before we replace this flexible arrangement that appears to work in the majority of instances, we should be sure that the perceived problems are indeed widespread and that the proposed solutions to them represent genuine improvements. ◀

Volume 3 | Number 1 | Spring 2007

Competition Policy International

Article 82 EC and Intellectual Property: The State of the Law Pending the Judgment in Microsoft v. Commission

Maurits Dolmans, Robert O'Donoghue, and Paul-John Loewenthal

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Are Article 82 EC and Intellectual Property Interoperable? The State of the Law Pending the Judgment in *Microsoft v. Commission*

Maurits Dolmans, Robert O'Donoghue, and Paul-John Loewenthal

The objectives of intellectual property rights (IPR) and competition law are essentially the same: both promote innovation to the benefit of consumers. IPRs are, however blunt instruments that strike the right balance in general, but in exceptional individual situations may not achieve (and may sometimes even obstruct) the innovation policy goal. Competition law is a useful tool to redress the balance in these situations, and the European Commission and EC courts have recognized that in exceptional cases the exercise of IPRs may infringe competition law. This article examines the extent to which Article 82 EC restricts the use of IPRs, pending the judgment of the CFI in Case T-201/04, *Microsoft v. Commission*.

Maurits Dolmans, Robert O'Donoghue, and Paul-John Loewenthal practice law with Cleary Gottlieb Steen & Hamilton LLP. The authors are grateful for the support from their colleague David Lyons. Cleary Gottlieb was counsel for IMS Health and represents third parties supporting the European Commission in the Court of First Instance proceedings in the *Microsoft EC* case, but the views expressed and arguments made in this paper do not necessarily reflect the views of the firm or its clients. This article is in part based on a paper written for the IBC Conference on Jun. 3, 2006.

I. Introduction

Despite the lack of complete harmonization in respects of intellectual property rights (IPR), EC law has made significant encroachments on the entitlements of IPR holders under the free movement of goods rules. More recently, it could be argued that competition law has also been used as a harmonization tool to take off the sharp edges of intellectual property law. This article examines the extent to which Article 82 of the EC Treaty, prohibiting abuse of a dominant position, restricts the use of IPRs and in particular to what extent it requires a firm to grant a compulsory license of its IPRs to third parties. When this article was originally planned, the authors expected to have the judgment of the European Court of First Instance in Case T-201/04, *Microsoft v. Commission*, which is expected to deal with the interface between intellectual property and competition, but the wheels of justice turn more slowly than expected. This article is, therefore, an overview of the current status, with particular reference to the arguments made in the *Microsoft EC* hearings and the European Commission's 2004 Decision (2004 MS Decision), that can be used as background when judgment is rendered.[1]

II. Competition Law and the Essential Function of IPR

There is considerable ongoing debate about the role of IPRs as engines driving innovation. The traditional goal of IPRs is perfectly summarized by Abraham Lincoln's statement that patent law "secured to the inventor, for a limited time, the exclusive use of his invention; and thereby added the fuel of interest to the fire of genius, in the discovery and production of new and useful things."[2] On the other hand, concern has been expressed that too many IPRs are being granted and for overly broad subject matters. There is testimony, for example in the context of the U.S. agencies' ongoing review of the interface between intellectual property (IP) and antitrust laws, that patent thickets can stifle innovation and increase costs.[3] There is also evidence that such strategies are pursued deliberately for the

1 Commission Decision of Mar. 24, 2004 [hereinafter 2004 MS Decision], Case COMP C-3/37.792, Commission v. Microsoft Corporation [hereinafter *Microsoft EC*]. Findings of fact and law by the Commission in its 2004 MS Decision are subject to dispute before the CFI.

2 A. Lincoln, *Lecture on Discoveries and Inventions*, *in* Collected Works of Abraham Lincoln (R. Basler, ed., 1953) (1858).

3 *See, e.g.*, U.S. Fed. Trade Comm'n, To Promote Innovation: The Proper Balance of Competition & Patent Law and Policy 165 (Oct. 2003), *available at* http://www.ftc.gov/os/2003/10/innovationrpt.pdf ("Many panelists and participants expressed the view that software and Internet patents are impeding innovation."). And also Shapiro:

> In short, our patent system, while surely a spur to innovation overall, is in danger of imposing an unnecessary drag on innovation by enabling multiple rights owners to tax new products, processes and even business methods. The vast number of patents currently being issued creates a very real danger that a single product or service will

sole purpose of excluding rivals, that a much larger number of patents have been granted in recent years, that the scope of such patents is broader than in the past, and that a greater number of patents receive unmeritorious protection.[4]

Similarly, the open source software community (which relies on the existence of copyright to create a framework within which its license provisions are enforceable) is scathing about the role of patents as potential threats to innovation.[5] Criticism is particularly pronounced in the United States: "While there is a formal process of patent examination, in practice the system seems more akin to a registration system: In many cases it appears that a determined patentee can get almost any award he seeks."[6]

---

footnote 3 cont'd

> infringe on many patents. Worse yet, many patents cover products or processes already being widely used when the patent issued, making it harder for the companies actually building businesses and manufacturing products to invent around these patents. Add in the fact that a patent holder can seek injunctive relief, i.e., can threaten to shut down the operations of the infringing company, and the possibility for hold up becomes all too real.

C. Shapiro, *Navigating the Patent Thicket: Cross Licences, Patent Pools and Standard Setting*, in Innovation Policy and the Economy (A. Jaffe et al. eds., 2001), *available at* http://faculty.haas.berkeley.edu/shapiro/thicket.pdf

4 Evidence submitted to the U.S. Federal Trade Commission suggested that companies sometimes reallocate significant portions of developers' resources to increase their patent portfolio for purely defensive reasons and that the engineers' time dedicated to assisting in the filing of defensive patents, which "have no...innovative value in and of themselves," could have been spent on developing new technologies (*id*. at 9).

5 For example, at the OSDL Enterprise Linux Summit held from January 31 to February 2, 2005, Linus Torvalds, the developer of the Linux kernel, stated:

> Are software patents useful? That's pretty clearly not the case. Software patents are clearly a problem and one that the open-source community has been aware of during the last five years. And proprietary vendors are starting to see it's a problem too.

Brian Behlendorf, co-founder of the Apache Web server software, opined:

> If you could not patent software algorithms or ideas, how much of the money spent on writing software would go away? How much innovation would disappear? How much investment in that innovation would disappear? I don't think any would disappear?

Mitch Kapor, chairman of the Mozilla foundation, referred to use of patents as an exclusionary weapon:

> We have to be concerned about [...] the use of patent WMDs. That will be the last stand of Microsoft [...]. If totally pushed to the wall because their business model no longer holds up in an era in which open source is an economically superior way to produce software, and the customers understand it, and it's cheaper and more robust, and you've got the last monopolist standing, of course they're going to unleash the WMDs. How can they not?

See also G. Ghidini, Intellectual Property & Competition Law: The Innovation Nexus (2006).

6 A. Jaffe & J. Lerner, Innovation & Its Discontents 11-21, 142 (2004). *See also* J. Cohen & M. Lemley, *Patent Scope and Innovation in the Software Industry*, 89 Cal. L. Rev. 1, 12-13 (2001); R. Merges, *As Many As Six Impossible Patents Before Breakfast: Property Rights for Business Concepts and Patent System Reform*, 14 Berkley Tech. L.J. 577, 589-91 (1999).

Some European courts appear to share this skepticism. Lord Justice Jacob wrote highly readable and controversial comments recently in his U.K. Court of Appeal judgment in the *Macrossan* case, rejecting business model and software patents:

> "18. ... people have been getting patents for these subject-matters in the USA. Since they can get them there, they must as a commercial necessity apply for them everywhere. If your competitors are getting or trying to get the weapons of business method or computer program patents you must too. An arms race in which the weapons are patents has set in. The race has naturally spread worldwide ... 19. ... Just as with arms, merely because people want them is not sufficient reason for giving them. 20. ... it is far from certain that they [software patents] have been what Sellars and Yeatman would have called a "Good Thing." *The patent system is there to provide a research and investment incentive but it has a price. That price (what economists call "transaction costs") is paid in a host of ways: ...the impediment to competition*, ... the cost of uncertainty, litigation costs and so on. There is, so far as we know, *no really hard empirical data showing that the liberalisation of what is patentable in the USA has resulted in a greater rate of innovation or investment in the excluded categories. Innovation in computer programs, for instance, proceeded at an immense speed for years before anyone thought of granting patents for them as such.* There is evidence, in the shape of the mass of US litigation about the excluded categories, that they have produced much uncertainty. (emphasis added)"[7]

As the reference to "the impediment to competition" suggests, IPRs and competition law at first sight appear to have divergent effects: IPRs grant a statutory monopoly or exclusive right, and the right to exclude others from using the subject matter of the right; competition law prevents, among other things, the exercise of monopoly power and the unlawful exclusion of competitors. On closer examination, however, the objectives of IPRs and competition law are essentially the same. Both sets of rules seek to promote innovation and investment to the benefit of consumers. This basic consistency has been recognized by the European Commission (Commission), the European Court of First Instance (CFI), and the European Court of Justice (ECJ). For example, in *Magill*—the first case dealing with the circumstances in which a refusal to license an IPR could be contrary to Article 82—Advocate General Gulmann stated that "it must not be forgotten ... copyright law—like other intellectual property rights—also serves to

7 *See* Aerotel Ltd. v. Telco Holdings Ltd & Ors, rev. 1, 2006 E.W.C.A. Civ. 1371 (Oct. 27, 2006), *available at* http://www.patent.gov.uk/2006ewcaciv1371.pdf (invention—a software-based online system which automated the completion of forms—could not be patented because it was "a computer program as such").

promote competition."[8] In other words, the common objectives of intellectual property and competition laws are to promote innovation and enhance consumer welfare.[9]

Notwithstanding this common goal of intellectual property and competition laws, the Commission and the EC courts have recognized that, in exceptional individual cases, IPRs can be too blunt an instrument, and the unrestrained exercise of IP may in these exceptional cases be found incompatible with the policy goals of competition rules. In recent years, the most controversial aspect concerns whether and in what circumstances a refusal to license an IPR may constitute an abuse of a dominant position contrary to Article 82. In cases such as *Volvo/Veng*,[10] *Renault*,[11] *Magill*,[12] *Ladbroke*,[13] and, most recently, *IMS*[14] and *Microsoft EC*, the Commission and the EC courts have developed a series of principles to address this question. These cases draw heavily on the essential facilities doctrine in U.S. law—which in exceptional cases requires firms to share facilities that cannot be duplicated by rivals—and the decisional practice and case law that have admitted

On closer examination the objectives of IPRs and competition law are essentially the same. Both sets of rules seek to promote innovation and investment to the benefit of consumers.

8 AG Opinion (Gulmann) of Jun. 1, 1994, Joined Cases C-241/91 P and C-242/91 P, Radio Telefis Eireann and Independent Television Publications Limited (RTE & ITP) v. Commission [hereinafter *Magill*], 1995 E.C.R. I-00743, at fn. 10.

9 This is also recognized under U.S. law. *See* Atari Games Corp.v. Nintendo of America, Inc., 897 F.2d 1572 (Fed. Cir.1990), in which the U.S. Court of Appeals for the Second Circuit held that:

> the aims and objectives of patent and anti trust laws may seem at first glance, wholly at odds. However, the two bodies of law are actually complementary, as both are aimed at encouraging, innovation, industry and competition.

10 Case 238/87, AB Volvo v. Erik Veng (UK) Ltd. [hereinafter *Volvo/Veng*], 1988 E.C.R. 6211.

11 Case 53/87, Consorzio italiano della componentistica di ricambio per autoveicoli and Another v. Renault, 1988 E.C.R. 6039.

12 *See* Case IV/31.851, Magill TV Guide/ITP, BBC and RTE, 1989 O.J. (L 78) 43; Case T-69/89, Radio Telefis Eireann (RTE) v. Commission, 1991 E.C.R. II-485; Case 70/89, The British Broadcasting Corporation and BBC Enterprises Ltd. (BBC) v. Commission, 1991 E.C.R. II-535; and Case T-76/89, Independent Television Publications Limited (ITP) v. Commission, 1991 E.C.R. II-575 (*aff'd in* ECJ Judgment of Apr. 6, 1995, Joined Cases C-241/91 P and C-242/91 P, Radio Telefis Eireann and Independent Television Publications Limited (RTE & ITP) v. Commission (*Magill*) [hereinafter *Magill* ECJ Judgment], 1995 E.C.R. I-00743).

13 Case T-504/93, Tiercé Ladbroke SA v. Commission, 1997 E.C.R. II-923.

14 NDC Health v. IMS Health, 2002 O.J. (L 59) 18.

exceptions to the rights of owners of physical property to refuse to deal.[15] The following sections discuss these principles.

It is a fundamental principle of EC law, enshrined in Article 295 (ex 222) of the EC Treaty and confirmed by the EC courts,[16] that the existence of national IPRs cannot be affected by the provisions of the EC Treaty. Since the existence of property is untouchable under Article 295 EC, the ECJ had to work its way around that provision. It did so by distinguishing "existence" from the "exercise" of IPRs, allowing the Commission and the Court to curb the latter where the use of IPRs could come into conflict with the policy goals of IPR and competition rules.[17] A similar principle—that curtailing the use of the right is not equivalent to eliminating it—is found also in other legal systems. The U.S. Court of Appeals for the DC Circuit stated in *United States v. Microsoft Corp* (*Microsoft III*): "The company claims an absolute and unfettered right to use its intellectual property as it wishes. ... That is no more correct than the proposition that use of one's personal property, such as a baseball bat, cannot give rise to tort liability."[18]

The existence/exercise dichotomy is helpful to get around Article 295 EC, but is not a useful balancing tool. The ECJ therefore developed the notion of the essential function of IPRs, to discern the essential policy objective of the IPR that free movement rules and competition law should respect. This policy goal, as indicated above, is to reward and encourage the initiative of creating the

15 *See* Port of Rødby, 1994 O.J. (L 55) 52; ACI - Channel Tunnel, 1994 O.J. (L 224) 28; European Night Services, 1994 O.J. (L 259) 20; Eurotunnel, 1994 O.J. (L 354) 66; IJsselcentrale, 1991 O.J. (L 28) 32; Irish Continental Group CCI Morlaix-Port of Roscoff, XXVth Competition Policy Report 43 (1996); Press Release, European Commission, IP/96/456, Port of Elsinore (May 1996); and, Case C-7/97, Oscar Bronner v. Mediaprint [hereinafter *Bronner*], 1998 E.C.R. I-7791.

Among the better articles on essential facilities are R. Subiotto *The Right to Deal with Whom One Pleases under EEC Competition Law: A Small Contribution to a Necessary Debate*, 6 Eur. Competition L. Rev. 234 (1992); K. Glazer & A. Lipsky, *Unilateral Refusals to Deal Under Section 2 of the Sherman Act*, 63 Antitrust L.J. 749 (1995); J. Temple Lang, *The Principle of Essential Facilities in European Community Competition Law—The Position Since Bronner*, 1 J. Network Indus. 375 (2000); and J. Temple Lang, *Defining Legitimate Competition: Companies' Duties to Supply Competitors, and Access to Essential Facilities*, 18 Fordham Int'l L.J. 437 (1994).

16 *See, e.g.,* Case 262/82, Coditel II, 1982 E.C.R. 3381, at para. 13 ("the existence of a right conferred by the legislation of a Member State in regard to the protection of artistic and intellectual property ... cannot be affected by the provisions of the Treaty."). *See also* Case 144/81, Keurkoop BV v. Nancy Kean Gifts BV, 1982 E.C.R. 2853, at para. 18; *Volvo/Veng*, *supra* note 10; and Case 53/87, Consorzio Italiano Della Componentistica Di Ricambio Per Autoveicoli (CICRA) and Maxicar v. Renault, 1988 E.C.R. 6039, at para. 10.

17 *See, e.g.,* Case 40-70, Sirena S.r.l. v. Eda S.r.l. and others, 1971 E.C.R. 69 and Cases 56 and 58/64, Consten and Grundig v. Commission, 1966 E.C.R. 429.

18 United States v. Microsoft Corp. [hereinafter *Microsoft III*], 253 F.3d 34 (D.C. Cir. 2001).

material and the investment in producing and marketing it.[19] If the exercise goes beyond what is necessary to fulfill the essential function, competition law may interfere.[20] If, on the other hand, an IPR owner is deprived of those rewards, or uses an IPR to stifle creative rivals, there is concern that the incentive to innovate may disappear. These principles are summarized as follows in a leading textbook on IPRs:

> "It can certainly be argued that this fencing off of intangible subject matter fulfils the economic function equivalent to that of ownership of physical property, because otherwise the incentive to optimise the value of information will be impaired or destroyed. Innovators will wait instead to be imitators and the dynamic processes which would have generated new ideas will disappear; in the end there will be little or nothing different to imitate."[21]

Thus, any interference with IPRs must be based on exceptional, clearly defined circumstances that do not materially affect incentives to innovate and therefore chill socially desirable innovation. Such circumstances may exist where an IPR is used in a manner not consistent with the essential function of IPRs, for instance, an exercise that cannot reasonably be deemed to maintain the IPR owner's research and development (R&D) incentives, especially if that exercise also stifles innovation by others in the industry.

There are indications that some industries such as the pharmaceutical and medical devices sectors may be more dependent on IPRs than others such as the information technology (IT) sector, where open source appears to have some measure of success in certain areas and non-IP intensive products such as the Internet have become ubiquitous. For the time being it appears that legally at least, all industries are treated equally, although it would be interesting to have better quantitative comparative analysis of the role of IPRs in different sectors.

This is not to say that all IPRs are also necessarily created equal. In many cases, IP law imposes conditions and limitations. For instance, the protection may be available only for a limited duration; copyrighted works must be original and only protect the expression of an idea, not its subject matter; patents must be innova-

19 *See, e.g.*, AG Opinion in *Magill*, *supra* note 8. *See also Magill* ECJ Judgment, *supra* note 12, at para. 28 (referring to the essential function of copyright as "to protect the moral rights in the work and ensure a reward for the creative effort").

20 *Id.* at para. 30.

21 W. Cornish, Intellectual Property: Patents, Copyright, Trade Marks & Allied Rights 353 (4th ed. 1999).

tive and novel with industrial application, etc. It is thought that where the legislature has struck a balance, competition authorities and courts should be reluctant in changing that balance absent exceptional circumstances. Some argue that antitrust agencies and courts have greater freedom in respect of property as to which the legislature has struck no balance—as is the case for trade secrets in the European Community. While even with trade secrets it is important to assess the impact on innovation before imposing licenses, in the absence of a unified body of trade secret law in the European Union, competition law may play a greater role in striking the balance. The *Microsoft EC* judgment will hopefully clarify that.

III. Precedent on Abusive Refusals to License

Article 82 bans "any abuse by one or more undertaking of a dominant position ... in so far as it may affect trade between Member States..." Article 82 provides no definition of "abuse", but lists four examples. It is settled case law that this list "is not an exhaustive enumeration of the abuses of a dominant position prohibited by the Treaty."[22] Article 82 and the relevant case law suggest that, broadly speaking, two types of abuse can be identified: exclusionary and exploitative abuses. The former includes conduct that limits rivals' production, markets, or technical development, discrimination that places rivals at a competitive disadvantage, and tying that creates barrier to entry in tied markets. The latter concerns excessive pricing, the imposition of unfair trading conditions and tying that imposes supplementary obligations on customers.

As regards exclusionary abuses, an overview of the case law viewed in light of Article 82(b) suggests that establishing an infringement of Article 82 requires evidence of the following four factors:

- Limitation of rivals' production, markets or technical development;
- Hindrance of maintenance or growth of competition;
- Prejudice to consumers; and
- Absence of objective, proportional justification.

These principles apply a fortiori to refusals to supply. There is, as a general rule of EC competition law, no duty on dominant companies to deal with or supply third parties. In the context of IPRs, there is also, as a general principle, no duty on dominant firms to license third parties.[23] Requiring a dominant company to contract with a third party against its will (whether by licensing arrangements or

22 Joined Cases C-395/96 P and C-396/96 P, Compagnie Maritime Belge Transport v. Commission, 2000 E.C.R. I-1365, at para. 112.

23 *Volvo/Veng*, *supra* note 10, at para. 8. See also J. Faull & A. Nikpay, The EC Law of Competition 157-8 (1999).

otherwise) is therefore an exceptional measure that should be used sparingly by competition authorities.

Each refusal to deal must be looked at on its merits in light of the specific circumstances of the market in question, including the degree of market power of the dominant firm, any applicable legislation or regulation and the types of consumer harm that might arise in that particular market setting. As Advocate General Jacobs recently stated in *Syfait*, "the factors which go to demonstrate that an undertaking's conduct in refusing to supply is either abusive or otherwise are highly dependent on the specific economic and regulatory context in which the case arises."[24]

This is not to say, however, that the conditions for an abusive refusal to deal are (or should be) open-ended or opaque. Given the vagaries of litigation and the factual peculiarities of potential exceptional circumstances (witness *Magill* and *IMS Health*), it is not possible to formulate an exhaustive list of all possible abusive refusals to deal. The core principles remain clear nonetheless. In essence, an abusive refusal to deal is one that risks eliminating effective dynamic competition or materially harms consumers in some other way (e.g., by preventing new kinds of products for which there is a clear and unsatisfied demand from coming on the market or foreclosing competition for an existing product that consumers wish to go on using). The essential point is that the refusal to deal would cause serious enough harm to dynamic competition and prejudice consumer interests to an extent sufficient to justify a duty to deal.

A. SHORT OVERVIEW OF THE DECISIONAL PRACTICE AND CASE LAW

1. Refusal to Supply Cases

As early as *Commercial Solvents*,[25] the ECJ recognized that it is an abuse for a dominant firm to cut off supplies of an essential input to an actual or potential

---

24 *See* AG Opinion (Jacobs) of Oct. 28, 2004, Case C-53/03, Syfait v. Glaxosmithkline [hereinafter *Glaxosmithkline*], 2005 E.C.R. I-4609, at para. 68 and the Commission in *Microsoft EC*:

> [T]here is no persuasiveness to an approach that would advocate the existence of an exhaustive checklist of exceptional circumstances and would have the Commission disregard a limine other circumstances of exceptional character that may deserve to be taken into account when assessing a refusal to supply.

2004 MS Decision, *supra* note 1, at para. 555.

25 Joined Cases 6/73 and 7/73, Istituto Chemioterapico Italiano S.p.A. and Commercial Solvents Corporation v. Commission [hereinafter *Commercial Solvents*], 1974 E.C.R. 223. Substantially the sameconclusion was reached in *Telemarketing*, which concerned the termination of supplies to an existing customer, with the intention of reserving another monopoly in an ancillary market to the dominant firm (Case 311/84, Centre Belge D'études De Marché Télémarketing v. SA Compagnie Luxembourgeoise De Télédiffusion & others [hereinafter *Telemarketing*], 1985 E.C.R. 3261). *See also* Hugin/Liptons, 1978 O.J. (L 22) 23, in which the Commission found that the refusal to continue to supply a customer with spare parts on the ground that the customer had established a business in servicing and the supply of spare parts in competition with the dominant supplier was abusive.

rival active in the downstream market for the final product. The basis for the refusal to supply was that the dominant firm was planning to vertically integrate in competition with its customer on the downstream market for the supply of the final product. The dominant firm was the only source of the input raw material in the European Community, such that its refusal to supply a rival on the downstream market would evict that rival from the market and preclude competition. The Court concluded that:

> "[A]n undertaking which has a dominant position in the market in raw materials and which, with the object of reserving such raw material for manufacturing its own derivatives, refuses to supply a customer, which is itself a manufacturer of these derivatives, and therefore risks eliminating all competition on the part of this customer, is abusing its dominant position within the meaning of Article [82]."[26]

The principles applicable to the termination of an existing course of dealing also apply to the duty to grant first-time access. In *British Midland/Aer Lingus*,[27] Aer Lingus had in the past cooperated with British Midland within the framework of an international multilateral agreement on interlining services. However, once British Midland commenced a competing route from London-Dublin, Aer Lingus terminated its past cooperation and refused to accept interchangeability of British Midland's tickets on the London-Dublin route. The Commission made clear that the outcome in that case would have been the same if British Midland had been a first-time customer. It stated that "both a refusal to grant new interline facilities and the withdrawal of existing interline facilities may, depending on the circumstances, hinder the maintenance or development of competition."[28]

Indeed, this was precisely the conclusion reached by the Commission in earlier cases in the same industry.[29]

26 *Commercial Solvents, id.* at 250.

27 British Midland/Aer Lingus, 1992 O.J. (L 96) 34. *See also* FAG-Flughafen Frankfurt/Main AG, 1998 O.J. (L 72) 30 (access to airport ground handling services).

28 *Id.* at para. 26.

29 *See* London European/Sabena, 1988 O.J. (L 317) 47. *See also* Amadeus Sabre, Twenty-first Competition Policy Report 73-4 (1991) (duty to give access to EU-wide computer reservation system).

More recently, in *Bronner*,[30] the ECJ clarified the conditions for an abusive refusal to deal. Advocate General Jacobs set out the requirement for a balancing test:[31]

> "[The] justification in terms of competition policy for interfering with a dominant undertaking's freedom to contract often requires a careful balancing of conflicting considerations. In the long term it is generally pro-competitive and in the interest of consumers to allow a company to retain for its own use facilities which it has developed for the purpose of its business. For example, if access to a production, purchasing or distribution facility were allowed too easily there would be no incentive for a competitor to develop competing facilities. Thus while competition was increased in the short term it would be reduced in the long term. Moreover, the incentive for a dominant undertaking to invest in efficient facilities would be reduced if its competitors were, upon request, able to share the benefits. Thus the mere fact that by retaining a facility for its own use a dominant undertaking retains an advantage over a competitor cannot justify requiring access to it."[32]

First, the input in question must be "indispensable to carrying on that person's business, inasmuch as there is no actual or potential substitute in existence...." Second, "the refusal ... [must be] likely to eliminate all competition in the [relevant market] on the part of the person requesting the service." And finally, the refusal must be "incapable of being objectively justified."[33]

Principles very similar to those described above have been repeatedly confirmed as applicable also in the context of intellectual property and related rights. This is where the essential function of IPRs comes in.

30 *See Bronner*, *supra* note 15.

31 The position under U.S. antitrust law is identical:

> If a patent or other form of intellectual property does confer market power, that market power does not by itself offend the antitrust laws. As with any other tangible or intangible asset that enables its owner to obtain significant supra-competitive profits, market power (or even a monopoly) that is solely 'a consequence of a superior product, business acumen, or historic accident' does not violate the antitrust laws. Nor does such market power impose on the intellectual property owner an obligation to license the use of that property to others.

See U.S. Dep't Justice & Fed. Trade Comm'n, Antitrust Guidelines for the Licensing of Intellectual Property § 2.2 (Apr. 1995).

32 *See Bronner*, *supra* note 15, at para. 57.

33 *Id*. at para. 41.

2. *Volvo/Veng*

Beginning with *Volvo*, the ECJ held that, while the refusal to license intellectual property is not an abuse in itself, the exercise of intellectual property rights may involve abusive conduct. Volvo held a U.K.-registered design for the front wing panels of Volvo series 200. Without Volvo's authorization, Veng imported imitations of Volvo's wing panels into the United Kingdom from other Member States. Volvo sought to prevent Veng from importing and marketing them in the United Kingdom and refused to license Veng even against a reasonable royalty. In its defense, Veng argued that Volvo's refusal to grant it a license for the registered design was an abuse. A U.K. court requested a preliminary ruling from the ECJ on whether this refusal amounted to an infringement of Article 82. The ECJ dismissed Veng's claim in the following terms:

> "[T]he exercise of an exclusive right by the proprietor of a registered design...may be prohibited under Article 8[2] if it involves, on the part of an undertaking holding a dominant position, certain abusive conduct such as the arbitrary refusal to supply spare parts to independent repairers, the fixing of the prices for spare parts at an unfair level or a decision no longer to produce spare parts for a particular model ... still in circulation. In the present case no instance of any such conduct has been mentioned by the national court."[34]

The judgment represents a careful compromise on the part of the ECJ. On the one hand, it recognized that a mere refusal to license could not, in itself, be an abuse. On the other hand, it left the door open for defining future situations in which Article 82 EC could prevail over the exercise of an IPR, where IPRs are used as a tool for, or where a compulsory license is an appropriate remedy for, some additional abusive conduct not consisting of a mere refusal to license.

IT LEFT THE DOOR OPEN FOR DEFINING FUTURE SITUATIONS IN WHICH ARTICLE 82 EC COULD PREVAIL OVER THE EXERCISE OF AN IPR, WHERE IPRS ARE USED AS A TOOL FOR SOME ADDITIONAL ABUSIVE CONDUCT NOT CONSISTING OF A MERE REFUSAL TO LICENSE.

3. *Magill*

It did not take the Commission long to find a case where there was an additional abusive conduct over and above a refusal to license—a case where copyright was used not to foster but to stifle innovation, in a manner inconsistent with the essential function of copyright. In *Magill*, three TV companies, RTE, BBC, and ITV, relied on their copyright in listings of TV programs to prevent Magill

34 *Volvo/Veng*, *supra* note 10, at para. 9.

from publishing a comprehensive weekly TV guide in Ireland and the United Kingdom.[35] At the time, each broadcaster published guides that only contained the listings for their own channels, with the result that consumers who wished to plan a comprehensive week's viewing had to purchase multiple guides. The Commission found that the broadcasters' refusal to disclose the copyright-protected listings information was abusive because it prevented the emergence of a new and much-needed product—a comprehensive TV listings guide—and enabled the broadcasters to leverage their monopoly in broadcasting activities into the downstream market for TV listings magazines.

On appeal, the EC courts sided with the Commission and found that "the exercise of an exclusive right by the proprietor may, in exceptional circumstances, involve abusive conduct."[36] The exceptional circumstances in that case were the following:[37]

- The information in question was indispensable to compete on the relevant downstream market, with the result that the refusal to share it would result in the elimination of competition on this market;
- The refusal would prevent the emergence of a new product on the downstream market—namely a composite TV listings guide, for which there was clear and unsatisfied demand (i.e., demand for a single, composite TV listings magazine); and
- There was no objective justification for the refusal.

The CFI and Advocate General in *Magill* did, but the ECJ did not, refer to essential function, and it has been suggested that the ECJ abandoned the essential function test as a relevant factor.[38] It is submitted that the combination of the new-product criterion as part of the exceptional-circumstances test is nothing but a restatement and application of the essential function test.[39] After all, the essential function of IPR is to foster the development of new products. The parties in the *Microsoft EC* case referred extensively to the essential function criterion in their pleadings, so it will be interesting to see whether the CFI will refer to it.

35 For the *Magill* cases, *see supra* note 12.

36 *Magill* ECJ Judgment, *supra* note 12, at para. 50.

37 This principles mirror the conditions of Article 82(b) of the EC Treaty, which prohibits a dominant undertaking from limiting innovation to the prejudice of consumers.

38 *See, e.g.,* U. Bath, *Access to Information v. Intellectual Property Rights*, 24 Eur. Intellectual Property Rev. 138 (2002) and L Prete, *From Magill to IMS: dominant firms' duty to license competitors*, Eur. Bus. L. Rev. (2004).

39 *See also* 2004 MS Decision, *supra* note 1, at para. 711.

4. *IMS Health*

The *Magill* principles were confirmed in the ECJ's judgment in *IMS Health*.[40] The case concerned IMS's copyright-protected data analysis structure in Germany. This structure, referred to as the "1860 Brick Structure", divides the German territory into 1,860 geographic bricks that are carefully designed to group doctors, patients, and pharmacies so as to allow the reporting of pharmaceutical sales data in a way that is useful for calculating the compensation of pharmaceutical company sales representatives.

In 2000, two companies established in Germany by former IMS personnel, NDC Health GmbH (NDC) and Azyx Deutschland GmbH (Azyx), entered the German market. It soon became apparent to IMS that the brick structures used by these companies' data services offerings infringed IMS's copyright in the 1860 Brick Structure. To prevent NDC and Azyx from further using its copyright, IMS obtained injunctions against these companies from the German courts.

On December 19, 2000, NDC complained to the Commission that IMS should be forced to license the 1860 Brick Structure to its competitors so that they can continue to use it to offer data services that compete with IMS's. On July 3, 2001, the Commission adopted an interim decision, which found that customers gave input in the development of the 1860 Brick Structure, and that that structure had become a de facto industry standard (Interim Decision).[41] The Interim Decision concluded that these factors made the 1860 Brick Structure an essential facility that must made available, on reasonable terms, for incorporation in competing NDC and Azyx services.

In the meantime, the German court requested a preliminary ruling from the ECJ in the main proceedings on whether IMS's conduct was compatible with Article 82 EC. IMS subsequently appealed the Commission's Interim Decision and the President of the CFI suspended the operation of the Decision.[42] The upshot of the President's Order was that the Interim Decision could not be enforced until IMS's main appeal was determined.[43]

On April 29, 2004, the ECJ issued its opinion in *IMS Health*. It confirmed the *Magill* criteria in holding that:

40 ECJ Judgment of Apr. 29, 2004, Case C-418/01, IMS Health v. NDC Health [hereinafter *IMS Health*], 2004 E.C.R. I-5039.

41 NDC Health/IMS Health: Interim Measures, 2003 O.J. L268/69.

42 Case T-184/01 R, IMS Health Inc. v. Commission, 2001 E.C.R. II-3193.

43 The CFI President's Order was confirmed on appeal by the President of the ECJ in Case C-481/01P(R), NDC Health v. IMS Health, 2002 E.C.R. I-3401.

> "[T]he refusal by an undertaking which holds a dominant position and owns an intellectual property right in a brick structure indispensable to the presentation of regional sales data on pharmaceutical products in a Member State to grant a licence to use that structure to another undertaking which also wishes to provide such data in the same Member State, constitutes an abuse of a dominant position within the meaning of Article 82 EC where the following conditions are fulfilled:
>
> The undertaking which requested the license intends to offer, on the market for the supply of the data in question, new products or services not offered by the owner of the intellectual property right and for which there is a potential consumer demand;
>
> The refusal is not justified by objective considerations;
>
> The refusal is such as to reserve to the owner of the intellectual property right the market for the supply of data on sales of pharmaceutical products in the Member State concerned by eliminating all competition on that market."[44]

These exceptional circumstances identified in *Magill* and reaffirmed in *IMS Health* appear to be the existence of the additional abuse itself, where the IPR is used as a tool for abusive restriction of innovation. This would mean that the mere refusal to supply a new customer who is a rival is normally competition on the merits. It would also mean that where an additional abuse inconsistent with the essential function of IPR is proven, there is no requirement to prove the additional exceptional circumstances.

5. *Microsoft EC*

The most recent application of these principles is the 2004 MS Decision.[45] Still subject to appeal at the time of writing, the Decision concerns two Commission findings of abusive conduct:

(1) a refusal to supply interoperability information, thus leveraging the desktop operating systems software (OS) monopoly to workgroup server OS products, and

(2) the tying of Windows Media Player to the desktop OS.

Since the latter does not concern IPRs it is not further discussed below.

44 *IMS Health*, *supra* note 40, at 52

45 2004 MS Decision, *supra* note 1.

The section of the 2004 MS Decision dealing with refusal to supply interoperability information is largely—but not exclusively—based on the criteria set out in the *Magill* and *IMS* cases. It identifies additional abuse consisting of exclusionary conduct in breach of Article 82(b) EC, where Microsoft's refusal to make essential interoperability information available hinders rival product development without noticeable contribution to Microsoft's own innovation incentive.[46] The Decision recognizes that a mere refusal to license IPRs is not an abuse (*Volvo/Veng*, *Magill*), but points out that Microsoft is not a case of mere refusal to supply (as was the case in *IMS*). Rather, Microsoft's refusal to supply essential interoperability information was found abusive and justified an obligation to license because of "exceptional circumstances". The Commission cited the following circumstances:[47]

- The need for interoperability,[48] which the Commission found to be essential for rival workgroup server OS producers to remain in the market in the long term. Interoperability information was of "significant competitive importance"[49] and there are no effective alternatives other than Microsoft providing this information;[50]
- The risk of elimination of competition on a secondary market.[51] The Commission proves this by showing that Microsoft is already dominant in workgroup server OS and market shares are growing, and showing that Microsoft's conduct tends to create a barrier to enter for work group server OS vendors,[52] while at the same time reinforcing barriers to entry in the PC operating system market (a monopoly maintenance theory).[53] Following *IMS Health*, it is determinative that two different stages of production may be identified and that they are interconnected;[54]
- The negative effect on innovation;[55]

46 *Id.* at paras. 693-701.

47 *Id*. at para. 712.

48 *Id*. at paras. 524, 637ff.

49 *Id*. at para. 586.

50 *Id*. at paras. 666 et seq.

51 *Id*. at paras. 585-692.

52 *Id*. at para. 524.

53 *Id*. at para. 769.

54 *IMS Health*, *supra* note 40, at paras. 44-6.

55 2004 MS Decision, *supra* note 1, at para. 693ff.

- The prejudice of consumers,[56] which the *Magill* and *IMS* cases did not discuss, including reduced choice of products, and consumer lock-in,[57] reduced innovation and thus reduction of future consumer choice,[58] and indirect harm by impairing competition;[59] and

- Absence of justification.[60] A disclosure requirement for interoperability information was consistent with EC legislation on the protection of software programs,[61] which establishes a policy encouraging interoperability. A duty to disclose the specifications did not adversely affect Microsoft's incentives to innovate, because source code information—which might allow competitors to develop clone products—would not be disclosed, and Microsoft's drive to develop interoperability technology would not be diminished since such technology makes its platforms more attractive.[62] Indeed, Microsoft's overall innovation incentives would increase as competitive alternatives become available.

Three legal observations can be made: First, again following *Magill* and *IMS*, the 2004 MS Decision finds an additional abuse over and above the mere refusal to supply. This includes in particular restriction of innovation in violation of Article 82(b) EC,[63] as well as disruption of past supplies.[64] Second, when discussing absence of justification, the Commission points out that Microsoft's uses its IPR claims in a manner that goes beyond what is necessary to fulfill the essential function of the IPR, by reducing innovation.[65] Third, the exceptional cir-

56 *Id*. at paras. 693-708.

57 *Id*. at para. 694.

58 *Id*. at para. 694ff.

59 *Id*. at para. 704 (*referring to* Case 85/76, Hoffmann-La Roche v. Commission [hereinafter *Hoffman-La Roche*], 1979 E.C.R. 461, at para. 125).

60 *Id*. at para. 709-78.

61 *Id*. at para. 743 et seq.

62 *Id*. at para. 714. Microsoft subsequently offered to make source code available, but this offer was not taken up since it carried with it the possibility of copyright suit for inadvertent copying.

63 *Id*. at para. 782.

64 *Id*. at paras. 587-8.

65 Quoting the 2004 MS Decision:

> The central function of intellectual property rights is to protect the moral rights in a right-holder's work and ensure a reward for the creative effort. But it is also an essential objective of intellectual property law that creativity should be stimulated for the general public good. A refusal by an undertaking to grant a licence may, under

footnote 65 cont'd on next page

cumstances are defined as the abuse itself,[66] suggesting that this criterion has no independent meaning. This is not to say that the Decision does not mention circumstances that could qualify as exceptional. The Commission mentions elsewhere a number of factors that it could have listed as exceptional, including:

- An exceptional level and duration of dominance,[67] reinforced by network effects.[68] Firms with substantial—let alone virtual monopoly—market power must be held to the strictest standard of conduct under Article 82 to ensure that their behavior in the marketplace does not have exclusionary effect.[69]

---

footnote 65 cont'd
exceptional circumstances, be contrary to the general public good by constituting an abuse of a dominant position with harmful effects on innovation and on consumers.

Id. at para. 711.

66 *Id*. at para. 712.

67 *Id*. at para. 471.

68 *Id*. paras. 459, 470.

69 *See* European Commission, DG Competition Discussion Paper on the Application of Article 82 of the Treaty to Exclusionary Abuses (Dec. 2005) [hereinafter Article 82 Discussion Paper], at 59 ("In general, the higher the capability of conduct to foreclose and the wider its application and the stronger the dominant position, the higher the likelihood that an anticompetitive foreclosure effect results.") and ECJ Judgment of Dec. 14, 2005, Case T-210/01, General Electric v. Commission, at para. 550 ("the greater the dominance of an undertaking, the greater is its special responsibility to refrain from any conduct liable to weaken further, a fortiori to eliminate, competition which still exists on the market.").

See also Advocate Fennelly in *CEWAL*:

> To my mind, Article 8[2] cannot be interpreted as permitting monopolists or quasi-monopolists to exploit the very significant market power which their superdominance confers so as to preclude the emergence either of a new or additional competitor. Where an undertaking, or group of undertakings whose conduct must be assessed collectively, enjoys a *position of such overwhelming dominance verging on monopoly*, [...] it would not be consonant with the *particularly onerous special obligation affecting such a dominant undertaking not to impair further the structure of the feeble existing competition* for them to react, even to aggressive price competition from a new entrant, with a policy [...] designed to eliminate that competitor [...]." (emphasis added.)

AG Opinion (Fennelly) of Oct. 29, 1998, Joined Cases C-395/96 P and C-396/96 P, Compagnie Maritime Belge Transport v. Commission (*CEWAL*), 2000 E.C.R. I-1365, at para. 137.

And in *Napp Pharmaceuticals*:

> We for our part accept and follow the opinion of Advocate General Fennelly in *Compagnie Maritime Belge* [...] that the special responsibility of a dominant undertaking is particularly onerous where it is a case of a quasi-monopolist enjoying "dominance approaching monopoly", "superdominance" or "overwhelming dominance approaching monopoly".

Napp Pharmaceutical Holdings Limited and Subsidiaries v. Director General of Fair Trading (*Napp Pharmaceuticals*), 2002 Comp.A.R 13, at para. 219. Although this judgment applied U.K. law, the relevant section of the U.K. Fair Competition Act is virtually identical to the wording of Article 82, and the Act requires that it is to be interpreted and applied in a manner consistent with EC competition law.

- A general pattern of exclusionary conduct, including another abuse (tying),[70] discrimination,[71] and the leveraging of dominance from a primary market (desktop OS) into a second product (workgroup server OS),[72] with the specific intent to foreclose specified rivals;[73]
- Deviation from a general industry practice of disclosure,[74] in which Microsoft originally participated, but from which it began to diverge when the company became powerful enough to do so, and the disruption of supply became profitable;[75] and
- Last, but not least, Microsoft's conduct reinforced its already dominant position in the PC OS market.[76]

B. THE CONDITIONS FOR AN ABUSIVE REFUSAL TO DEAL

The cases discussed above indicate that where the abuse consists of a (constructive) refusal to supply or license a rival, the mere refusal to license absent some other abuse cannot give rise to liability, with one exception. If (a) there is a refusal to license; (b) the IPR is essential and required for rivals to be or remain commercially viable in a downstream market; (c) the refusal to share the information or input creates a serious risk of elimination of all effective competition in the downstream market (even though the IPR does not apply to the downstream product or is only a component of it); and (d) the refusal to deal lacks objective, proportionate justification, the IPR owner must not unjustifiably discriminate between its own integrated downstream business and third parties competing with it. Even then, there are arguments that a compulsory license may be imposed only if the refusal is a tool for another abuse, or inconsistent with the essential function of IP, such as the "limitation of technical development to the prejudice of consumers" in violation of 82(b).

---

70 2004 MS Decision, *supra* note 1, at § 5.3.1.1.3.1, para. 531 et seq.

71 *Id.* at para. 574.

72 *Id.* at paras. 697-700.

73 *Id.* at paras. 774-8 (*especially* the quote from Mr. Gates at 778).

74 *Id.* at paras. 730 et seq.

75 *Id.* at paras. 587-8:

> The value that [rivals'] products brought to the network also augmented the client PC operating system's value in the customers' eyes and therefore Microsoft—as long as it did not have a credible work group server operating system alternative—had incentives to have its client PC operating system interoperate with non-Microsoft work group server operating systems [...] Once Microsoft's work group server operating system gained acceptance [...] Microsoft's incentives changed and holding back access to information relating to interoperability with the Windows environment started to make sense.

76 *Id.* at para. 769.

There is some discussion as to whether the exclusion must be in a downstream or secondary market distinct from an upstream market for the IPR, for an abuse to be found in these circumstances. This so-called "two markets" requirement seems a necessity for essential facilities cases such *IMS Health*, but even in that case the ECJ seems to recognize that:

> "It is sufficient that a potential market or even a hypothetical market can be identified. Such is the case where the products or services are indispensable in order to carry on a particular business and where there is an actual demand for them on the part of the undertakings which seek to carry on the business for which they are indispensable."[77]

This condition appears to be met where it makes economic sense for the IPR owner to license the IPR or provide the interoperability information but for the advantage the owner gains in excluding effective competition in, and monopolizing, the downstream market.

Whether there still is a need to show exceptional circumstances over and above the abuse in question remains to be seen. The "exceptional" circumstances in *Magill*, *IMS Health*, and *Microsoft EC* were effectively defined as the abuse itself. Arguments could be made that given the nature of IPR as a means to encourage competition through innovation, any remedy involving IPR in dynamic markets—those characterized by innovation—should be imposed only in the exceptional situation where the imposition of a compulsory license results in greater overall innovation incentives (for the entire industry including the IPR owner) than are maintained if the refusal is recognized.

A reading of the decisional practice and case law, as confirmed by the Commission's Article 82 Discussion Paper, suggests the following application of the exceptional circumstances in practice.

1. A Refusal to Deal

The concept of a refusal to deal has an expansive meaning under Article 82 EC, covering not only actual refusals, but also constructive refusals to deal.[78] In *Deutsche Post*, the Commission stated that "the concept of refusal to supply covers not only outright refusal but also situations where dominant firms make sup-

77 *IMS Health*, *supra* note 40, at para. 44.

78 Article 82 Discussion Paper, *supra* note 69, at paras. 62, 209, 219 and 225.

ply subject to objectively unreasonable terms."[79] The latter includes requests that are not met with a positive response without undue delay.[80] For example, a response by a dominant firm that was "entirely negative and consisted of raising difficulties"[81] is tantamount to a refusal to deal. So too is a dilatory attitude towards a request by one customer in circumstances where the dominant firm adopts a cooperative attitude towards another[82] (i.e., discrimination, generally applied delaying tactics),[83] or where the dominant company has established a clear pattern of refusing access to indispensable information and it therefore makes no sense for independent developers to request such information.

2. The Input or Information in Question Is Indispensable for Competition

Indispensability implies that the input or information in question is essential for the exercise of a viable activity on the market for which access is sought.[84] The test is whether the creation of substitute inputs or information is impossible or extremely difficult;[85] in other words, whether there are "technical, legal or economic obstacles capable of making it impossible or at least unreasonably difficult"[86] to create alternatives, or to create them within a reasonable timeframe.[87] Thus, it must be shown that the cost of duplicating the allegedly essential facil-

79 Deutsche Post AG, 2001 O.J. (L 331) 40, at para. 141.

80 GVG/FS, 2004 O.J. (L 11) 17, at para. 123.

81 *See* Sea Containers v. Stena Sealink (Interim measures), 1994 O.J. (L 15) 8, at para. 71.

82 *See* Commission Decision of Jun. 4, 2004, Case COMP/38.096, Clearstream (Clearing and Settlement) (not yet published) [hereinafter *Clearstream*], *available at* http://ec.europa.eu/comm/competition/antitrust/cases/decisions/38096/en.pdf, at paras. 293 et seq.

83 Article 82 Discussion Paper, *supra* note 69, at paras. 209 and 225.

84 *See* Case T-504/93, Tiercé Ladbroke SA v. Commission, 1997 E.C.R. II-923, at para. 130 (live pictures of French races not indispensable to compete in the relevant Belgian market).

85 *See* AG Opinion of May 28, 1998, *Bronner*, *supra* note 15, at 7813-4.

86 *See* IMS Health, *supra* note 40, at para. 28:

> It is clear from paragraphs 43 and 44 of Bronner that, in order to determine whether a product or service is indispensable for enabling an undertaking to carry on business in a particular market, it must be determined whether there are products or services which constitute alternative solutions, even if they are less advantageous, and *whether there are technical, legal or economic obstacles capable of making it impossible or at least unreasonably difficult* for any undertaking seeking to operate in the market to create, possibly in cooperation with other operators, the alternative products or services..." (emphasis added)

87 *See* Case T-374/94, European Night Services v. Commission [hereinafter *European Night Services*], 1998 E.C.R. II-3141, at para. 209, fn. 34.

ity constitutes a barrier to entry such that there are no viable alternatives to the dominant firm's input,[88] or the cost of such alternatives is "prohibitively expensive and would not make any commercial sense."[89]

In the case of intellectual property rights similar considerations apply. Because of the legal restrictions, the test is whether competitors can turn to any workable alternative technology or workaround the right in question in such a way that they can remain effective competitors without the supply.

This arose in the *Microsoft EC* case, where Microsoft argued that interoperability information (albeit not complete) was available in part through it and through other sources, including reverse engineering, and further information is not indispensable to be in the market. The Commission and its supporters argued that this is not a defense, since interoperability information is technically necessary and without it, rival servers cannot effectively communicate with Windows, Outlook and Office on a level playing field with Microsoft's own servers.[90] Second, there are no workarounds that offer any realistically workable alternative without prohibitive time lag. The Commission found in *Microsoft EC* that reverse engineering is not a viable alternative because of the time and expense involved, as well as the fact that Microsoft can simply make a strategic change to its code base to eliminate or substantially weaken any interoperability achieved.[91] Moreover, if the partial interoperability information Microsoft has made available in the past were sufficient for a workaround, then it would not have been faced with the complaints that led to the 2004 MS Decision, since rivals could have developed fully interoperable products. The Commission's Article 82 Discussion Paper states the indispensability requirement "would likely be met where the technology has become the standard or where interoperability with the rightholder's IPR protected product is necessary for a company to enter or remain on the product market."[92] This is the case for any interoperability information that may be protected by IPRs in regard to products that have become de facto standards or where interoperability is necessary to compete in the market.

88 *See European Night Services*, *id.* at para. 209 and *Clearstream*, *supra* note 82, at para. 227 ("Clearstream a *de facto* monopolist and unavoidable trading party for primary clearing and settlement services in Germany").

89 *See* GVG/FS, 2004 O.J. (L 11) 17, at paras. 109, 120, and 148.

90 It was argued that the information meets the definition of an essential facility given by Advocate General Jacobs in *Bronner* in that independent development "is impossible or extremely difficult..." (*see* AG Opinion in *Bronner*, *supra* note 85 and *IMS* Health, *supra* note 40, at para. 28).

91 2004 MS Decision, *supra* note 1, at paras. 685-7.

92 Article 82 Discussion Paper, *supra* note 69, at para. 23.

Indispensability is not required for an abuse not involving a (constructive) refusal to license a rival, where a compulsory license may be an appropriate remedy, but where dominance in addition to some other abusive behavior may be enough for application of Article 82.

3. The Refusal Risks Substantially Eliminating Effective Competition on the Relevant Market

Elimination of effective competition generally. The refusal to share the indispensable input must entail the "elimination or substantial reduction of competition to the detriment of consumers in both the short and the long term."[93] This is a higher standard than the distortion of competition that must be proven if the abuse involves tying, discrimination, imposing unfair terms and conditions, or standards manipulation. This condition is the corollary of the condition that the dominant firm's input is indispensable for competition: if the input is not indispensable, the refusal to share would not have substantial effects on competition. Conversely, if an input is essential for competition, it would, ultimately, allow the firm or firms that own or control it to exclude all competition on the relevant downstream market in which the input is used. The Commission has explained this underlying policy rationale for imposing a duty to deal in the following terms:

> "The duty to provide access to a facility arises if the effect of the refusal to supply on competition is objectively serious enough: if without access there is, in practice, an insuperable barrier to entry for competitors of the dominant company, or if without access competitors would be subject to a serious, permanent and inescapable competitive handicap...."[94]

No need to wait for actual exit. There should, however, be no requirement to show that the rival who wishes to have access to the information is already excluded from the market before the refusal to supply can be found to risk substantially eliminating competition.[95] Any such requirement would deprive the remedy of its useful effect. Rather, it should be enough to show that if the information continues to be unavailable, then (as the product and demand evolve) there is a serious risk of elimination of competition. Thus the Commission con-

93 *See* AG Opinion in *Bronner*, supra note 85, at para. 61.

94 *See* Organisation for Economic Co-operation & Development, The Essential Facility Concept 94 (1996), *available at* http://www.oecd.org/dataoecd/34/20/1920021.pdf.

95 Article 82 Discussion Paper, *supra* note 69, at paras. 27 and 58.

cluded in *Microsoft EC* that the relevant legal test is not whether each and every competitor has irreversibly exited, but whether there is some present basis for identifying a "serious risk of foreclosing competition and stifling innovation."[96]

This reflects the EC courts' view that Article 82 is not only concerned with actual anticompetitive effects, but also potential or likely anticompetitive effects.[97] This makes sense, since, otherwise, competition authorities and courts would have to stand idly by and wait for actual exclusion and anticompetitive effects to materialize before they could act, even where the long-term harm caused by exclusion would be serious, or even irreversible, due to very high barriers to re-entry. Moreover, in a monopoly maintenance case—which the Commission found that *Microsoft EC* case is, in part, because the denial of interoperability raises interoperability barriers to entry and thus reinforces Microsoft's PC operating systems monopoly—the anticompetitive effect is not the mere exclusion of competitors, but consumer harm from the continuation of a substantial degree of market power and reduction of product diversity.

THIS REFLECTS THE EC COURTS' VIEW THAT ARTICLE 82 IS NOT ONLY CONCERNED WITH ACTUAL ANTICOMPETITIVE EFFECTS, BUT ALSO POTENTIAL OR LIKELY ANTICOMPETITIVE EFFECTS.

Marginalized competition is the same as no effective competition. A test based on the elimination of all competition could also be open to abuse. A dominant firm could always allow one or two small rivals to remain on the market as marginalized competitors (sometimes referred to as "bonsai"). But the mere presence of a competitor does not mean that no elimination of competition has occurred. Especially in markets where significant investments are required to compete through innovation, effective competition does not mean the mere presence of one or more niche rivals. It implies a meaningful process of competition whereby firms have an effective opportunity to compete on the merits on the basis of price, quality, and innovation. Indeed, it is well established in the economics literature that there is a significant risk of falsely concluding that no harm to competition has occurred merely because rivals have not fully exited.[98] Competitors that are marginalized in dynamic markets and that are unable—or deprived of

96 2004 MS Decision, *supra* note 1, at recital 842. *See also* Article 82 Discussion Paper, *supra* note 69, at para. 22 ("An abuse may only arise when the termination is likely to have a negative effect on competition in the downstream market.").

97 *See* Case T-219/99, British Airways plc. v. Commission, 2003 E.C.R. II-5917 and Case T-203/01, Manufacture française des pneumatiques Michelin v. Commission, 2003 E.C.R. II-4071.

98 *See* T. Krattenmaker & S. Salop, *Anticompetitive Exclusion: Raising Rivals' Costs To Achieve Power over Price*, 96 YALE L.J. 209 (1986) and S. C. Salop & D. T. Scheffman, *Cost-Raising Strategies,* 36 J. INDUS. ECON. 19 (1987). *See also* Article 82 Discussion Paper, *supra* note 69, at para. 231 ("An abuse only may arise when the exclusion of competitors is likely to have a negative effect on competition in the downstream market. This should however not be understood to mean the complete elimination of all competition.").

further incentives—to engage in viable competitive innovation are effectively the same as no competition in those areas.[99]

4. Limiting Innovation to the Prejudice of Consumers

It bears emphasis that prejudice to consumers can occur in a variety of factual settings. The EC courts have confirmed that no exhaustive list of criteria applies.[100] Thus, each refusal to deal or instance of non-disclosure must be reviewed on its merits in light of the details of the market under consideration, the scope for harm to consumers in that market, and possible proportionate justifications.

In particular, there is no requirement that the refusal must always prevent the emergence of a product that has not existed before in any form. The situation where consumers are deprived of a specific new product for which they have present unsatisfied demand, as occurred in *Magill*, is but one example of a limitation of innovation to the prejudice of consumers. No such requirement is mentioned in the judgments in *Bronner* or in earlier cases such as *Commercial Solvents* and *Télémarketing*. Moreover, the examples cited in *Volvo* do not, by definition, involve new products, and yet the ECJ was willing to recognize those as examples of abuse where a compulsory license might be an appropriate remedy. Indeed, in *Ladbroke*, which concerned copyright, the CFI indicated that the new-product test could justify imposition of a duty to deal under Article 82, but that other criteria could also justify such a duty.[101] This was confirmed again in *IMS Health*, where the new-product criterion was mentioned as merely one of several sufficient conditions, thereby suggesting, implicitly but clearly, that this criterion (together with the other elements) is sufficient but not necessary.[102]

The new-product test applied in *IMS* must be understood as a proxy to identify conduct that stifles innovation and reduces consumer welfare, or that "limit[s] pro-

99 This interpretation of the law is also consistent with the decisional practice of the EC courts. In both *Commercial Solvents* and *Télémarketing,* for example, the EC courts found a breach of Article 82 where there was risk of "eliminating all competition *from that customer*" [emphasis added] not of eliminating all competition. *See Télémarketing*, *supra* note 25, at paras. 25 and 26 and *Commercial Solvents*, *supra* note 25, at para. 25.

100 *See* AG Opinion in *Glaxosmithkline*, *supra* note 24, at para. 68. *See also* 2004 MS Decision, *supra* note 1, at recital 555:

> [T]here is no persuasiveness to an approach that would advocate the existence of an exhaustive checklist of exceptional circumstances and would have the Commission disregard a limine other circumstances of exceptional character that may deserve to be taken into account when assessing a refusal to supply.

101 *See* Case T-504/93, Tiercé Ladbroke SA v. Commission, 1997 E.C.R. II-923, at para. 131.

102 *See* L. Gyselen, *Le titulaire d'un droit de propriété intellectuelle doit-il fournir le produit de son droit à un concurrent*, 2 Concurrences 24, 27 (2005). *See also* ECJ President Vesterdorf's obiter dictum in *Microsoft EC* (not yet reported), at para. 206.

duction...or technical development to the prejudice of consumers" within the meaning of Article 82(b).[103] This thinking appears to underpin the following (sometimes controversial) statement in the Commission's Article 82 Discussion Paper:

> "A refusal to licence an IPR protected technology which is indispensable as a basis for follow-on innovation by competitors may be abusive even if the licence is not sought to directly incorporate the technology in clearly identifiable new goods and services. The refusal of licensing an IPR protected technology should not impair consumers' ability to benefit from innovation brought about by the dominant undertaking's competitors."[104]

There are arguments that this comment is more liberal than the EC courts' interpretation of the new-product requirement. In *Magill*, the Court required proof of unsatisfied consumer demand and not merely the prospect of future innovation, and assessed the relevant market in which the follow-on innovation would compete. On the other hand, the holding in *Magill* is not necessarily exhaustive. Consumers can of course be harmed in many ways other than the narrow case of suppression of existing new products. One example of consumer harm is where rival software vendors lack access on equal terms to essential interoperability information and cannot offer products (even better or more functional or more innovative products) that have full interoperability with a virtual monopoly standard. Interoperability is a policy goal designed to provide users the freedom to combine best-of-breed components of a system or network in any way they wish. The non-disclosure of essential information in such a case not only deprives users of that freedom, but also is an artificial handicap to rivals' products that otherwise a) could evolve in innovative ways, creating product diversity or b) could directly or indirectly foster innovation that challenges the dominant firm's monopoly. Thus, in *Microsoft EC*, the Commission found that the key element of prejudice to consumers was the lack of interoperability between Microsoft's monopoly Windows operating system software and server software that limited competitors' innovation, including their scope for developing new products:

103 *See* F. Lévêque, *Innovation, Leveraging and Essential Facilities: Interoperability Licensing in the EU Microsoft Case*, 28 World Competition 71 (2005); M. Leistner, *Intellectual Property and Competition Law: The European Development from Magill to IMS Health Compared to Recent German and U.S. Case Law*, 2 ZWeR 138, ¶¶ 150-2 (2005); M. Stopper, *Der Microsoft-Beschluss des EuG,* 1 ZWeR 87 ¶ 102 (2005).

104 Article 82 Discussion Paper, *supra* note 69, at para. 240.

> "Due to the lack of interoperability that competing work group server operating system products can achieve with the Windows domain architecture, an increasing number of consumers are locked into a homogeneous Windows solution at the level of work group server operating systems. This impairs the ability of such customers to benefit from innovative work group server operating system features brought to the market by Microsoft's competitors. In addition, this limits the prospect for such competitors to successfully market their innovation and thereby discourages them from developing new products."[105]

Microsoft argued that the new-product requirement must satisfy potential demand by meeting the needs of consumers in ways that existing products do not. That is, a new product must exist that will expand the market significantly by bringing in consumers who were not satisfied before. This is clearly relevant but seems too limited, since, as noted, there may be other situations of consumer harm. Microsoft's argument would mean that if the relevant products are so important that all relevant consumers effectively require them and buy them whether or not they are good enough, consumer harm could not be found even if improvements are smothered. Restriction of innovation and lack of interoperability can prejudice consumers even if there are no new products yet, but incentives and opportunity to innovate are stifled to such an extent that rivals who in the past have shown a propensity to innovate are being cut out of the market.

In any event, the Commission and the interveners argued that fully interoperable third-party products fit the new-product criterion mentioned as being sufficient in *IMS*. There is unsatisfied consumer demand for third-party products with full interoperability with Windows and Office. There is substantial consumer prejudice in particular where: a) the rivals' activities could directly or indirectly foster innovation that challenges the dominant firm's monopoly; or b) rivals' products themselves can be expected to evolve in innovative ways, creating product diversity. In that case, there will be little scope for innovation—except, possibly, innovation coming from Microsoft, and even Microsoft's incentives are reduced in the absence of pressure from rivals. Thus, the scope for competitive harm in cases of denial of interoperability is far greater than in any previous case that involved only one new type of product (e.g., *Magill*). The CFI will now have to rule on the

THUS, THE SCOPE FOR COMPETITIVE HARM IN CASES OF DENIAL OF INTEROPERABILITY IS FAR GREATER THAN IN ANY PREVIOUS CASE THAT INVOLVED ONLY ONE NEW TYPE OF PRODUCT.

105 2004 MS Decision, *supra* note 1, at 694.

importance of interoperability to enable the emergence of multiple complementary new products and other forms of innovation by all competitors.

5. Objective Justification and Proportionality

Objective and proportionate efficiencies or other justification can immunize conduct from liability.[106] The elements to be proven for an objective justification analysis under Article 82 are four-fold. As applied to refusal to license cases, it is up to the dominant undertaking to show that:

- The refusal seeks to attain a legitimate goal. The range of acceptable justifications for a refusal to deal will vary from case to case depending on the facts. Examples include capacity limitations, quality degradation, and security.[107] In the case of IPRs, the desire to recover past R&D expenses and to underpin investments in future innovation may be provided as a legitimate goal;

- The conduct is effective, in that it is reasonably capable of achieving that legitimate goal; the objective must not be a theoretical or a sham or subterfuge for exclusionary intent;

- The conduct is necessary to achieve the pro-competitive goal. If this is convincingly alleged, the plaintiff must show there are less restrictive and effective alternatives;

- The use of the IPR is proportionate in light of the pro-competitive goal and the anticompetitive effect (called the balance-of-interest test); this test should focus on the essential function of IPRs, that is, to foster innovation. If the IPR is used in a way that reduces overall innovation, the balance of interest should arguably fall in favor of compulsory licensing.

This rule of reason type inquiry is similar to the analysis applied in the United States to Section 2 Sherman Act offenses (including the *Microsoft III* proceedings in the United States).[108]

106 *See also Magill*, *supra* note 12; *Bronner*, *supra* note 15 ("incapable of being objectively justified"); *Telemarketing*, *supra* note 25 ("without objective necessity"); United Brands v. Commission, 1978 E.C.R. 207, at paras. 189-90, 184 ("an undertaking must be conceded the right to take such reasonable steps as it deems appropriate to protect its said interests [although] such behaviour cannot be countenanced if its actual purpose is to strengthen this dominant position and abuse it..." (emphasis added)); *Hoffmann-La Roche*, *supra* note 59, at para. 90; and Case 322/81, Michelin v. Commission, 1983 E.C.R. 3461, at paras. 73 and 85.

107 *See, e.g.*, Commission Notice on the application of the competition rules to access agreements in the telecommunications sector, 1998 O.J. (C 265) 2, at para. 91.

108 In its review of the *Microsoft III* decision (*supra* note 18), the U.S. Court of Appeals for the DC Circuit states a rule of reason test very close to the EC proportionality test.

In *Microsoft EC*, Microsoft did not invoke a specific efficiency objective that it claimed to pursue through the refusal to disclose full interoperability information. Instead, Microsoft invoked general efficiencies and innovation incentives associated with the freedom to contract and protection of intellectual property:

> "The major objective justification put forward by Microsoft relates to Microsoft's intellectual property over Windows. However, a detailed examination of the scope of the disclosure at stake leads to the conclusion that, on balance, the possible negative impact of an order to supply on Microsoft's incentives to innovate is outweighed by its positive impact on the level of innovation of the whole industry (including Microsoft). As such, the need to protect Microsoft's incentives to innovate cannot constitute an objective justification that would offset the exceptional circumstances identified. Microsoft's other objective justification, which is that it has no incentive to engage in anti-competitive conduct with respect to interoperability, is not supported, and in fact is largely contradicted, by the evidence in this case."[109]

In assessing Microsoft's incentives to innovate, the Commission distinguished between interoperability technology and general OS technology, and concluded that the disclosure of interoperability information (externals) does not affect Microsoft's incentives to innovate OS internals.[110] Since no source code or internal code is disclosed, the Commission found that there is no risk of cloning.[111] Because of time lag and disadvantages, "Microsoft's competitors will have to provide additional value to the customer, beyond the mere interoperability of their products ... if such products are to be commercially viable."[112] In fact, because of the difficulty of implementing specifications designed for another system: rivals will have to be more efficient to benefit from the disclosure obligation.[113] Nor is Microsoft's incentive to innovate foreclosed since, according to the Commission, "there is ample scope for differentiation and innovation [by Microsoft] beyond the design of interface specifications."[114] The

109 2004 MS Decision, *supra* note 1, at para. 783.

110 *Id.* at para. 698.

111 *Id.* at paras. 713-22.

112 *Id.* at para. 722.

113 *Id.* at paras. 721-33.

114 *Id.* at para. 698.

Commission also noted that the U.S. remedies (the Microsoft Communications Protocol Program) did not reduce incentives to innovate either.[115] Conversely, the prospect of exclusion would reduce third parties' incentives to innovate, as well as Microsoft's incentives to innovate operating systems.[116] Ultimately, application of the balance-of-interest test led the Commission to the following conclusion:

> "[O]n balance, the possible negative impact of an order to supply on Microsoft's incentive to innovate is outweighed by its positive impact on the level of innovation of the whole industry (including Microsoft). As such, the need to protect Microsoft's incentives to innovate cannot constitute an objective justification that would offset the exceptional circumstances identified."[117]

As a threshold matter, the Commission suggests that it is for Microsoft to disclose what valid intellectual property rights it claims in any interoperability information that it would be required to provide to third parties. If the information in question is only or mainly trade secrets, many argue that the sanctity of intellectual property rights cease to be as clear or relevant.[118] Should some intellectual property be implicated, the interveners submitted that the limited disclosure of essential interoperability information strikes an appropriate and proportionate balance between the interests of a system of undistorted competition as laid down in Article 3(g) EC and respect for property rights.

First, after *Magill*, there is no general principle under Article 82 EC that a dominant firm can put forward a defense in a duty to deal case merely because intellectual property rights are at issue. Intellectual property laws do not create economic monopolies that can be defended in all circumstances and at all costs. Intellectual property laws coexist with antitrust law and accommodate antitrust law discipline. Intellectual property laws are blunt instruments that cannot bal-

115 *Id.* at para. 728.

116 *Id.* at para. 725.

117 *Id.* at para. 783

118 Trade secrets are not exclusive rights granted by the law and therefore do not deserve the same level of protection as patents, copyright, or trademarks, all of which are recognized and established property rights created by the legislator. In any event, where the violation of competition law consists precisely in keeping secret essential interoperability information, potential trade secrecy of such information must give way to proportionate antitrust remedies imposed in the public interest.

ance innovation incentives in all cases or regulate exhaustively and purely by themselves all possible economic and legal conflicts.

Second, it is said that the very purpose of IP rights is to grant a reward to the owner by restricting competition, in return for the benefits that valuable innovations bring to society.[119] But the same general justification can be advanced for physical property: the nature, scope, and duration of protection are the result of a legislative consensus that property rights confer net benefits to society in the form of desirable investment activity. Furthermore, it is well established that there are limits to the right to (physical) property that can be imposed in the common interest (e.g., on land use planning or environmental grounds).

INTELLECTUAL PROPERTY LAWS ARE BLUNT INSTRUMENTS THAT CANNOT BALANCE INNOVATION INCENTIVES IN ALL CASES OR REGULATE EXHAUSTIVELY AND PURELY BY THEMSELVES ALL POSSIBLE ECONOMIC AND LEGAL CONFLICTS.

Third, the interference with any intellectual property should be limited and proportionate and should not materially affect its wider innovation incentives. Microsoft's rivals argued they already have their own competing products with different features and functionality and have no desire to clone Microsoft's products; indeed, their competitive strategy is based on innovation and product differentiation. Any disclosures would be strictly limited to information that is essential to allow their products to have the same degree of interoperability with the virtual-monopoly Windows products as Microsoft affords to its own business. In particular, source code would not be required. Microsoft's rivals do provide, and will have to continue to provide, additional value to the customer beyond the mere interoperability of their products if such products are to be commercially viable. Interoperability is essential but it certainly does not in itself guarantee rivals' commercial success.

IV. Compulsory License on FRAND Terms

As explained, Article 82 EC applies to IPRs only if an IPR is used as a tool for an abuse, and in such a case, a compulsory license may be an appropriate remedy. As a rule, any remedy imposed by the Commission for abusive conduct should be proportionate. According to the ECJ in *Magill*: "[T]he burdens imposed on undertakings in order to bring an infringement of competition law to an end must not exceed what is appropriate and necessary to attain the objective sought, namely re-establishment of compliance with the rules infringed."[120]

119 *See* L. Kaplow, *The Patent-Antitrust Intersection: A Reappraisal*, 97 HARVARD L. REV. 1813, 1817 (1984).

120 *Magill* ECJ Judgment, *supra* note 12, at para. 30.

In the case of IPRs, this requires an evaluation of the impact of the remedy on overall innovation, the essential function of IPRs. This means that if a less-burdensome remedy can be found that effectively addresses the competition concerns of a refusal to license, the Commission should not resort to an order for compulsory licensing. If the holder of an IPR has several effective ways to eliminate an abuse, a choice should be allowed.

Once a choice has been made for compulsory licensing, the main difficulty facing a regulatory agency is implementing the remedy with appropriate speed and determining the terms at which it should be set.[121]

The Commission is by now acutely aware that timely implementation is especially important in dynamic markets, such as IT. It is striking that three years after the 2004 MS Decision was adopted, the remedy is still not effective. In the mean time, the complainants allege, the exclusionary effects on competition continue and new products are coming to market, such as Vista and Longhorn, giving rise to disputes as to whether and to what extent interoperability information must be disclosed for these products. This delay is perhaps understandable for a precedent case such as *Microsoft EC*, but it bodes ill for the useful effect of the compulsory licenses in complex cases. If the remedy is not implemented timely in a forward-looking manner, there is a risk that rivals' products are condemned to interoperability with old products that have been superseded in the mean time.

Another lesson learned from the *Microsoft EC* remedy is that the Commission needs to exercise vigilance when considering the terms and conditions for the compulsory license. The IPR owner may have incentives to deprive the remedy of useful effect. In *Microsoft EC*, the Commission ordered Microsoft to release its interoperability information on reasonable and non-discriminatory terms (RAND).[122] Microsoft responded by demanding significant royalties for the Workgroup Server Protocol Program that, according to the complainants, exceed the royalties charged for entry-level server operating systems.[123] If that is confirmed, price-squeeze concerns arise. In addition, there may be a temptation to pack lengthy license agreements with complicated and restrictive terms and conditions, contrasting with the one-page licenses that are employed in other cases.[124]

121 *See, e.g.,* IBM 1984 Undertaking, 1991 O.J. (L 122) 42.

122 2004 MS Decision, *supra* note 1, at paras. 1005-8.

123 *See* Microsoft Work Group Server Protocol Program, *at* http://www.microsoft.com/about/legal/intellectualproperty/protocols/wspp/wspp.mspx (accessed Feb. 13, 2007). he Commission obtained a commitment from IBM to reveal interface information to competitors on new IBM products.

124 *See* Microsoft's Royalty Free Protocol License Agreement for specific client-server protocols implemented in Windows, *at* http://msdn.microsoft.com/library/default.asp?url=/library/en-us/randz/protocol/royalty_free_protocol_license_agreement.asp (accessed Feb. 13, 2007).

Some guidance can be found in the practice of setting industry standards. In standard setting, licensing on fair, reasonable, and non-discriminatory (FRAND) terms has long been commonplace. In that context, the requirement of RAND terms is understood by most participants to mean that the prices charged must not be excessive, exclusionary, or anticompetitive, basically the same criterion as laid down in Article 82(a). A reasonable price is a moderate one, bearing some rational relation to objective assessment of the innovative value added by the technology protected by the IPR, rather than a monopolist's desire to maximize its profits. In addition, in the words of the 2004 MS Decision, "restrictions should not create disincentives to compete with Microsoft, or unnecessarily restrain the ability of the beneficiaries to innovate."[125]

The 2004 MS Decision contained a useful limiting criterion: the royalties and terms and conditions "should not reflect the 'strategic value' stemming from Microsoft's market power in the client PC operating system market or in the work group server operating system market," which means that it can charge, at most, for the value of innovation proven to be included in the documentation. Even that may be too much: where the abuse is exclusionary, the licensee may have been deprived of the minimum efficient scale of operation that would have allowed it to support RAND royalties in a competitive environment. If that is the case, there is an argument that the royalty should be less until conditions of competition have been restored in the leveraged market, to ensure that the remedy has a useful effect.

But assuming that some innovative value is conveyed, and that price squeezing is avoided, what should the price be? The economic theory seems relatively clear. In competitive conditions, if the technology to be licensed is equivalent to alternative available technology, there is no reason to believe that the IPR owner, absent its monopoly, would find a buyer or be able to charge a positive price for it. Indeed, in a competitive and non-collusive environment, royalties for equivalent and competitive technical solutions would tend towards marginal costs, which is often close to zero in the case of IT. Where technologies are not equivalent, the fee for the lesser solution would tend to approach zero, with the owner of the better solution being able to charge no more than the incremental value that the licensee expects from the use of the better solution (for instance, because it saves costs, leads to expansion of demand, or allows the licensee to charge higher prices to end users). The fee for the better solution is no higher than the opportunity cost that the licensee would incur if it used the next best alternative.

Unfortunately, the economic theory appears difficult to apply in practice. It is perhaps most useful as a framework of reference that can be used to validate and verify the results of alternative pricing methods. Several methods might be employed. These methods have often been used in excessive pricing cases, but

125 2004 MS Decision, *supra* note 1, at para. 1008.

each has its own benefits and drawbacks. First, recourse can be had to Article 82's traditional criteria for determining reasonable prices in the context of the Article 82(a) case law and decisional practice on excessive pricing.[126] A useful starting point is a comparison of the price charged and the historical or long-run incremental cost of R&D. Another alternative is to focus on profits and not prices, and lower the price until a profit is achieved commensurate to the normal return on investment in competitive conditions in this industry. However, these calculations are fraught with difficulties in ordinary industries, and raise even more concerns in dynamic markets such as IT. Moreover, a focus on profits ultimately penalizes success where excessive pricing is absent.

> UNFORTUNATELY, THE ECONOMIC THEORY APPEARS DIFFICULT TO APPLY IN PRACTICE. IT IS PERHAPS MOST USEFUL AS A FRAMEWORK OF REFERENCE THAT CAN BE USED TO VALIDATE AND VERIFY THE RESULTS OF ALTERNATIVE PRICING METHODS.

A fallback would be to conduct a consistent comparison with the prices of similar products charged by the licensor in competitive markets, charged by licensor to its own downstream business, or charged by rivals for similar technology. Interestingly, it is argued, the type of information at issue in the *Microsoft EC* case is generally made available in the industry for free or for a nominal fee, and Microsoft itself makes similar information available for free where it suits its strategic goals.[127] Absent its monopoly position, the complainants argue that Microsoft would have an inherent interest in making the information at issue available for free since this would drive sales of its software products, in particular its desktop operating systems. Finally, it is argued that Microsoft is fully remunerated for the creation of the licensed information through the sale of client and server operating systems. The Commission is currently reviewing these arguments, and the result may have important precedent effect for future cases.

---

126 *See, e.g.,* Case 26/75, General Motors v. Commission, 1975 E.C.R. 1367; Case 27/76, United Brands, *supra* note 106; Case 30/87, Bodson v. Pompes Funèbres des Régions Libéréés, 1988 E.C.R. 2479; Case 395/87, Ministère Public v. Tournier, 1989 E.C.R. 2521, at para. 38; and Deutsche Post, 2002 O.J. (L 331) 40.

127 For instance, for Webservices Specifications (WSTX), Microsoft:

(a) provides a reasonable, royalty-free copyright license to the specifications with relatively few restrictions; and

(b) provides a royalty-free license to any patents considered essential to implement the specifications.

See Microsoft License Agreement, *at* http://download.microsoft.com/download/8/e/5/8e59dce62b27-4fc3-bd00-0531c5514ae3/WSS_LicenseAgreement.pdf (accessed Feb. 13, 2007). *See also* Press Release, Microsoft Corporation, Microsoft Announces Availability of Open and Royalty-Free License For Office 2003 XML Reference Schemas (Nov. 17, 2003), *available at* http://www.microsoft.com/presspass/press/2003/nov03/11-17XMLRefSchemaEMEAPR.mspx.

The recent FTC remedy decision in *Rambus* is an interesting example.[128] When the FTC found that Rambus set a "patent ambush", it set a royalty of 0.5 percent for the patents in question (going to zero after three years), where Rambus had asked for a permanent royalty of 2.5 percent. In determining the terms of such a RAND license, the FTC noted the inherent difficulties attendant to reconstructing marketplace conditions that would have prevailed in the absence of anticompetitive conduct. The FTC held, however, that antitrust defendants should not be allowed to avoid appropriate remedies because determining the but-for world is challenging in practice.[129] The FTC found that a RAND license requires a royalty rate no higher than the *ex ante* value of the technology, which "is the amount that the industry participants would have been willing to pay to use a technology over its next best alternative prior to the incorporation of the technology into a standard."[130] That amount, the FTC found, takes proper account of the value of the technology to the IPR holder. To determine the specific royalty rate, the FTC turned to "real-world examples of negotiations involving similar technologies."[131]

Two comments should be kept in mind, however. First, even if the IP owner (or indeed anyone else) identified comparable technologies licensed for a fee, it does not mean that the IP owner should be allowed to charge an equivalent fee. It may be that the owner of the comparable technology is able to charge a fee only because the technology market is not competitive or because the IPR owner refuses to license the equivalent technology in the first place. In sum, comparables should be reviewed, but this should be done on a consistent basis and without allowing the IPR owner to charge a monopoly rent, which would be equivalent to a constructive refusal to license. General licensing practices in the industry may provide guidance also, on condition that they are properly applied, and result in a royalty no greater than justified by the extent to which the IPR owner's innovation constitutes part of the overall technology used in the product.[132] If it is at the core of the rival's product and very innovative, then it justifies a greater royalty than mere interoperability information at the edge of a rival's product.

---

128 *See* FTC Final Order of February 2, 2007 and Opinion of the Commission on Remedy, In the Matter of Rambus Incorporated, *available at* http://www.ftc.gov/os/adjpro/d9302/070205opinion.pdf. *See also* Dissenting Opinion (Harbour), *available at* http://www.ftc.gov/os/adjpro/d9302/070205harbourstmnt.pdf and Dissenting Opinion (Rosch) (remedy should be royalty-free), *available at* http://www.ftc.gov/os/adjpro/d9302/070205roschstmnt.pdf.

129 Op. at 16-19.

130 Op. at 17.

131 Op. at 18.

132 *See, e.g.,* R. Goldscheider, New Companion to Licensing Negotiations: Licensing Law Handbook ¶ 7.02[8][b] (4th ed. 2002/3).

Second, for a remedy involving a compulsory licensing scheme to work, access must be set at a price low enough for an equally or more efficient licensee to compete effectively. In order for a remedy to have a useful effect and achieve its goal (elimination of the abuse as well as restoration of competitive conditions), in some cases, it may require the dominant firm to lower its fees to a sustainable level until competitive conditions have been restored, and further pricing can be left to the market.

As regards non-discrimination, differential treatment should be allowed only if it is justified by proportional objective considerations. This requires that the differential treatment

(1) attain a legitimate objective,

(2) that it is effective in attaining that objective,

(3) that it is necessary to obtain the objective (there is no less restrictive alternative), and

(4) a weighing of the interests of the parties involved (balance-of-interest test).

For example, a cross-license may justify a royalty readjustment if it is agreed to at arms' length and fair value is paid on both sides. Ultimately, the royalty system should ensure a level playing field between all participants in the market when dealing with the licensor. For instance, it should not discriminate between development models (proprietary versus open source models) or insiders and outsiders.

V. Conclusion

While IPRs confer exclusive rights, and a mere refusal to license is not an abuse, IPRs do not provide complete immunity from application of competition law. The use of IPRs—and indeed any other asset—as a tool for an abuse other than a refusal to license (such as unjustified discrimination, tying, exclusionary pricing and price squeezing, the unjustified disruption of supplies, restriction of innovation, standard manipulation or breach of FRAND promises given to a standard setting organization, unjustified refusal to allow rivals access to essential facilities, and even excessive pricing) can give rise to liability under Article 82 EC and equivalent provisions of national competition laws. Even then, two points should be kept in mind: first, the fact that the abuse involved a refusal to license an exclusive IPR may be invoked as a justification. It is submitted that the proper balancing tools to evaluate such a defense and distinguish between legitimate and abusive exercise of IPR is the essential function test, always keeping in mind the policy goal of IPRs to provide innovation incentives, and the proportionality test. Second, a compulsory license may be imposed to remedy the abuse only if such a remedy is appropriate and proportionate to redress the abuse, and regen-

erates innovation incentives more than it restricts them. In practice, these two points require the same balancing exercise. Interoperability and standards cases are arguably special in this respect. Encouraging interoperability with monopoly platforms is one of the situations where this balancing exercise suggests a compulsory license is appropriate. Denying interoperability with a monopoly platform raises barriers to entry in the markets for complementary products, as well as (crucially) the market for the platform itself. Interoperability with monopoly platforms normally increases innovation incentives for third parties as well as those of the owner of the platform with which interoperability is sought.

In the circumstances of *Magill* and (according to the 2004 MS decision) *Microsoft EC*, copyrights and trade secrets (and, in future, patents) were used in a fashion inconsistent with the essential function of IPRs, namely, to suppress rivals' incentives and capability to innovate, without countervailing benefit: the broadcasters' incentives to improve their program guides and Microsoft's incentives to advance its interoperability protocols were not dependent on continued protection of exclusive rights. To the contrary, the reduction of competition may have increased the funds they had available for innovation, but reduced their incentives. Thus, IPRs were used in a manner at odds with their goals, or at least in a manner that was not proportionate or necessary to maintain innovation incentives.

In *Magill*, *IMS*, and *Microsoft EC*, reference is consistently made to the exceptional-circumstances test, but it would seem that this test has no independent meaning. The exceptional circumstances were defined in terms of the abuse addressed in those cases (restriction of innovation to the prejudice of consumers). It may well be that cases of abuse are exceptional, but the test seems to impose no additional burdens or requirements on the plaintiff in specific cases. It is hoped that the CFI in *Microsoft EC* will clarify this.

At this stage, one conclusion can be drawn: There is a dearth of quantitative information about the actual contribution of IPRs to innovation. Much of the support of IPRs is based on a general understanding that exclusive rights encourage investment in innovation, and that they therefore benefit consumer welfare and society overall. This is a matter of common sense and almost religious belief, but cannot be used as a hard and fast rule. Thorough and independent quantitative studies in different industry sectors would be very welcome. The very existence of the open standards-based Internet and the work done by the open source community indicates that at least in some sectors, innovation is not dependent on exclusive IPRs. Similarly, even in industries where IPRs are needed, IPRs are blunt instruments, and not always well-adapted to the specific situations in which they are invoked. There are cases where IPRs are perverted to achieve commercial objectives antithetical to the policy goals the legislature sought to achieve. In the rare case that involves a restriction of competition, experience suggests that competition law can be used as a balancing tool. It remains to be seen, however, what the CFI will decide in *Microsoft EC*.

Ultimately, for a compulsory licensing scheme to work, remedies must be implemented in a timely and effective manner. Unwilling defendants should not be allowed to fix the licensing terms and conditions, because incentives are to deprive the remedy of useful effect. Access must be set at a price low enough for the licensee to compete effectively and restore conditions of competition. Where the abusive conduct has deprived the victim of economies of scale and market opportunity, it may not be able to sustain a level of royalties that it could have readily borne had the market remained competitive. This may in some cases mean that royalties should be lowered below that level until competitive conditions have been restored, and further pricing can be left to the market. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Introduction to the Symposium

Jean-Charles Rochet and Jean Tirole

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1.
Competition Policy International is a free publication. To order or download additional copies, visit eSapience.org.

Introduction to the Symposium

Jean-Charles Rochet and Jean Tirole

It is our pleasure to introduce this special issue of *Competition Policy International*, dedicated to the Two-Sided Markets Symposia organized in May 2006 at University College London and June 2006 at MIT in Cambridge, Massachusetts. The contributions presented in this volume are a good illustration of the incredible richness and depth of the challenges posed by multi-sided industries. Although some convergence can be acknowledged, there is still some debate among economists, lawyers, and regulators about several important issues. As a trivial illustration, several contributors to this special issue criticize the terminology itself: Evans and Schmalensee suggest that the denomination "two-sided markets" is misleading because the word "market" is not used in the antitrust sense and, of course, many platforms have more than two sides. The multi-sided platforms (or MSPs) nomenclature they and others propose is likely to become the new standard.

The contributions presented in this symposium show that the paradigm of MSPs is applicable to a growing number of industries. A first reason is that new (multi-sided) business models sometimes become successful in formerly one-sided industries. Professor Andrei Hagiu of Harvard Business School has pointed out that Japanese convenience store Lawson and the railway commuter card Suica entered new markets by going from one-sided to two-sided businesses. A second reason is that many existing two-sided platforms are expanding into other two-sided industries. For example, latest generation videogame consoles (e.g., PS2, Xbox, GameCube) offer DVD playing, Internet browsing, and computer capabilities. They have been termed the "Trojan horses" of the digital homes. Similarly, Brito and Pereira analyze how the development of mobile virtual network operators is bound to reduce considerably the costs of entry in the mobile telephone industry.

The authors are Professors of Economics at the Toulouse School of Economics.

Another reason for the wider applicability of the MSP model is that it is now well-accepted that traditional network industries like telecommunications should, in fact, be viewed as two-sided. Even if, ex ante, two distinct sides cannot always be identified (in the sense that most people use their telephones both to call and to receive calls), any given call is initiated by a caller and that the receiver's utility is influenced (positively or negatively) by the call. Thus the "usage externality" model that we developed for payment cards also applies to telecommunications.[1] Even in mature networks where membership is almost universal (for example, almost everybody now has a debit card) the structure of usage pricing matters. This becomes particularly important in the context of expanding MSPs, which are going to lead to generalized multi-homing. For example, more and more people will have several devices that can provide payment services in their pockets. Similarly, more and more homes will be equipped with multiple devices allowing the access to music or movies through the Internet. In such a context, it is important to give the right price signal to the party that is in the driver's seat (i.e., who chooses which device to use). This shows clearly that relative prices matter.

Waverman gives an excellent illustration of the two-sidedness of the mobile telephone industry by showing that the different developments of this industry in Europe and in the United States can be explained by the use of different price structures. For historical reasons, European mobile operators essentially used the caller pays model while the United States, from the start, adopted a more balanced model where caller and receiver share the costs of each call. More generally, skewed pricing (that is, when one side pays most or all of the costs) is a fascinating and recurrent theme in two sided industries. As discussed by Bolt, theoretical models predict that skewed pricing is more likely to be the norm than the exception for MSPs. Surprisingly, skewed pricing has sometimes been used by competition authorities in completely opposed ways. In the case of payment cards, for example, skewed pricing has sometimes been viewed as evidence that dominant platforms distort the price structure. This incorrect view (small platforms adopt price structures that are more skewed than larger ones) results from insufficient attention paid to efficiency considerations related with usage externalities. By contrast, Wotton shows that media markets have sometimes been wrongly classified as one-sided because, in these industries, the bulk of revenues are often extracted from one side, the advertisers, only. This fails to recognize that absent readers (or viewers) to the newspaper (or TV channel), no advertiser would ever pay anything for access.

In any case, the views of competition authorities are changing rapidly. Initially, while acknowledging the inadequacy of the traditional antitrust doctrine to

1 *See* J.-C. Rochet & J. Tirole, *Cooperation among Competitors: The Economics of Payment Card Associations*, 33 RAND J. Econ 549-70 (2002) and J.-C. Rochet & J. Tirole, *Platform Competition in Two-Sided Markets*, 1 J. Eur. Econ. Ass'n 990-1029 (2003).

MSPs, they also criticized economists for not offering applicable and empirically tested alternative models. This has now changed, thanks in particular to the empirical work of Rysman and collaborators which is reviewed in his contribution to this volume. Rysman recalls how he was able to establish empirically the reality of indirect network externalities in several industries (such as yellow pages directories and payment cards). Moreover, this empirical work has put forward a fundamental distinction between potential and effective multi-homing which might reveal itself to be of crucial importance in the assessment of inter-platform competition. Competition authorities now put emphasis on the possibilities to "enrich" the traditional antitrust analysis, to use the title of the contribution to this volume by Park and Rooney. Fletcher suggests for example that the traditional predation test could be adapted to MSPs by using the notion of "opportunity cost". Similarly, Hesse argues that the U.S. Department of Justice found a way to adapt the traditional SSNIP test to the payment card industry to define the relevant product market of PIN-debit network services in a recent merger case. These arguments are well-taken for one category of MSPs, that Evans and Schmalensee call the "transaction systems". These are the industries that can be described by the "usage externality model", where the notions of transaction volume and total transaction price can be identified. Therefore, the SSNIP test and the predation test can easily be adapted (with two-sidedness remaining important for analyzing price structure). However other MSPs, like advertised supported media, do not fit within this category, since there is no natural notion of volume or total price. This calls for further research by economists.

A particularly interesting case is the real-estate industry discussed by Brown and Yingling. They argue that real estate agents perform two distinct functions: searching for clients and facilitating transactions (close to what Evans and Schmalensee call "building audiences"). In the absence of the first function, the restrictive practices that U.S. realtors have adopted in their cooperative management of the multiple listings platforms (in which they pool their information about their clients) could easily be viewed as anticompetitive. However, if the "building audiences" activities of realtors are taken into account, these restrictive practices might appear as a necessary ingredient for providing realtors with appropriate incentives to attract customers. Here again, further research is needed.

This special symposium is a must read for anyone (including business executives, lawyers, and economists) wanting to better understand the fascinating world of "multi-sided platforms". ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

The Industrial Organization of Markets with Two-Sided Platforms

David S. Evans and Richard Schmalensee

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

The Industrial Organization of Markets with Two-Sided Platforms

David S. Evans and Richard Schmalensee

Many diverse industries are populated by businesses that operate "two-sided platforms." These businesses serve distinct groups of customers who need each other in some way, and the core business of the two-sided platform is to provide a common (real or virtual) meeting place and to facilitate interactions between members of the two distinct customer groups. Platforms play an important role throughout the economy by minimizing transactions costs between entities that can benefit from getting together. In these businesses, pricing and other strategies are strongly affected by the indirect network effects between the two sides of the platform. As a matter of theory, for example, profit-maximizing prices may entail below-cost pricing to one set of customers over the long run and, as a matter of fact, many two-sided platforms charge one side prices that are below marginal cost and are in some cases negative. These and other aspects of two-sided platforms affect almost all aspects of antitrust analysis—from market definition, to the analysis of cartels, single-firm conduct, and efficiencies. This paper provides a brief introduction to the economics of two-sided platforms and the implications for antitrust analysis.

David S. Evans is Chairman of eSapience, Ltd. in Cambridge, MA, Managing Director of the Global Competition Policy Practice at LECG, Cambridge, MA and Executive Director of the Jevons Institute for Competition Law and Economics and Visiting Professor at University College London. Richard Schmalensee is Professor of Economics and Management at the Massachusetts Institute of Technology (MIT) and the John C Head III Dean of the MIT Sloan School of Management, Cambridge, MA.

I. Introduction

Many diverse industries are populated by businesses that operate "two-sided platforms." These businesses serve distinct groups of customers who need each other in some way, and the core business of the two-sided platform is to provide a common (real or virtual) meeting place and to facilitate interactions between members of the two distinct customer groups. Two-sided platforms are common in old-economy industries such as those based on advertising-supported media and new-economy industries such as those based on software platforms and web portals. They play an important role throughout the economy by minimizing transactions costs between entities that can benefit from getting together.

In these businesses, pricing and other strategies are strongly affected by the indirect network effects between the two sides of the platform. As a matter of theory, for example, profit-maximizing prices may entail below-cost pricing to one set of customers over the long run and, as a matter of fact, many two-sided platforms charge one side prices that are below marginal cost and are in some cases negative. These and other aspects of two-sided platforms affect almost all aspects of antitrust analysis—from market definition, to the analysis of cartels, single-firm conduct, and efficiencies.[1]

This paper provides a brief introduction to the economics of two-sided platforms and the implications for antitrust analysis.

Two-sided platforms were first identified clearly in pioneering work by Jean-Charles Rochet and Jean Tirole, which began circulating in 2001.[2] A significant theoretical and empirical literature quickly emerged, and the subject has become a very active area of research in economics.[3] For the purposes of this paper, it is helpful to clarify some terminology that is used in the economics literature and which sometimes causes confusion. Rochet and Tirole used the term "two-sided markets" to refer to situations in which businesses cater to two interdependent groups of customers. The term "market" was meant loosely and does not refer to how that term is often used in antitrust. This paper refers to "two-sided platforms" but it is synonymous with "two-sided markets" as used in much of the economics literature. How to determine what market a two-sided platform competes

1 *See* David S. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325 (2003) and Julian Wright, *One-Sided Logic in Two-Sided Markets*, 3 REV. OF NETWORK ECON. 44 (2004).

2 Jean-Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, 1 J. EUR. ECON. ASS'N 990 (2003). Some of the key issues were identified in the context of payment cards in an important contribution WIliam F. Baxter, *Bank Exchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541 (1983). There are also literatures for particular industries that also provide precursors.

3 *See* Conference on Competition Policy in Two-Sided Markets (Institute d'Economie Industrielle, U. Toulouse) (Jun. 29 - Jul. 1, 2006), *available at* http://idei.fr/doc/conf/tsm/programme.pdf.

in, from an antitrust perspective, is one of the questions considered here.[4] Two-sided platforms often compete with ordinary (single-sided) firms and sometimes compete on one side with two-sided platforms that serve a different second side.

II. Economic Background on Two-Sided Platforms

A heterosexual, singles-oriented club offers some intuition on the economics of two-sided platforms. A nightclub, such as Bungalow 8 in Manhattan, provides a platform where men and women can meet and search for interactions and potentially dates. The club needs to get two groups of customers on board its platform to have a service to offer either one: it needs to get both men and women to come. Moreover, the relative proportion of men and women matters. A singles club with few women will not attract men, and a club with few men will not attract women. Pricing is one way to get the balance right. The club might want to offer women a break if they are in short supply (through a lower price or free drinks). Or it might want to ration the spots to ensure the appropriate number of women; popular clubs typically have queues waiting outside, and women are picked out of line disproportionately.

The dating club example motivates the informal definition of a two-sided platform that we introduced in the beginning paragraph. There are two groups of customers—men and women. Members of each group value members interacting with members of the other group. And the platform provides a place for them to get together and interact. By doing so it enables members of these two groups to capture various benefits from having access to each other.

In their 2006 paper, Rochet and Tirole have proposed a formal definition:

> "A market[5] is two-sided if the platform can affect the volume of transactions by charging more to one side of the market and reducing the price paid by the other side by an equal amount; in other words, the price structure matters, and platforms must design it so as to bring both sides on board."[6]

4 Although, for the most part, we will use the term two-sided platform the reader should note that some platforms have more than two distinct groups of customers. Digital media platforms, for example, often have four: users, developers, hardware makers, and content providers.

5 Note that the word market below is being used in the loose manner that is the custom among economists and not in the antitrust sense. The Rochet-Tirole definition would be more precise if it said "A two-sided platform business exists if …."

6 Jean-Charles Rochet & Jean Tirole, *Two-Sided Markets: A Progress Report,* RAND J. Econ. (Autumn 2006).

To satisfy this definition, "the relationship between end-users must be fraught with residual externalities" that customers cannot sort out for themselves.[7] That is clear in the case of the dating environment. In contrast, in the textbook wheat market there are no externalities connecting buyers and sellers, and the price structure doesn't matter: a tax on wheat levied on buyers has the same effect on quantity as the same tax levied on sellers.

In addition, it must not be possible for the two sides to arbitrage their way around the price structure chosen by the platform. Men and women, for example, want to be able to search for dates among a large number of opposites. It is hard to conceive of a practical mechanism for women to reward men who come to a singles club but who they reject. Likewise, for the other two-sided platform industries we consider it is difficult, if not impossible, for customers on one side to make side payments to customers on the other side. As a result the platform owner can institute a pricing structure to harness indirect network effects, and it is not feasible for customers to defeat this pricing structure through arbitrage. Generally, one can think of two-sided platforms as arising in situations in which there are externalities and in which transactions costs, broadly considered, prevent the two sides from solving this externality directly. The platform can be thought of as providing a technology for solving the externality in a way that minimizes transactions costs.

THINK OF TWO-SIDED PLATFORMS AS ARISING IN SITUATIONS IN WHICH THERE ARE EXTERNALITIES AND IN WHICH TRANSACTIONS COSTS, BROADLY CONSIDERED, PREVENT THE TWO SIDES FROM SOLVING THIS EXTERNALITY DIRECTLY.

It is helpful to review four different types of two-sided platforms: exchanges, advertiser-supported media, transaction devices, and software platforms.[8]

A. EXCHANGES

Exchanges have two groups of customers, who can generally be considered "buyers" and "sellers." The exchange helps buyers and sellers search for feasible contracts—that is where the buyer and seller could enter into a mutually advantageous trade—and for the best prices—that is where the buyer is paying as little as possible and the seller receiving as much as possible. (In organized exchanges,

7 As a result a necessary condition for a market to be two-sided is that the Coase theorem does not apply to the transaction between the two sides. For more details, *see* Rochet & Tirole (2006), *id*.

8 For discussion, see David S. Evans, Andrei Hagiu, & Richard Schmalensee, Invisible Engines: How Software Platforms Drive Innovation and Transform Industries, ch. 3 (MIT Press 2006). We refer there to software platforms more generally as shared input facilities. Armstrong uses the term "competitive bottlenecks" to refer to certain shared-input facilities. Although his discussion is analytically sound, his term is pejorative and has a meaning in competition law that differs from the one he assigns to it. *See* Mark Armstrong, Competition in Two-Sided Markets (EconWPA, working paper, 2005).

such as the New York Stock Exchange, it is often more useful to think of the two sides as liquidity providers—specialists or market-makers who quote prices to both buyers and sellers and thus bring liquidity to the market—and liquidity consumers—ordinary customers who accept liquidity providers' offers.[9]) We use the term buyers and sellers here loosely. The term, "exchanges," covers various matchmaking activities such as dating services and employment agencies. It also covers traditional exchanges such as auction houses, internet sites for business-to-business, person-to-business, and person-to-person transactions, various kinds of brokers (insurance and real estate) and financial exchanges for securities and futures contracts. Finally, exchanges include a variety of businesses that provide brokerage services. These include publishers (readers and authors), literary agents (authors and publishers), travel services (travelers and travel-related businesses), and ticket services (people who go to events, and people who sponsor events).

Exchanges provide participants with the ability to search over participants on the other side and the opportunity to consummate matches. Having large numbers of participants on both sides increases the probability that participants will find a match. Depending on the type of exchange, however, a larger number of participants can lead to congestion. That is the case with physical platforms such as singles clubs or trading floors. Moreover, participants may derive some value from having the exchange prescreen participants to increase the likelihood and quality of matches.

Some exchanges charge only one side. For example, only sellers pay directly for the services provided by eBay. This is also true for real-estate sales in the United States. Other exchanges charge both sides, although the prices may bear little relation to side-specific marginal costs. Internet matchmaking services charge everyone the same, for instance, while, as we mentioned, physical dating environments sometimes charge men more than women. Auction houses charge commissions to buyers and sellers. Insurance brokers historically charged both insurance customers and insurance providers in some types of transactions (some have agreed not to charge both as a result of settlements of lawsuits brought by the New York State Attorney General).

B. ADVERTISING-SUPPORTED MEDIA

Advertising-supported media such as magazines, newspapers, free television, and web portals are based on a two-sided business model. The platform either creates content (newspapers) or buys content from others (free television). The content is used to attract viewers. The viewers are then used to attract advertisers. There is a clear indirect network effect between advertisers and viewers—advertisers value platforms that have more viewers; the extent to which viewers value adver-

9 Bernhard Friess & Sean Greenaway, *Competition in EU Trading and Post-Trading Service Markets*, 2 Competition Pol'y Int'l (2006).

tisers is the subject of more debate but we suspect that viewers value advertisers more than they might admit.[10]

Most advertising-supported media earn much of their revenues—and probably all of their gross margin—from advertisers.[11] Print media are often provided to readers at something close to or below the marginal cost of printing and distribution.[12] In some cases—such as yellow page directories and some newspapers—they are provided for free. Free television is just that. And most web portals—Google and Yahoo for example—receive revenue only from advertisers.

C. TRANSACTION SYSTEMS

Any method for payment works only if buyers and sellers are willing to use it. Humans switched from barter when they were agreed on a standard medium for exchange—such as metallic coins or seashells. Governments facilitated this by ensuring the integrity of coins (to various degrees) and by using government-issued coinage for buying and selling. Cash, which has no intrinsic value in most modern economies, provides a payment platform because buyers and sellers expect that other buyers and sellers will use it. Of course the government facilitates this with various laws and through its own buying and selling activities.

For-profit transaction systems are based on the same principles although they have challenges that governments—which at least in principle can create a platform by fiat—do not necessarily have. Although bank checks and travelers' checks are also examples of for-profit transaction systems, we focus on payment cards, which have been the subject of significant competition policy scrutiny in many countries.

---

10 *See, e.g.,* James M. Ferguson, *Daily Newspaper Advertising Rates, Local Media Cross-Ownership, Newspaper Chains, and Media Competition*, 26 J.L. & ECON. 637 (1983) ("Readership studies show that advertising, especially retail advertising, is considered as important as, or more important than, editorial content.") and R.D. Blair & R.E. Romano, *Pricing Decisions of the Newspaper Monopolist*, 59 SOUTHERN ECON. J. 731 (1993) ("circulation demand rises with increases in the quantity of advertising").

Other studies have shown that, unlike Americans, readers in certain European countries are averse to advertising. *See, e.g.*, Nathalie Sonnac, *Readers' Attitudes Toward Press Advertising: Are They Ad-Lovers or Ad-Averse?*, 13 J. MEDIA ECON. 249 (2000). On the other hand, TiVo and other related products that permit ad avoidance and deletion are very popular currently, with one study citing that TiVo viewers skip about 60 percent of commercials. *See A Farewell to Ads?*, THE ECONOMIST, Apr. 15, 2004.

11 In a two-sided platform there is no rigorous way to define the profit "earned" by one side or the other. Not only are there typically costs that are common to both sides (the floor of the New York Stock Exchange, for instance), outlays that build business on one side of the market (via product enhancement, say) will also tend, via the externality, to build business on the other side. By "gross margin" we mean the difference between revenue and the variable costs, if any, that depend entirely on the volume on only one side of the market. The cleanest examples of such a cost would be the manufacturing costs of videogame consoles or the marginal printing costs of newspapers or yellow page directories.

12 Blair & Romano, *supra* note 10.

Diners Club started the first two-sided payment system in 1950. Before then stores issued payment cards to their customers for use only at their stores. Diners Club began by getting a set of restaurants to agree to take its card for payment; that is to agree to let Diners Club reimburse the restaurant for the meal tab and then in turn collect the money from the cardholder. It also persuaded individuals to take its card and use it for payment. Starting with a small base in Manhattan it grew quickly throughout the United States and other countries.

Diners Club initially charged restaurants seven percent of the meal tab; cardholders had to pay an annual fee, which was offset in part by the float they received as a result of having to pay their bills only once a month. As a result Diners Club earned most of its revenue—and most likely all of its gross margin—from merchants. Other entrants into the charge and debit card businesses have followed this same approach. Determining who pays in the case of credit cards is a bit more complicated since that product bundles a transaction feature (for which the cardholder pays little) and a borrowing feature (for which the cardholder incurs finance charges). However, it is safe to say that merchants are the main source of revenue for credit cards held by people who do not revolve balances.

American Express, Discover, and, until its recent absorption into MasterCard, Diners Club, set prices to merchants—the merchant discount, which gives rise to a positive variable transaction price—and to cardholders—annual fees and various rewards which may give rise to negative variable transaction prices. Card associations such as MasterCard and Visa have been examples of cooperative two-sided platforms. For a transaction to be consummated there has to be an agreement on the division of profits and the allocation of various risks between the entity that services the merchant and the entity that services the cardholder. Most card associations set this centrally as, in effect, a standard contract between the businesses that service the two sides. Typically, they agree that the entity that services the merchant pays a percentage of the transaction—the "interchange fee"—to the entity that services the cardholder. This fee ultimately determines the relative prices for cardholders (issuers obtain a revenue stream which they compete for) and merchants (acquirers pass the cost of the interchange fee onto merchants). This centrally set fee has been the subject of litigation and regulatory scrutiny, as we discuss below.[13]

D. SOFTWARE PLATFORMS

A software platform provides services for applications developers; among other things, these services help developers obtain access to the hardware for the computing device in question. Users can run these applications only if they have the same software platform as that relied on by the developers; developers can sell

13 David S. Evans & Richard Schmalensee, The Economics of Interchange Fees and Their Regulation: An Overview (MIT Sloan, Working Paper, 2005), *in Interchange Fees in Credit and Debit Card Industries* 73-120 (Federal Reserve Bank of Kansas City, 2005).

their applications only to users that have the same software platform they have relied on in writing their applications.

Software platforms are central to several important industries. These include personal computers (e.g., Apple, Microsoft); personal digital assistants (e.g., Palm, Treo); 2.5G+ mobile telephones (e.g., Vodafone, DoCoMo); video games (e.g., Sony PlayStation, Xbox); and digital music devices (e.g., Creative Zen Micro, Rio Carbon). With the exception of video games, the software platform owners make most of their revenue, and all of their gross margin, from the user side; developers generally get access to platform services for free, and they obtain various software products that facilitate writing applications at relatively low prices. Videogame console manufacturers, on the other hand, typically receive most of their gross margin from licensing access to the software and hardware platforms to game developers; they sell the videogame console at close to or below manufacturing cost.

Software platforms facilitate a market for applications by reducing duplicative costs. Application programs need to accomplish many similar tasks. Rather than each application developer writing the code for accomplishing each task the software platform producer incorporates code into the platform. The functions of that code are made available to application developers through an application program interface (API). The user benefits from this consolidation as well since it reduces the overall amount of code required on the computer, reduces incompatibilities between programs, and reduces learning costs.[14] An important consequence of this reduction in cost is an increase in the supply of applications for the platform, an increase in the value of the software platform to end users, and positive feedback effects to application developers.

E. METHODS FOR MINIMIZING TRANSACTIONS COSTS

The fundamental role of a two-sided platform in the economy is to enable parties to realize gains from trade or other interactions by reducing the transactions costs of finding each other and interacting. Two-sided platforms do this by matchmaking, building audiences, and minimizing costs. Different platforms engage in these activities to different degrees. Software platforms are mainly about minimizing duplication costs, advertising-supported media in mainly about building audiences, and exchanges are mainly about matchmaking. But they all seem to engage in each to some degree. All platforms help reduce costs by providing a virtual or physical meeting place for customers. We will see that these platforms all minimize transactions costs by through matchmaking, audience-making, and cost minimization through the elimination of duplication.[15]

14 *See* Evans, Hagiu, & Schmalensee, *supra* note 8.

15 *See* David S. Evans & Richard Schmalensee, Catalyst Code: The Strategies Behind the World's Most Successful Companies (Harvard Business School Press 2007).

MySpace provides an example of how a two-sided platform engages in all three functions. It is a popular internet site where young people can post their profiles and develop networks of friends. It provides matchmaking between the people who sign up as well as the advertisers who would like to meet them. It builds audiences for advertisers as well as members—particularly musicians—who want to make themselves known. And it reduces the costs to people of getting together by providing a common meeting place.

III. Economic Principles

The theoretical economics literature on two-sided platforms is relatively new. Economists have derived many results based on stylized models that apply to some of the industries described above. The precise results are sensitive to assumptions about the economic relationships among the various industry participants. Even for these special cases it has turned out to be challenging to derive results without making further assumptions about the precise nature of the demand, cost, and indirect network effects relationships.[16] Nevertheless, several principles have emerged that seem to be robust. They appear to depend only on the assumptions that the platform has two groups of customers, that there are indirect network externalities, and that the customers cannot solve these externalities themselves.

A. PRICING

To see the intuition behind pricing consider a platform that serves two customer groups ***A*** and ***B***. It has already established prices to both groups and is considering changing them.[17] If it raises the price to members of group ***A*** fewer ***A***s will join. If nothing else changed the relationship between price and the number of ***A***s would depend on the price elasticity of demand for ***A***s. Since members of group ***B*** value the platform more if there are more ***A***s fewer ***B***s will join the platform at the current price for ***B***s. That drop-off depends on the indirect network externality which is measured by the value that ***B***s place on ***A***s. But with fewer ***B***s on the platform, ***A***s also value the platform less leading to a further drop in their demand. There is a feedback loop between the two sides. Once this effect is taken into account, the effect of an increase in price on one side is a decrease in demand on the first side because of the direct effect of the price elasticity of demand and on both sides as a result of the indirect effects from the externalities.

A few equations will make this point more sharply for readers familiar with the concept of elasticity. The situation described just above can be summarized by two demand functions: $Q^A = D^A(P^A,Q^B)$ and $Q^B = D^B(P^B,Q^A)$. The first of these gives

16 That is, the models are based on assuming particular functional forms—e.g. linear—for relationships.

17 To keep matters simple we consider the case where each side is charged a membership fee as in Mark Armstrong, Competition in Two-Sided Markets (EconWPA, Working Paper, 2005). More generally, platforms are natural businesses for two-part tariffs involving an access fee and a usage fee.

participation by members of group **A** as a function of the price charged to group **A** and participation by group **B**, and the second gives participation by members of **B** similarly. Let $e^I = -(\partial D^I/\partial P^I)(P^I/Q^I)$, for $I = A,B$. These are the own-price elasticities for each group, holding constant *participation* by the other (i.e., ignoring the externalities linking the two groups). Let $\theta^I_J = (\partial D^I/\partial Q^J)(Q^J/Q^I)$ for $I,J = A,B$ and $I \neq J$. These elasticities measure the strengths of the externalities connecting the two groups. In the normal two-sided case, both would be expected to be positive. Finally, let $E^I = -(dQ^I/dP^I)(P^I/Q^I)$ for $I = A,B$. These are the ordinary own-price elasticities, computed assuming other prices remain constant but allowing participations (quantities) to vary. Differentiating both demand functions totally with respect to either price, and solving, yields:

$$E^I = e^I/(1 - \theta^I_J \theta^J_I); I,J = A,B; I \neq J.$$

Even if the **A**s are not particularly price-sensitive, and as long as the externalities between the groups are strong (in either direction!), participation by group **A** may be highly sensitive to the price its members are charged, and similarly for group **B**. Even a small response by group **A** to a price change will trigger a response by group **B**, which in turn will produce a response by **A**, and so on. (The equation above assumes that these response sequences converge.)

The platform of course would like to find the prices that maximize its profits by taking these same sorts of considerations into account. For a single-sided business that would occur by selecting the output at which marginal revenue equals marginal cost and then charging the corresponding price for this quantity from the demand curve. (This equilibrium is often described by the Lerner formula that says that the price marginal-cost margin equals the inverse of the own-price elasticity of demand.) For two-sided platforms three results appear to be robust:

1) The optimal prices depend in a complex way on the price sensitivity of demand on both sides, the nature and intensity of the indirect network effects between the two sides, and the marginal costs that result from changing output of each side.

2) The profit-maximizing, non-predatory price for either side may be below the marginal cost of supply for that side or even negative.

3) The relationship between price and cost is complex, and the simple formulas that have been derived for single-sided markets do not apply.

For many platforms it is possible to charge two different kinds of prices: an access fee for joining the platform and a usage fee for using the platform. Although these are interdependent, one can think of the access fee as mainly affecting how many customers join the platform and the usage fee as mainly affecting the volume of interactions between members of the platform. Most software platforms charge access fees to users—they have to license the software platform but then can use it as much as they want—and do not charge access or usage fees to developers. Videogame console vendors, though, charge a usage fee to

game developers—a royalty based on the numbers of games that are sold; users pay this usage fee indirectly through their purchase of games for the console. Payment card systems generally charge merchants a usage fee but no access fee. Cardholders may pay an access fee (the annual card fee); they often pay either no usage fee or a negative one (to the extent they receive rewards based on transactions volume).

The profit-maximizing reliance on access versus usage fees depends on many factors including the difficulty of monitoring usage and the nature of the externality between the two sides. Cardholders care about card acceptance, for instance, while merchants care about usage. It thus seems sensible not to charge merchants for access and not to charge consumers for usage.

The empirical evidence suggests that prices that are at or below marginal cost are common for two-sided platforms. Table 1 summarizes some relevant evidence.

Table 1[18]

Examples of two-sided pricing structures

Industry	Side	Access	Usage
Heterosexual Dating Clubs	Men	√	√
	Women	√	√
DoCoMo i-Mode	User	√	√
	Content-Provider	Ø	√
U.S. Real Estate Brokers	Seller	Ø	√
	Buyer	Ø	Ø
Magazines	Reader	√ (≤MC)	Ø
	Advertiser	Ø	√
Shopping Malls	Shopper	–	Ø
	Store	√	Ø
PC Operating Systems	User	√	Ø
	Developer	√ (<MC)	Ø
Video Game Consoles	Player	√ (≤MC)	Ø
	Game Developer	√ (<MC)	√
Payment Card Systems	Merchant	Ø	√
	Cardholder	√ (<MC)	Ø

Note: √ and Ø indicate that the entity either pays or does not pay, respectively, for either access or usage of the two-sided platform. Items in parentheses indicate where marginal cost or below marginal cost pricing is prevalent for a particular side of a two-sided platform.

18 This table shows pricing structures that are common in these industries. In many cases, fees will differ from these pricing structures. For example, some clubs offer free entry to women, some magazines offer free subscriptions, some videogame players pay fees for on-line play, and some payment cardholders do not pay fees for their cards and/or get usage based rewards. For dating clubs, usage fees

footnote 18 cont'd on next page

B. DESIGN DECISIONS

Two-sided platforms are in the business of encouraging customers to join their platforms and stimulating them to interact with each other once they have joined. They design their platforms with this in mind. This can lead to decisions that in a narrow sense harm one side.

A simple example is a shopping mall. Shoppers would prefer to get to stores in the least amount of time. Merchants would like to maximize the amount of foot traffic outside their stores and therefore the number of potential shoppers. Shopping malls are sometimes designed to encourage shoppers to pass by many stores (e.g., by putting the up and down escalators at different ends of the mall).

Advertising-supported media are another obvious example. Viewers would like to gain access to the content—and perhaps even the advertisements of their choice—in the most convenient way. Some magazines are laid out to make it difficult even to find the table of contents or to find the continuation of an article without thumbing through many advertisements. Television watchers might benefit from having advertisements clustered at the beginning or the end of each program, but television providers (in the United States, at least) typically intersperse the advertisements and precede them perhaps with a cliffhanger to discourage viewers from taking a long break.

Two-sided platforms may also bundle features that directly benefit side ***A*** but harm side ***B*** (putting aside the indirect externalities from increasing the participation of side ***A***).[19] All software platforms include features for example that do not benefit most users. However, some developers value each of these features and in particular value knowing that any user of the software will have that feature and therefore be able to run its applications. All payment card systems require merchants that take their cards for payment to take any of their cards for payment, regardless of who presents it or which entity issued it. Some merchants would benefit from being selective—taking cards only from people who lack cash, for example. But this would reduce the confidence that cardholders have that their cards will be taken at stores that display the acceptance mark. (We will see later that special cases of these requirements, linking acceptances of credit and debit cards, have given rise to tying claims. This paragraph is not meant to suggest that tying

footnote 18 cont'd
for men and women refer to fees for drinks in the club. For real estate, the usage fee for sellers refers to the fee for selling a house; there is typically no fee for using the system to list or show a house. For shopping malls, the negative usage fee for shoppers refers to the free parking that is commonly available. For videogame consoles, players do not pay a fee for using the console, although they do pay for video games to the game developer (which in some cases is the same firm that makes the console and in other cases pays a royalty to the console manufacturer). For payment cards, cardholders are also subject to penalty fees, such as for exceeding credit limits or for late payments; we have not included these fees in the table.

19 *See* Rochet & Tirole (2006), *supra* note 6.

could not be used in an anticompetitive way by two-sided platforms but rather to point out that there is an additional efficiency explanation for at least one aspect of this practice that does not arise in one-sided businesses.)

C. RULES AND REGULATIONS

Given that platforms promote interactions between customers and seek to harness indirect network externalities it should come as no surprise that two-sided platforms have an incentive to devise rules and regulations that promote these externalities and limit negative externalities between customers. The most sophisticated rules and regulations may be those employed by exchanges. All exchanges have rules against "front-running," for instance. This practice occurs when a broker receives a large purchase order from a customer, first buys on his own account, and then executes the customer order, which drives the price up slightly, and then sells on his own account and pockets the resulting profit. Banning this practice directly harms brokers, but it makes buyers and sellers more confident that they are getting the best price possible, and thereby boosts volume on the exchange.

COOPERATIVE TWO-SIDED PLATFORMS HAVE FURTHER NEED FOR RULES AND REGULATIONS BECAUSE THE BEHAVIOR OF THEIR MEMBERS CAN AFFECT THE VALUE OF THE TWO-SIDED PLATFORM AS A WHOLE.

Cooperative two-sided platforms have further need for rules and regulations because the behavior of their members can affect the value of the two-sided platform as a whole. Visa, for example, has rules that govern the appearance of cards issued by members, to provide some uniformity for the common brand, as well as to prevent members from using the brand inappropriately. The system also has rules that address disputed transactions. Acquirers would have an incentive to favor their customers (merchants) in a dispute while issuers would favor their customers (cardholders). The system's rules attempt to find a balance between these competing interests, to increase the attractiveness of the system as a whole.

IV. Industrial Organization of Markets with Two-Sided Platforms

Casual empiricism shows that industries with two-sided platforms are quite diverse. We explain some of the basic determinants of this heterogeneity from a theoretical perspective and then document aspects of it by surveying industries in which two-sided platforms are central.

A. DETERMINANTS OF PLATFORM SIZE AND STRUCTURE

Five fundamental factors determine the relative size of competing two-sided platforms. Table 2 summarizes the factors we discuss below and their effect on size (with a "+" indicating that there is a positive association between size and the factor).

Cause	Effect on Size/Concentration
Indirect network effects	+
Scale economies	+
Congestion	–
Platform differentiation	–
Multi-homing	–

Table 2

Determinants of industry structure

1. Indirect Network Effects

Indirect network effects between the two sides promote larger and fewer competing two-sided platforms. Platforms with more customers of each group are more valuable to the other group. For example, more users make software platforms more valuable to developers and more developers make software platforms more valuable to users. These positive-feedback effects make platforms with more customers on both sides more valuable to both sets of customers. To take another example, a payment card system whose cards are taken at more merchants is more valuable to card users—that is why we see card systems touting their acceptance ("MasterCard: No card is more accepted.") in consumer advertisements.

If there were no countervailing factors, we would expect that indirect network effects would lead two-sided platforms to compete ***for*** the market. First movers would have an advantage, all else being equal. We would have the familiar story that the firm that obtains a lead tends to widen that lead as a result of positive-feedback effects and therefore wins the race for the market.[20] Other firms could compete with this advantage only if they offered consumers on either side something that offset the first mover's size advantage.

Indirect network effects may decline with the size of the platform. For example, the probability of finding a match increases at a diminishing rate with the number of individuals on either side (buyers or sellers, men or women).[21] At some point positive externalities from more participants may turn into negative externalities in the form of congestion as discussed below.

20 *See, e.g.,* David S. Evans & Richard Schmalensee, *A Guide to the Antitrust Economics of Networks*, 10 Antitrust Mag. 36 (1996) and Carl Shapiro & Hal R. Varian, Information Rules: A Strategic Guide to the Network Economy (Harvard Business School Press 1999).

21 *See* Evans, *supra* note 1.

2. Economies and Diseconomies of Scale

For many two-sided platforms there would appear to be significant fixed costs of providing the platform. This should lead to scale economies over some range of output. For example, card payment systems have to maintain networks for authorizing and settling transactions for cardholders and merchants (and for their proxies—issuers and acquirers—in the case of association-based payment systems such as MasterCard). The costs of developing, establishing, and maintaining these networks are somewhat independent of volume. To take another example, there is a fixed cost of developing a software platform but a low marginal cost of providing that platform to developers and end users. In some cases the scale economies may mainly operate on one side. For example, there are scale economies in providing newspapers to readers (there is a high fixed cost of creating the newspaper and a relatively low marginal cost of reproducing and distributing it) but not in providing space to advertisers. Lastly, some physical platforms such as trading floors and singles clubs have scale economies at least in the short run, up to their capacity levels.

Diseconomies may set in at some point for various reasons on one or both sides. For example, to persuade existing end users to replace (i.e., upgrade) their existing software platforms software, platform vendors have to add features and functionality. Many of these improvements may be designed to encourage application developers to write new or improved applications for the platform that in turn benefit end users. However, as software platforms have gotten larger and more complex, it has become more expensive and time consuming to add features and functionality. The most recent version of the Apple OS took four months longer to develop than the previous version.[22] Microsoft's Vista operating system has also been plagued with very long delays.

3. Congestion and Search Optimization

Several design issues tend to limit the size of two-sided platforms. Physical platforms such as trading floors, singles clubs, auction houses, and shopping malls help customers search for and consummate mutually advantageous exchanges. At a given size expanding the number of customers on the platform can result in congestion that increases search and transaction costs.[23] It may be possible to reduce congestion by increasing the size of the physical platform, but that in turn may increase search costs. Indeed, to optimize searching for partners, two-sided platforms may find that it is best to limit the size of the platform and prescreen

22 For Apple OS release dates, see Jason Snell, *Jaguar unleashed: Mac OS X 10.2 Arrives*, MACWORLD, Sept. 1, 2002; Sarah Stokely, *Apple Sets Panther Release Date*, IDG DATA, Oct. 10, 2003.; and, Steven Musil, *This Week in Tiger: Apple releases Mac OS X 10.4,* CNET NEWS, Apr. 29, 2005.

23 For a general discussion on matching, search, and congestion see, for example, Robert Shimer & Lones Smith, *Matching, Search, and Heterogeneity*, 1 ADVANCES IN MACROECONOMICS (2001) and Mark Rysman, *Competition Between Networks: A Study of the Market for Yellow Pages*, 71 REV. ECON. STUDIES 483 (2004b).

the customers on both sides to increase the probability of a match. One might argue that singles-type clubs do this explicitly (deciding who can get into an "exclusive" club) or implicitly (compare church-oriented singles groups and Club Med resorts). We will return to this subject below in discussing platform differentiation. Congestion may arise on one side alone. For example, increasing the volume of advertising in a newspaper may not only crowd out the content that attracts the readers but also result in a cacophony of messages that reduces the effectiveness of any particular advertisement.

4. Platform Differentiation and Multi-Homing

Platforms can differentiate themselves from each other by choosing particular levels of quality (what is known as "vertical differentiation") with consumers choosing the higher or lower quality of platform depending on the income and relative demand for quality. There are, for example, upscale and downscale malls. Platforms can also differentiate themselves from each other by choosing particular features and prices that appeal to particular groups of customers (what is known as "horizontal differentiation"). Thus there are numerous advertising-supported magazines that appeal to particular segments of readers and advertisers (e.g., *Cape Cod Bride* or *Fly Fisherman*).

Horizontal differentiation can result in customers choosing to join and use several platforms—a phenomenon that Rochet and Tirole have called "multi-homing". Customers find certain features of different competing platforms attractive and therefore rely on several. Payment cards are an example of multi-homing on both sides. Most merchants accept credit and debit cards from several systems, including ones that have relatively small shares of cardholders. Many cardholders carry multiple cards, although they may tend to use a favorite one most often.[24] Advertising-supported media also has multi-homing on both sides—advertisers and viewers rely on many differentiated platforms. Other two-sided platforms have multi-homing only on one side. Most end-users rely on a single software platform for their personal computers, for instance, while many developers write for several platforms.

B. EMPIRICAL EVIDENCE ON TWO-SIDED INDUSTRY STRUCTURE

It is possible to see some regularities across industries in which two-sided platforms appear to be the dominant form of organization. Table 1 above and Table 3 reveal several features:

- It is relatively uncommon for industries based on two-sided platforms to be monopolies or near monopolies. Some industries based on two-sided platforms have several large differentiated platforms, while others have many small platforms that are differentiated by location as well as along other dimensions.

24 Mark Rysman, An Empirical Analysis of Payment Card Usage (Boston University Department of Economics, Working Paper, 2004).

Table 3

Presence of multi-homing and largest competitor share of selected two-sided platforms

Multi-Sided Platform	Sides	Presence of Multi-homing	Largest Competitor Share in the United States
Residential Property Brokerage	Buyer Seller	Uncommon: Multi-homing may be unnecessary, since a multiple listing service allows the listed property to be seen by all member agencies' customers and agents.	Fifty largest firms have a 23% share. (2002)
Securities Brokerage	Buyer Seller	Common: The average securities brokerage client has accounts at three firms. Note that clients can be either buyers or sellers or both.	Four largest firms accounted for 37% of in securities brokerage and 16% in financial portfolio management. (2002)
Newspapers and Magazines	Reader Advertiser	Common: In 1996, the average number of magazine issues read per person per month was 12.3. Also common for advertisers: for example, AT&T Wireless advertised in the New York Times, The Wall Street Journal, and Chicago Tribune, among many other newspapers, on Aug. 26, 2003.	Wall Street Journal had a 28% share of the five largest newspapers. (2001)
Network Television	Viewer Advertiser	Common: For example, viewers in Boston, Chicago, Los Angeles, and Houston, among other major metropolitan areas, have access to at least four main network television channels: ABC, CBS, FOX, and NBC. Also common for advertisers: for example, Sprint places television advertisements on ABC, CBS, FOX, and NBC.	U.S. law forbids broadcasters from owning TV stations reaching more than 35% of the nation's television audience.
Operating System	End User Application Developer	Uncommon for users: Individuals typically use only one operating system. Common for developers: As noted earlier, the number of developers that develop for various operating systems indicates that developers engage in significant multi-homing.	Microsoft has a 96% share of revenue of client operating systems. (2004)
Video Game Console	Game Player Game Developer	Varies for players: The average household (that owns at least one console) owns 1.4 consoles. Common for developers: For example, in 2003, Electronic Arts, a game developer, developed for the Nintendo, Microsoft, and Sony platforms.	Sony PS1 and PS2 had a 63% share of console shipments in North America. (2003)
Payment Card	Cardholder Merchant	Common: Most American Express cardholders also carry at least one Visa or MasterCard. In addition, American Express cardholders can use Visa and MasterCard at almost all places that take American Express.	The Visa system had a 45% share of all credit, charge, and debit purchase volume. (2004)

Source: Adapted from David S. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325 (2003). Industry share data from United States Census Bureau, 2002 Economic Census, available at http://www.census.gov/econ/census02/guide/INDSUMM.HTM; "Top 20 U.S. Daily Newspapers by Circulation," Newspaper Association of America (2001), at http://www.naa.org/info/facts01/18_top20circ/index.html (accessed Feb. 21, 2007); Stephen Labaton, *U.S. Backs Off Rules for Big Media*, NY TIMES, Jan. 28, 2005; A. Gillen & D. Kusnetzky, *Worldwide Client and Server Operating Environments 2004-2008* Forecast, IDC MARKET ANALYSIS, No. 32452 (Dec. 2004); Schelley Olhava, *Worldwide Videogame Hardware and Software 2004-2008* Forecast and Analysis, IDC MARKET ANALYSIS, No. 31260 (May 2004); THE NILSON REPORT, No. 828 (Feb. 2005); THE NILSON REPORT, No. 833 (May 2005).

- Multi-homing on at least one side is common. Horizontal product differentiation tends to be the norm.
- Asymmetric pricing is relatively common. Many two-sided platforms appear to obtain the preponderance of their operating profits (rev-

enues minus direct costs) from one side. A nontrivial portion of two-sided platforms appear to charge prices that are below marginal cost or below zero.

V. Overview of Antitrust Cases Involving Two-Sided Markets

Many antitrust cases have involved two-sided platforms. A few—including several important ones—seem to have touched on two-sided issues before economists began to address them formally. And some are based on analyses of markets and practices that, putting aside whether they led to the correct outcome or not, are analytically wrong from the perspective of the two-sided literature.

Table 4 presents an overview of antitrust cases in the European Community and the United States that concern two-sided platforms. We have not done a systematic review of cases but have rather listed cases that have had a high profile in these

	Case	Case Type
Media	Times Picayune	Monopolization
	Magill	Refusal to supply
	BT Yellow Pages	Monopolization
	Lorain Journal	Exclusive dealing
Exchanges	Sotheby's–Christies	Cartel
	Marsh McLennan	Cartel
	Stock Exchanges	Merger
	Mobile operators	Excessive Pricing

	Case	Case Type
Transaction Systems	NaBanco	Cartel
	Wal–Mart	Tying
Software Platforms	Microsoft–Browser	Monopolization, Tying
	Microsoft–Media Player	Tying
	Nintendo	Exclusivity

Table 4[25]

Summary of leading cases by two-sided platform type

25 United States v. Times-Picayune Publishing Co., 345 U.S. 594 (1953); Joined Cases C-241/91 P and C-242/91 P, RTE, BBC, and ITP v. Commission of the European Communities (*Magill*), 1995 E.C.R. I-00743 (Apr. 6, 1995); U.K. Competition Commission, Classified Directory Advertising Services (1996); U.K. Office of Fair Trading, Classified Directory Advertising Services: Review of Undertakings Given by BT to the Secretary of State in July 1996 (2001); United States v. Lorain Journal Co., 342 U.S. 143 (1951); United States v. Taubman, 297 F.3d 161 (2d Cir. 2002); State of New York v. Marsh & McLennan Companies, Inc., et al., Complaint filed October 14, 2004, Index No. 04-403342; U.K. Competition Commission, A report on the proposed acquisition of London Stock Exchange plc by Deutsche Börse AG or Euronext NV (2005); U.S. Department of Justice, Department of Justice Antitrust Division Statement on the Closing of its Two Stock Exchange Investigations (Nov. 16, 2005); U.K. Office of Communications, Wholesale Mobile Voice Call Termination (Jun. 1, 2004); National Bancard Corp. v. Visa U.S.A., Inc., 779 F.2d 592, 602 (11th Cir. 1986); In re Visa Check/MasterMoney Antitrust Litigation, 192 F.R.D. 68 (E.D.N.Y 2000); United States v. Microsoft, 87 F. Supp. 2d 30 (D.D.C. 2000); Commission of the European Communities v. Microsoft, Case COMP/C-3/37.792/Microsoft; Atari Games Corp. v. Nintendo, 975 F.2d 832 (Fed. Cir. 1992).

jurisdictions with which we are generally familiar.[26] The cases span all of the major categories of two-sided platforms and involve the spectrum of competition policy issues. This section summarizes some key issues that arose in several of these cases.

A. *NABANCO*

In *NaBanco v. Visa*, the federal district court and the U.S. Court of Appeals for the Eleventh Circuit recognized several of the key features of what have become known as two-sided platforms. Visa was (and is) a cooperative of banks that issued cards and acquired those card transactions from merchants. It established a rule for governing the situation in which an individual whose card was issued by bank A paid with that card at a merchant acquired by bank B, where A and B are different banks. Although those banks could have a bilateral agreement, Visa established a default rule that among other things determined the allocation of the profits and risks of the transaction. This rule provided that given the various allocations of risks and costs that the bank that acquired the transaction (B) had to pay the bank (A) that issued the card a percent of the transaction amount; this percent is known as the interchange fee, and it was initially set at 1.95 percent.

NaBanco argued that the interchange fee violated Section 1 of the Sherman Act because it was a price set collectively by competitors. Visa argued that unlike classic price-fixing, the ability to set an interchange fee was a mechanism to allocate costs between the issuing and acquiring sides of the business and enhanced output by, among other things, limiting opportunistic behavior by individual members and avoiding the chaos of bilateral negotiations among thousands of member banks. The Eleventh Circuit concluded:

> "Another justification for evaluating the [interchange fee] under the rule of reason is because it is a potentially efficiency creating agreement among members of a joint enterprise. There are two possible sources of revenue in the VISA system: the cardholders and the merchants. As a practical matter, the card-issuing and merchant-signing members have a mutually dependent relationship. If the revenue produced by the cardholders is insufficient to cover the card-issuers' costs, the service will be cut back or eliminated. The result would be a decline in card use and a concomitant reduction in merchant-signing banks' revenues. In short, the cardholder cannot use his card unless the merchant accepts it and the merchant cannot accept the card unless the cardholder uses one. Hence, the [interchange fee] accompanies "the coordination of other productive or distributive efforts of the parties"

26 *See* J. Wotton's article in this issue (John Wotton, *Are Media Markets Analyzed as Two-Sided Markets?*, (3)1 Competition Pol'y Int'l 237–47 (2007)).

> that is "capable of increasing the integration's efficiency and no broader than required for that purpose."[27]

Professor William Baxter worked for Visa on this matter. His 1983 article in the *Journal of Law and Economics* presented many of the key concepts of two-sided markets within the context of the determination of interchange fees.[28] The modern literature now recognizes that the interchange fee is at least partly a device for determining the pricing structure for the card system.[29] Some regulators and antitrust authorities, while recognizing the two-sided nature of the business, have argued in recent years that the interchange fee is set at a level that encourages the overuse of cards.

B. STOCK EXCHANGE MERGERS

In recent years, stock exchanges have increasingly looked to merge with each other. In December 2004, Euronext and Deutsche Börse, respectively the second and third largest stock exchanges in Europe by value of trading, made bids to take over the London Stock Exchange, the largest stock exchange in Europe. Both bids were referred to the U.K.'s Competition Commission for investigation under U.K. competition law—they did not qualify for investigation by the European Commission under EU law. In its report, the Competition Commission expressed concerns about the ownership of clearing services by the Euronext or Deutsche Börse that was likely to result post merger. It was believed that ownership of clearing services by the London Stock Exchange's parent company would act as a barrier to potential competitor exchanges to the London Stock Exchange that needed access to same clearing service to be competitive. Both Euronext and Deutsche Börse made commitments that satisfied the concerns of the Competition Commission but as a result of business rather than regulatory reasons, neither deal went through.

In the United States, in 2005 the New York Stock Exchange agreed to merge with Archipelago, an electronic stock exchange, and the NASDAQ Stock Exchange agreed to merge with Instinet, also an electronic stock exchange. The U.S. Department of Justice approved both mergers, in part because it believed that there were no likely anticompetitive effects given the planned and likely entry of other firms. In 2006, the New York Stock Exchange and Euronext

27 National Bancard Corp. v. Visa U.S.A., Inc., 779 F.2d 592, 602 (11th Cir. 1986).

28 Baxter, *supra* note 2.

29 *See, e.g.*, Richard Schmalensee, *Payment Systems and Interchange Fees*, 50 J. Indus. Econ. 103 (2002); Jean-Charles Rochet & Jean Tirole, *Cooperation among Competitors: Some Economics of Credit Card Associations*, 33 Rand J. Econ. 549 (2002); *See* Rochet & Tirole, *supra* note 2; *See* Wright, *supra* note 1; David S. Evans & Richard Schmalensee, Paying with Plastic: The Digital Revolution in Buying and Borrowing (MIT Press 2005); and Evans & Schmalensee (2005), *supra* note 13.

announced they had agreed to merge. As of this writing, the transaction has recently received antitrust and regulatory approval in the United States and Europe, but has not yet been consummated.

Stock and other exchanges exhibit significant network effects. Fundamentally, more trading activity on the part of providers and consumers of liquidity tends to reduce spreads between bid and ask prices and to make markets more liquid, so that large blocks of stocks, options, or commodities can be bought or sold rapidly without a price penalty. And, of course, smaller bid-ask spreads and more liquidity tend to attract more trading. The more investors that come to a market, the more attractive that market becomes to liquidity providers, and the more liquidity providers are present, the more attractive the market is to investors.[30]

Traditionally, stock exchanges have tended to be local monopolies, due in large part to these network effects, to regulations that restricted cross-border trading and, historically in the United States, to communications costs that created a niche for regional exchanges like the Boston Stock Exchange. As these restrictions have been relaxed and communications costs have fallen, competition has increased generally, and many exchanges have abandoned their traditional non-profit, cooperative structures and become for-profit firms. In the United States, regional stock exchanges have had trouble competing with the NYSE, but competition between the NYSE and NASDAQ has intensified. There are now six competitive equity options exchanges in the United States; they are linked electronically so that investors are guaranteed the best available price, and the largest market shares hover below 40 percent. Stock exchanges have been ordered to provide such linkage; this is expected to happen in the first half of 2007 and may have a major effect on the competitive landscape.

In Europe, on the other hand, there has thus far been very little direct competition between the London Stock Exchange and other European exchanges, such as Euronext and Deutsche Börse. One key question in mergers between stock exchanges is whether network effects will continue to limit the scope for competition or whether falling communications costs and the computerization of the securities business will make global competition—of one sort or another—inevitable.

C. MICROSOFT MEDIA PLAYER

The European Commission found that Microsoft had abused a dominant position in operating systems by including media player technologies in Windows.[31] It argued that there were indirect network effects between the use of media play-

30 *See* Friess & Greenaway, *supra* note 9.

31 For contrary views on this case, *see* Maurits Dolmans & Thomas Graf, *Analysis of Tying Under Article 82 EC: The European Commission's Microsoft Decision in Perspective*, 27 World Competition 225 (2004). *See also* David S. Evans & A. Jorge Padilla, *Tying Under Article 82 EC and the Microsoft Decision: A Comment on Dolmans and Graf*, 27 World Competition 503 (2004).

ers and the provision of content. If more people have a particular media player, content providers will tend to encode content in that format. If more content is available in the format for a particular media player, users will tend to use that media player. The Commission argued that content providers would standardize on Windows Media Player because this player was available on most personal computers, which of course included Windows. In effect, the Commission argued that the existence of network effects would result in the "media player market" tipping to Windows Media Player.[32]

For its part Microsoft has agreed that there are indirect network effects but that the existence of such effects is not sufficient to tip a market to a single platform. In particular, it has argued that media players are horizontally differentiated products and that most content providers and many users engage in multihoming. Who is right on this score depends on factual disputes between the Commission and Microsoft that we do not consider here.

D. *MAGILL*

Magill is a leading EC case involving the compulsory licensing of intellectual property. What makes it interesting from a two-sided standpoint is that it involved several interlinked two-sided platforms. The defendants in the case were three television networks (RTE, BBC, and ITV) whose broadcasts were received in Ireland. RTE and ITV were two-sided platforms, receiving revenues from advertisers. RTE was also supported by licenses paid by consumers for having television sets. The BBC received similar revenues from licenses for television sets in the United Kingdom (but not Ireland). The BBC did not allow advertising and was not a two-sided platform. All three networks published an advertising-supported television guide that contained their own weekly listings; these were two-sided platforms. In addition they each provided their daily listings to newspapers—other two-sided platforms—that combined the listings.

Magill TV Guide (Magill) wanted to publish a weekly advertising-supported guide that contained the listings of the three networks. The networks complained that this violated their copyrights. The Commission and ultimately the EC courts concluded that there would be a market—in the antitrust sense—for a weekly television guide and that the refusal to supply the copyrighted information prevented the emergence of the weekly guide product. As it turns out, the weekly newspapers were the main beneficiaries of this decision since they started weekly television supplements included in the Sunday newspapers. Magill never made a successful go of it.

We will return to these issues when we discuss the analysis of market definition and market power. The key point is that the analysis by all the parties

32 CFI Order of Dec. 22, 2004, Case T-201/04 R 2, Microsoft Corporation v. Commission, at paras. 365 and 388, *available at* http://curia.eu.int/en/content/juris/index_form.htm.

(including the television networks) ignores a key side of the two-sided industry here—the advertisers who were the likely source of much of the revenue and profits—as well as the link between the guides and the television business.

VI. Antitrust Implications of Two-Sided Platform Economics

Whether the economics of two-sided platforms can assist in determining whether a merger or business practice is anticompetitive is, like many aspects of economics, an empirical question. As with market power generally two-sidedness is a matter of degree. Sometimes the two-sided nature of the business is critical for the analysis. Other times it is an interesting aspect of the industry that should be thought about but is not ultimately determinative. And still other times an industry may have two-sided aspects that are too insubstantial to matter.

A. MARKET DEFINITION AND MARKET POWER

The analysis of market power, and the associated issue of the definition of the relevant market are typically a central component of antitrust cases, although the reasons for this vary somewhat across antitrust matters. In most cases it is crucial to determine whether the defendants have or could obtain significant market power and thus, by definition, maintain or raise prices above the competitive level. The determination of whether a firm or group of firms has market power can also be important because entities that have significant market power are more likely to have the ability and incentive to engage in business practices that could foreclose competition. Moreover, entities that obtain significant market power as a result of a business practice may be able to recoup costs they incur from investing in anticompetitive activities such as predatory pricing and vertical foreclosure. Business practices engaged in by entities that either lack market power or are unlikely to acquire it are often presumed benign (except of course for naked price-fixing and related cartel practices).

The economics of two-sided platforms provides several insights into analysis of market power.

(1) The link between the customers on the two-sides affects the price elasticity of demand and thus the extent to which a price increase on either side is profitable. It therefore necessarily limits market power all else equal. Consider two sides ***A*** and ***B***. An increase in the price to side ***A*** reduces the number of customers on side ***A*** and therefore reduces the value that customers on side ***B*** receive from the platform. That in turn reduces the price that side ***B*** will pay and the number of customers on side ***B***. The reduction in the number of customers on side ***B*** in turn reduces the demand on side ***A*** and thus the price that customers on side ***A*** will pay. These positive feedback effects may take some time to work themselves out, but, as we demonstrated above,

even if, say, customers on side **A** are not very sensitive to price, all else (including the behavior of those in side **B**) equal, demand from side **A** may nonetheless end up being very price-sensitive indeed when these feedback effects work themselves out.

(2) For two-sided platforms it can be important to recognize that competition on both sides of a transaction can limit profits. Suppose in a market without multi-homing that there is limited competition on side ***A*** because customers cannot easily switch between vendors of that side, but there is intense competition on side ***B*** because customers can and do switch between vendors based on price and quality. Then if competitors on side ***B*** cannot differentiate their products and otherwise compete on an equal footing, the ability to raise prices on side ***A*** will not lead to an increase in profits. Any additional profits on side ***A*** will be competed away on side ***B***. This is different from a simple multi-product setting, since the platform cannot stop serving side ***B*** without leaving the business entirely. This point is especially relevant for assessing incentives and recoupment. It is also worth noting that the possibility of multi-homing on side ***B*** will permit positive profits, since it reduces the intensity of competition.

(3) Price equals marginal cost (or average variable cost) on a particular side is not a relevant economic benchmark for two-sided platforms for evaluating either market power, claims of predatory pricing, or excessive pricing under EC law. As we saw above, the non-predatory, profit-maximizing price on each side is a complex function of the elasticities of demand on both sides, indirect network effects, and marginal costs on both sides. Thus it is incorrect to conclude, as a matter of economics, that deviations between price and marginal cost on one side provide any indication of pricing to exploit market power or to drive out competition.[33]

The constraints on market power that result from interlinked demand also affect market definition. Market definition assists in understanding constraints on business behavior and assessing the contours of competition that are relevant for evaluating a practice. In some cases, the fact that a business can be thought of as two-sided may be irrelevant. That could happen either because the indirect network effects though present are small or because nothing in the analysis of the practices really hinges on the linkages between the demands of participating groups. In other cases, the fact that a business is two-sided will prove important both by identifying the real dimensions of competition and focusing on sources of constraints.[34]

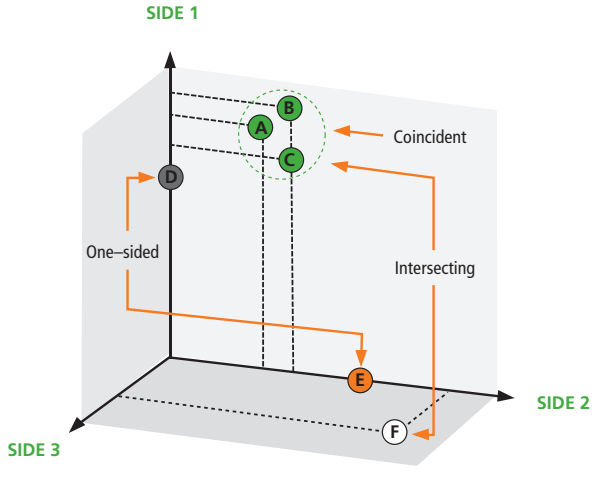
33 For the two-sided platform as a whole, a formula similar to the standard Lerner index emerges in the Rochet-Tirole model. This is not a general result, and it thus suggests that the overall price-cost margin is somewhat less relevant than in single-sided businesses for evaluating overall market power.

34 *See* David S. Evans & Michael Noel, *Defining Antitrust Markets When Firms Operate Multi-Sided Platforms*, 3 Colum. Bus. L. Rev. 667 (2005).

Figure 1 shows potential sources of competitive constraints for a two-sided platform denoted by **A**. It faces competition of some degree from other differentiated two-sided platforms that serve the same customer groups (e.g., the newspapers in a city). It also faces competition from single-sided businesses that provide competitive services to one side only (e.g., billboards). And it faces competition from other two-sided platforms that provide a product that competes mainly with one side but not the other (e.g., advertising-supported television). Again, the existence of these constraints does not mean they are important, only that they need to be looked at.

Figure 1

Types of differentiated platform competition



	Coincident			One–sided		Intersecting
	A	B	C	D	E	F
SIDE 1	X	X	X	X		
SIDE 2	X	X	X		X	X
SIDE 3						X

B. COORDINATED PRACTICES

The key insight of the economics of two-sided platforms in the oligopoly context is that to be successful cartels may need to coordinate on both sides. Consider the situation in which there are several competing two-sided platforms. If they agree to fix prices on one side only the cartel members will tend to compete the supra-competitive profits away on the other side. This observation has two corollaries.

The first is that it is harder to form an effective cartel in an industry with two-sided platforms than in single-sided industries, all else equal. The cartel requires more agreements and monitoring because of the additional side. The second is that if an authority finds evidence of a price fix on one side it should probably look carefully for evidence on the other side. This was relevant, as we note above, in the price-fixing case involving Sotheby's and Christie's.

The economics of two-sided platforms is also relevant for evaluating the practices of cooperatives and joint ventures as we saw from the discussion of the *NaBanco* case. Payment card systems, financial exchanges, and music collecting societies are examples of two-sided platforms that are sometimes organized as not-for-profit cooperatives. The two-sided platforms adopt various rules and regulations for the members and take charge of certain centralized functions. The economics of two-sided platforms is useful for assessing whether there is an efficiency rationale behind an agreement over prices. In *NaBanco*, as we noted, the court found that the collective setting of the interchange fee helped balance the demands between cardholders and merchants (it helped internalize an externality) and eliminated the need for bilateral negotiations (it reduced the transactions cost of internalizing the externality).

C. UNILATERAL PRACTICES

In trying to assess whether unilateral practices are anticompetitive the special economic features of two-sided platforms need to be considered.

1. Predatory and Excessive Pricing

Our review of pricing showed that a robust conclusion of the economics literature is that profit-maximizing two-sided platforms may find that it is profitable overall to price the product offered on one side below average variable cost, below marginal cost, or even below zero. The empirical evidence indicates that such below-cost pricing is common, occurs in stable market equilibrium, and is therefore not designed mainly for the purpose of foreclosing competition. Therefore, any presumption that below-cost pricing by two-sided platforms is anticompetitive is simply not valid. Of course, it is certainly possible for two-sided platforms to engage in predatory pricing by setting its price on one side so low as to deny other platforms access to this side of the market. It is also possible for a two-sided platform to engage in two-sided predatory pricing, charging below cost overall on both sides with the purpose of foreclosing competitors. Cost-based tests make some sense in the latter case, but it is hard to see how they could be used to analyze an allegation of one-sided predation.

Under Article 82 of the EC Treaty a dominant firm can be found to have made an abuse by charging "unfair purchase or selling prices." Just as a below-cost price on one side can emerge in long-run market equilibrium so can an above-cost price on the other side. Indeed, such below-cost/above-cost prices will come

together. This issue has come up in a series of cases in Europe in which regulatory authorities have found mobile telephone operators to have charged fixed-line carriers excessive prices for terminating calls on their networks; the authorities recognize that the profits from these excessive prices are competed away in part through low prices for handsets and call origination. Indeed, the U.K.'s Office of Communication (OfCom) recognized that mobile telephone platforms were highly competitive (on the mobile subscriber side at least) and did not overall earn supracompetitive returns.[35] Although they did not accept that this was a two-sided business, and did not apply two-sided analysis, OfCom did provide an "indirect network externality" kicker to the regulated price it imposed on the mobile termination side.[36]

2. Tying

Under a rule of reason analysis[37] the economics of two-sided platforms can provide an explanation for certain tying practices that seem to reduce consumer choice and harm consumers. As we discussed above, the platform provider designs the platform—including the constellation of services and features—to harness internalized externalities, minimize transactions costs between the customers and both sides, and maximize the overall value of the platform. As part of harnessing externalities this platform provider wants to increase positive indirect network effects while limiting negative indirect network effects. As a consequence, the two-sided platform may impose requirements on side ***A*** that do not benefit them directly and which customers on that side might even reject after comparing private benefits and costs. But such requirements may benefit side ***B***. And if the demand increases on side ***B***, these requirements

UNDER A RULE OF REASON ANALYSIS THE ECONOMICS OF TWO-SIDED PLATFORMS CAN PROVIDE AN EXPLANATION FOR CERTAIN TYING PRACTICES THAT SEEM TO REDUCE CONSUMER CHOICE AND HARM CONSUMERS.

35 *See, e.g.*, U.K. OFFICE OF TELECOMMUNICATIONS, DISCONTINUING REGULATION: MOBILE ACCESS AND CALL ORIGINATION MARKET §1.2 (2003), *available at* http://ofcom.org.uk/static/archive/oftel/publications/eu_directives/2003/discon1103.pdf ("no mobile network operator, either individually or in combination with one or more other mobile network operators, has [significant market power] in that market."). No provider has a share exceeding 28 percent. *See, e.g.*, ECONOMIST INTELLIGENCE UNIT, UNITED KINGDOM: TELECOMS AND TECHNOLOGY BACKGROUND (2005).

36 U.K. OFFICE OF COMMUNICATIONS, WHOLESALE MOBILE VOICE CALL TERMINATION 163-72 (2004), *available at* http://www.ofcom.org.uk/consult/condocs/mobile_call_termination/wmvct/wmvct.pdf. *See* Armstrong, *supra* note 8.

37 Economists and legal scholars generally agree that tying should be considered under a rule of reason analysis rather than a *per se* test. That is not the state of the law in the United States or the European Community, both of whose highest courts have adopted something closer to a *per se* test of liability. However, both courts admit that efficiencies can at least play a limited role in the analysis (in the United States through the separate product test and in the European Union through the possibility of "objective justification" of the practice).

may increase the value placed on the platform on side **A**—and in fact could increase value so much that the feature provides a net benefit to side **A.**[38]

The honor-all-cards rule for payment cards is a possible example. Card systems generally require that merchants that agree to take the system's branded cards agree to take all branded cards that are presented by shoppers. Thus, merchants that have a contract to take American Express (Amex) cards cannot decide to take payment by Amex corporate cards but not Amex personal cards, or to take payment from visibly wealthy travelers but not from locals. For at least some merchants the private cost of this requirement outweighs its benefits (generally we would expect that merchants would privately want a choice to take whatever card they wanted).[39] However, this rule makes the system's branded card more valuable to its cardholders, who have the assurance that their card will be accepted for payment at merchants that display the system's acceptance mark. By increasing the number of cardholders it makes the card a more valuable payment device for merchants to accept.[40]

3. Exclusive Dealing

The potential for profits on the other side provides a possible incentive for exclusive contracts in two-sided platforms. One of the main Chicago School observations about exclusive contracts is that a consumer is always free not to agree to exclusivity. The conclusion is that exclusivity in contracts must reflect consumers' judgment that the benefits (lower prices or efficiencies) outweigh the costs of only dealing with one firm. For two-sided platform businesses, it is at least possible that there is an externality; exclusive contracts on one side might help a platform gain market power on other sides. The consumers agreeing to the exclusive contracts on one side might, at least in the short run, gain from or be indifferent to exclusivity, but they may not take into account the costs to consumers on the other sides from decreased platform competition. Some recent work suggests that it is at least theoretically possible for a two-sided platform to use exclusive contracts to exclude competitors, although the welfare consequences of these contracts are not clearly harmful.[41]

38 *See* Rochet & Tirole (2005), *supra* note 6.

39 For a discussion of this issue, *see* Robert E. Litan & Alex J. Pollock, The Future of Charge Card Networks (AEI-Brookings Joint Center for Regulatory Studies, Working Paper, 2006).

40 A class of merchants claimed that Visa and MasterCard had illegally tied by requiring merchants that accepted their credit cards to also accept their debit cards. The card associations agreed to end this practice after a federal district court judge applied the *per se* tying test and ruled that the associates failed several prongs of this test as a matter of law. In re Visa Check/MasterMoney Antitrust Litigation, 192 F.R.D. 68 (E.D.N.Y 2000). American Express has been sued by a class of merchants for illegally tying its corporate and personal cards. *See* Lavonne Kuykendall, *Merchants Suing Amex Add Citi, MBNA as Defendants*, 170 Am. Banker (2005).

41 *See* Mark Armstrong & Julian Wright, *Two-Sided Markets, Competitive Bottlenecks and Exclusive Contracts,* Econ. Theory (forthcoming 2006).

As with exclusivity in one-sided markets, however, this can only be a concern if one firm has exclusivity over most or all of the market and if the exclusivity is persistent and durable. For example, consumers on the nonexclusive side could respond by moving to a competing platform, thus exerting pressure on consumers on the exclusive side to end exclusivity. Moreover, in markets with significant buyer concentration, the buyers would be reluctant to agree to exclusivity if there is some expectation that it will lead to dominance by that platform, as that will likely result in higher prices in the future for all sides. As with one-sided markets, one needs to consider whether the efficiencies from exclusive contracts—for example, in helping to create a platform that might not otherwise exist for the benefit of consumers—offset possible costs from reducing competition.

VII. Qualifications and Conclusions

The indirect network effects between customer groups served by a single business are strong in many important industries. Businesses in these industries operate two-sided platforms. The economics of two-sided platforms provides insights into how these businesses and industries behave that are relevant for competition analysis including market definition, coordinated practices, unilateral practices, and the evaluation of efficiencies. The economic literature provides robust results—that is, ones that are not dependent on only fragile assumptions—that can assist in this analysis. These results include the consequences of interlinked demand between customer sides for prices; prices do not, contrary to the standard model, have a tight relationship with cost.

As with almost any application of economics to policy several cautions are prudent. First, many of the theoretical results in the literature to date are, like those in other areas of industrial organization, based on quite abstract models of how industries operate and special assumptions of demand and cost. Second, to date there has been little rigorous empirical research on two-sided platforms or competition among them. Third, the theoretical and empirical work to date suggests that how two-sided businesses work is highly dependent on the specific institutions and technologies of an industry. One must be careful generalizing.

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Comments on Evans & Schmalensee's "The Industrial Organization of Markets with Two-Sided Platforms"

Janusz A. Ordover

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Comments on Evans & Schmalensee's "The Industrial Organization of Markets with Two-Sided Platforms"

Janusz A. Ordover

A freshman student in economics or a Nobel prize-winning macroeconomist who has lately stumbled across a journal or two in industrial organization economics may be somewhat perplexed or confused by many references to two-sided markets. Surely, is it not the case that all markets have two sides, namely buyers and sellers? Consequently, to the uninitiated, the concept of a two-sided market offers little, if any, additional analytical insight. Some of that confusion is perhaps dispelled by a more informative description, namely: markets with two-sided platforms. So for the rest of this paper, we shall forget about two-sided markets and speak of two-sided platforms (2SP) and of markets in which these 2SPs compete. Professors David Evans and Richard Schmalensee (hereinafter E&S), who have done more than most of the thinking about the economics of 2SPs and advocating the importance of the idea to other academics, lawyers, policymakers, and business people, offer a highly accessible survey of the state of play in their excellent contribution to this volume.[1]

1 This paper is a comment on a paper by Evans & Schmalensee (E&S) published in this issue, 3(1) Competition Pol'y Int'l 151–179 (2007), and also forthcoming as D. Evans & R. Schmalensee, *The Industrial Organization of Markets with Two-Sided Platforms*, *in* Issues in Competition Law & Policy (W.D. Collins ed., 2007). All E&S page cites refer to the version published in this issue.

It is perhaps worth noting that much of the impetus behind the outpouring of research on 2SPs can be attributed to antitrust litigations in several industries, especially electronic payment networks, and to (generally) misguided regulatory initiatives pertaining to these networks in the European Community and in Australia. I have advised American Express with respect to regulatory and other issues relating to interchange fees.

The author is a Professor of Economics at New York University and a Senior Consultant at Competition Policy Associates.

It is important to recognize that, perhaps without knowing so, economists have studied 2SPs for quite some time (i.e., way before the term was invented) as have competition authorities and the courts. And, of course, business people, who generally are ahead of theoreticians in such matters, have also intuitively understood the specific pricing and marketing challenges that have to be solved in order to launch a 2SP and make it prosper. These challenges arise because (by definition) a 2SP links two (or more) distinct groups of consumers whose demands are interrelated in that each group confers (perhaps up to a point) a positive external benefit on the other. These effects are generally referred to as indirect network effects, in distinction to more standard network effects that realize themselves among the same customer group. When there are indirect network effects, a business strategy that stimulates demand on side A of the platform will, when properly implemented, stimulate demand on side B of the platform, which in turn creates a positive feedback to side A, and so on. Because of this interdependence, a 2SP entrepreneur must solve two problems: first, how to get both sides on board[2] and second, how to structure prices to the two sides.[3]

E&S offer a wealth of examples on how these two problems have been solved by a wide range of 2SPs in a variety of markets, such as dating clubs, newspapers, credit card networks, and video game consoles. Surprisingly, they do not discuss Global Distribution Systems (GDS), which offers an excellent case study of how platform operators respond to changes in the competitive and regulatory environment in which they operate (i.e., changes in the relative importance of attracting the two sides to the platform, inter-platform competition, and changes in platform ownership).[4] For example, loosely speaking, pricing on these platforms has flipped from the initial arrangement whereby travel agents paid for each booking (and were offered incentives to join the platform) and airlines paid to join the platform (and the membership fee was determined by the level of display preference) to the current arrangement where airlines pay for each booking while travel agents receive a per booking financial assistance.[5] This rebalancing of fees is consistent with the predictions of the theoretical literature on 2SPs but also reflects the changing structure of GDS ownership as well as the fact that travel agents no longer receive per booking payments from the airlines. Because many travel agents multi-home—subscribe to more than one GDS—and can

2 This is the typical chicken-and-egg problem with respect to which it is worth recalling Marshall McLuhan's adage that chicken is simply egg's idea for getting more eggs.

3 In fact, according to the definition first offered by J.-C. Rochet and J. Tirole, a platform (or a market) is two-sided when the volume of transactions on the platform depends on both the level of the total price and on the structure of prices charged to the two sides, holding total price constant. *See* J.-C. Rochet & J. Tirole, *Platform Competition in Two-Sided Markets*, 1 J. Eur. Econ. Ass'n 990 (2003).

4 GDSs were formerly known as Computerized Reservations Systems (CRSs).

5 However, the recent trend is for GDSs to aggressively discount the fees to the airlines in exchange for agreements to provide full fare information.

also bypass the GDS altogether, the platform vendor now must offer more powerful inducements to travel agents to use its platform in order to keep the airlines willing to pay per the booking fee to the platform owner. Importantly, these inducements entail not only direct payments, but also (costly) contracts with airlines for attractive content and massive investments in platform capability.[6]

Given the ubiquity and importance of 2SPs in modern economies—as demonstrated by E&S—it is important to ask whether competition policy (antitrust and regulation) has to be retooled to better capture the special features of 2SPs and whether public and private decision-makers have been led astray by failing to account for this two-sidedness in their analyses of business conduct. Here, again E&S deliver by providing the reader a comprehensive review of the lessons from 2SP economics for competition policy. E&S claim—correctly in my view—that important analytical and policy errors can result when policymakers take a one-sided view of markets with 2SPs. I am less convinced, however, that the extent of needed reassessment of competition policy in light of this new learning is as profound as that triggered by, for example, the developments in economics of vertical relationships in production and distribution. Invoking a two-sided nature of the business will not get one off the hook in an antitrust case and, in some situations may make the predicament even worse. Thus—like free-riding or network effects were before—2SPs may be a passing concept which calls for analytical vigilance but does not require a policy revolution. Let us consider a few examples.

INVOKING A TWO-SIDED NATURE OF THE BUSINESS WILL NOT GET ONE OFF THE HOOK IN AN ANTITRUST CASE AND, IN SOME SITUATIONS MAY MAKE THE PREDICAMENT EVEN WORSE.

Consider first the matter of predation. As we have seen, an important insight from 2SP literature is that structure of prices matters for the profitability of the platform and that changes in market conditions can prompt the platform owner to rebalance prices, possibly in a rather drastic fashion. Indeed, in many settings, a price to one side is less than the marginal cost of serving it (assuming that such marginal cost is even a meaningful concept). Of course, the proposition that price to one group of buyers may be below the direct marginal cost of serving these buyers is not new: some 25 years ago, Professor Robert Willig and I proposed the Ordover-Willig test for predatory pricing by a multi-product firm in which the correction term in the standard formula accounts for all the pertinent cross-elasticities.[7] Admittedly, given how difficult it is to implement even the

6 Since most or all airlines multi-home on all the GDSs, the platforms cannot effectively distinguish themselves based on membership (unlike credit card networks, for example) but can and do distinguish themselves based on the depth and quality of price and other information provided from the airlines.

7 J. Ordover & R.D. Willig, *An Economic Definition of Predation: Pricing and Product Innovation*, 91 YALE L.J. 8-53 (1981).

standard test for predation, the need to account for (inter-side) cross-elastic effects only exacerbates the challenge. However, if predation is alleged, the challenge has to be met somehow. One sensible avenue might be to invoke the finding that low (or even negative price) makes sound business sense for 2SPs irrespective of its impact on competition. The next step might be to show that the price structure delivers a per transaction price that exceeds the pertinent measure of marginal cost, as in the Ordover-Willig test, for example. Or perhaps the analysis might focus on a comparison of incremental revenues versus incremental costs defined over packages of goods or services that serve the interests of customers on both sides of the platform.[8] In no case, however, is being a 2SP likely to offer immunity from a claim of predation. At the same time, the alleged 2SP predator has a wealth of economics and business rationales on its side when defending such claim: namely, the need to balance or get on board the demand on both sides.

More interesting then predation is the set of competitive problems engendered by the question of access to the platform. E&S briefly address this topic in their discussion of the EC's investigation of Microsoft's integration of its media player into its Windows operating software and in connection with the *Magill* case that was also litigated in the European Community.[9] In the United States, the issues of access featured prominently in the *Microsoft* case and in the litigation brought by the United States against Visa and MasterCard that dealt, among other matters, with the rules that restricted participation of Visa and MasterCard issuing banks in competing credit card networks.[10] And, although 2SP terminology was nowhere on the horizon yet, the IBM cases of the 1970s that focused on access by stand-alone manufacturers of peripherals to IBM's CPU platform, stimulated inquiry into the same economic issues that access to the 2SPs raises, only more so. For example, the research on 2SPs has demonstrated that the ability and ease with which the members of the two sides can (and do) multi-home, effects inter-platform competition and both the level and structure of prices, often in the manner that conduces to overall social welfare.[11] Hence, in markets populated by 2SPs exclusivity could be quite adverse to social welfare. For example, for a new

8 In my testimony in United States v. American Airlines I suggested that a route is a better object of analysis than a single flight and that analysis of profitability of a route should reflect the flow traffic (i.e., network effects), but that care is needed to avoid double-counting. *See* United States v. American Airlines, 743 F.2d 1114 (5th Cir. 1984).

9 *See also* E&S, at §§ C.2 and 3 for a discussion of EC's investigation of Microsoft (Commission of the European Communities v. Microsoft, Case COMP/C-3/37.792) and *Magill* (ITP v. Commission of the European Communities, 1995 E.C.R. I-00743).

10 I have acted as a consultant to American Express in connection with U.S. Department of Justice (DOJ) investigation of Visa/MasterCard practices. *See* United States v. Visa U.S.A. et al., 98 Civ. 7076 (S.D.N.Y.).

11 For a very interesting summary of the extent of multi-homing on selected platforms, *see* E&S, Table 3 at 167.

platform the ability to attract participants on both sides could be a matter of life and death, given the importance of inter-side externalities (and the concomitant scale economies). On the other hand, the literature on 2SPs has also amply demonstrated that the success of a platform—including its ability to get off the ground—requires that the owner be able to design the platform and the rules of access that get both sides on board (with the least amount of regulatory interference and including the governance rules), promote effective balancing of demand on both sides, and give the platform enough flexibility to meet the challenges of changes in competitive environment. From the latter perspective, exclusivity as well as "technological tying", as it was called a while back, makes especially good, pro-competitive sense. Hence, in the context of 2SPs, access crystallizes the difficult trade-offs between openness and exclusion. There has been much written on this vexing problem in general terms but, as I noted, the economics of 2SPs raises further the analytical challenge.

In general, the issue of access and exclusivity is linked to the question of market power and to the forensic tools for detecting it in the data. E&S correctly point out that because the structure of prices matters, the platform operator sets the price to each side in a manner that reflects the indirect network effects. An increase in price on one side above some initial level reduces participation on that side and sets off a chain of adverse effects that bounce back and forth between the sides. E&S state that these indirect effects "limit market power, all else being equal." It is not obvious what the standard caveat means here: surely, it is also true that by improving quality on one side of the platform (charging a low price), the platform operator reduces elasticity of demand on the other side of the platform and thus, at least in principle, lessens the adverse effects of a price increase to the participants on that side. Like an owner of a mundane single-sided business (say a tobacco company) who uses advertising and other means to reduce the elasticity of demand facing it so as to get its Lerner index up, the owner of a 2SP can use prices (and other tools) to affect the various elasticities that are pertinent to the platform's profitability.[12] In fact, some commentators on market power possessed by 2SPs have made the opposite point of E&S, namely: because customers on one side must participate on every platform, each platform has more market power than competitive analysis (e.g., counting the number of rival platforms) might suggest.

SOME COMMENTATORS HAVE MADE THE OPPOSITE POINT, NAMELY: BECAUSE CUSTOMERS ON ONE SIDE MUST PARTICIPATE ON EVERY PLATFORM, EACH PLATFORM HAS MORE MARKET POWER THAN COMPETITIVE ANALYSIS MIGHT SUGGEST.

E&S are also on point when they note that "competition on both sides of the platform limits profits." This claim is uncontroversial but I am not sure whether

12 I say "various" because the owner of the platform is both concerned about membership and usage and thus has to pay attention to both intensive and extensive margins on both sides.

complete dissipation of incremental profits from the less competitive side to the more competitive side is a reasonable benchmark, as E&S seem to suggest. More likely such dissipation is imperfect but could be sufficient to undermine incentives for anticompetitive unilateral or coordinated conduct, given the costs of the conduct and the possible penalties. This seems to be the key public policy takeaway from this feature of 2SPs. Another takeaway might be that the operators of 2SPs may have enhanced incentives to engage in business strategies that lessen competition on that side of the platform from which the feedback effects (i.e., the inter-side network effects) are the most pronounced. Thus, the flip side of the finding that competition on one side of the platform affects profits on both sides is that reduction of competition on that side where dissipation is particularly potent could be especially profitable because of the possibility of recoupment on both sides.

Staying with the issue of competition, E&S could have noted that increased competition among platforms may have a rather surprising impact on price structure. As an example, increased competition among credit card networks for issuers have led to an increase in interchange fees; and a similar phenomenon was observed in PIN debit networks where intensified competition for exclusive bugging of PIN debit cards by issuing banks also had a similar effect.[13] Thus, in typical one-sided markets increased competition predictably leads to lower prices, but this need not be the case in markets with 2SPs. This of course does not mean that increased competition is somehow harmful but only that invigorated competition can have a complex impact on the different sides (groups) of platform customers. However, the impact on price (or prices) is, of course, only a part of the story. The other key part is the impact on the quality that the platform delivers to each side. Thus, a reduction in price on the side where profits are being dissipated may be nothing more than a partial corrective for the reduction in quality on the other side caused by price elevation. Indeed, a theme that runs deeply through the E&S paper is that quality of the platform, as gauged by the depth and breadth of services that it offers and the quality of participants on both sides, is an important dimension of competition analysis.[14]

E&S close their discussion of competition and market power by noting that "price equals marginal cost ... on a particular side is not a relevant economic benchmark ... for evaluating market power ...." This is surely true because the literature has amply demonstrated that 2SP's pricing to each side depends on a complex web of intra-side elasticities and inter-side cross-elasticities. Moreover,

13 For more on this point, *see* B. Klein et al., 73 Antitrust L.J. 571 (2006) (credit card networks) and R. Hesse & J. Soven, *Defining Relevant Markets in Electronic Payment Network Antitrust Cases*, 73 Antitrust L.J. 709 (2006) (PIN debit networks). I served as an expert witness for the DOJ in United States v. First Data Corp., 03 Civ. 02169 (D.D.C. 2003).

14 This theme comes to the fore in the important body of research by Andrei Hagiu at Harvard Business School.

pricing strategies deployed by the operator can often be quite complex. For example, the operator may charge some combination of membership fees (fee for joining the platform) and usage fees (on a per transaction basis). Such pricing arrangements make sense given that the platform has to bring the right mix of participants to the two sides of the platform and then make them use it efficiently. In such an environment, comparisons of price to marginal cost are apt to be misleading, at least in some situations but not all. For example, percentage commissions charged by real estate brokers relative to the expected costs of making a successful match may be a reasonable basis from which to measure market power in real estate brokerage services. On the other hand, to take credit card platforms as an example, it does not make any economic sense—as some regulators insist on doing—to first allocate the various buckets of platform costs to each side and then to compare fees that each side pays to these potentially highly, arbitrary measures of costs. Clearly, many of these costs are joint and common and perhaps more importantly, the expenditure of costs on one side ultimately benefits both cardholders and merchants (e.g., development of intelligent systems for fraud detection at the point of sale).

Surely, the implication from the literature is not that 2SPs cannot have market power but, rather, that a great deal of caution has to be exercised in inferring such market power from standard indicia of market power. E&S do not suggest that once a firm is found to be a 2SP it should get a free pass from the strictures of competition policy. However, they do not point to alternative measures of market power that stem from the literature that could be used to make the requisite findings or how the traditional measures should be modified to account for two-sidedness. Following on my earlier remark, we have a general idea how to adjust the Lerner index to account for the cross-elastic effects in a variety of settings.[15] While this may not be enough to capture all the complexity, it is a start.

It is also a start that can help define the relevant antitrust market(s). The market definition step has lately had some tough times, what with some antitrust commentators calling for its jettisoning altogether as an unnecessary distraction from the ultimate task of antitrust analysis, which is the assessment of competitive effects from unilateral or multi-firm business strategies (including mergers and so on). Without getting entangled in this debate, I want to comment briefly on the issue of product market definition in industries populated by 2SPs.

The main point I want to make is that there is no need to despair at the task. As E&S note, "...the fact that a business can be thought of as 2SP may be irrelevant [to the market definition step]. ... In other cases, the fact that a business is a 2SP will prove important both by identifying the real dimensions of competi-

15 This point is well-illustrated in G. Parker & M. Van Alstyne, *Two-Sided Network Effects*, 51 Mgmt. Sci. 1494 (2005).

tion and focusing on sources of constraint."[16] The question is, in those situations where two-sidedness matters a great deal, whether the traditional tools that economists now use for market definition should be jettisoned or merely adapted to deal with complications like those depicted in Figure 1 in E&S.[17] In particular, can the small but significant and non-transitory increase in the price (SSNIP) methodology for market definition—which has earned its place in the global antitrust toolbox—be used in defining markets in industries with 2SPs? During the hearing in the *First Data* case, one of the experts for the defendants concluded that the SSNIP test could not be readily used to gauge the scope of the relevant market in which PIN debit networks competed and should be abandoned.[18] Unfortunately, he failed to offer an alternative approach that would address the apparent inappropriateness of the SSNIP test.[19]

The obvious problem for the SSNIP test is that it is typically applied to one price (or to a collection of prices of putative substitutes). In a 2SP market, an increase in the price on one side has implications for demand on the other side and thus for the overall profitability of the platform and impact of the price increase itself. This is not an unfamiliar complication: for example, a hypothetical monopolist supplier of tennis rackets has to factor in the effects of a SSNIP on tennis rackets on its tennis ball business, for example. This suggests that one way to implement the SSNIP test in the example would be to inquire whether a monopolist of tennis equipment system would be able to elevate profitably by five percent the price of the system or would it would suffer enough loss in demand to other sporting pursuits as to render the increase unprofitable. The answer to this question is neither obvious nor simple (in terms of data requirements) but, conceptually at least, it is not impossible to address.

CONSEQUENTLY, THE EMPIRICAL ASSESSMENT OF HOW THE TWO SIDES WILL RESPOND TO A HYPOTHETICAL INCREASE ON ONE SIDE IS THAT MUCH MORE COMPLICATED.

In the case of 2SPs the feedback effects that reflect inter-side network effects are, of course, likely to be much more complicated that in the tennis playing system example. For starters, in the example above, the same consumers are generally purchasers of both tennis rackets and tennis balls, but this is not the case with 2SPs where participants on the two sides are distinct groups of consumers. Consequently, the empirical assessment of how the two sides will respond to a hypothetical increase on one side is that much more complicated. Perhaps even

16 *See* E&S, at 174, footnote omitted.

17 *See id.*, Figure 1 at 175.

18 United States v. First Data Corp., 03 Civ. 02169 (D.D.C. 2003).

19 For more detail on how the DOJ used the SSNIP methodology in that litigation, *see* Hesse & Soven, *supra* note 13. *See also* E. Emch & S. Martin, *Market Definition and Market Power in Payment Card Networks,* 5 REV. NETWORK ECON. 45 (2006).

more complicated is the formulation of an optimal price strategy by a hypothetical monopolist relative to the prevailing strategies: this is because the hypothetical 2SP must not only find the optimal price level but also the optimal price structure. If the structure is invariant to the degree of market power then the SSNIP test would proceed on the assumption proportional increase in prices on both sides. In other situations, a hypothetical SSNIP can be applied to one side while holding the other price(s) constant. If this is profitable, then factoring in (a downward) price adjustment on the other side should only improve profitability of the SSNIP because it will neutralize some of the inter-side externality. Of course, it is not necessarily obvious which side is a more attractive candidate for the proposed price elevation—should a hypothetical shopping mall monopolist increase its take of stores' revenues or get rid of free parking? But in some instances it may be readily apparent which side to apply the SSNIP given industry dynamics, evidence from the industry participants, or so on.

As should be clear from this brief reaction to the E&S paper, there is still much to be done on the topic of competition in markets with two-sided platforms. The paper gives an excellent introduction to the topic (despite being cryptic here and there) and should serve as a launch pad for further explorations in both the realm of policy and the realm theoretical modeling.

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Two-Sided Platform Markets and the Application of the Traditional Antitrust Analytical Framework

Renata B. Hesse

Published in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Two-Sided Platform Markets and the Application of the Traditional Antitrust Analytical Framework

Renata B. Hesse

It only takes working through a single matter that involves a two-sided market to recognize that the antitrust analysis can be a bit more complicated than with standard one-sided markets. The principle reason for the complication is evident from the descriptive moniker given these markets: they have two sides or, put more practically, they have two sets of independent customers. Generally, two-sided markets are characterized by

(1) the presence of two distinct classes of customers for a vendor's product or service, both of which are necessary for the existence of the product or service, and

(2) indirect positive externalities between different classes of customers, meaning that the value of the product or service to one class of customer increases with the level of usage by the other customer class, at least up to a point.[1]

1 *See* D. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. REG. 325, 332 (2003); R. Roson, *Two-Sided Markets: A Tentative Survey*, 4 REV. NETWORK ECON. 142 (2005) ("In two-sided markets, two (or more) parties interact on a platform, and the interaction is affected by special 'indirect' network externalities.").

The author is a partner in the Washington, DC office of Wilson Sonsini Goodrich & Rosati. Prior to joining Wilson Sonsini, the author was Chief of the Networks & Technology Enforcement Section of the Antitrust Division of the U.S. Department of Justice, where she supervised the Antitrust Division's lawsuit to enjoin First Data Corp.'s proposed acquisition of Concord EFS, the Antitrust Division's lawsuit against Oracle Corporation in connection with its proposed acquisition of PeopleSoft, and the remedial portion of United States v. Microsoft Corp.

But why do these features make a difference in terms of the application of standard antitrust principles to these markets? Or, more colloquially, why is everyone talking about two-sided markets?

Two-sided markets do present certain unique practical problems. Not surprisingly, the complexity primarily arises from the presence of two unique, but interdependent, classes of customers. In a traditional market, the analysis centers around the responses of a single set of customers to changes in supply (either price or output) and the responses of the vendors to changes in demand. In a two-sided market the analysis becomes multi-dimensional. The analysis needs to account for

(1) the responses of two sets of customers to the vendors,

(2) the vendors' responses to two sets of customers, and

(3) the responses of one class of customers to changes in the others' behavior and vice versa.

This multi-dimensionality affects each step of standard antitrust analysis, from product market definition, to entry and efficiencies. It does not, however, dictate abandoning the typical tools that one applies in the analysis of single-sided markets.

I. Defining Relevant Product Markets

The standard technique for defining markets is the hypothetical monopolist test set forth in the U.S. Federal Trade Commission and U.S. Department of Justice Horizontal Merger Guidelines.[2] The test, however, is designed to examine the reactions of one set of customers, not two, to changes in price. The test has no direct mechanism to account for the two sets of customers involved in two-sided markets, or the reactions of one class of customers to price changes imposed on the other. For example, even though a hypothetical monopolist would profitably impose a small but significant and non-transitory increase in the price (SSNIP) on one side of the market in isolation, the other side of the market might respond to the SSNIP by reducing demand for

THE TEST HAS NO DIRECT MECHANISM TO ACCOUNT FOR THE TWO SETS OF CUSTOMERS INVOLVED IN TWO-SIDED MARKETS, OR THE REACTIONS OF ONE CLASS OF CUSTOMERS TO PRICE CHANGES IMPOSED ON THE OTHER.

2 *See* U.S. Dep't of Justice & Federal Trade Comm'n, Horizontal Merger Guidelines (1992, revised 1997) *available at* http://www.ftc.gov/bc/docs/horizmer.htm. The test takes the smallest possible group of competing products and asks whether a hypothetical monopolist that sells those products could profitably impose a small (5-10 percent) but significant and non-transitory price increase, commonly referred to as a SSNIP. *Id.* at § 1.11.

the product, rendering the SSNIP unprofitable.[3] If this effect is not taken into account, the analysis could yield an improperly small relevant product market. Consequently, some have argued that the hypothetical monopolist test is not the appropriate market definition tool for two-sided markets.

Despite these challenges, both scholarship and recent public and private antitrust litigation have demonstrated that it is possible to apply the SSNIP test in two-sided markets. Most recently, the U.S. Department of Justice (DOJ) applied the SSNIP test to define a relevant product market of PIN debit network services in *United States v. First Data Corp. and Concord EFS*.[4]

II. Evaluating Barriers to Entry

The interdependency of the two customer groups also impacts the analysis of the likelihood and success of new entry in two-sided markets. First, because both sides of the market are needed for the product or service to function (i.e., the provider must get both sides of the market on board), new entrants face a form of the chicken-and-egg problem. This problem is probably fairly easy to overcome in some two-sided markets, but quite difficult in others. For example, the owner of an attractive new nightclub may find it relatively easy to get the necessary critical mass of both men and women customers. In contrast, a new payment network likely would find it considerably more difficult to obtain the required critical mass of both issuers and merchants.

The difficulty of entry is further increased in some two-sided markets because of the presence of indirect network effects (i.e., the value of the product or service to one class of customers often increases directly with the level of usage by the other customer class). Thus, not only must the new entrant simultaneously convince both sets of customers to purchase its product, but it must also overcome the challenge that for many customers the value of purchasing the product or service from the established provider is likely significantly greater than from purchasing from the start-up.

Obtaining the information needed to analyze these issues is often complex. For example, what critical mass of both sides of the market does a new entrant need to compete effectively? Does conduct by incumbents designed to get both sides of the market on board (e.g., a payment network signing bonuses to issuers) increase the difficulty of entry, and potentially constitute unlawful exclusionary conduct?

3 In the electronic payment network context, for example, it is possible (but unlikely) that while merchants would reduce their demand only slightly in response to a SSNIP imposed by a hypothetical payment network monopolist, issuer demand for the payment service would be so sensitive to even a modest decline in merchant volume that it would be sufficient to make the merchant SSNIP unprofitable.

4 United States v. First Data Corp., 03 Civ. 02169 (D.D.C. 2003).

Answering these types of questions is difficult, but it can be done through careful focus on the two-sided nature of the market. The DOJ's case in *United States v. Microsoft Corp.*[5] was built in part on its conclusion that network effects present in the two-sided operating system market made both new entry and expansion by existing market participants very difficult.

III. Assessing Competitive Effects

Finally, the characteristics of two-sided markets increase the difficulty of analyzing the competitive effects of mergers and other conduct. For example, a merger may slightly reduce competition among vendors on one side of the market, but produce substantial pro-competitive gains from efficiencies for the customers on the other side of the market. Deciding how to balance these offsetting effects is not easy. A related problem is that it is possible to confuse vigorous competition for one set of customers for the exercise of market power against the other. Payment networks have long argued that increases in interchange fees for merchants are largely due to intense competition for issuers.[6]

THE CHARACTERISTICS OF TWO-SIDED MARKETS INCREASE THE DIFFICULTY OF ANALYZING THE COMPETITIVE EFFECTS OF MERGERS AND OTHER CONDUCT.

Nevertheless, the existence of a two-sided market has not prevented proof of competitive harm in litigated cases. For example, the DOJ successfully demonstrated harm to competition in both *United States v. Microsoft Corp.* and *United States v. Visa*,[7] each of which involved assessing harm to competition in the context of a two-sided market.

5 United States v. Microsoft, 253 F.3d 34 (D.C. Cir. 2001).

6 In *United States v. First Data Corp.*, the defendants (through their economic experts) asserted that the Division's application of the hypothetical monopolist test was faulty because it purportedly ignored competition between PIN and signature debit networks for the business of card issuers. *See* Transcript of Hearing (Dec. 5, 2003) at 97:12 to 98:12, United States v. First Data Corp., 03 Civ. 02169 (D.D.C. 2003) (testimony of Professor Michael Katz, expert for the defendants), *available at* http://www.usdoj.gov/atr/cases/f201900/201902a.htm. The parties maintained that the increase in interchange rates was the result of head-to-head competition between PIN and signature debit networks for issuer customers rather than a reflection of the exercise of market power by PIN debit networks against merchants. *See id.* at 102:12-22.

7 United States v. Visa U.S.A. et al., 98 Civ. 7076 (S.D.N.Y. 1998).

IV. In Summary

The dialogue over two-sided markets has been fueled in part by a growing scholarship that has increased understanding of these markets, combined with a number of significant antitrust cases that involved two-sided markets. This dialogue will continue. The greater complexity associated with analysis of two-sided markets and the potential for mistakes of consequence to the overall outcome of a matter should increase the care and diligence that goes into analyzing these markets.

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

The Empirics of Antitrust in Two-Sided Markets

Marc Rysman

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

The Empirics of Antitrust in Two-Sided Markets

Marc Rysman

Recent theoretical research on the implications of two-sided markets is gaining recognition for its implications in antitrust.[1] However, the role of empirical analysis in antitrust cases for two-sided markets has been unexplored thus far. Empirical tools of economics are playing an increasingly large role in antitrust litigation.[2] At the same time, there have been several recent attempts to bring empirical analysis to two-sided markets. To the extent that this empirical work on two-sided markets bares similarities to common empirical tools of antitrust, it can provide a template for how the empirics of antitrust cases will proceed in two-sided markets.

This paper studies several issues in which empirical contributions can impact antitrust in the context of two-sided markets. For each issue, I discuss recent empirical research that exemplifies my point. The first issue I discuss is the implementation of market simulations. Market simulations have an important role in determining relevant markets and the price effects of horizontal coordination.[3]

1 Overviews of the research literature in economics appear in J.-C. Rochet & J. Tirole, *Two-Sided Markets: A Progress Report*, RAND J. Econ. (Autumn 2006) and M. Armstrong, *Competition in Two-Sided Markets*, RAND J. Econ. (Autumn 2006). For discussions of the role in antitrust, see D. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20(2) Yale J. on Reg. (2003) and D. Evans & M. Noel, *Defining Antitrust Markets When Firms Operate Two-Sided Platforms*, 3 Colum. Bus. L. Rev. 667 (2005).

2 *See, e.g.,* J. Baker & D. Rubinfeld, *Empirical Methods in Antitrust Litigation: Review and Critique*, 1 Am. L. & Econ. Rev. 385 (1999) and R. Epstein & D. Rubinfeld, *Merger Simulation with Brand-Level Margin Data: Extending PCAIDS with Nests*, 4(1) Advances in Econ. Analysis & Pol'y (2004).

3 G. Werden & L. Froeb, *The Antitrust Logit Model For Predicting Unilateral Competitive Effects*, 70(1) Antitrust L.J. 257 (2002).

The author is Associate Professor at Boston University. He thanks David Evans for providing motivation and encouragement to write this paper. Participants at the Antitrust for Two-Sided Markets conference in Cambridge, MA, June 2006 provided valuable feedback.

However, in the context of two-sided markets, the investigator must specify substantially more demand parameters, and results can depend on small changes in certain parameters. This feature raises the issue of where these parameters come from, whether they are estimated from data or simply reflect informed guesses about industry features. I turn to the research described in my 2004 paper on the yellow pages market to provide a helpful example.[4]

The second tool taken up in this paper is price regressions. Price regressions can provide direct evidence on the relationship between market structure and pricing and has played an important role in some litigation. A prominent example is the merger of Office Depot and Staples.[5] Price regressions could potentially be applied in a two-sided context as well. There are naturally at least two prices in a two-sided market and the measure of market structure must account for possibly different market structures on each side of the market. For an example of how this method might proceed, I refer to my paper on sports card conventions (with Professor Ginger Jin).[6]

Whereas the first two issues represent examples of standard tools being adjusted for two-sided markets, the final part of the paper addresses new questions that arise in two-sided markets for which empirical research might be important. Naturally, this discussion is open-ended but I focus on two questions that seem important and potentially testable in data. The first is the basic question of whether or not a market is two-sided. Showing that a market is not two-sided may be difficult as there is no firm agreement on the definition, and some definitions are quite broad. However, markets that exhibit positive feedback loops (or indirect network effects) are two-sided under any definition.[7] Establishing such a feedback loop would be strong evidence in favor of the relevance of two-sided markets. A second question that can be important is whether or not agents multi-home, that is, whether they interact with more than one intermediary. In forthcoming papers, Professors Rochet and Tirole and Professor Armstrong establish the importance of multi-homing in determining pricing structure.[8] If a group of agents single-home, the intermediary has market power over access to its agents. That can lead to relatively high prices for the other side of the market and very competitive pricing for the single-homing agents. In an upcoming

4 M. Rysman, *Competition between Networks: A study of the Market for Yellow Pages*, 71(2) Rev. Econ. Stud. 483 (2004).

5 For discussion, *see* S. Dalkir & F.R. Warren-Boulton, *Prices, Market Definition, and the Effects of Merger: Staples-Office Depot*, *in* The Antitrust Revolution: Economics, Competition, & Policy 52-72 (J. Kwoka, Jr., & L. White eds., 2004).

6 G. Jin & M. Rysman, Platform Pricing at Sports Card Conventions (2006).

7 *See* J. Farrell & G. Saloner, *Standardization, Compatibility, and Innovation*, RAND J. Econ. 70 (1985).

8 Rochet & Tirole (2006), *supra* note 1 and Armstrong, *supra* note 1.

paper, I test for both of these issues in a detailed data set covering the payment card industry.[9]

The list of issues covered here is not meant to be exhaustive. The general point is rather that empirical research on the economics of two-sided markets is relevant in antitrust settings. The theoretical literature on two-sided markets is new and typically, empirical work lags behind theory. While that may be the case, empirical research has progressed far enough to provide models for how empirical analysis should proceed in antitrust litigation when issues associated with two-sided markets are important.

I. Market Simulations

Market simulations provide a method for assessing the anticompetitive impacts of mergers and horizontal collusion. More detailed descriptions appear in Werden and Froeb's 2002 paper and Epstein and Rubinfeld's 2004 paper, but the standard analysis specifies a demand system for a set of products and an ownership structure.[10] Specifying demand means determining own-price and cross-price elasticities. The investigator must also specify how firms interact. Formally, the interaction is a game theoretic equilibrium solution concept, and typical examples are to assume that firms set prices simultaneously or that they set quantities simultaneously. Given these assumptions, the investigator can map observed market shares and prices into implied marginal costs, that is, the marginal costs that rationalize the observed market outcome. With these elements in hand, the investigator can specify alternative market structures, such as one in which one product exits the market or a set of products switch from one firm to another. The investigator calculates prices and quantities under the new market structure and may be interested in whether prices rise by more than 5 perecent or consumer surplus significantly changes.

There has been little research explicitly validating these models ex post. Also, to my knowledge, simulation models have not been presented as evidence in an antitrust proceeding. However, simulation models are increasingly popular at competition authorities as a screening tool for determining which mergers should be challenged.[11] Arguably, the appearance of simulations in court is not far away.

One important tension in this approach is where the parameters come from, particularly the elasticities. The economics literature provides numerous examples of estimation from data using econometric techniques. However, competition

9 M. Rysman, *An Empirical Analysis of Payment Card Usage*, J. Indus. Econ. (forthcoming 2007).

10 Werden & Froeb (2002), *supra* note 3 and Epstein & Rubinfeld, *supra* note 2.

11 G. Werden & L. Froeb, *An Introduction to the Symposium on the Use of Simulation in Applied Industrial Organization*, 7(2) Int'l J. Econ. Bus. 133 (2000).

authorities rarely have the data or time available to rigorously pursue these techniques. Rather, these investigators are often in the position of having to make educated guesses at these parameters, and presenting results for a range of parameters. For instance, Werden and Froeb say that "We do not view high quality and elaborate econometrics as prerequisites ... based out of necessity of just informed guesses and intuition."[12]

COMPETITION AUTHORITIES RARELY HAVE THE DATA OR TIME AVAILABLE TO RIGOROUSLY PURSUE THESE TECHNIQUES. RATHER, THESE INVESTIGATORS ARE OFTEN IN THE POSITION OF HAVING TO MAKE EDUCATED GUESSES AT THESE PARAMETERS, AND PRESENTING RESULTS FOR A RANGE OF PARAMETERS.

Two-sided markets bring up several new challenges. Firstly, there are normally at least two markets interacting. Naturally, that implies the investigator must provide own and cross-price elasticities for each market. Crucially, the investigator must also specify how the two markets interact. For instance, the videogame console market is thought of as two-sided because game producers will develop games for a console if consumers purchase the console and consumers purchase the console if there are a large variety of games to choose from. To provide a simulation of the console market, the investigator must specify the standard price responses: how consumers respond to prices of different consoles, and how developers respond to developer fees. However, it is also necessary to specify the strength of the consumer response to an increase in games, and the strength of the response of game producers to consumer adoption. These network-effect parameters can be crucial to the predicted outcome, but estimating them requires data on two markets and is still often subject to questions about endogeneous determination of the outcomes in complementary markets. Further, guessing at these parameters is difficult. Investigators are likely to have some experience with guessing price elasticities in different markets and market participants have good incentives to learn price elasticities relatively accurately. However, network-effect parameters fall outside of the experience of most investigators and market participants are unlikely to know them beyond a general sense that network effects are strong or weak.

A second problem is that dynamics are typically very important in two-sided markets. Most discussions of simulations in merger contexts only discuss static models. Naturally, they must be applied to industries for which dynamics do not play too important a role. But two-sided markets are often characterized by tipping and aggressive penetration pricing, for which a dynamic model is more appropriate. Conceptually, it is feasible to introduce dynamics into simulation.[13]

12 *Id.*

13 *See, e.g.*, A. Pakes & P. McGuire, *Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model*, RAND J. ECON. 555 (1994). For airlines, *see* L. Benkard, *A Dynamic Analysis of the Market for Wide-Bodied Aircraft*, 71(3) REV. ECON. STUD. 581 (2004). For a general model of network effects, *see* M. Mitchell & A. Skrzypacz, *Network Externalities and Long-Run Market Share*, 29(3) ECON. THEORY 621-48 (2006).

But in practice, doing so is a major computational undertaking and will often not be a reasonable option as part of the merger review process.

A third issue to keep in mind is that the link between prices and quantities is often more ambiguous in two-sided markets. For instance, it is often possible for consumers and sellers to rationally not utilize an intermediary if the other side does not, even if price is low. In that case, the traditional focus of the U.S. merger guidelines on price effects (in particular the SSNIP test)[14] may be misguided.

To see the importance of these issues, consider my 2004 paper which studies the market for yellow pages.[15] In their 2006 paper, Kaiser and Wright take a similar approach to study price-cost markups in the magazine industry.[16] Yellow pages are a two-sided market because consumers value a directory based on how much advertising is in the directory and advertisers demand advertising based on consumer usage, leading to a positive feedback loop. A publisher determines the price and quantity of advertising (and other features) taking into account how readers will respond. I model the two-sided market as two simultaneous equations, one to represent consumer demand and one to represent advertiser demand. Stripped to essentials, the model is as follows. In the paper, I specify consumer usage of book j as a function of how much advertising appears in books j and in the competitors of the book, indexed as $-j$:

$$Usage_j = f(Advertising_j, Advertising_{-j}, X_j^U) \quad (1)$$

Naturally, one would expect $Usage_j$ to increase in $Advertising_j$, which represents the first half of the network effect. Also, $Usage_j$ should decrease in $Advertising_{-j}$. Here, X_j^U refers to consumer demographics, such as education level and income.

Advertiser demand is specified as follows:[17]

$$Advertising_j = g(Price_j, Usage_j, X_j^A) \quad (2)$$

This equation states that the quantity of advertising at directory j is a function of the price of advertising at directory j and the consumer usage of directory j. Advertising should increase in usage, which represents the second part of the network effect. The relationship between advertising and price represents the

14 SSNIP is an abbreviation for a small but significant non-transitory increase in price.

15 Rysman (2004), *supra* note 4.

16 U. Kaiser & J. Wright, *Price structure in two-sided markets: Evidence from the magazine industry*, 24(1) Int'l J. Indus. Org. 1 (2006).

17 In fact, in my paper I specify Equation 2 with price on the left-hand side and quantity of advertising on the right-hand side. Doing so has some technical advantages for purposes of estimation, but I believe that seeing the Equation 2 with quantity on the left-hand side is more intuitive. (Rysman, *supra* note 4).

standard relationship between quantity and price in any demand curve and should be downward sloping. Generically, the quantity at directory j should be a function of prices at all competing directories, but I argue that this effect can be assumed away in the yellow pages market, and I test this assumption. Also, X_j^A represents consumer demographics that affect advertising demand, such as income.

Finally, in the paper I specify a third equation that determines how a publisher sets prices. Following standard oligopoly theory, I assume that publishers set marginal revenue equal to marginal cost taking their comepetitors' choices as given:

$$\textit{Marginal Revenue}(\textit{Advertising}_j, \textit{Advertising}_{-j}, \textit{Usage}_j, X_j^U, X_j^P) = MC_j \qquad (3)$$

This equation is largely for purposes of identifying marginal cost, which contributes to later calculations.

Note that studying the yellow pages market simplifies a number of issues. Consumers do not pay to use yellow pages directories, which not only implies that there are no prices in Equation (1) but also eliminates the equation that determines the price on the consumer side. Another simplification that is realistic for yellow pages is that consumers value all advertising. In many media markets, consumer valuation of advertising is ambiguous. For instance, newspaper consumers may attach positive value to local and classified advertisements but negative value to national advertisements, and then there is the further valuation of editorial content. In his 1970 paper, Rosse specifies a model with five equations to study newspapers.[18] Finally, the yellow pages market is relatively mundane and established, at least when compared to many of the markets that might be considered two-sided. It is reasonable in this case to specify a static model and ignore such issues as consumer learning over time.

In my 2004 paper, I estimate this model on a data set of 419 directories in a several metropolitan statistical areas. The model considers diary data of consumer usage as well as a number of prices and the number of pages in each directory, which proxies for quantity. The result is a statistically and economically significant positive feedback loop, both that advertising affects usage in Equation (1) and that usage affects advertising in Equation (2).

As an application, I consider what would happen if the number of directories were to increase exogenously. For these purposes, I turn to simulation in the spirit of Werden and Froeb.[19] Because the estimation procedure finds that directories are close substitutes from the point of view of consumers, switching from monopoly to duopoly leads to massive price decreases and advertising increases as a way

18 J. Rosse, *Estimating Cost Function Parameters Without Using Cost Data: Illustrated Methodology*, 38(2) Econometrica 256 (1970).

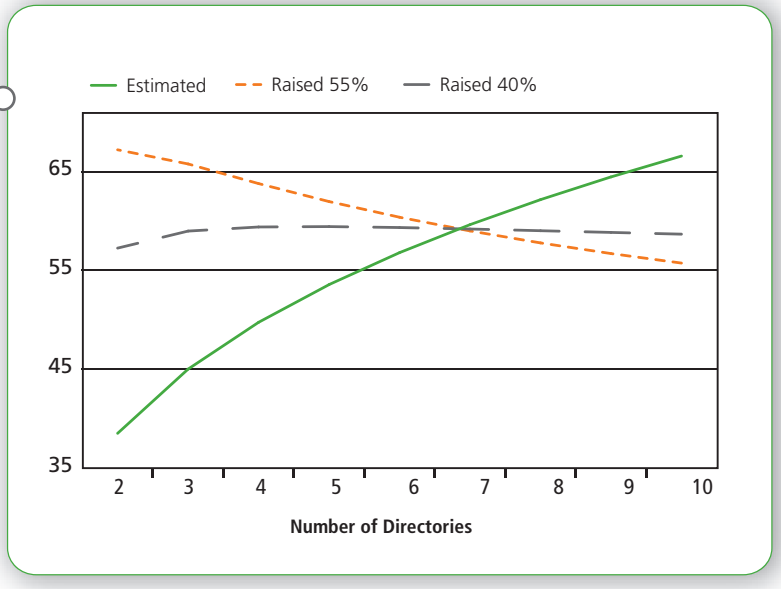
19 Werden & Froeb (2000), *supra* note 11.

to attract consumers. Hence, welfare rises although the network is broken up among two directories instead of one.

One insight from my paper that is particularly germane to the points being made here is the fragility of the results in the face of small parameter changes. Figure 1 presents total surplus calculated for different numbers of competitors for three parameterizations. The parameterizations differ in the treatment of the effect of usage on advertiser demand. The solid line represents the estimated parameter. The other two lines represent cases where this effect is 40 percent and 55 percent larger, respectively. What is interesting is what the result would have been if the network effect has been estimated to be larger. These experimental network parameters are larger than what was found but not unreasonably so, and probably could not be ruled out just on a priori common knowledge of the industry. For the estimated parameter, surplus increases as the number of firms increases. But strikingly, a 40 percent larger network parameter leads to a slightly humped shaped total surplus curve and a 55 percent larger parameter leads to a downward slope. Why downward sloping? A typical merger simulation could find surplus decreasing in the number of competitors because there are economies to have production concentrated at a single firm. That is not the case here as the simulation assumes that marginal cost is constant across quantities and publishers, and there are no fixed costs at all. Instead, the reduction in surplus is coming entirely from the increased number of competitors breaking up the network of consumers and advertisers and thereby reducing the benefits of network effects.

Figure 1

Surplus for different network parameters



The lesson for the use of simulations in antitrust is that the strength of the feedback between two sides of a market are crucially important in determining the outcome of the study. If one is to use guesswork rather than estimation to determine these parameters, one must consider a range of reasonable parameters to test the sensitivity of results. The fragility of these results would seem to support the use of estimation based on representative data rather than even well-informed guesswork. But one should not be unrealistic about what estimation approaches can deliver. Estimation procedures are driven by their own assumptions that can also be subject to sensitivity analysis. Furthermore, confidence intervals may play an important role. For instance, my 2004 paper uses statistical tests to reject the possibility that welfare decreases in the number of competitors but cannot reject the possibility of a hump shape. The paper concludes that the data argue in favor of moderate levels of competition but are silent on further increases.

II. Price Regressions

A direct test of market power is to show that prices increase in markets with less competition. A popular approach in the antitrust literature is to regress a market price on the number of competitors in a market and control for other market characteristics through observable variables. Like market simulations, price regressions play an important role in the determination of which cases the government pursues. In addition, price regressions have actually been introduced as evidence in court and seemed to play an important role in the results. In the discussion in Dalkir and Warren-Boulton's 2004 paper about the Office Depot-Staples merger,[20] the regression is essentially:

$$\text{Price} = f(N_{Competitors}, \textit{market variables})$$

The focus was on whether price dropped by more than five perecent in markets with an additional competitor.

Extending this sort of regression to a two-sided market brings up several issues. First, a two-sided market implies that there are at least two prices to check. One could imagine just looking at one price in isolation, but that may be misleading in the context of two-sided markets. Rather, a more useful approach will often be to specify two regressions, one predicting a price on each side of the market. Second, the number of competitors may differ across the two sides of the market. Therefore, the investigator must determine the relevant market in two markets rather than one. Furthermore, the market structures on both sides of the market determine each of the prices. Hence, there would be two measures of competition in each regression, necessarily complicating the analysis.

20 *See, e.g.,* the discussion in Dalkir & Warren-Boulton, *supra* note 5.

As an example, I discuss (very) preliminary work detailed in my 2007 paper (with Professor Ginger Jin), that studies sports card conventions, typically baseball cards.[21] Convention organizers must attract both collectors and dealers, which has implications for how they price. We observe prices on both sides of the market for around 50,000 conventions in the early 1990's. At the height of the market, there were up to 2,000 conventions a month in the United States so consumers and dealers often had a choice of conventions to attend, bringing conventions into competition with each other.

Crucial to the analysis is the determination of the number of conventions that compete for dealers and consumers. Based on discussions with industry sources, we argue that conventions on the same weekend that are within a particular distance compete for dealers but not consumers. A reasonable distance is one hundred miles. Consumers are unlikely to travel one hundred miles for a sports card convention whereas dealers would travel this distance. Conversely, conventions in the same town but on different days or adjacent weekends compete more strongly for consumers. Dealers will likely turn out for each of the conventions (multi-home) whereas consumers will go to only one, if only because the same dealers with the same collections will be at each. With these thoughts in mind, we specify the following regression system:

$$P_{Dealer} = f(N_{Dealer}, N_{Consumer}, \textit{market variables})$$

$$P_{Consumer} = g(N_{Dealer}, N_{Consumer}, \textit{market variables})$$

The goal of our 2007 paper is to test recent theories of two-sided markets, such as Rochet and Tirole present in their 2003 paper.[22] We expect N_{Dealer} to have a more negative effect than $N_{Consumer}$ on P_{Dealer}, and vice versa. In fact, we present a theoretical model in which $N_{Consumer}$ has a positive effect on P_{Dealer} (and vice versa). Certainly, it would be hard to justify such a result without appealing to explanations based on two-sided markets. The larger point is that we have gone from focusing on one parameter in the Office Depot-Staples merger case to four parameters in our 2007 paper.

This example may oversimplify many of the issues that would arise in typical antitrust examples. First of all, it may be difficult to characterize markets with a single price and finding methods for representing price schedules can raise complications.[23] Second, sports card conventions are attractive for research purposes not only because of their simple pricing but also because the vast number of them lends the industry to statistical analysis. Standard examples of two-sided markets,

21 Jin & Rysman, *supra* note 6.

22 J.-C. Rochet & J. Tirole, *Platform Competition in Two-Sided Markets*, 1 J. Eur. Econ. Ass'n 990 (2003).

23 For an attempt at this, see M. Busse & M. Rysman, *Competition and Price Discrimination in Yellow Pages Advertising*, RAND J. Econ 378 (2005).

such as website portals or videogame console manufacturers, likely generate more ambiguous prices and much, much fewer prices that may be proprietary secrets. While these problems are true even in the case of single-sided markets, they are magnified in the case of two-sided markets where we require data on two sides.

III. Other Questions

The previous two sections focused on tools that already have a role in antitrust analysis. However, two-sided markets bring up a number of questions that have not arisen previously, for which empirical analysis can be relevant. This section is very open-ended but I focus on two questions that seem particularly important. The first is the basic question of whether or not a market is two-sided. The second is whether market participants single-home or multi-home. I discuss both cases in the context of the analysis of the payment card industry in my forthcoming paper.[24]

One can easily imagine an antitrust case turning on the question of whether or not a market is two-sided. For example, the interchange fee set by Visa and MasterCard has been heavily litigated, and one of the principal defenses has been that the interchange fee is crucial in achieving the optimal level of transactions on both sides of the payment card market.[25] Testing for two-sidedness requires detailed data on both sides of the market, which is often a daunting task. Also, it would often be unclear what to test for as there is no widely agreed on definition of two-sidedness and some definitions are quite broad.

In my forthcoming paper, I address these issues in payment card industry. In order to test for two-sidedness, I test for a positive feedback loop between consumer usage and merchant acceptance, which can be thought of as an indirect network effect. There is some confusion as to the relationship between the two-sidedness of a market and whether a market exhibits an indirect network effect. An indirect network effect exists when consumers value a product based on how much of some complementary product is provided, and the amount of the complementary product depends on consumer purchases of the first good. This positive feedback loop between consumer purchases and the provision of complementary products has a similar flavor to the idea of getting both sides on board associated with two-sided markets. However, Rochet and Tirole, in their forthcoming paper, suggest a definition of two-sidedness that is somewhat broader

24 Rysman (forthcoming 2007), *supra* note 9.

25 In fact, the decision in favor of the interchange fee in the *NaBanco* case seemed to be based more on joint venture issues rather than two-sided arguments. We can logically separate whether a single, collectively set interchange fee is necessary for a payment card association to exist, in which case it might be legal under the standard treatment of joint ventures, from whether such an interchange fee is necessary to optimally provide a two-sided service, which would break new legal ground. NaBanco v. Visa, 779 F.2d 592 (1986).

than network effects.[26] In my paper, I rely on the fact that the presence of indirect network effects implies two-sidedness under all definitions of two-sidedness. While it is arguable whether the lack of indirect network effects implies a lack of two-sidedness, the presence of indirect network effects is surely sufficient.

It is obvious that there must be at least some network effect because consumers would not hold a card if no merchant accepted it. However, one may wonder if network effects are still detectable in a mature market with firms as large as Visa and American Express.[27] To establish two-sidedness, I rely on data from the *Payment Systems Panel Study* (from Visa International) that records consumer usage from 1998 to 2001. For one month out of each quarter, consumers record how they make every monetary transaction for the month. I observe whether the consumer uses cash or a payment card (or many other options) and the brand of the payment card. In addition, a separate data set, the Visa Transactions Database, records the dollar value of transactions on the Visa network for all merchants. I have these data monthly from 1998 to 2001. Because some charges for the other networks (MasterCard, American Express, and Discover) appear on the Visa network, I have proxies for network acceptance by month for each major network. Both data sets indicate the zip code, either of the household or the merchant, which allows me to establish regional correlations.

I use the panel survey to establish the favorite network of each household (the networks are Visa, MasterCard, American Express, and Discover). I then estimate a multinomial logit model of how consumers make this choice, which includes household demographics as explanatory variables and in particular, counts of how many merchants transact on each network in the household's 3-digit zip code. I interpret the counts as (noisy) measures of the extent of merchant acceptance. The results show a strong correlation between my measure of merchant acceptance and consumer usage of the payment network for Visa, American Express, and Discover. Interestingly, high merchant acceptance of Visa is not correlated with less consumer usage of MasterCard, and vice versa. This result is not surprising given that true merchant acceptance is practically identical for MasterCard and Visa and suggests that my proxies for merchant acceptance capture their intended effects well.

Even with the very detailed data, the study has some important limitations. In particular, it is difficult to establish causality. That is, I do not take a stand on whether the correlation between consumer usage and merchant acceptance is

26 Rochet & Tirole (2006), *supra* note 1.

27 For instance, Michael Katz writes: "There is an argument made by some analysts that implies that 'mature' payment networks might reasonably be treated as one-sided platforms at the margin." (M. Katz, What Do We Know about Interchange Fees and What Does It Mean for Public Policy?, Remarks at Interchange Fees in Credit and Debit Card Industries, Federal Reserve Bank of Kansas, May 5, 2005) *in* Proceedings - Payments System Res. Conf., May 2005, at 121.

caused by consumer usage affecting merchant acceptance or merchant acceptance affecting consumer usage or both.[28] However, it is sufficient to imply that the market is two-sided in any of these cases.

Given that a market is two-sided, one may then ask whether agents practice multi-homing or single-homing. That is, do buyers or sellers participate in multiple platforms or just one. The answer has important implications for market power. If one side of a market practices single-homing, then the only way for the other side to reach those agents is through their preferred platform. That is, a platform has substantial market power over access to subscribers that single-home, but much less so if they multi-home. Theoretical models such as Armstrong's predict intense competition between platforms on the single-homing side of the market and almost non-existent competition on the multi-homing side.

If one side of a market practices single-homing, then the only way for the other side to reach those agents is through their preferred platform. That is, a platform has substantial market power over access to subscribers that single-home, but much less so if they multi-home.

Finally in this paper, I characterize the level of single-homing in the payment card market. In particular, I use the panel survey to establish the extent to which consumers hold cards from different networks and the extent that they use cards from different networks. I find that the question of whether consumers multi-home has a more complex answer than commonly envisioned. With regards to usage, few consumers regularly use multiple networks. Most consumers put a great majority of their payment card purchases on a single network. The level of concentration varies only slightly with the choice of network or with consumer characteristics such as income, education, and spending. However, with regards to ownership, most consumers do maintain cards from different networks, which would allow them to take advantage of different networks quickly if they chose to do so. These results suggest that consumers prefer single-homing but are willing to use a less-preferred payment network to purchase a product for which there is no sufficiently close substitute. A merchant in a highly competitive environment most likely must associate with multiple payment networks or risk a real decrease in sales.

28 In fact, it is possible that there is some omitted heterogeneity that drives the correlation between usage and acceptance, although I try to rule that out by focusing on how merchant acceptance and usage at one network are correlated relative to the correlation at another network, rather than some absolute level of correlation.

IV. Conclusion

The paper provides an overview of recent empirical research in the economics of two-sided markets from the perspective of antitrust enforcement. Whereas the empirical tools developed in the study of industrial organization have found an increasingly important role in antitrust litigation, the much more recent empirical research on two-sided markets have yet to make an impact. However, this is likely to change in the near future. The two main empirical tools in antitrust, market simulations and price regressions, have natural corollaries for two-sided markets. Adopting these tools to two-sided markets brings up several problems. In particular, there is extra work in correctly calibrating or estimating how outcomes on one side of the market affect the other side of the market, and the requirement to learn about both sides of the market brings up associated data constraints. However, just as in one-sided markets, empirical tools can provide valuable information to antitrust enforcers in two-sided markets. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

The Two-Sided Market Literature Enriches Traditional Antitrust Analysis

William H. Rooney and David K. Park

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

The Two-Sided Market Literature Enriches Traditional Antitrust Analysis

William H. Rooney and David K. Park

The term "two-sided market" sounds strange to the antitrust lawyer's ear. Antitrust markets typically are not described as having "sides." They consist of a relevant product or set of products, cover a geographic area, and include transactions between buyers and sellers at a particular level of distribution (e.g., manufacturing, wholesale, retail). Although most market participants buy inputs and sell outputs, they usually buy in the market for the input and sell in the market for the output, not compete in a two-sided market.

Still, the growing and informative literature on two-sided platforms, businesses, and markets has much to offer antitrust law. That literature emphasizes that the demand for otherwise distinct products or services may in fact be linked and that a competitive-effects analysis cannot myopically ignore that linkage. To the extent that the two-sided market literature improves competitive-effects analysis, it improves the fundamental purpose of antitrust law.

This essay briefly discusses the importance of acknowledging linked demand for, and relationships among, otherwise distinct products or services, as recommended by the two-sided market literature, with respect to competitive-effects assessments and market definition. We also observe that recognizing linked demand and interrelationships among products or services facilitates the application of legal rules in antitrust cases.

William H. Rooney is a partner in and David K. Park is special counsel to the law firm of Willkie Farr & Gallagher LLP.

I. Recognizing Linked Demand and Interrelationships among Products or Services Improves Competitive-Effects Analysis and Market Definition

Commentators writing about two-sided markets characterize certain businesses that depend on (and facilitate) the interdependent demand of two or more discrete groups of constituents as two-sided platforms or markets. Rochet and Tirole, for example, define two-sided markets as follows:

> "A market is two-sided if the platform can affect the volume of transactions by charging more to one side of the market and reducing the price paid by the other side by an equal amount; in other words, the price structure matters, and platforms must design it so as to bring both sides on board."[1]

Another commentator summarizes the necessary conditions for the emergence of a platform business or market as follows:

(1) there are two or more distinct groups of customers;

(2) there are externalities associated with customers A and B becoming connected or coordinated in some fashion; and

(3) an intermediary is necessary to internalize the externalities created by one group for the other group.[2]

THE OBSERVATION THAT THE PRESENCE OF ONE SET OF CONSTITUENTS MAY AFFECT THE DEMAND OF ANOTHER SET OF CONSTITUENTS IS BOTH TYPICAL OF THE TWO-SIDED MARKET LITERATURE AND USEFUL TO A COMPETITIVE-EFFECTS ASSESSMENT.

Although the two-sided market literature sometimes uses market in more of a business sense than in a technical antitrust sense, the observation that the presence of one set of constituents may affect the demand of another set of constituents is both typical of the two-sided market literature and useful to a competitive-effects assessment.

1 J.-C. Rochet & J. Tirole, *Two-Sided Markets: A Progress Report*, RAND J. ECON. (Autumn 2006).

2 David S. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325, 331-34 (2003).

Newspaper competition provides a simple example. Newspapers have (at least) two sets of buyers: readers, who buy news, and advertisers, who buy page space. Both readers and advertisers pay the newspaper a fee for the product that they are buying. Advertisers pay more to newspapers with more readers. Readers, on the other hand, do not pay more to newspapers with more advertisers but with better content, which represents a cost to the newspaper. The newspaper may determine that it can maximize revenue by charging readers very little even for expensive content to maximize the number of readers and attract advertisers. A newspaper may also conclude that advertisers in the aggregate are willing to pay more for page space than readers are willing to pay for content. Asserting that a subscription rate is below the cost of providing content, for example, would omit the important revenue that the newspaper receives from advertisers and overlook that publishers seek to maximize aggregate revenues from readers and advertisers alike.

Capturing competitive effects in a dynamic analytical paradigm also has important implications for market definition. Demand for apparently distinct products may be linked in a way that requires the products to be included in the same competitive venue—or relevant market—for their competitive dynamics to be understood properly. Although courts and agencies typically include in a relevant market products that are substitutes for one another,[3] cluster markets have been defined to include complementary products that respond to linked consumer demands and offer sellers economies of scope. Examples of cluster markets have included such complementary or related product groupings as:

(1) general acute care inpatient hospital services;[4]

(2) commercial banking services;[5]

(3) accredited central station service alarms (including burglar and fire alarm services);[6] and

(4) small business loans and depository services.[7]

---

3 *Compare e.g.*, U.S. Department of Justice and Federal Trade Commission Horizontal Merger Guidelines (the "Horizontal Merger Guidelines") §1.11 (1992, revised 1997) (own-price and cross-product elasticities, and "practical indicia") *with* Brown Shoe Co. v. United States, 370 U.S. 294, 325 (1962) (reasonable interchangeability of use, cross-product elasticities of demand, and practical indicia) and *with* United States v. E.I. du Pont de Nemours & Co., 351 U.S. 377, 395, 400 (1956) (reasonable interchangeability for the same purposes and cross-elasticity of demand across products).

4 United States v. L.I. Jewish Med. Cen., 983 F.Supp. 121, 140 (E.D.N.Y. 1997).

5 United States v. Phil. Nat'l Bank, 374 U.S. 321, 356 (1963).

6 United States v. Grinnell Corp., 384 U.S. 563, 572 (1966).

7 *See* Robert L. Webb, *Divestiture: A Prescription for Healthy Competition,* The Regional Economist, Jan. 2001, *available at* http://stlouisfed.org/publications/re/2001/a/pages/economic-backgnd.html.

Further to the cluster-market authorities, the U.S. Court of Appeals for the Ninth Circuit upheld two all-parts markets in *Kodak II* that included all replacement parts for Kodak photocopiers and for Kodak micrographics equipment, respectively.[8] The court in *Kodak II* rejected defendants' argument that "because no two parts are interchangeable, the relevant markets for parts consist of the market for each individual part for Kodak photocopiers and each single part for Kodak micrographics equipment."[9] The Ninth Circuit explained that "Kodak's market definition focuses exclusively on the interchangeability of the parts although ignoring the 'commercial realties' faced by ISOs and end users."[10] The Ninth Circuit cited *Grinnell* and *Philadelphia National Bank* as examples of cases where, after analyzing the commercial realities, the Supreme Court "has held that groups of non-interchangeable products and services may be aggregated to form a single relevant market."[11]

Consumer demand for the deposit, withdrawal, and use of funds provides another example of linked demand for a compound product that is comprised of otherwise apparently distinct components. A bank customer can obtain a compound product—the deposit of funds and the withdrawal funds—from a single supplier (i.e., a bank). That product, however, increases in value when it is provided by multiple suppliers cooperating with one another (i.e., a bank and its network of ATMs). Multiple and diverse suppliers (i.e., a bank, network ATMs, and network merchants) can collaborate to provide consumers with an even more valuable compound product—the deposit and withdrawal of funds and the use of those funds to purchase goods and services from merchants throughout the economy).

Whether all services that facilitate the deposit, withdrawal, and use of funds, and all of the providers of those services, compete in one network market or in multiple markets consisting of only portions of those services poses an interesting question under antitrust law. Although we do not propose to answer that question here, principles from the two-sided market literature imply that the market-definition inquiry may be affected by the practice in question and its competitive context. Such principles suggest that the examination of the competitive objective of the particular practice at issue and the consumer demand to which the practice is intended to respond help identify the venue in which the competitive effects of the practice should be assessed. They further suggest that competitive objective and consumer demand bear on the qualitative analysis of competitive impact, including whether the relevant competition is properly described as intrabrand or interbrand.

8 Image Technical Services, Inc. v. Eastman Kodak Co., 125 F.3d 1195, 1203 (9th Cir. 1997).

9 *Id.*

10 *Id.*

11 *Id.* at 1204.

II. Acknowledging Linked Demand Improves the Application of Legal Rules

In *BMI*,[12] the U.S. Supreme Court examined whether the collective pricing of blanket licenses offered to the copyrighted works of songwriter members of ASCAP and BMI constituted per se price-fixing.[13] If the practice were viewed as an agreement among otherwise competing songwriters as to the terms of their respective licenses, the agreement may have been properly viewed as per se illegal. The Court, however, examined the blanket license in competitive context and recognized that the license responded to the demand of radio stations for a bundle of related services and that the collaborating songwriters could not have provided the same product themselves:

> "[h]ere, the whole is truly greater than the sum of its parts; [the blanket license] is, to some extent, a different product. The blanket license has certain unique characteristics: It allows the licensee immediate use of covered compositions, without the delay of prior individual negotiations, and great flexibility in the choice of musical material. ... Thus, to the extent the blanket license is a different product, ASCAP is not really a joint sales agency offering the individual goods of many sellers, but is a separate seller offering its blanket license, of which the individual compositions are raw material. ASCAP, in short, made a market in which individual composers are inherently unable to compete fully effectively."[14]

By recognizing that the songwriter members of ASCAP and BMI were collaborating to supply a compound product in response to a linked demand by radio stations, the U.S. Supreme Court declined to declare the blanket license per se illegal.[15] *BMI* has provided important guidance in the last 25 years by instructing courts to review the substance of an arrangement, not its form, and to assess whether alleged co-conspirators are in fact collaborating to satisfy consumer demand more effectively than any one participant could on its own.

The U.S. Supreme Court recently elaborated on its holding in *BMI* to clarify that the legal capacity in which market participants act is determined by the sub-

12 Broadcast Music, Inc. v. Columbia Broadcasting System, Inc. ("BMI"), 441 U.S. 1 (1979).

13 *Id.* at 4.

14 *Id.* at 21-23 (footnotes omitted).

15 *Id.* at 24.

stance and objective of their concerted activity. In *Dahger*,[16] Texaco Inc. and Shell Oil Co. formed a joint venture called Equilon Enterprises to market their respective gasoline in the western part of the United States.[17] Texaco and Shell Oil maintained their respective brands of gasoline but set the prices of their gasoline jointly through Equilon.[18] Plaintiffs claimed that Equilon provided a vehicle through which Texaco and Shell Oil had engaged in per se illegal price-fixing.[19] The Court, however, characterized the price setting by the joint venture as "little more than price setting by a single entity—albeit within the context of a joint venture [Equilon]—and not a pricing agreement between competing entities with respect to their competing products."[20]

BMI HAS PROVIDED IMPORTANT GUIDANCE IN THE LAST 25 YEARS BY INSTRUCTING COURTS TO REVIEW THE SUBSTANCE OF AN ARRANGEMENT, NOT ITS FORM, AND TO ASSESS WHETHER ALLEGED CO-CONSPIRATORS ARE IN FACT COLLABORATING TO SATISFY CONSUMER DEMAND MORE EFFECTIVELY THAN ANY ONE PARTICIPANT COULD ON ITS OWN.

Further to its holding in *BMI*, the U.S. Supreme Court forcefully rejected the application of the per se rule of illegality to the pricing conduct of Equilon:

> "When "persons who would otherwise be competitors pool their capital and share the risks of loss as well as the opportunities for profit...such joint ventures [are] regarded as a single firm competing with other sellers in the market." *Arizona v. Maricopa County Medical Soc.*, 457 U.S. 332, 356 (1982). As such, though Equilon's pricing policy may be price fixing in a literal sense, it is not price fixing in the antitrust sense. See *Broadcast Music, Inc. v. Columbia Broadcasting System Inc.*, 441 U.S. 1, 9 (1979) ("When two partners set the price of their goods or services they are literally 'price fixing,' but they are not *per se* in violation of the Sherman Act")."[21]

16 Texaco Inc. v. Dagher, 126 S.Ct. 1276 (2006).

17 *Id.* at 1278.

18 *Id.*

19 *Id.* at 1279.

20 Id. at 1280.

21 *Id.* (alterations in original).

Although the U.S. Supreme Court formally limited its holding to rejecting the application of the per se rule of illegality,[22] it implied that Equilon was a single market participant and that, following the formation of Equilon, Texaco and Shell Oil acted through Equilon not as competitors but as shareholders.[23] Indeed, the Court clarified that the pricing conduct at issue was sufficiently close to the core of the collaboration between Texaco and Shell Oil that such conduct could not be considered a restraint subject to the ancillary-restraints analysis that is fundamental to the application of Section 1 of the Sherman Act: "We agree with petitioners that the ancillary restraints doctrine has no application here, where the business practice being challenged involves the core activity of the joint venture itself—namely, the pricing of the very goods produced and sold by [the joint venture]."[24] The Court therefore found that the price setting neither was per se illegal price-fixing nor should be assessed under the ancillary restraints doctrine of Section 1 of the Sherman Act.[25]

BMI and *Dahger* reflect the importance of examining the context and objective of the conduct at issue to determine the capacity in which the parties were acting—whether the parties to the restraint were acting as conspiring competitors or collaborating suppliers that formed a single market participant. That examination is also critical to determining whether any residual or incidental competition among the relevant sellers was intrabrand (competition in the sale of the same product) or interbrand (competition in the sale of substitute products). The U.S. Supreme Court has emphasized for almost 30 years, and most recently in *Volvo Trucks North America, Inc. v. Reeder-Simco GMC, Inc.*,[26] that interbrand competition is the primary concern of antitrust enforcement.[27]

In *SCFC ILC, Inc. v. Visa USA, Inc.*,[28] the U.S. Court of Appeals for the Tenth Circuit considered a Visa rule that did not allow Discover (or American

22 *Id.* at n.2 (noting that "Respondents have not put forth a rule of reason claim.").

23 *Id.*

24 *Id.* at 1281.

25 *Id.*

26 126 S.Ct. 860, 872-73 (2006).

27 The U.S. Supreme Court defined interbrand and intrabrand competition in Continental T.V. Inc. v. Sylvania, 433 U.S. 36, 52 n.19 (1977): "Interbrand competition is the competition among the manufacturers of the same generic product . . .. In contrast, intrabrand competition is the competition between the distributors . . . of the product of a particular manufacturer."

28 36 F.3d 958 (10th Cir. 1994).

Express) to issue Visa cards.[29] The Tenth Circuit properly began its assessment by identifying the fundamental competitive objective of the market participants and thus the inter-brand competition at issue:

> "In this lawsuit, Sears and Visa USA stipulated "the relevant market is the general purpose charge card market in the United States." Presently, the only participants in this market are Visa USA, MasterCard, American Express, Citibank (Diners Club and Carte Blanche), and Sears (Discover Card). Competition among these five firms to place their individual credit cards into a customer's pocket is called *intersystem*. "Interbrand competition is the competition among the manufacturers of the same generic product...and is the primary concern of antitrust law." *Continental T.V., Inc. v. GTE Sylvania, Inc.*, 433 U.S. 36, 52 n.19 (1977)."[30]

The *SCFC* court further explained that competition among those collaborating to form the Visa network was properly understood as intrabrand competition.[31] Although issuers and acquirers may compete with each other in the issuance of Visa cards and the acquisition of transactions, that competition was intrabrand when viewed within the context of the primary competitive objective of permitting cardholders to use funds in depository accounts to purchase goods and services throughout the economy:

> "[T]o the extent that Visa USA is in the market, it operates in the *systems* market, not the *issuer* market. Its *members* issue cards, competing with each other to offer better terms or more attractive features for their individual credit card programs. This is *intrasystem* competition."[32]

29 The district court notes the language Visa added to its bylaws: "[T]he corporation shall not accept for membership any applicant which is issuing, directly or indirectly, Discover cards or American Express cards, or any other cards deemed competitive...." SCFC ILC, Inc. v. Visa U.S.A., Inc., 819 F.Supp. 956, 964 (D. Utah 1993). Under the bylaw "non-VISA members who develop a successful proprietary card would be prohibited from joining the VISA system and current VISA members would be expelled from the system if they developed such a card." *Id.* at 966.

30 SCFC ILC, Inc., 36 F.3d at 966 (emphasis added) (internal citations omitted).

31 *Id.* at 967.

32 *Id.* (emphasis added).

III. Conclusion

The two-sided market literature enriches antitrust analysis by illustrating how consumer demand can require product or service compilations and supplier collaborations that, in other contexts, may present concerns under the Sherman Act. Identifying the competitive objective of the suppliers and the consumer demand to which the suppliers are responding permits a more accurate competitive-effects assessment and market definition and facilitates the application of legal rules, including those prohibiting price-fixing and preserving interbrand competition. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Predatory Pricing in Two-Sided Markets: A Brief Comment

Amelia Fletcher

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Predatory Pricing in Two-Sided Markets: A Brief Comment

Amelia Fletcher

Over the past few years, there has been a burgeoning literature on two-sided markets and economic understanding of such markets has improved hugely. Less attention has, however, been paid to how competition policy should be applied in two-sided markets.

This short note does not attempt to provide a comprehensive analysis of this issue, but merely presents a brief comment on the implications of two-sided market theory for one possible abuse of a dominant position under Article 82 of the EC Treaty: predatory pricing.

I. Pricing in Two-Sided Markets

A key finding of two-sided market theory is that the prices charged on one side of the market need not reflect the costs incurred to serve that side of the market. Rather, the price structure in a two-sided market will typically be designed to get both sides of the market on board.

If we define one side of the market as the buyer side and the other as the seller side, then the price charged to one side (say, the buyer side) will tend to be lower when either:

- each additional buyer generates significant extra revenue on the seller side; or
- it is difficult to persuade buyers to join the platform.

The author is Chief Economist, U.K. Office of Fair Trading. The views expressed here are my own, and not necessarily those of the Office of Fair Trading. The author would like to thank Peter Lukacs and Mark Armstrong for useful comments on earlier versions of this note.

In their 2006 paper, Rochet and Tirole analyze this situation more formally and show that the standard Lerner formula for monopoly profit maximization can be applied to two-sided markets.[1] That is, within a given market, a monopoly platform will price such that:

$$\frac{\text{price} - \text{'cost'}}{\text{price}} = \frac{1}{\text{elasticity of demand}}. \quad (1)$$

The key difference in a two-sided market context relates to how one interprets the cost term in this equation. Under the standard Lerner formula, this is marginal cost. In a two-sided market, the cost term needs instead to be interpreted as a form of opportunity cost, which comprises the marginal cost of serving the buyer side of the market minus any extra revenue that the extra sales on the buyer side of the market generate on the seller side of the market, either through extra usage charges or by being able to increase sellers' membership fees.

II. Implications for Predatory Pricing

What does this mean for the assessment of predatory pricing in two-sided markets? The first point to make is that we might expect to often observe:

- pricing below cost on one side of the market; and
- pricing well above cost on the other.

Thus, if looked at in isolation, there is a risk that a supplier could be accused of predatory pricing on one side of the market. This issue has been highlighted by a variety of commentators, for example, Wright in his 2004 paper.[2]

Application of the simple *Akzo*[3] test for predation, under which a presumption of abuse is formed if price lies below a cost benchmark, could clearly give erroneous results in such circumstances.[4] When applied in a simple one-sided market,

1 For an excellent recent summary of the latest literature on two-sided markets, *see* J.-C. Rochet & J. Tirole, *Two-Sided Markets: A Progress Report*, RAND J. Econ. (Autumn 2006).

2 J. Wright, *One-sided Logic in Two-sided Markets*, 3(1) Rev. Network Econ. (2004).

3 Case 62/86, AKZO v. Commission, 1991 E.C.R. I-3359.

4 In the United States, this test is more usually known as the Areeda-Turner rule. The test has historically used an average variable cost benchmark, although many commentators have argued that average avoidable cost would be a more relevant benchmark, and this view now seems to have been accepted by the European Commission. *See* European Commission, DG Competition discussion paper on the application of Article 82 of the Treaty to exclusionary abuses (Dec. 2005), *at* http://ec.europa.eu/comm/competition/antitrust/others/discpaper2005.pdf.

this test provides a way of assessing whether a particular price level is likely to be anticompetitive in both intent and effect. In a two-sided market, however, prices on one side of the market may well lie below cost without the pricing structure having either anticompetitive intent or effect. This is clearly something that the competition authorities need to be aware of when assessing predation.

Does this mean, though, that predation will never occur in two-sided markets? The answer must be no. Firstly, predation can clearly occur where a platform prices its total service at a level that fails to cover its avoidable costs of providing the total service, taking revenues from both sides of the market into account. In such a case, a competing platform may be unable to make a positive profit, regardless of how it structures its pricing, and therefore may be excluded from the market.

Secondly, and more subtly, it may be possible in some circumstances for a dominant platform to predate through asymmetric pricing between the two sides of the market. This can potentially occur even where the platform is covering its avoidable costs of supply overall, taking into account all revenue streams.

This potential concern seems to have received minimal coverage in the literature on two-sided markets to date. Most current models appear to take market structure as given; *n* firms compete and they all compete on both sides of the market. By contrast, the issue here is whether a given pricing structure can affect market structure, and specifically whether low pricing on one side of a market can prevent entry into both sides.

THE ISSUE HERE IS WHETHER A GIVEN PRICING STRUCTURE CAN AFFECT MARKET STRUCTURE, AND SPECIFICALLY WHETHER LOW PRICING ON ONE SIDE OF A MARKET CAN PREVENT ENTRY INTO BOTH SIDES.

This is unlikely to be a feasible exclusion strategy where firms are entirely symmetric. In such a situation, if one firm can gain incremental revenues on one side of a market when it wins extra business on the other side, and prices accordingly, then the same opportunities and pricing incentives will apply to its competitors.

But what if firms are not symmetric? In particular, what if some firms have less ability than the dominant incumbent to turn extra business on one side of the market into incremental revenues on the other? One might, for example, expect this to be true of smaller firms, or newer firms. Such firms could find it hard to compete against a very asymmetric pricing structure, and therefore may be excluded from both sides of the market. This in turn may restrict or eliminate competition between platforms.

In this context, it is worth noting that two-sided markets can tip easily. Buyers will tend to prefer (all other things equal) the platform that offers access to the most sellers, and sellers will tend to prefer the platform that offers access to the most buyers. Such network effects can tip the market towards being served by

just one or two platforms.[5] There is a risk that the asymmetric pricing structure described above could further increase the likelihood of such tipping occurring.

III. Policy Implications

The above discussion suggests that asymmetric pricing between the sides of a two-sided market can potentially constitute predatory pricing and merit competition policy intervention. The question is how to distinguish between low pricing that is predatory and low pricing that is merely the optimal pricing response in a two-sided market.

One possible option, which would merit further consideration, is to adjust the simple *Akzo* test for predation for two-sided markets to employ an opportunity cost benchmark, as described above, rather than the more usual average variable (or avoidable) cost benchmarks.

In applying such a test, it would clearly be important to ensure that the incremental revenues that are generated on the other side of the market—and feed into this opportunity cost calculation—relate directly to the general volume increasing impact of the lower prices on the side of the market where the predation is alleged and do not simply equate to the monopoly profits of recoupment associated with exclusion. However, so long as consideration is given to this point, such a test may have merit.

5 Such tendencies towards tipping may be ameliorated to the extent that there is platform compatibility (for example, such that buyers using one platform can access sellers using another), or that users are able to multi-home (for example, such that buyers are able to switch readily between platforms in order to reach different sellers).

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Antitrust and Real Estate: A Two-Sided Approach

Thomas P. Brown and Kevin L. Yingling

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Antitrust and Real Estate: A Two-Sided Approach

Thomas P. Brown and Kevin L. Yingling

When John Jacob Astor died in 1848, he was the wealthiest man in the United States. Like so many people since, Mr. Astor made his fortune speculating on real estate, specifically undeveloped land on the fringe of the city then growing on the island of Manhattan. Mr. Astor did not start out in the industry. He turned to it only after a shift in fashion diminished the prospects for his fur trading business. On his deathbed, his only regret was that he had not bought more.[1] Over the last decade, Americans have taken Mr. Astor's regret to heart. From 1996 to 2005, the residential real estate industry witnessed the greatest run-up in prices ever seen. In 2005, sales of existing homes hit an all-time high of 7 million units.[2] This should have been the best of times for people in the business of buying and selling houses, but to hear most residential real estate agents tell it, the boom passed them by.

1 Edwin G. Burrows & Mike Wallace, Gotham: A History of New York City to 1898 449 (1999) (noting that Astor is reported to have said just before he died that "[c]ould I begin life again, knowing what I now know, and had money to invest, I would buy every foot of land on the Island of Manhattan").

2 National Association of Realtors, *Existing Home Sales* (2006), *available at* http://www.realtor.org/Research.nsf/files/EHSreport.pdf/$FILE/EHSreport.pdf.

Thomas P. Brown is a partner and Kevin L. Yingling is counsel in the international law firm of O'Melveny & Myers LLP. Both authors are grateful to Christina Brown and Julia Stahl for excellent research assistance. They would like to thank David Brownstein, Robert Pitofsky, Grace Shoet, and the participants in the symposium on competition policy and two-sided markets hosted by eSapience in June 2006 for their comments and suggestions. The views expressed in this paper are the authors' own, and any errors are their responsibility.

According to most residential real estate agents, there was simply too much competition.[3] Real estate may be the easiest profession to enter. Real estate agents do not need to go to college, let alone graduate school. In virtually every state, anyone with a clean criminal record can get a license to sell real estate by spending a few hours in a class and passing a short exam. As a result, in many parts of the country, the annual growth in the number of brokers has outpaced year-over-year increases in the total value of real estate sold. Even over the period that witnessed the greatest price increase in the history of the industry, the expected income for real estate agents in some of the more torrid U.S. markets actually declined.[4] The industry is also remarkably unconcentrated—the top one hundred residential brokerage firms account for just 17 percent of all sales.[5]

But neither the Antitrust Division of the U.S. Department of Justice (DOJ) nor the U.S. Federal Trade Commission (FTC) subscribes to the view that there is too much competition in the real estate industry. At the moment, both agencies are pursuing cases, with the Antitrust Division focusing on the buyer side and the FTC pursuing the seller side.[6] Although the agencies have different targets, they are advancing a common theory. They regard the real estate industry as a poorly functioning cartel, and they claim that real estate brokers are fixing the price of their services at an artificially high level.[7]

The agencies have not climbed out on a limb in reaching this view. Scorn for the industry is not merely conventional wisdom; it is a universally held belief. Economists regard the real estate brokerage industry with the same skepticism as the DOJ's Antitrust Division and the FTC. Even Steven D. Levitt, co-author of *Freakonomics* and holder of the John Bates Clark medal (an award bestowed once every two years to the top U.S. economist under 40), describes the industry as "a cross between a cartel and a mafia" and has put it on the endangered species list.[8]

---

3 *See, e.g.*, National Association of Realtors: Research Division, *Structure, Conduct, and Performance of the Real-estate Brokerage Industry* at 1 (Nov. 2005), *available at* http://www.realtor.org/Research.nsf/files/Structure%20Paper%20FINAL%2011-28-05.pdf/$FILE/Structure%20Paper%20FINAL%2011-28-05.pdf.

4 *Id.* at 16.

5 *Id.* at 2.

6 *See* Press Release, U.S. Department of Justice, Justice Department Sues National Association of Realtors for Limiting Competition Among Real-estate Brokers (Sept. 8, 2005), *available at* http://www.usdoj.gov/atr/public/press_releases/2005/211008.htm; Press Release, U.S. Federal Trade Commission, FTC Charges Real-estate Groups with Anticompetitive Conduct in Limiting Consumers' Choice in Real-estate Services (Oct. 12, 2006), *available at* http://www.ftc.gov/opa/2006/10/realesstatesweep.htm.

7 *See id.*

8 Stephen J. Dubner & Steven D. Levitt, *Endangered Species*, N.Y. Times, Mar. 5, 2006, § 6 (magazine), at 24.

The public seems to agree. As Levitt and Dubner observe, "it is hard to think of an occupation that garners less goodwill these days than the real estate agent."[9]

Given the animosity directed at the industry, we thought that it would be interesting (if slightly foolhardy) to examine the industry somewhat sympathetically for this symposium. The real estate industry shares some characteristics with another industry that we know fairly well, the payment card industry. That industry has also seen a steady stream of antitrust litigation. Like the antitrust litigation against the real estate industry, antitrust litigation involving the payment card industry operates on the premise that the industry is a poorly functioning cartel. We regard the attacks on the payment card industry as misguided. As our colleague and former FTC Chairman Tim Muris has explained, the cases against the payment card industry fail to appreciate the economics of the industry, namely the economics of operating a business in a platform or two-sided industry.[10] With that background, we wonder whether the cases against the real estate industry might be similarly flawed.

WE WONDER WHETHER THE CASES AGAINST THE REAL ESTATE INDUSTRY MIGHT BE SIMILARLY FLAWED.

We have not reached a definite view about the cases against the real estate industry. We do, however, have some preliminary thoughts. The real estate industry does seem to be a two-sided industry. The cases against the industry, the current set as well as the many preceding rounds of litigation, generally do not take into account how the economics of operating a two-sided industry might shape the real estate market. We think that the increasingly familiar concept of a two-sided market provides an interesting perspective on the chronic antitrust issues.

I. Real Estate Brokers Compete in a Two-Sided Market

The concept of a two-sided market is, at this point, well understood. Two-sided markets have three characteristics:

(1) they involve two distinct groups of users;

(2) an intermediary connects one group of users to the other; and

(3) demand for the service provided by the intermediary on one side of the market increases as the number of participants on the other side increases (i.e., demand is interdependent).[11]

9 *Id.*

10 Timothy J. Muris, *Payment Card Regulation and the (Mis)Application of the Economics of Two-Sided Markets*, 2005 COLUM. BUS. L. REV. 515 (2005).

11 *Id.* at 517-18.

Newspapers, payment card systems, and computer operating systems all compete in two-sided markets. Newspapers connect advertisers to subscribers. Payment card systems connect merchants to cardholders. Computer operating systems connect users of programs with developers of programs. In each case, the value of the service provided to one of the sides increases as the number of participants on the other side increases. A payment card system, for example, becomes more valuable to cardholders as the number of merchants accepting the card increases.

The residential real estate business is not as clearly a two-sided market as these more classic examples. The first two criteria, multiple parties and an intermediary, are easy to spot. In order for a house to sell, there must be two parties, a seller and a buyer, and real estate agents clearly connect the two. The third criteria, interdependent demand, is a bit trickier. The demand of a given buyer for the services provided by her agent does not obviously increase with the number of home sellers. The interdependence in the industry arises from how agents on both sides interact with each other and their respective clients.

The dominant feature of the residential real estate market in the United States is the local multiple listing service (MLS). Real estate agents control access to the MLS for both sellers and buyers. Real estate agents use MLSs to pool their property listings.[12] By posting a house on an MLS, an agent representing a seller can communicate with all agents, thus increasing the pool of potential buyers beyond those who happen to be known to the listing agent. Buyer's agents, meanwhile, gain access to the entire inventory of houses.[13] More sellers mean more demand for access by buyers, and more buyers mean more demand for access by sellers.

Most of the information posted on the MLS comes from the agreement signed between the seller and the seller's agent, known in the trade as a listing agreement.[14] Listing agreements identify the property up for sale, the seller's asking price, and the agent's commission. Agents post this information along with the portion of the commission that they are willing to share with the buyer's agent on the MLS.

12 *See* Owen R. Phillips & Henry N. Butler, *The Law and Economics of Residential Real-estate Markets in Texas: Regulation and Antitrust Implications*, 36 Baylor L. Rev. 623, 626-27 (1984).

13 William C. Erxleben, *In Search of Price and Service Competition in Residential Real-estate Brokerage: Breaking the Cartel*, 56 Wash L. Rev. 179, 184 (1981).

14 *Id.* at 181.

II. The Real Estate Industry Reacts to the Internet

MLS services are obviously powerful tools for increasing liquidity in local real estate markets. Not surprisingly, they lie at the heart of the antitrust issues afflicting the real estate industry. Local real estate brokerages usually own MLS services on a cooperative basis, and they have changed the rules of access as the Internet has disrupted traditional ways of doing business. The DOJ's Antitrust Division and the FTC have taken issue with some of these changes.[15]

By connecting virtually everyone at all times, the Internet has posed real challenges for people operating in traditional two-sided businesses. Newspapers, phone companies, and even convention centers have all had to react as their customers have found ways to bypass their services. People who want to make long distance phone calls no longer need to rely on a traditional long distance carrier to make a circuit available. So long as the people on both ends of the call have an Internet connection, a microphone, and an ear piece, they can talk to one another through a virtual circuit supported by the Internet.

The ability of the Internet to disrupt two-sided businesses has not been limited to telecommunications. Prior to the Internet, buyers and sellers of specialized products—like specialized books and collectibles—struggled to find each other. The Internet has enabled them to overcome geographic separation. Instead of relying on classified ads, catalogs, and conventions, they can gather virtually on eBay, creating larger markets than was possible before the Internet.

From the days that the Internet first opened to commercial traffic, the real estate industry has kept a wary eye on it. Theoretically, buyers and sellers of real estate could use the Internet to bypass real estate agents and the listing services in the same way that buyers and sellers of baseball cards now skip Beckett's card guide in favor of eBay. And a number of firms have created websites encouraging them to do precisely that. To this point, however, most buyers and sellers of real estate have not abandoned real estate agents and the traditional customs of the industry (i.e., MLS listings, commissions, open houses) in favor of web sales.

Although the Internet has not supplanted MLSs as the preferred meeting place for buyers and sellers of residential real estate, it has put pressure on some traditional business practices. Historically, the vast majority of residential real estate

15 *See* Press Release, U.S. Department of Justice, Justice Department Sues National Association of Realtors for Limiting Competition Among Real-estate Brokers (Sept. 8, 2005), *available at* http://www.usdoj.gov/atr/public/press_releases/2005/211008.htm; Press Release, U.S. Federal Trade Commission, FTC Charges Real-estate Groups with Anticompetitive Conduct in Limiting Consumers' Choice in Real-estate Services (Oct. 12, 2006), *available at* http://www.ftc.gov/opa/2006/10/realesstatesweep.htm.

sales have taken place under exclusive right-to-sell arrangements.[16] Under an exclusive right-to-sell agreement, a real estate agent collects a commission if the house sells within a set period of time, typically 60 to 90 days, regardless of whether the agent generates the sale. Home sellers have sought to capitalize on the development of the Internet by pushing for exclusive agency agreements.[17] Under an exclusive agency agreement, an agent typically collects an upfront fee but does not collect a commission unless their actions yield a sale.

Real estate agents in some communities have tried to combat this trend by changing the rules of access to MLSs. In the past, although most sales have involved right-to-sell agreements, some MLSs have allowed agents to post listings regardless of the nature of the agency relationship. The move toward exclusive agency agreements prompted a change in policy in at least seven communities around the country. In those communities, MLS boards decided to limit MLS posts to listings secured under exclusive right-to-sell arrangements.[18]

Similar issues have arisen on the buyer side. Relationships on the buyer side tend to be less formal than relationships on the seller side. Buyers generally do not sign contracts with the agents representing them. Nevertheless, agents representing buyers have traditionally required buyers to visit their offices before providing MLS listings. When the Internet opened a new channel of communication, technology-savvy agents responded by making listings available to buyers who visited their websites. Some offered listings to anyone who visited their sites. Others password protected the listings. As the availability of listings on the Internet became more widespread, a few agents began offering commission rebates to prospective buyers who agreed to access listings through their websites.

Again, MLS owners have tried to limit the practice. The rules on the buyer side have been a bit more subtle than those on the seller side. Basically, the National Association of Realtors (NAR)—a national trade association of real estate agents that controls 80 percent of the nation's MLSs—created a special opt-out right for Internet distribution of MLS listings.[19] Historically, all MLS listings have been available to all participating agents. Under the policy adopted by NAR, a real estate agent who posts a listing on the MLS can forbid another agent from distributing that listing on the Internet.

16 *See* Federal Trade Comm'n, Residential Real-estate Brokerage Industry 30 n.17 (Dec. 1983) ("Most MLSs will accept and disseminate information relating *only* to exclusive right-to-sell listings.") (emphasis in original).

17 Press Release, U.S. Federal Trade Commission, FTC Charges Real-estate Groups with Anticompetitive Conduct in Limiting Consumers' Choice in Real-estate Services (Oct. 12, 2006), *available at* http://www.ftc.gov/opa/2006/10/realesstatesweep.htm.

18 *Id.*

19 Press Release, U.S. Department of Justice, Justice Department Sues National Association of Realtors for Limiting Competition Among Real-estate Brokers (Sept. 8, 2005), *available at* http://www.usdoj.gov/atr/public/press_releases/2005/211008.htm.

III. The Reactions Have Led to a New Round of Antitrust Litigation

As noted above, the FTC and the DOJ's Antitrust Division have taken issue with these industry developments. The FTC is pursuing the seller side, and the Antitrust Division has sued NAR. The cases, although they challenge different practices, are carbon copies of one another. They are also of piece with nearly five decades worth of antitrust litigation.

The FTC's administrative complaints against the local real estate boards contain essentially three allegations. The FTC alleges the following:

(1) some real estate agents were posting listings collected under exclusive agency relationships;

(2) local real estate agents acted collectively to stop the practice by changing the rules of access to their MLSs; and

(3) the change in practice will lead to higher prices for the services provided with no apparent offsetting efficiency rationale.[20]

The Antitrust Division's complaint against NAR is longer than the barebones administrative complaints filed by the FTC. But the Division's theory of the case, as reflected in the complaint and its opposition to the defendant's motion to dismiss, is quite similar. The Division advances a three-pronged argument:

(1) new brokerage business models have begun to communicate listings information to their customers through the Internet, rather than traditional ways such as in person or by mail or fax, but

(2) the new NAR rules allow traditional brokers to withhold their MLS listings information from the websites of these new competitors, although no such rule limits traditional brokerage models, and

(3) this undercuts competition from these new brokers, which offer innovative service at lower cost.[21]

These straightforward claims of anticompetitive conduct against the real estate industry are not new. Real estate agents have faced nearly fifty years of litigation over their practices. In 1950, the U.S. Supreme Court declared one real estate board's code of ethics, which provided that brokers should not deviate

20 *See, e.g.*, Complaint, *In re* Realcomp II Ltd., Doc. No. 9320, FTC File No. 061 0088 (issued Oct. 10, 2006); Complaint, *In re* MiRealSource, Inc., Doc. No. 9321, FTC File No. 061 0266 (issued Oct. 10, 2006).

21 Complaint, United States v. Nat'l Ass'n of Realtors, No. 05C-5140 (N.D. Ill. filed Sept. 8, 2005); Memorandum of the United States in Opposition to Defendant's Motion to Dismiss, United States v. Nat'l Ass'n of Realtors, No. 05C-5140 (N.D. Ill. filed Feb. 6, 2006).

from standard commission rates, to be per se illegal price-fixing.[22] Since then, a number of courts have viewed conduct by the real estate industry as price-fixing. Most recently, the U.S. Court of Appeals for the Ninth Circuit found a per se violation against a group of real estate associations that set support fees for a common MLS.[23] In addition to price-fixing theories, another line of complaints has alleged group boycotts by members of the MLSs. These cases have ranged from the mundane to the sinister. Two cases from the U.S. Court of Appeals for the Fifth Circuit are illustrative. In one case, the court found that the MLS's membership requirements, including maintaining a real estate office with regular business hours, were unreasonable.[24] In the other, the court sustained a jury verdict finding a group boycott where a flat-fee broker was subject to punitive commission splits, refusals to show his listings, disparaging remarks, and having his customers harassed by anonymous phone calls.[25]

On the surface, the cases against the real estate industry seem to be very straightforward. Otherwise competing real estate agents set the rules for the jointly owned listing services. The agreement, which is so often the hurdle in a U.S. Sherman Act § 1 case, is a given. The only challenge is demonstrating that the particular practice threatens to increase the price of the services that real estate agents or the MLS offer. This, too, may be relatively easy to establish. In fact, with regard to the restrictions that triggered the current wave of scrutiny, they seem to have been designed with this outcome in mind. Viewed in this way, the industry seems certain to lose.

So far anyway, the real estate industry seems to be dealing with these cases at that level. Neither NAR nor the local MLS boards have made any public effort to defend the practices.[26] In fact, the litigation strategy of NAR seems deliberately designed to change the subject. In moving to dismiss the Antitrust Division's complaint, NAR basically argues that the case is premature because the policy has been suspended pending resolution of the litigation.[27] As the Antitrust Division points out in its opposition, this seems like a strange argument given that NAR suspended the policy only after it was threatened with a lawsuit chal-

22 United States v. Nat'l Ass'n of Real-estate Bds., 339 U.S. 485 (1950).

23 Freeman v. San Diego Ass'n of Realtors, 322 F.3d 1133 (9th Cir. 2003).

24 United States v. Realty Multi-List, Inc., 629 F.2d 1351 (5th Cir. 1980).

25 Park v. El Paso Bd. of Realtors, 764 F.2d 1053 (5th Cir. 1985).

26 *See* Decision and Order, *In re* MiRealSource, Inc., Docket No. 9321, FTC file No. 061 0266 (issued Feb. 5, 2007) (announcing consent order putting an end to the challenged conduct).

27 Memorandum of Law in Support of Defendant's Motion to Dismiss, United States v. Nat'l Ass'n of Realtors, No. 05C-5140 (N.D. Ill. filed Dec. 6, 2005).

lenging it.[28] The current approach, however, is in line with prior defensive tactics. Following the first case striking down as per se illegal agreements to fix commission rates, the industry spent thirty years arguing that real estate was a local business and, thus, not subject to the Sherman Act.

IV. Perhaps There Is Room for a Different View

The striking thing about the real estate industry when viewed through the two-sided market lens is the lack of any competition at the platform level. Two-sided industries are generally characterized by competition among platforms: American Express, Visa, MasterCard, and Discover all battle it out in the payment card industry; Sony, Microsoft, and Nintendo vie for preeminence in the videogame console industry; the New York Times, ESPN, and TV Guide all compete for allegiance among advertisers and subscribers. In the real estate industry, by contrast, only one platform seems to exist—the MLS.

THE STRIKING THING ABOUT THE REAL ESTATE INDUSTRY WHEN VIEWED THROUGH THE TWO-SIDED MARKET LENS IS THE LACK OF ANY COMPETITION AT THE PLATFORM LEVEL.

The U.S. antitrust agencies do not quibble with the absence of platform competition. In their complaints, they concede the efficiency of the MLS system. Indeed, the agencies have premised their attacks on how irreplaceable the MLS system is. The agencies argue, essentially, that the MLS system is so efficient that rational buyers and sellers of real estate will not attempt to circumvent it. Yet, the cases do not push for the creation of an alternative platform. They seek to reduce the price that buyers and sellers pay for access to the existing system. The theory seems to be that competition among real estate agents is a substitute for competition among platforms.

However, the residential real estate industry exhibits a couple of features not ordinarily associated with a lack of competition. Concentration is low, and entry is easy. The industry attracted so many new agents that mean compensation for agents declined even in a period of skyrocketing home prices. The problem, if one exists, lies not with the amount of competition among agents but rather with the nature of that competition. Historically, agents have not competed for listings or buyers by offering to reduce their commissions. Instead, agents have competed on the basis of what they describe as service and what others, more pejoratively, criticize as glad-handing.

The U.S. antitrust agencies have taken a rather formalistic approach to changing the nature of competition among real estate agents. In bringing these cases, the agencies use to their advantage the fact that they are attacking a collaborative enterprise. Although the gap in treatment has narrowed, joint ventures and

28 Memorandum of the United States in Opposition to Defendant's Motion to Dismiss, United States v. Nat'l Ass'n of Realtors, No. 05C-5140 (N.D. Ill. filed Feb. 6, 2006).

other legitimate collaborations of competitors remain subject to different rules than their more traditionally organized competitors.[29] Consequently, the agencies attack on horizontal conspiracy grounds practices that, but for the collaborative ownership, look like garden-variety vertical restraints.

If MLSs were independently owned, it would be more difficult to argue that the restraints at issue pose a significant antitrust problem. The ban on wholesale distribution of listings via the Internet makes considerable sense from the standpoint of the upstream owner of such listings. Internet distribution of listings makes it difficult for the downstream agents to differentiate themselves from one another, reducing the incentive that agents have to collect listings. The other two restraints, the ban on exclusive agency contracts and the ban on discounted buyer-side commissions, simply combat discounting among distributors of the service created by the MLS. Although *Dr. Miles* remains good law,[30] Ben Klein and others have shown that vertical price restraints are both rational from the standpoint of the upstream party and welfare-enhancing.[31]

The U.S. antitrust agencies also seem to ignore the two-sided nature of the industry. The agencies accept as true the criticism that real estate agents earn lots of money for doing very little work. To be sure, at the level of particular transactions, this criticism seems valid. Real estate agents do seem to collect far more money on the sale of particular homes than the work put into that sale warranted. On this view, real estate agents just increase the transaction costs associated with the transfer of real estate, and consumers should benefit from the effort to reduce those costs.

> If their search for clients actually leads people who would not otherwise buy or sell homes to enter the active market, then the structure of the industry might actually benefit consumers.

There is, however, another way to look at the role of the real estate agent. For an industry marked by few repeat players, the residential real estate industry in the United States seems remarkably liquid. The question becomes whether real estate agents have anything to do with the apparent liquidity in the residential real estate market. Ironically, one of the more well-worn criticisms of the industry suggest that they do: the observation, by Levitt and others, that real estate agents spend nearly all of their time looking for clients and relatively little actually working on par-

29 *See* Texaco v. Dagher, 126 S. Ct. 1276, 1280 (2006) ("As a single entity, a joint venture, like any other firm, must have the discretion to determine the prices of the products that it sells . . .").

30 With the U.S. Supreme Court having granted certiorari in Leegin Creative Leather Products, Inc. v. PSKS, Inc., No. 06-480 (Dec. 7, 2006), *Dr. Miles* may be headed for the chopping block.

31 *See* Benjamin Klein, *Exclusive Dealing as Competition for Distribution "On the Merits,"* 12 Geo. Mason L. Rev. 119 (2003); Benjamin Klein, *Competitive Price Discrimination as an Antitrust Justification for Intellectual Property Refusals to Deal*, 70 Antitrust L.J. 599 (2003).

ticular transactions. If this is right and if their search for clients actually leads people who would not otherwise buy or sell homes to enter the active market, then the structure of the industry might actually benefit consumers.

This view, if it were adopted, would create a real challenge for the U.S. antitrust agencies. The impact is particularly easy to see with the alleged efforts to restrict the ability of buyer side agents to rebate their commissions to clients. Viewed solely from the standpoint of buyers, this restraint seems unambiguously bad. It increases the cost of buying a house. Sellers, however, may take a different view. Sellers want to maximize the pool of potential buyers. They could reasonably conclude that offering larger commissions to agents will do more to increase the pool of potential buyers than providing a small discount. On this view, the practice seems eminently reasonable.

V. Conclusion

As we noted at the outset, we have not made up our minds about this industry or the cases that the U.S. antitrust agencies have filed. Our observations may fall short of a ringing defense. The claim that the current structure of the industry may benefit consumers contains a heroic assumption, namely that marketing efforts by real estate agents expand the set of willing buyers and sellers. We have simply teed up the empirical question that should be at the heart of the current round of cases—whether consumers would be better off if the agencies were to prevail. We are not, however, optimistic that the cases will answer this question. The agencies appear to have assumed that the answer to this question is yes, and the industry, at least thus far, has ignored it altogether. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Are Media Markets Analyzed as Two-Sided Markets?

John Wotton

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1.
Competition Policy International is a free publication. To order or download additional copies, visit eSapience.org.

Are Media Markets Analyzed as Two-Sided Markets?

John Wotton

This paper sets out to describe how, in practice, the U.K. competition authorities analyze competition in those markets in the media sector, which may have the characteristics of two-sided markets. The typical characteristics of media markets are described and several decisions in the media sector, taken over a period of a number of years, are analyzed. Conclusions are then drawn on the extent to which the two-sided characteristics of the relevant markets have been taken into account.

The cases that are analyzed have been chosen because they are considered to be of some importance in the context of the subject under consideration and do not represent a comprehensive list of media cases decided by the U.K. competition authorities over any particular period. The author advised principal parties in a number of these cases. The information contained in this paper is, however, taken from public sources and the views expressed are those of the author alone.

I. The Typical Characteristics of Media Markets

A very straightforward terminology is used in this paper to identify the various actors concerned in media markets. The term "owner" is used to mean the owner or operator of the media asset under consideration, for example a print title, radio station or television channel. The term "consumer" is the reader or viewer who uses the medium as a source of information or entertainment. The term "advertiser" is the individual or business that pays the owner for the inclusion of advertising or promotional material in the medium. The focus is on the economic relationships between advertisers and owners and between owners and consumers. Relationships between owners in different capacities, for example content

The author is a Partner in Allen & Overy LLP's London office.

providers and platform operators, as important as they are in competition terms, are not addressed in this paper.

A number of the typical characteristics of media markets are mentioned below. The reason for drawing attention to these features of media markets is that they are among the factors that work by other authors suggests may be of relevance to the economic analysis of two-sided markets.[1]

A. OWNERSHIP

Media assets are most frequently under single ownership, with occasional joint ventures and minority investments. As a result of this, the competition authorities' analysis of media markets has differed from that of the multi-bank owned payment systems, which are considered in detail elsewhere in this issue. In particular, the exhaustive analysis of payment systems under Article 81 of the EC Treaty and Chapter I of the U.K. Competition Act 1998 finds no close parallel in the media sector. Recent U.K. media cases under Chapter I[2] are not very relevant to a discussion concerned with two-sided markets and issues raised by the joint selling of broadcasting rights[3] are not the subject matter of this paper. The cases discussed are a mixture of merger reviews, market investigations, and behavioral inquiries.

B. FUNDING

The owner's funding model may vary from 100 percent advertiser funding to 100 percent consumer funding. The broadcasting media have tended to polarize towards predominantly advertiser or consumer funding, whereas the print media display a wide range of funding mixes.

C. MULTI-HOMING

Multi-homing by advertisers and consumers is prevalent in the media sector. Advertisers may use a variety of different media for an advertising campaign in order to achieve the required impact and over time may switch their expenditure significantly from one set of media to another. The extent to which advertisers regard different media as substitutes for one another has been the subject of considerable analysis by the U.K. Office of Fair Trading (OFT) and the U.K. Competition Commission in both merger and behavioral cases. Consumers also

1 *See, e.g.,* J.-C. Rochet & J. Tirole, *Platform Competition in Two-Sided Markets*, 1 J. Eur. Econ. Ass'n 990 (2003) and M. Armstrong, Competition in Two-Sided Markets (2005) (mimeo, University College London) (on file with author).

2 *See, e.g.,* OFT Decision of May 24, 2005, No. CA98/03/2005, TV Eye Limited.

3 The joint selling of U.K. football rights was recently investigated by the European Commission under Article 81 of the EC Treaty. *See* Commission Decision of Mar. 22, 2006, Case COMP/C-2/38.173, Joint selling of the media rights to the FA Premier League.

use a wide variety of different media, in which they may be exposed to advertising, but the extent to which consumers regard the media they use as complements, or as substitutes, has been less fully investigated in published decisions of the U.K. authorities.

D. THE RELATIONSHIP BETWEEN USE AND VALUE

The relationship between consumer use of a media property and the value to the advertiser of such use, in terms of impacts made or sales leads generated is an intrinsically complex one and may be hard to measure accurately. Viewing and readership data is prone to errors and omissions in consumers' reporting of their media use. Consumers may not always give accurate information to advertisers who seek to monitor the media source which gives rise to each sales lead, although advertisers endeavor to devise systems to overcome this. By contrast with the cardholder who uses a payment system solely for the purpose of making transactions with merchants, a consumer may not use a media product mainly (or at all) in order to gain information about, still less to make transactions with, advertisers. Whereas a product containing only directional, classified advertising may generally only be used by consumers to find potential suppliers of goods or services, the consumer of a product containing a mixture of advertising and editorial matter may have no prior interest in and may neither see nor absorb the advertising it contains.

E. CONSUMERS DO NOT PAY FOR TRANSACTIONS WITH ADVERTISERS

The consumer may or may not pay the owner for the use of the medium, but any such payments (e.g., subscription or cover price) will not be related to the consumer's transactions (if any) with advertisers in the medium. Advertisers generally do not pay the owner directly for sales or sales leads generated by their advertising, although the price agreed for an advertisement may be dependent on the circulation, readership, or audience of the publication or program in which it is inserted. The owner does not regulate, or generally become concerned with, any transactions between advertisers and consumers that are generated by advertising.

F. PRICE DISCRIMINATION

In principle the owner can discriminate between advertisers when setting prices for advertising. The owner may not always, however, have the information necessary to do so in a profit-maximizing way. Discrimination in pricing to consumers is also possible, within limits of practicality, although most consumer sales are no doubt at published cover or subscription prices.

G. NETWORK EFFECTS FAVOR INCUMBENTS

It is generally observed that it is not easy for a new entrant to switch established consumer use of an incumbent media property to the new entrant's own publica-

tion. Nevertheless, new entrants may over time gain a substantial share of consumer use from a long-established incumbent (as is illustrated, for example, by the shift in television viewing which has taken place in the United Kingdom from ITV to more recent entrants) and product innovation (for example web-based advertising) can create more rapid shifts in use and, thus, in revenue.

II. Some Decisions by U.K. Authorities in Media Cases

In this section a number of OFT and Competition Commission decisions and opinions concerning media markets are discussed, noting whether or not the two-sided nature of the market concerned has been considered and whether the authority's treatment of apparently comparable cases has been consistent in this respect. It is notable that, at the date of the University College London's "Competition Policy for Two-Sided Markets Colloquium," a search of the Competition Commission's website for the phrase two-sided markets produced a reference to only two media cases, namely *Archant/Independent News and Media*[4] and *Classified Directory Services*[5]: A search for the same phrase on the websites of the OFT and the U.K. Office of Communications produced no references to media cases.

A. NEWSPAPERS, MAGAZINES, AND JOURNALS

The U.K. competition authorities have had occasion to consider competition in markets for newspapers, magazines, and journals in a number of different contexts over recent years, as merger activity has proliferated and complaints of ineffective competition and anticompetitive behavior have abounded. An early acknowledgement of the two-sided nature of a publishing market is to be found in 2001 in *Reed Elsevier/Harcourt*.[6] The Competition Commission investigated a proposed merger between two major publishers of scientific, technical, and medical (STM) journals and concluded (by a majority of the group of members conducting the inquiry) that the merger would not operate against the public interest, notwithstanding that it raised concerns about access and pricing. The Commission noted that the market for STM journals is largely circular, with the same members of the academic community writing the articles, peer-reviewing

4 U.K. Competition Commission, Archant Limited and the London newspapers of Independent News and Media Limited: A report on the acquisition by Archant Limited of the London newspapers of Independent News and Media Limited (2004).

5 Press Release, U.K. Competition Commission, OFT Refers Classified Directory Advertising Services to Competition Commission (Apr. 5, 2005).

6 U.K. Competition Commission, Reed Elsevier plc and Harcourt General Inc: A report on the proposed merger, Cm 5186 (2001).

them and (through their institutions' libraries) purchasing the journals.[7] In considering barriers to entry, the Commission observed that it is difficult for a journal to become established and secure a strong reputation. Researchers greatly prefer to publish in established journals, where their article will be peer-reviewed and edited by leading figures in the discipline. Publication in a leading journal confers status on the author, ensures wide readership and thus the prospect of wide citation, and is also influential in funding allocation. All this creates an environment in which leading journals in a field enjoy a prestige that it is difficult for others to challenge.[8] This effect is more succinctly described by the OFT, in a report[9] of an investigation of the STM market, undertaken as a result of the Competition Commission's inquiry, as a "virtuous circle".[10] Such a careful review of the two or more demands for each print media product has not, however, featured consistently in subsequent decisions of the U.K. authorities.

A CAREFUL REVIEW OF THE TWO OR MORE DEMANDS FOR EACH PRINT MEDIA PRODUCT HAS NOT FEATURED CONSISTENTLY IN SUBSEQUENT DECISIONS OF THE U.K. AUTHORITIES.

An example of a case in which only one side of the market was considered is *Aberdeen Journals*.[11] The case concerned a complaint of predatory pricing against Aberdeen Journals, the incumbent in a local newspaper market which was found, by both the OFT and U.K. Competition Appeal Tribunal (CAT), to include both paid-for and free titles. The OFT and CAT both considered in great detail the attributable variable costs and revenues of Aberdeen Journals' *Herald & Post* title, in respect of which the predatory pricing allegation was made. This free title had cut its advertising rates and improved its quality in response to new market entry by a free title. All the *Herald & Post*'s costs were treated as attributable to advertising when considering the test of predation, even though some costs related to editorial material. Neither the OFT nor the CAT considered in their analysis the question of competition for consumers, notwithstanding that the single revenue streams of the two free titles most closely concerned could be sustained only through evidence of their readership and that they competed in the same market with at least one paid-for title.

Although it might be argued that the nature of the issue in *Aberdeen Journals*, predatory pricing, justified the OFT and CAT in confining their attention to

7 *Id.* at para. 2.63.

8 *Id.* at paras. 2.43-45.

9 U.K. OFFICE OF FAIR TRADING, THE MARKET FOR SCIENTIFIC, TECHNICAL AND MEDICAL JOURNALS, OFT 396 (2002).

10 *Id.* at para. 6.6.

11 OFT Decision of Sept. 16, 2002, No. CA98/14/2002, Predation by Aberdeen Journals Limited (*aff'd* CAT Judgment of Jun. 23, 2003, No. 1009/1/1/02, Aberdeen Journals v. Director General of Fair Trading).

costs and advertising revenues, the same point cannot be made in respect of a merger review. It is therefore striking that in *Archant/Independent News and Media* neither the OFT nor the Competition Commission considered explicitly the question of competition for consumers. This case concerned a completed merger between two local newspaper publishers in the London area and was the first newspaper merger to be considered by the U.K. authorities under the substantial lessening of competition (SLC) test.[12] The Commission identified two overlapping areas in which potential competition concerns arose, due to the combined share of circulation (aggregating both paid-for and free titles) that was held by the merging parties. The Commission applied the SLC test in these two areas by reference to advertising alone and decided to clear the merger, despite the parties' high local market shares, and notwithstanding the Commission's assessment of incumbency advantages and barriers to entry in these markets. They did so for a number of reasons, including: residual competition from both local newspapers and certain other local print media; survey evidence of advertiser behavior; lack of concern about the merger on the part of advertisers; and the inability of the merged group to practice systematic advertising price discrimination. The Commission did not expressly consider the effect of the merger on consumers and it must be inferred that they presumed that there was no likelihood of harm in this case.[13] The approach taken by the OFT in their merger reference decision[14] did not expressly limit the SLC concerns to advertising markets, although the potential adverse effects identified all related to advertising.

The Competition Commission's conclusions in *Archant/Independent News and Media* followed a long series of investigations of newspaper mergers in which competition between local newspapers had been analyzed in detail, but essentially in terms of competition for advertising and effects on advertisers.[15] Harm to consumers as a result of reduced competition was not generally identified in these investigations. This may go some way to explain the Commission's appar-

12 Newspaper mergers in the United Kingdom had previously been considered by reference to a public interest test under which not only competition, but also matters such as the accurate presentation of news and free expression of opinion were taken into account.

13 It is instructive to compare the Competition Commission's report in *Archant/Independent News and Media* to *Newsquest plc and Independent News and Media* (*see supra* note 4 and U.K. Competition Commission, Newsquest (London) Limited and Independent News & Media PLC: A report on the proposed transfers, Cm 5951 (2003)). This inquiry, concerning an alternative merger proposal for the same target and published only a few months earlier, was conducted under the public interest test, not the SLC test. The Commission necessarily considered effects on the accuracy of news and free expression of opinion, finding no harm in either case. In considering competition, the Commission considered only the effect on advertisers, not consumers, and cleared the merger subject to certain divestments.

14 OFT Decision of Apr. 29, 2004, Completed acquisition by Archant Ltd of the London Regionals Decision of Independent News & Media.

15 *See, e.g.*, U.K. Competition Commission, Newsquest (London) Limited and Independent News & Media PLC: A report on the proposed transfers, Cm 5951 (2003) and U.K. Competition Commission, Johnston Press plc and Trinity Mirror plc: A report on the proposed merger, Cm 5495 (2002).

ent presumption that the only issues in the case concerned competition for advertisers. It appears, however, that in considering competition cases involving national newspapers, the OFT and Commission have taken greater cognizance of the dual demand from consumers and advertisers and the interaction between the dual revenue streams on which the publisher's business depends. There is recognition of this at the descriptive, if not the analytical level in *National Newspapers*[16], although that inquiry chiefly concerned distribution and therefore paid little regard to advertising. It would appear also to have been recognized by the OFT in its consideration of a series of complaints concerning the cover pricing and subscription pricing of certain newspapers over a period of several years, starting in 1994. Analysis of the OFT's approach in these cases is hampered by the lack of any reasoned decision published by the OFT. However, from the slight information which has been published,[17] it may be inferred that the OFT has recognized the publisher's dual revenue streams from advertising and cover/subscription price and has considered the relationships between circulation and cover price, between circulation/readership and advertising revenue, and between cover price and multi-homing (in the sense of multiple purchases) by consumers.

More recently the OFT has published a draft advisory opinion on national newspaper and magazine distribution,[18] in order to provide guidance to the industry on the assessment of whether current exclusive distribution agreements between publishers or distributors and wholesalers, which typically confer absolute territorial protection on the wholesaler, infringe Chapter I of the Competition Act 1998. A detailed description of the OFT's reasoning would go beyond the scope of this paper and it is sufficient to observe that the OFT state that newspapers and magazines operate in two-sided markets in which each title competes to attract readers, on the one hand, and advertisers, on the other; and publishers take account of the interaction between these two customer groups when determining their pricing strategy. The example instanced is that, when determining the retailer's margin, the publisher will take into account the impact of additional retail sales on its advertising income.[19]

It would therefore seem that the principle of two-sided market analysis is now established for this sector, so far as the OFT is concerned. This is to some extent confirmed by the OFT's reference decision on a proposed merger between two

16 U.K. Competition Commission, The supply of national newspapers: A report on the supply of national newspapers in England and Wales, Cm 2422 (1993). *See, in particular,* ch. 3.

17 *See, e.g.*, Press Release, U.K. Office of Fair Trading, Newspaper Pricing: News International gives assurances (May 21, 1999).

18 U.K. Office of Fair Trading, Newspaper and Magazine Distribution; Public consultation on the draft opinion of the Office of Fair Trading, OFT 851 (2006).

19 *Id.* at para. 1.36.

consumer magazine publishers, *Future/Highbury House I*[20], which would have given the merged group a very high share of one type of special interest consumer magazine, namely computer games magazines. The OFT considered separately the relevant advertising and readership markets, concluding that there would be no SLC in advertising, due to the market power of media buyers and the inability of the owner to price discriminate systematically against captive advertisers, but that there would be an SLC in the readership market, as the owner would be able to raise prices or reduce quality.[21] Although the OFT noted the existence of incumbency advantages and barriers to entry, in this as in other media markets, it made no specific connection between the two sides of the market and in this respect cannot be said to have recognized fully the principles of two-sided market analysis. The OFT did not, for example, consider the effect which raising cover price or reducing quality would have on advertising revenue and, therefore, whether such conduct would overall be profit-enhancing to the owner.

B. TELEVISION AND RADIO

The pattern that may be discerned from the above summary of print media cases, namely a recognition of the two-sided nature of media markets in some cases, but not others, is also to be found in broadcast media cases. It is instructive to contrast in this respect *Carlton Communications/Granada*[22] with *Capital Radio/GWR Group*[23], that both concern mergers in free-to-air media. *Carlton Communications/Granada* represented a merger between the two largest free-to-air television broadcasters in the United Kingdom (representing between them almost the whole of the ITV channel) and was cleared by the Secretary of State, on the recommendation of the Competition Commission, subject to complex behavioral remedies concerning advertising sales.[24] The Commission recognized the two-sided nature of the market, in terms of the need for the owner to attract large numbers of consumers in order to sell airtime to advertisers, the competition between broadcasters for audience, and the need to maximize the attractiveness of the audience to advertisers. The Commission identified incumbency advantages and network advantages on the part of ITV and considered programming benefits to be delivered by the merger, finding that they did not outweigh the

20 OFT Decision of Apr. 14, 2005, Anticipated acquisition by Future plc of Highbury House plc.

21 The OFT also identified a third aspect of the market, namely the demand from computer games manufacturers for owners to publish official magazines, under license.

22 U.K. Competition Commission, Carlton Communications Plc / Granada Plc: A report on the proposed merger, Cm 5952 (2003).

23 OFT Decision of Dec. 22, 2004, Anticipated acquisition by Capital Radio Plc of GWR Group plc.

24 This case was dealt with under the general public interest test, in practice confined largely to competition matters, that applied to all non-newspaper mergers until it was replaced by the SLC test on Jun. 20, 2003.

competitive detriments to advertisers. The Commission did not, however, use two-sided market terminology in its analysis and appears to have considered separately the various relevant markets that it identified.[25]

A year after *Carlton Communications/Granada*, the OFT reviewed *Capital Radio/GWR Group*, a proposed merger to create the largest commercial radio broadcaster in the United Kingdom, whose analogue and digital stations were all free-to-air and (with one exception) local. The OFT cleared the merger under the SLC test on the basis of: a lack of significant local overlap (except in one area in which an appropriate divestment remedy was offered); the ability of advertisers to "buy-around" the merged group by choosing alternative stations; the lack of ability or incentive on the part of the merged group to bundle or tie its stations; and a recognition of the buyer power of the major media buying agencies. The OFT applied the SLC test in relation to competition for advertising, but not for audience, and gave no explicit consideration to the interdependence of advertising revenue and audience, or the effect of the merger on consumers.[26] Just as with *Archant/Independent News and Media*, therefore, two-sided market considerations were wholly absent from this merger review.[27]

THE OFT APPLIED THE SLC TEST IN RELATION TO COMPETITION FOR ADVERTISING, BUT NOT FOR AUDIENCE, AND GAVE NO EXPLICIT CONSIDERATION TO THE INTERDEPENDENCE OF ADVERTISING REVENUE AND AUDIENCE; THEREFORE, TWO-SIDED MARKET CONSIDERATIONS WERE WHOLLY ABSENT FROM THIS MERGER REVIEW.

C. CLASSIFIED DIRECTORIES

Since April 2005 the Competition Commission has been conducting a market investigation into the supply of classified directory advertising services (CDAS), following an earlier inquiry completed in 1996.[28] The Commission's provisional findings[29] were published in June 2006 and expressly adopt two-sided market ter-

25 These were, in addition to advertising, programming, and bidding for broadcasting licenses.

26 In addition to the OFT's review, a very limited review of public interest issues was undertaken by the Department for Trade & Industry, under media-specific powers. The sector regulator, the U.K. Office of Communications, conducted a station-by-station review of the effect of the merger on the group's broadcasting services. Neither of these reviews concerned competition.

27 The same may be said of *Scottish Radio Holdings / GWR Group / Galaxy Radio*. *See* U.K. COMPETITION COMMISSION, SCOTTISH RADIO HOLDINGS PLC AND GWR GROUP PLC AND GALAXY RADIO WALES AND THE WEST LIMITED: A REPORT ON THE MERGER SITUATION, Cm 5811 (2003).

28 U.K. COMPETITION COMMISSION, CLASSIFIED DIRECTORY ADVERTISING SERVICES, Cm 3171 (1996).

29 U.K. COMPETITION COMMISSION, CLASSIFIED DIRECTORY ADVERTISING SERVICES: PROVISIONAL FINDINGS REPORT, ISBN 0117025119 (2006) [hereinafter Provisional Findings] and U.K. COMPETITION COMMISSION, CLASSIFIED DIRECTORY ADVERTISING SERVICES: FINAL REPORT, ISBN: 0117037373 (2006) [hereinafter Final Report].

minology in analyzing competition in the market and barriers to entry. The Commission found that CDAS providers operate in a two-sided market, in which success depends on their ability to attract both users and advertisers. The interdependence of advertiser and user demand gives rise to a network effect or virtuous circle, as a directory with high usage and advertising is more attractive to new advertisers and users.[30] The Commission noted that this network effect appears to give Yell (the largest and longest-established CDAS operator) a significant advantage over smaller providers and any new entrant, making it difficult for a new provider to compete[31] and provisionally concluded that Yell has market power. In these findings the Commission adopted two-sided market analysis more fully than in previous media cases which they have investigated.

III. Provisional Conclusions

The survey of cases in the previous section is necessarily superficial, but allows certain provisional conclusions to be drawn. First, it is only very recently that the OFT and Competition Commission have expressly adopted two-sided market analysis in their published decisions. It is noteworthy that the Commission has done this for the first time in relation to classified directories that contain only directional classified advertising and in which, therefore, use of the directory and value to advertisers is very closely linked. It might be expected that two-sided market effects are particularly relevant to this example.

Second, it may be said that the potential use of two-sided market analysis in media cases has now been established by both the OFT (in national newspaper and magazine distribution) and the Commission (in classified directories), although neither of these cases has reached its final conclusion, at the time of writing.

Third, setting aside these current cases, the U.K. authorities' analysis has tended to focus separately on the one or more sides of the relevant market for which advertisers or consumers make payments to the owner, without looking at competition issues in the context of the media product or platform as a whole.

Fourth, in certain recent cases, including *Aberdeen Journals*, *Archant/Independent News and Media*, and *Capital Radio/GWR Group*, the authority's competitive analysis has been confined to the effect on advertisers, even though, in the first two of these cases, the relevant markets included paid-for as well as free media. In other, apparently comparable and no less recent cases (e.g., *Future/Highbury House* and *Carlton Communications/Granada*), however, effects on both advertisers and consumers have been considered.

30 Provisional Findings, *id.* at para. 18; Final Report, *id.* at paras. 5.9, 6.2, and 6.122.

31 Provisional Findings, *id.* at paras. 36 and 44; Final Report, *id.* at paras. 6.112 and 6.123.

Fifth, though very tentatively, in behavioral cases where the owner enjoys two revenue streams from the media property, the U.K. authorities appear to have accepted that the competitive effect of a pricing decision is properly to be judged by taking into account the costs and revenues of the media property as a whole, without separating editorial from advertising costs, or consumer revenue from advertising revenue.

Finally, it seems clear that two-sided market analysis needs to be applied more systematically and consistently by the U.K. authorities, than has been the case until now, in both behavioral and merger cases involving the media.

It would go beyond the scope of this short paper to attempt to analyze whether the use of such analysis in any particular cases would have produced a different result, and more detailed work would need to be undertaken in order to draw any such conclusion. The author therefore does not suggest that any case mentioned in this paper has been wrongly decided, for want of two-sided market analysis, merely that the absence of consideration of one side of the market, or of the interdependence of two sides of the market, may have led to competitive factors relevant to the authorities' analysis having been left out of consideration. It would also be instructive to undertake a similar analysis of recent decisions of the European Commission on media cases. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Two-Sided Telecom Markets and the Unintended Consequences of Business Strategy

Leonard Waverman

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Two-Sided Telecom Markets and the Unintended Consequences of Business Strategy

Leonard Waverman

A two-sided market is one where two different parties are connected to each other through a third-party platform. Examples are many: nightclubs and dating clubs are platforms that bring together people wishing to meet other people; newspapers are platforms providing advertising and content to readers. In this brief paper, I examine the two-sided nature of telecommunications. It is clear that a traditional telecom is a platform allowing a calling party (C) to connect to a receiving party (R). However, it is, in a sense, too easy to label economic activity as two-sided. Without clear limits, most activities appear to be of a two-sided nature. Therefore, I begin by examining whether telecoms does meet the conditions of two-sidedness as defined by Tirole and Rochet in their 2007 paper.[1]

I then turn to examining briefly the history of pricing in fixed-line and mobile telecoms. The pricing structure we see today in many markets is a result of historical business models. In most countries, the calling party pays all the costs of the call, while caller and called pay for access to the network. I show how the pricing structures first developed in fixed-line telecoms had unintended consequences on subsequent developments in new mobile telephony. Since pricing structures and not just the level of prices are important in two-sided markets, these unintended consequences need to be recognized, and dealt with, if possible. I then turn to the brave new world—telecom operators providing content and being the platform for IP services and applications.

1 J.-C. Rochet & J. Tirole, *Two-Sided Markets: A Progress Report*, RAND J. Econ. (Autumn 2006).

The author is Professor and Chair of Economics, London Business School and Director, LECG.

There is a danger that the original pricing model developed when telecoms was a circuit switched voice call will be carried over to the new IP world. When the platform connects multiple parties and provides more than a conversation between the caller and the receiver, the pricing model need not be that the calling party pays. In the world of free over-the-air (FOTA) broadcasting, advertisers pay for content and for the costs of building and running the platform. In the IP-based telecom world, cost contributions could come from content providers, advertisers, and users, as well as service and application providers. The burgeoning literature on two-sided markets indicates that simple cost allocation rules no longer need to dictate. That is, because of the existence of positive externalities in a market on another side of the platform, prices can be below attributable costs.

I. Is Telecoms Two-Sided?

For a market to be two-sided, Tirole and Rochet cite two conditions that hold; consider telecommunications where C is the calling party, and *R* the receiving party to a call:

(a) The structure of prices matters:

Consider usage prices (z_1C, z_2R).

- Definition: market is two-sided if volume *V* depends on the structure and not only on the level of aggregate price z; $z = z_1C + z_2R$, Otherwise, the market is one-sided.

(b) For a market to be two-sided, the Coase theorem must not apply

- Definition: Coase theorem: If C and *R* bargain efficiently, then they (1) maximize the size of the pie (which depends only on $z_1C + z_2R$) and (2) share it.

Consider a voice call between two people C and *R*. Condition (b) above clearly holds, with millions of possible connections, the caller and called parties cannot negotiate each time a call is attempted. Is it obvious that condition (a) holds: that the structure of the division of the price of the call—z—will affect the volume of calls? Both parties usually benefit from the voice call. C should always benefit—otherwise why originate the call? *R* will usually benefit but not if the call is an unwanted sales call, spam, etc. Let us assume that *R* always benefits; then having C bear all the costs of the call is sub-optimal as C is subsidizing *R*'s benefit. The sub optimality would be C undertaking too few calls. Similarly, having *R* bear all the costs is sub-optimal—and *R* will want to receive fewer calls than if C contributed to the costs. Normally, if society wanted to force one party to bear all the costs/price—z—we consider it superior that C pays all incremental costs since as the initiating party, C knows the purpose of the call. However, having C bear all costs is inefficient, and there are no measures that I know of to quantify the magnitude of the social loss imposed by this pricing scheme.

Empirical evidence would require lab experiments or natural experiments where the same people faced price—z—but under different sharing rules. I know of no such data. The data used to suggest that telecoms is a two-sided market is the very different levels of cell phone ownership and penetration in the United States and in Europe. In the United States, both *R* and C pay part of the call costs while in Europe only C pays. Hence it is conjectured that receiving parties kept their phones turned off in the United States, diminishing the externality value of cell phones, hence limiting adoption. Thus in the early days of mobile calling, a far lower percentage of the population had mobile phones in the United States than in Europe. This is shown in Table 1 where mobile subscribers per one hundred inhabitants are given for the United States, Canada, the European Community (EC) 15 and the EC 25. Until 1999, a greater percentage of people subscribed to cell phones in the United States than in Europe (and until 1997 in Canada). But beginning in 1999, far more people have mobile phone subscriptions in Europe than in North America.

Table 1

Mobile phone subscribers per 100 inhabitants

Year	USA	Canada	EU15	EU25
1995	12.69	8.81	5.77	5.42
1996	16.35	11.77	9.00	8.57
1997	20.29	13.99	14.09	13.68
1998	25.09	17.68	23.85	23.35
1999	30.84	22.66	40.69	40.15
2000	38.90	28.35	63.24	59.09
2001	45.03	34.20	74.02	69.95
2002	48.88	37.73	79.20	76.08
2003	54.58	41.65	84.82	81.83
2004	62.11	46.72	92.12	89.86

Source: ITU World Telecommunications Indicators, 2006.

These data however cannot be used to support the conjecture that the sharing of costs of calling in the United States and Canada lowered the desirability of owning a mobile phone, as many circumstances differed between the United States and Europe. In the United States, mobile numbers were similar to landline phone numbers—an area code and 7 digits. In Europe, mobile phones were given a distinct national numbering plan with 8 digits, unrelated to the city or area.

Thus in the United States, it is not obvious that the phone number that one is calling is a mobile number while in Europe it is obvious. Hence, making C pay all costs in the United States was thought to be unfair, since only after the call was made and the bill received would C know that he/she called a mobile phone. In Europe the caller knows it is a mobile phone that is being called. Other important distinctions exist as well. In most jurisdictions in the United States, a local call is free (I discuss this below), hence if mobile was to compete with free local calling, then C could not be asked to pay all of the z costs. In addition, in the United States there are a number of competing technologies available to mobile subscribers—analogue (AMPS), two kinds of Time Division, GSM (the European standard), and CDMA. Few papers examine this technological difference between the United States and Europe and its impact on diffusion and calling.[2]

II. Pricing in Telecoms

Does it matter if we ignore the two-sidedness of voice calls? In fixed-line calling, the charging model has always been that the calling party pays all costs (i.e., z). I ignore free local calls here. When the call was national or international long distance, the calling party paid. In some cases the receiving party countries levied huge taxes on incoming international calls. These taxes caused the U.S. regulator, the Federal Communications Commission, to unilaterally limit the termination fee charged by outside countries to U.S. callers. Clearly if the receiving party paid for termination, then taxing callers by raising termination fees is not possible.

There are other examples of pricing systems that shed light on the two-sided nature of telephone calls. In much of the United States and Canada local calls are free (i.e., the price of a local call is zero). This pricing system dates to the beginning of last century when the Bell system was engaged in fighting for dominance of telephony against independent competitors. The Bell system's strategic advantage was its ownership of long distance lines and by refusing to interconnect with independent telecoms and by pricing local calls at zero while charging (tolling) for long distance calls, it was able to achieve dominance.[3] Even when the Bell system became a regulated monopoly, the practice of free local calls (i.e., bundled with the access subscription) was maintained. This, however, impacted mobile networks. Because of the charging model for fixed lines, using a mobile for a local call was costly compared to free fixed-line calls. And when mobile receiving parties share part of the costs of call, mobile subscription lagged in the United States and Canada (as seen in Table 1). To overcome this lag in adoption, AT&T Wireless introduced bundles—a monthly

2 N. Gandal et al., *Standardization versus Coverage in Wireless Telephone Networks*, CoRR: Computers & Soc'y (2001).

3 If we consider local and long distance calls as two sides of a market, then two-sided pricing could have been used for foreclosure purposes.

fixed-fee option to pay for access, as well as for all calls incoming and outgoing, local and national. This bundle effectively priced incoming terminating and outgoing local calls (as well as outgoing national calls) at zero within the bundle, effectively matching the zero price for fixed-line outgoing local calls, and for all incoming fixed-line calls. Other mobile operators quickly matched AT&T Wireless. As a consequence, revenues per minute in mobile systems are now lower in the United States than in Europe.

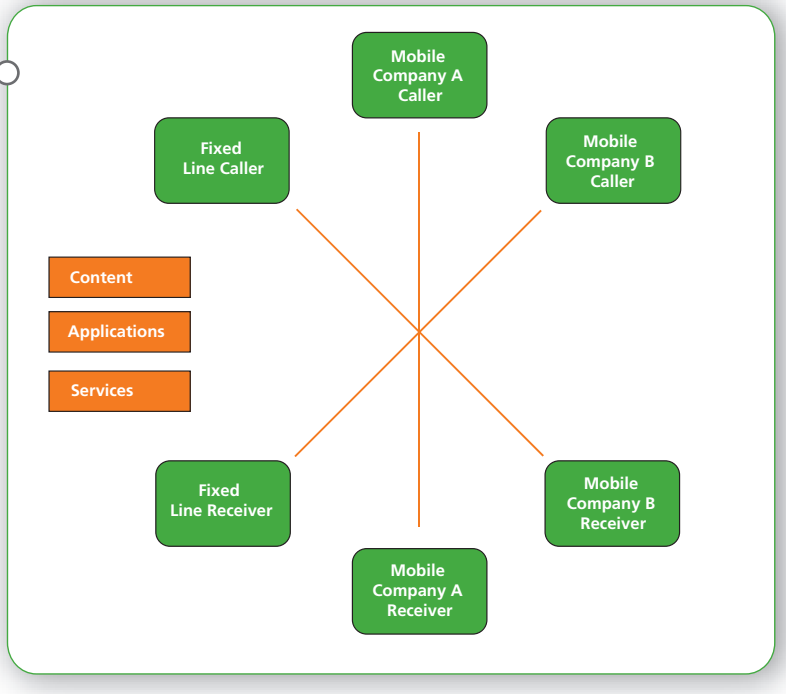
DECISIONS ON HOW TO SPLIT THE COST OF A TRANSACTION BETWEEN TWO PARTIES MAKING A VOICE CALL OVER A FIXED LINE HAVE HAD UNINTENDED CONSEQUENCES ON COMPLEMENTARY PRODUCTS AND ON SUBSEQUENT DIFFUSION.

Hence, decisions on how to split the cost of a transaction (a call) between two parties making a voice call over a fixed line (the calling party and the receiving party) have had unintended consequences on complementary products and on subsequent diffusion.

III. Mobile Markets

A fixed-line call can involve up to two parties (where both sides are using fixed-line telecoms). A mobile phone call can involve up to four potential parties if both the caller and receiver are using mobile networks and the networks are competing (see Figure 1).

Figure 1



Also included in Figure 1 are three boxes labeled content, entertainment, and services. Content can be data or information; entertainment can include movies, blogs, videos, games, etc.; and services can be dating or employment agencies, restaurants, GPS (positioning), etc. The network providers are then platforms and the market is multi-sided.

Traditional telecoms have a lot to learn about pricing in multi-sided markets. Some six years ago, wireless application protocol (WAP) was touted as the means of offering services to the mobile phone customer. Network operators in most countries did not recognize that it was necessary to get both sides on board in order for the WAP market to form. The operators thought since they "owned" the customers, that WAP service and application providers needed to pay operators (or receive very little revenue) in order to access the operators' customers. The inevitable happened—WAP failed. Similarly, poor recognition of the multi-sidedness of markets also enveloped much of the 3G service rollout in Europe. But the early rollout and acceptance of 3G in Japan (primarily by DoCoMo) showed a workable two-sided market model.

Unlike WAP and 3G in Europe, DoCoMo's approach allowed easy entry to its large accepted list of service application providers. To be accepted meant submitting basic financial data and plans, and having acceptable material. DoCoMo took 9 percent of application service revenues as its share and let consumer choice dictate providers' location on menu selections. That is, DoCoMo did not choose or sell the right to be first on the menu of, say, ring tone providers. Instead such providers competed to be first on the list. The list ranked providers according to popularity.

DoCoMo also understood two other aspects of business models for emerging two-sided markets. As customers could not foresee how much calling time or data charges they would use in accessing new services, DoCoMo initiated three significant controls that had never been used elsewhere (although now, many years later, they are becoming commonplace). First, DoCoMo limited the price that service providers could charge end users. Second, customers could see their bill on their phone in real time, with details of spending since their last bill, the last day, the last hour. Third, DoCoMo implemented controls on applications that could use a lot of network time. For example: four or five years ago, a fishing game became fashionable among company executives where the phone could be used to catch fish. This turned out to be fairly addictive and DoCoMo insisted that the game developer have the fishermen fall asleep after an hour. These controls by DoCoMo showed an understanding of the pricing and usage requirements to ensure that markets formed and were used optimally.

Many European 3G providers did not learn from the Japanese experience. They selected services and content they thought their customers would want (i.e., walled gardens). The pricing to customers is not simple to understand, nor can costs of accessing content be calculated as it is based on megabits of down-

loads. The price charged to content owners is not known, but given that access to the menu is tightly controlled, operators likely attempt to acquire significant revenue shares. As a result, provider ranking on menu selections is dictated by the telecom operator, not by customers. Third, most mobile telecoms have not introduced real-time bill information accessible on the device. Thus the business models for 3G services of a number of European telecom operators do not recognize the two-sided nature of service markets.

IV. Content

All telecoms, fixed and wireless, see the provision of content as new revenue sources—new multi-sided businesses. Different pricing models co-exist in these markets. For example, competing content platforms—newspapers, magazines, and broadcast television—have third-party advertisers that elect to pay part of the costs of the content and platform.

Take as an example, FOTA broadcasting. Since its inception in the 1950's, viewers pay to acquire their own devices (e.g., TV receivers) and advertisers pay for the provision of content and the platform over which the content is delivered (e.g., the costs of the broadcasters). Thus, the costs of both the content and of the platform are paid for by advertisers. Broadcasting has shifted its business model so that there is now both advertiser-supported content and programming (i.e., free to the viewer), as well as subscriber-paid content. The subscriber-pay model is via both a fixed monthly fee and pay-per-view. The subscriber-pay models include charges for the platform.

IT IS REASONABLE TO PROJECT THAT NEW CHARGING MODELS WILL EVOLVE IN TELECOMMUNICATIONS AS THE BUSINESS CONVERGES. HENCE IT IS REASONABLE TO EXPECT THE CURRENT TELECOM PRICING MODEL WILL ALSO EVOLVE.

As telecoms move into platform provision of content, more sides than the traditional calling parties of a voice call are added to the business model. Telecoms have high fixed and sunk costs. Other platforms are eroding the once fortress-like hold that telecoms had over the voice market. One such platform is IP-based, peer-to-peer file sharing platforms such as Skype or voice over IP (VoIP). As of November 2006, VoIP accounted for 20 percent of all voice traffic in France. Another platform threat to traditional telecoms is Wi-Fi and WiMAX. Google is experimenting by offering free Wi-Fi in San Francisco, California. These so-called free calling services such as VoIP or Wi-Fi generate revenues in ways other than charging the calling or the called party. Google is an advertising-based model. Hence, its free Wi-Fi experiment is one whereby the cost of calling (i.e., the platform) is paid for, all or in part, by advertisers.

The FOTA broadcasting model has evolved into a situation today where there are multiple price charging mechanisms for ensuring that all sides are on board.

It is reasonable to project that new charging models will evolve in telecommunications as the business converges from offering a voice channel platform to two parties to a business platform providing access to communications and content services. Hence it is reasonable to expect the current telecom pricing model will also evolve.

There is a current debate in the United States as to whether communications carriers can discriminate among content services. This net neutrality debate is not one of whether the Internet is free, but about who will pay for the high fixed and sunk costs of Internet communications networks. It is inefficient and incorrect to regulate that future multi-sided communications markets should charge according to the model established accidentally by the fixed-line Bell system a century ago (i.e., calling party pays). Forcing all costs of next-generation networks and fiber upgrades on subscribers is inefficient. Broadcasting has moved from FOTA broadcasting to multi-charging business models. Communications firms need the ability to allocate costs across all sides in a manner that maximizes network effects for all. Thus, pricing in telecoms may migrate from calling party pays to receiving party pays to FOTA to perhaps FOTP, or free-over-the-platform, where free really means that other sides to the market pay.

Hence it is time to understand the multi-sided nature of communications markets and the platform role of infrastructure providers. All parts of the ecosystem—telecoms, content and application providers, and service providers as well as politicians and regulators—need to account for two-sidedness in their policies and in their pricing decisions. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Retail Payments and Card Use in the Netherlands: Pricing, Scale, and Antitrust

Wilko Bolt

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Retail Payments and Card Use in the Netherlands: Pricing, Scale, and Antitrust

Wilko Bolt

Efficient payment systems are essential components of well-functioning economies and financial markets, facilitating the exchange of goods, services, and assets. The speed and ease with which payments can be processed and executed will in general affect economic activities, output, and price levels. Therefore it is important that payment systems satisfy some basic principles of economic efficiency. The payment landscape is changing rapidly, with the fast growth of credit and debit card payment systems in many developed economies as perhaps one of the most striking examples. Data from a 2004 paper by Zinman show that in the United States alone, in 2002, consumers used their debit and credit cards in 33.4 billion transactions to charge around USD 2.3 trillion in total.[1] Furthermore, data from Krueger's 2001 paper and the Bank for International Settlements (BIS) illustrate that in ten industrial countries the use of debit and credit cards rose from over nine billion transactions in 1987 to about 51 billion transactions in 2002.[2] In particular, in the Netherlands, the enormous upswing in the usage of debit cards has been the main driver for the rapid developments in non-cash payments. Debit card payments in the Netherlands exceed-

1 J. Zinman, Why Use Debit Instead of Credit? Consumer Choice in a Trillion Dollar Market (Federal Reserve Bank of New York, Staff Report No. 191, 2004).

2 M. Kreuger, Interchange Fees in the Line of Fire (2001) (unpublished, Institute for Prospective Technological Studies, Seville, Spain) (on file with author).

The author is senior economist in the Research Department, De Nederlandsche Bank, Amsterdam, The Netherlands. The views expressed in this paper are those of the author and do not necessarily represent those of De Nederlandsche Bank, or the European System of Central Banks.

ed EUR 56 billion in 2004 (more than 12 percent of GDP) with a volume of around 1.25 billion debit card transactions (50 times higher than in 1990), and they are still growing rapidly.

Payment systems and payment services are not free. They impose considerable resource costs on society. To illustrate, Humphrey, Pulley, and Vesala have estimated that in 1995, the United States spent three percent of its GDP just to make payments.[3] For the Netherlands, Brits and Winder calculate that the social costs involved with all point of sale (POS) payments amount to 0.65 percent of Dutch GDP in 2002.[4] Similarly, the Belgian National Bank recently estimated that the total social cost of various POS payment instruments (cash, debit cards, credit cards, and stored-value cards) amounts to 0.75 percent of Belgian GDP in 2003, of which 0.6 percent involves the use of cash.[5] Hence, there is much to be gained in designing payment systems efficiently.

The analysis of pricing structures and competitiveness in payment markets warrants special attention. The widely observed shift from the use of cash toward electronic modes of payment has undoubtedly led to an increase in the overall efficiency of (retail) payment systems. Still, non-transparent—and possibly inefficient—pricing arrangements and market power potential of card schemes in the card payment markets have attracted controversy and triggered antitrust scrutiny. Recently, a federal antitrust lawsuit brought by U.S. retailers against Visa and MasterCard regarding their debit card pricing practices resulted in an out-of-court settlement involving compensation payments of some USD 3 billion. In addition, the European Commission has devoted considerable attention to interchange fees and the rules set by the members of credit card associations.[6] In Germany recently, a lively debate has come to the fore on the adoption of a multilateral interchange fee for all debit card payments. Further, in the Netherlands, retailers expressed their dissatisfaction with some parts of the Dutch payment system, especially drawing attention to current pricing and acquiring arrangements for debit card services. Many of the Dutch retailers' complaints involved alleged monopolistic behavior by Interpay—the central routing switch in the nationwide debit card network—in terms of pricing policies, transparency, and delivered quality of services. Clearly, pricing issues are central to the analysis of card payment services.

3 D. Humphrey et al., *The Check's in the Mail: Why the United States Lags in the Adoption of Cost-Saving Electronic Payments*, 17 J. Fin. Services Res. 17-39 (2000).

4 H. Brits & C. Winder, *Payments Are No Free Lunch*, 3(2) Occasional Stud. (2005).

5 Belgian National Bank, Costs, Advantages and Disadvantages of Different Payment Methods (2005).

6 Press Release, European Commission, Commission Plans to Clear Certain Visa Provisions, Challenge Others (Oct. 16, 2000).

But what economic principles should guide such payment pricing? Indeed, appropriate pricing arrangements for payment instruments are a complex matter, since payment networks give rise to strong usage and network externalities. Until recently, no structural theoretical analysis of price determination in (electronic) payment networks was available. The situation has changed just over the last years by observing that the payment industry is a two-sided market, stressing the fact that in setting the prices for payment instruments, banks need to get both consumers and retailers on board by pricing both sides of the market in an effective way.[7] The theoretic analysis of two-sided markets has increased our understanding of payment pricing, social welfare of payment systems, network competition, and antitrust issues.

APPROPRIATE PRICING ARRANGEMENTS FOR PAYMENT INSTRUMENTS ARE A COMPLEX MATTER, SINCE PAYMENT NETWORKS GIVE RISE TO STRONG USAGE AND NETWORK EXTERNALITIES.

This paper adds to the surging literature on payment economics—a term first coined by Edward Greene during a 2004 conference, hosted by the Federal Reserve Bank of Kansas City, on the economics of payments—and attempts to bridge the gap between observed Dutch payment patterns on the one hand, and empirical and theoretic models based on two-sided logic on the other hand. Specifically, in Section II our analysis provides a rationale for why Dutch debit card prices are completely skewed towards retailers, in the sense that the price markup for retailers is much higher than for consumers. In Section III, we will analyze the effects of transaction-based pricing on the adoption of card payments, and the role that antitrust authorities may play. Section IV attempts to quantify possible payment scale economies arising from payment transaction growth and consolidation of processing centers. For competition authorities, though, huge scale benefits may be at odds with competitiveness when increased consolidation efforts reduce the number of players in the payment processing market. Section V concludes the analysis, but we turn first, in Section I, to a simple model of two-sided markets to set the stage.

I. A Simple Two-Sided Market Model for Debit Cards

This section describes a model of a monopolistic platform that supplies network services. This simple model is a fairly good representation of the Dutch debit card market, where there is only one nationwide debit card network with only

7 The recent general literature on two-sided markets began around 2002 with seminal papers by Jean-Charles Rochet and Jean Tirole and early versions of an important related paper by Mark Armstrong. *See* J.-C. Rochet & J. Tirole, *Cooperation Among Competitors: The Economics of Payment Card Associations*, 33 RAND J. ECON. 549-70 (2002); J.-C. Rochet & J. Tirole, *Platform Competition in Two-Sided Markets*, 1 J. EUR. ECON. ASS'N 990-1029 (2003); and, M. Armstrong, Competition in Two-Sided Markets (2005) (mimeo, University College London) (on file with author).

one processor. Moreover, in the Netherlands there is hardly any competition from credit cards and checks at the point of sale: it is only cash or debit cards.

The model features potential gains from trade which are created by transactions between two different groups of end-users, whom we will call buyers (subscript b) and sellers (subscript s). Such transactions are mediated and processed by the monopoly platform. To provide these (network) services, the platform charges buyers and sellers positive transaction fees, denoted by t_b and t_s, with the total price labeled $t_T = t_b + t_s$. The pricing structure denotes the allocation of the total price t_T over t_b and t_s. For simplicity, we abstract from any fixed periodic fees for end-users to connect to the platform. In performing its tasks, the platform incurs joint marginal costs $c > 0$ per transaction. There are no fixed costs.

Buyers and sellers that transact on the platform enjoy positive benefits of usage. We assume that buyers and sellers are heterogeneous in the benefits they receive from a transaction, i.e. $b_i \in [\underline{b}_i, \infty]$, $\underline{b}_i > 0$, $i = b,s$. The probability density function of these benefits is labeled $h_i(\cdot)$, with cumulative density $H_i(\cdot)$, $i = b,s$. To illustrate, in case of a debit card transaction, a buyer (or consumer) who wants to buy a good or service from a seller (or retailer) at price p, prefers to use his debit card whenever he gets positive benefits from using the card relative to other payment instruments, say cash. A transaction using the debit card takes place if, at the same time, the seller prefers accepting the debit card payment to accepting cash.

The model logic is simple. Only buyers with benefits b_b larger than incurred fees t_b will transact on the platform. Formally, the fraction of buyers connecting to the platform is given by:

$$q_b = D_b(t_b) = Pr(b_b \geq t_b) = 1 - H_b(t_b). \tag{1}$$

Analogously, the fraction of sellers which connects to the platform is equal to:

$$q_s = D_s(t_s) = Pr(b_s \geq t_s) = 1 - H_s(t_s). \tag{2}$$

Assuming independence between b_b and b_s, the total expected fraction of transactions processed by the platform amounts to:

$$q = D(t_b,t_s) = D_b(t_b)\, D_s(t_s). \tag{3}$$

Further, assume that the monopoly platform operates in a price region such that the price elasticities of quasi demand, $\varepsilon_i(t)$, $i = b,s$, exceed 1 for both sides of the market. Finally, for simplicity the total number of transactions is exogenously fixed, both on and off the platform, at N. So the total demand for platform services on the platform is given by $ND(t_b,t_s)$. A profit maximizing monopolistic platform operator will maximize:

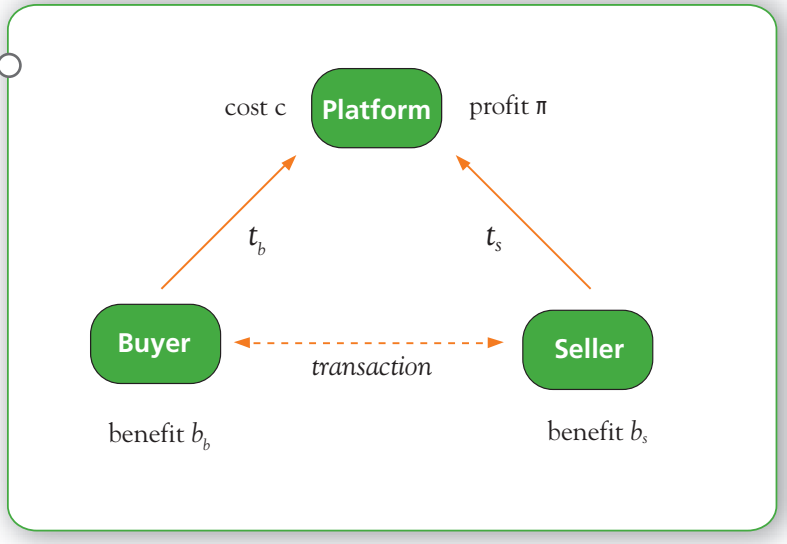
$$\pi(t_b,t_s,c) = N(t_b + t_s - c)\, D(t_b,t_s). \tag{4}$$

The two-sidedness is clear. Cardholders attach value to their payment card only to the extent that these are accepted by the retailers that they visit for shopping; in turn, affiliated retailers benefit from a widespread diffusion of cards among consumers. By setting its transaction fees, the monopolistic payment network must make sure that both sides of the market get on board. In particular, under two-sidedness, the platform chooses a total price for their payment services and also chooses an optimal pricing structure. As Evans states in a 2003 paper, in two-sided industries the product may not exist at all if the business does not get the pricing structure right.[8]

Figure 1 schematically depicts the model.

Figure 1

The monopoly platform



II. Skewed Pricing of Dutch Debit Card Payments

Over the last decade the Netherlands has shown a huge increase in the usage of debit cards. In 1990, the Netherlands started with one debit card transaction per person per year, but rose to 77 in 2004, a 33 percent annual growth rate. In 2005 its volume totaled 1.25 billion transactions and its value EUR 56 billion. Consumer participation is complete, with virtually all consumers over age 18 carrying a debit card. Usage density on the retailers' side is lower, with about 56 percent of all Dutch retailers accepting debit card payments. However, virtually all large retailers accept debit cards.

8 D. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325-382 (2003).

Dutch debit card pricing is completely skewed to the retailers' side. On average, in 2004 retailers paid t_s = EUR 0.05 per debit card transaction, while t_b = EUR 0.0 for consumers. And although Dutch merchants are allowed to quote different prices for cash versus card payments, this right to surcharge is rarely exercised, probably because of transaction costs. Recently, Dutch retailers expressed their dissatisfaction with the current pricing strategy of Interpay, perceiving the high retailer fee as a form of abuse of market power. Indeed, the skewed nature of prices in the debit card industry triggered antitrust scrutiny and led to an in-depth investigation by the Netherlands Competition Authority (NMa). As a result, Interpay was penalized for EUR 30 million and the participating commercial banks were fined some EUR 17 million. However, questions were raised as to whether the NMa fully took notice of the two-sided nature of the debit card market and its economic consequences. Indeed, in 2005 the penalty for Interpay was fully withdrawn (as well as a reduction of EUR 3 million for commercial banks) by taking the position that further research is necessary to fully determine whether merchant discounts are indeed excessive for Dutch debit cards.

In their 2003 paper, Bolt and Tieman show that under constant elasticity of demand, the side of the market that is sufficiently more elastic is kept to a minimum fee, while the other side pays a relatively high fee.[9] Mathematically, this result is characterized by a corner solution. The economic intuition underlying our skewed pricing result is that the most elastic side of the market is effectively subsidized by the other side, so as to boost the demand for services supplied by the platform. Indeed, every agent on the high elasticity, low price side of the market will connect to the platform. Because it benefits from full participation on one side, the other side is therefore also encouraged to join. However, since this side is more price-inelastic, the platform is able to extract higher prices. In particular, assuming that buyers are more elastic than sellers, the authors show that profit-maximizing fees are equal to:[10]

$$t_b^M = \underline{b}_b \text{ and } t_s^M = \frac{(c - \underline{b}_b)\varepsilon_s}{\varepsilon_s - 1}. \qquad (5)$$

Recent empirical analysis has shown that consumers are quite sensitive to price changes of payment services.[11] At the same time, retailers often complain that due to competitive pressures they are forced to facilitate debit card services. Retailers cannot afford to say no to their customers. At the same time, they do not see many payment alternatives. Hence, retailers may be assumed to be much

9 W. Bolt & A.F. Tieman, Pricing Debit Card Payment Services: an IO Approach (International Monetary Fund, Working Paper No. 202, 2003).

10 *Id.*

11 *See, e.g.*, D. Humphrey et al., *Realizing the Gains from Electronic Payments*, 33 J. Money, Credit, & Banking 216-34 (2001).

less price-elastic in their demand for debit card services than consumers. And this might explain why the pricing structure in the Dutch debit card market is so heavily skewed towards the retailers' side of the market, as predicted by the above skewed pricing result.

These results potentially have important bearings on antitrust issues. In antitrust analysis, high markups raise concerns of abuse of market power. However, traditional antitrust logic should be reconsidered in two-sided markets.[12] The fact that benefits and costs arise jointly on the two sides of the market effectively means that there is no direct economic relation between price and cost on either side of the market. It is generally not possible to examine price effects on one side of a market in isolation, i.e., without considering the resulting feedback effects from the other side. In particular, with skewed pricing strategies that may hold in a social optimum as well, one will always observe a non-negligible gap between the consumers' and retailers' price. Retailers might mistakenly perceive the resulting markup on their side of the market as a consequence of abuse of market power by the payment platform.

RETAILERS MIGHT MISTAKENLY PERCEIVE THE RESULTING MARKUP ON THEIR SIDE OF THE MARKET AS A CONSEQUENCE OF ABUSE OF MARKET POWER BY THE PAYMENT PLATFORM.

III. The Effect of Transaction Pricing on Card Payment Use

The production of electronic payments by banks often cost from one-third to one-half as much as paper-based equivalents or cash[13] Banks and merchants are interested in shifting users to electronic payments to save costs, as are some government policymakers who seek to improve the cost efficiency of their nation's payment system. Historically, banks have recouped their payment costs through:

(1) interest earned on payment float (from delaying availability of funds credited to accounts and debiting accounts prior to bill payment value dates);

(2) maintaining a spread between market rates and the rate paid on deposits; and

(3) charging flat monthly fees or imposing balance requirements.

12 For a discussion of the main fallacies that can arise from using conventional wisdom from one-sided markets in two-sided industries, *see, e.g.,* J. WRIGHT, ONE-SIDED LOGIC IN TWO-SIDED MARKETS (AEI-Brookings Joint Center for Regulatory Studies, Working Paper No. 10, 2003). For antitrust implications of pricing in two-sided markets, *see, e.g.,* Evans, *supra* note 8.

13 *See, e.g.,* D. Humphrey et al., *Benefits from a Changing Payment Technology in European Banking*, 30 J. BANKING & FIN. 1631-52 (2006).

In contrast to business users, consumers face very few payment services that are priced on a per transaction basis and so have little incentive to choose the lowest cost instrument either at the point of sale or for bill payments.

Banks are well aware that transaction pricing can speed up the shift to electronic payments, but are reluctant to lose deposit market share by being the first (and perhaps only) bank to implement explicit prices differentiated according to underlying costs. While this problem is mitigated if most (or all) banks implement pricing at about the same time, antitrust authorities are unlikely to view such coordination as being in the public interest, unless the social benefits from pricing are significant and the result is a compensating reduction in payment float, a higher interest rate paid on deposits, or a reduction in flat fees or balance requirements. Indeed, float reduction was the trade-off when banks coordinated the timing of when they would implement pricing in Norway (there was no coordination in the prices to be charged and initially some were zero).

In my 2005 paper joint with Humphrey and Uittenbogaard, we use the experience of Norway (which directly priced its payment services to consumers) and the Netherlands (which did not) over the time period 1990 to 2004 to try to determine what the incremental effect of transaction pricing may be on the adoption of debit cards versus withdrawing cash from an ATM, and on the adoption of electronic giro transactions (credit transfers and direct debits) over paper giros.[14] Specifically, we compare payment instrument use per person in Norway in response to the prices being charged, the availability of terminals, and the level of real consumption with the experience of the Netherlands, which also adopted electronic payments but did not price. Figure 2 shows debit card usage and ATM usage in Norway and the Netherlands from 1990 to 2004, along with its deployment of terminals. Figure 3 depicts the per-transaction Norwegian prices for debit card transactions and ATM withdrawals.

Differences between Norway and the Netherlands are used to try to explain per capita use of debit cards, ATM cash withdrawals, and electronic and paper giro payments. The main influences on payment use and composition are differences in the number of EFTPOS and ATM terminals per million population, the prices being charged in Norway (positive) and the Netherlands (zero), and differences in the level of real per capita consumption. In a two-sided context, EFTPOS terminal availability is a proxy for acceptance of debit cards by retailers, since their prices are not known. Our four-equation country difference model spanned 15 years—the limit of the available data. The model is estimated in a systems equation framework using levels data and robustness as illustrated by estimating models in a first difference and error correction framework.

14 Bolt et al., The Effect of Transaction Pricing on the Adoption of Electronic Payments: A Cross-Country Comparison (De Nederlandsche Bank, Working Paper No. 71, 2005).

Figure 2

Debit card usage and ATM withdrawals in Norway (orange line) and the Netherlands (green line), 1990-2004

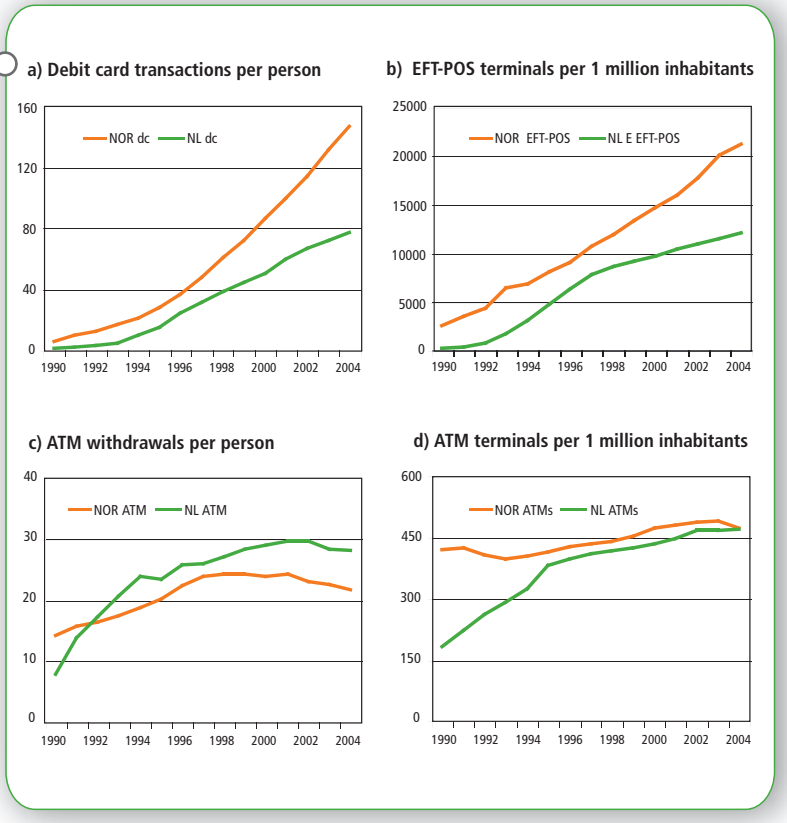


Figure 3

Norwegian prices (in euro) per debit card transaction (orange line) and ATM withdrawal (green line

The effects of pricing differ depending on which instruments are being considered. Overall, pricing has a smaller effect on shifting consumers from ATM cash withdrawals to debit card use than it does in shifting use from paper to electronic giro transactions. The reason for this difference seems to be that there are non-price benefits associated with debit card use (e.g., convenience, security) that consumers value such that the availability of terminals needed for debit card transactions has a stronger effect on debit card use than prices—as evidenced by the fact that the debit card price elasticity is smaller than the terminal elasticity. Debit cards also substitute for costly checks and the high price on these instruments in Norway was associated with their virtual elimination. (Although the same thing happened in the Netherlands, which did not price.) While terminal availability appears to have a stronger effect on debit card use than does pricing, the shift to cards can be sped up when pricing is combined with terminal availability. Using our estimated elasticities and the actual changes in prices and terminals, the predicted relative rise of debit card use over ATMs was eight percent from terminal effects alone but rose to 10.4 percent with pricing, an increase of over 20 percent.

The effect of pricing on electronic giro use was greater than it was for debit cards since the electronic giro price elasticity is larger and the percent change in price experienced was greater. Reasons for this difference are the above-mentioned non-price convenience and security attributes of debit cards along with the fact that for one-third of our time period the absolute price of a debit card transaction was higher than the weighted average price of an ATM cash withdrawal. In contrast, the price of an electronic giro was always absolutely lower than the paper giro price. Even though the relative prices of debit cards and electronic giros were both falling over the entire period, the higher absolute price of a debit card transaction versus an ATM would be expected to dull the overall price response being measured for the entire period since there is no strong reason to believe that the price response is symmetric (and symmetry was not imposed in our model) since the non-price attributes of debit cards and ATMs are different. Thus, if pricing is implemented, it will likely be more successful if the absolute price of the less expensive instrument is always absolutely lower per transaction than the price of the more expensive instrument.

THE SOCIAL BENEFITS OF ELECTRONIC PAYMENTS ARE QUITE LARGE AND MAY CONVINCE ANTITRUST AUTHORITIES TO ALLOW THE COORDINATION OF THE TIMING OF THE IMPLEMENTATION OF PRICING TO SPEED UP THIS TRANSITION.

As both Norway and the Netherlands are well on their way to realizing the full potential gains from electronic payments, the issue of pricing or not pricing is seemingly more a policy topic for developed countries that are not as far along in the substitution process or for most developing countries that are just in the initial stages of thinking about how to improve the efficiency of their payments sys-

tem. The social benefits of electronic payments are quite large and may convince antitrust authorities to allow the coordination of the timing of the implementation of pricing (but not, of course, the prices to be charged) to speed up this transition. Pricing could become a reality even in countries that have largely shifted to electronic payments since, with low or falling interest rate margins, this may facilitate the recoupment of bank payment costs. At the same time, inefficient cross-subsidization practices would be removed.

IV. Measuring Payment Scale Economies

Debit cards have largely replaced checks in many European countries (with France and the United Kingdom being the exceptions) and they continue to replace cash for medium value transactions. However, debit card costs have hindered their use for the replacement of cash for small value payments. This has led banks and other suppliers to offer potentially lower cost stored value cards for small value transactions, a solution that required consumers and merchants, or just merchants, to adopt two different payment technologies. Consumer adoption and use of stored value cards seems stalled at a relatively low level of market penetration.[15] Although data are incomplete, stored value cards account for only EUR 1.2 billion in payments across eleven European countries in 2004. In contrast, the value of debit card transactions is estimated to be EUR 1,146 billion while the value of cash withdrawals (a proxy for cash use) was EUR 2,189 billion. Overall, card payments comprised 34 percent of the total, cash withdrawals accounted for 66 percent, while stored value payments were only 0.04 percent. As stored value payments are used to replace small value cash transactions, and thus would be expected to be a small portion of overall POS payments, their current small share is almost entirely due to their low level of market penetration.

An important drawback of stored value cards is that consumers may have to carry two cards to replace cash—a debit card plus a stored value card—and the latter requires filling at terminals while the former does not. While convenience is enhanced if both technologies are on a single card and if merchants have a single terminal that can handle both types of transactions, banks often charge an extra fee to handle a stored value transaction. In 2002, the average total (fixed plus variable) bank plus merchant cost of a cash transaction at the point of sale in the Netherlands was EUR 0.30 while a debit card transaction was EUR 0.49 and that of a stored value transaction was EUR 0.93.[16] The hope was that stored value transaction volume would rapidly expand and substantially lower average fixed costs since average variable costs for stored value transactions are the lowest of the three at EUR 0.033 per transaction versus EUR 0.176 for cash and EUR 0.197 for debit cards. This has not happened. Debit cards are the more

15 L. Van Hove, *Why electronic purses should be promoted*, 2 BANKING & INFO. TECH. 20-31 (2006).

16 *See* H. Brits & C. Winder, *supra* note 4, at Table 4.3.

mature product, already have a significant market penetration, and do not require consumers to access terminals to refill them. If debit card costs could be lowered sufficiently, they could further reduce cash use and replace stored value cards for small value transactions.

As the replacement of cash by debit cards for smaller value transactions is importantly influenced by unit costs and unit costs are largely dependent on transaction volume, the goal is to try to determine payment scale economies in the Netherlands and other European countries, especially for debit cards. Estimates of scale economies, when combined with expected transaction growth within a country or the consolidation of card processing operations across countries, permit future card unit costs to be approximated and the likelihood of debit cards replacing small value cash transactions assessed. Payment scale economies are considered to be the main economic driver behind the creation of a Single European Payments Area (SEPA), which entails the harmonization and standardization of retail payment instruments (especially payment cards, direct debits, and credit transfers) across the European Union. SEPA aims to improve the efficiency of cross-border payments and "to develop common instruments, standards, and infrastructures in order to foster substantial economies of scale."[17] The obvious question is whether it is possible to quantify these payment scale economies.

In my 2006 paper joint with Humphrey, we estimate payment scale economies with European data using a panel of payment and banking data for eleven European countries over 18 years.[18] Specifically, we relate bank operating (not total) costs to measurable physical characteristics of banking output associated with payment processing and service delivery levels and mix. In this manner we focus on those activities and expenses directly associated with the provision of payment services. Interest expenses paid to depositors and with a markup charged to borrowers are functionally separable from these activities. This approach allows us to determine how the level and mix of payment activities, along with the number of ATMs and bank branches, are directly associated with the size of a bank and its labor, capital, and materials operating cost from which scale economies may be approximated. In this regard our approach represents an alternative and more specific way to identify the likely effect on costs from technical change in banking. As POS and bill payment transactions are jointly processed in the deposit accounting function, while aspects of service delivery are jointly produced via branches and ATMs, these two activities can be considered functionally separable.

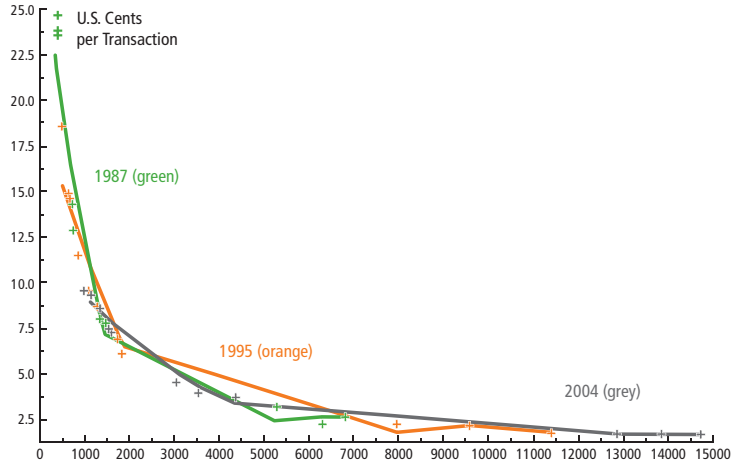
17 European Central Bank, Towards a Single Euro Payments Area, Fourth Progress Report (2006).

18 W. Bolt & D. Humphrey, Payment Scale Economies and the Replacement of Cash and Stored Value Cards (De Nederlandsche Bank, Working Paper No. 122, 2006).

Our results show that average scale economies for all payments across eleven countries using an estimated translog function is 0.27, indicating that substantial scale benefits would be expected as payment volume rises. Doubling of payment volume would only increase total costs by 27 percent. Not surprisingly, payment cards and bill payment instruments show bigger scale economies than ATMs and bank branches. Figure 4 shows how an approximation to unit payment cost varies by the total number of payment transactions. Although the curves in the figure are not identical to average cost curves, the slopes give a fair reflection of how payment unit costs change with payment volume.

Figure 4

Scale effects of European payment markets (smoothed data)



Translog: Predicted Unit Payment Costs vs. Payment Transaction Volume
(11 countries, PPP exchange rates, POS and bill payments)

These results provide preliminary scale economy information that may be helpful in outlining possible benefits from SEPA arising from the consolidation of electronic payment processing centers across the European Union. If this approach were pursued, the experience of the United States, which consolidates card processing across states, may serve as a useful example concerning the realized costs and benefits, as well as likely implementation issues of cross-border consolidation. However, for competition authorities increased consolidation and positive scale effects in payment processing may be at odds with the competitive potential of the market. It remains to be seen whether cost reductions arising from scale benefits are passed onto the end-users—in this case both consumers and retailers.

FOR COMPETITION AUTHORITIES INCREASED CONSOLIDATION AND POSITIVE SCALE EFFECTS IN PAYMENT PROCESSING MAY BE AT ODDS WITH THE COMPETITIVE POTENTIAL OF THE MARKET

V. Concluding Remarks

The Dutch retail payment market went through turbulent times during the last decade. The Netherlands observed a rapid shift from cash and paper-based payment instruments toward electronic payment instruments. Banks are well aware that transaction pricing can speed up the shift to low-cost electronic payments. But payment pricing is a complex matter, due to strong usage and network externalities. Recently, theoretic models of two-sided markets have provided useful insights in the complexity of the multi-player problems that payment activities pose, regarding efficient payment pricing, payment network competition, and antitrust consequences.

This paper showed how heavily skewed pricing of debit card payments can be rationalized in a simple two-sided model, and how the implementation of transaction-based pricing affects adoption rates of electronic payments. In addition, it briefly examined the impact of payment scale economies, which will be a main driver for the economic success of SEPA. At the same time, payment systems and payment arrangements feature a natural tension between cooperation and consolidation on the one hand and competitiveness on the other. This natural tension mixed with sharp two-sided ingredients, causes traditional one-sided economic logic to break down, and requires antitrust practitioners and competition authorities to look at the world with a drastically different view. ◀

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Mobile Virtual Network Operators: Beyond the Hyperbolae

Duarte Brito and Pedro Pereira

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Mobile Virtual Network Operators: Beyond the Hyperbolae

Duarte Brito and Pedro Pereira

The diffusion of mobile telephony has been very fast. In the European Union, in 2005 426 million people, almost 93 percent of the population, had mobile telephones, whereas in 1998 only 18 percent of the population did.[1] Although the market is maturing and the number of subscribers is stabilizing, new service providers, defined broadly, continue to enter the market. Between 2004 and 2005, the number of service providers increased from 166 to 214 in the European Union.[2] Value-added 3G services and personalized services are becoming increasingly important. About five years ago, a particular type of mobile telephony firm was launched in the United Kingdom, Sweden, and Finland: mobile virtual network operators (MVNOs).

In this paper, we discuss briefly, and in non-technical terms, some aspects related to the entry of MVNOs. These aspects range from issues related to the entry process itself, such as barriers to entry and exclusionary practices, to the effects of entry on prices and product differentiation.

The paper is organized as follows. Section I presents some useful definitions. Section II discusses entry barriers and describes the role of MVNOs. Section III discusses exclusionary practices. Section IV analyzes the impact on prices of the

1 European Commission, 11th Report on the Implementation of the Telecommunications Regulatory Package (2005).

2 *Id.* There is some of heterogeneity across Europe, ranging from the United Kingdom, with more than 50 service providers, to Cyprus, Malta, and Slovakia with only two GSM (global system for mobile communications) network operators.

Duarte Brito is Assistant Professor at Universidade Nova de Lisboa. Pedro Pereira is Senior Economist at Autoridade da Concorrência.

entry of MVNOs. Section V discusses the choice of product characteristics by MVNOs.[3] Finally, Section VI concludes.

I. Basic Concepts

We begin with a few definitions. A mobile network operator (MNO) is a firm that owns a public mobile telephony network; A service provider (SP) is a firm that resells minutes purchased from an MNO. An MVNO is a firm that offers mobile telephony services without holding a license to use the radio-electric spectrum, and therefore without a mobile radio access network, but that issues its own branded SIM cards, has its own unique mobile network code, and operates a physical network infrastructure comprising as a minimum:

(1) a mobile switching center,

(2) a home location register, and

(3) an authentication center.[4]

An MVNO my also have:

(1) an equipment identity register and associated signaling capabilities, and

(2) an intelligent network platform to provide its customers with its own value-added services.

This definition is sometimes referred to as that of a full MVNO. However, there is no universally accepted definition, and the term MVNO is used to designate firms that range from an SP to a full MVNO.

There are significant differences in fixed costs and level of autonomy from the host between an SP or thin MVNO.[5] A full MVNO owns a comprehensive mobile telephony network. This implies that, compared with an SP, it has both higher fixed costs and more autonomy from the host in the design of products in the definition of price plans and in the introduction of new services.

Fixed costs and level of autonomy from the host are complements. Since it has high fixed costs, a full MVNO needs a large customer base to benefit from

3 Sections III and IV are based on D. Brito & P. Pereira, Access to Bottleneck Inputs under Oligopoly: A Prisoners' Dilemma? ,Autoridade da Concorrência Working Paper No 16, (2006). Section V is based on D. Brito & P. Pereira, Product differentiation when competing with the Supplier of a Bottleneck Input (2006) (mimeo, Universidade Nova de Lisboa) (on file with authors).

4 A radio access network consists of the masts, the base stations, and the frequencies. Access to the radio access network requires, at least, roaming privileges. Roaming is the ability of a customer of a mobile telephony firm to use its handset to automatically access service from another MNO.

5 A host is an MNO that gives access to its network to an MVNO.

economies of scale. Consequently, it needs to target broad market segments. To do so, a full MVNO needs to have considerable autonomy from the host. On the contrary, because it has small fixed costs, an SP can afford to target only niche market segments, for which no considerable autonomy is required.[6]

Since it needs to target broad market segments, the entry of a full MVNO may have a significant impact in the retail market, which may threaten the host MNO. On the contrary, because an SP can afford to target only niche market segments, entry by an SP may be welcomed by the host MNO. Hosting an SP can be a way of earning revenue from excess capacity, without increasing competition significantly.

From hereinafter, when we refer to an MVNO we mean a full MVNO.

II. Entry Barriers and the Role of MVNOs

To operate a mobile network a firm has to be licensed to use the radio-electric spectrum. Since spectrum is scarce, this means that only a few firms will be licensed. The large investments required to deploy a mobile telephony network limit the number of MNOs that the market can accommodate. In addition, consumer inertia under the form of switching costs, network effects, or brand effects, makes entry difficult. This is particularly true now that mobile telephony markets are reaching their saturation levels, as illustrated in Table 1.

The relative importance of the various entry barriers is unclear. However, there is at least a natural concern that the number of licensed firms may be smaller than the number of firms that would emerge in free entry equilibrium, in particular because mobile telephony markets are typically very concentrated, as illustrated in Table 2.

The entry of SPs, or any other type of firms that offers a limited range of services, can help promote competition, at least in some dimensions like price. However, to the extent that these firms have no autonomy from their hosts in terms of pricing policies, entry of SPs cannot replicate the full range of services offered by MNOs and cannot develop new innovative services, so their ability to compete with MNOs is limited.

In this regard, MVNOs are fundamentally different. MVNOs make possible the entry of firms that offer consumers a portfolio of services indistinguishable from those provided by MNOs, without requiring the allocation of additional radio-electric spectrum. MVNOs allow for a free-entry equilibrium to be attained.

6 Examples of these are the firms targeting the ethnic minorities.

Table 1

	Penetration rate	Growth in percentage points
	Oct-05	(2004/05)
Austria	99%	5
Belgium	83%	5
Cyprus	99%	11
Czech Republic	105%	7
Denmark	96%	6
Estonia	104%	16
Finland	98%	3
France	76%	5
Germany	90%	8
Greece	89%	7
Hungary	90%	8
Ireland	96%	8
Italy	111%	9
Latvia	79%	14
Lithuania	117%	37
Luxembourg	150%	17
Malta	81%	6
Netherlands	94%	10
Poland	71%	16
Portugal	106%	10
Slovakia	80%	5
Slovenia	90%	–
Spain	94%	6
Sweden	101%	7
United Kingdom	103%	10
EU15	92%	–
EU25	91%	–

Source: European Commission, 11th Report on the Implementation of the Telecommunications Regulatory Package (2005).

III. Exclusionary Practices

A. MARKET FORECLOSURE

In order to operate, an MVNO needs to obtain access to the radio access network of an MNO. The host and the entrant negotiate over several dimensions, such as the prices of origination and termination traffic, the elements of the host's network that the entrant will hire, and the capacity that the entrant

Table 2

	Operators with licenses, Sep-05				Market shares based on customers, Oct-05	
	DCS or GSM	DCS and GSM	UMTS and GSM/DCS	UMTS	Leading operator	Main competitor
Austria	1	3	4	1	40%	25%
Belgium		3	3		47%	32%
Cyprus		2	2		93%	7%
Czech Rep		3	3		NA	NA
Denmark		3	2	1	31%	20%
Estonia	1	3	3		46%	32%
Finland		3	3		NA	NA
France		3	3		47%	36%
Germany	4		4	1	38%	37%
Greece	1	3	3		NA	NA
Hungary		3	3		45%	34%
Ireland		3	2	1	49%	40%
Italy		3	3	2	40%	32%
Latvia		3	4		NA	NA
Lithuania		3	0		37%	32%
Luxembourg	1	2	3		58%	29%
Malta	2		2		52%	48%
Netherlands	2	3	5		36%	24%
Poland		3	3		36%	34%
Portugal		3	3		NA	NA
Slovakia		2	2		56%	44%
Slovenia	1	2	1		74%	20%
Spain		2	3	1	48%	28%
Sweden		4	2	1	NA	NA
United Kingdom	4		4	1	25%	24%
EU15					42%	31%
EU25					43%	32%

Source: European Commission, 11th Report on the Implementation of the Telecommunications Regulatory Package (2005).

expects to use. Typically, a contract between a host and an entrant involves non-linear price schedules, with payments flowing in both directions. The entrant might be compensated by the host if it brings new customers to the network, if it increases the total network traffic, or if it makes the use of the network more evenly divided throughout the day.

In principle, both parties can negotiate freely a mutually beneficial agreement, whereby the MNO concedes access to its network to the MVNO.[7] However, some wonder whether MNOs will voluntarily negotiate agreements with MVNOs, since the services the latter provide compete with the MNOs' own retail services. The regulation on MVNOs varies greatly across EU member states. In countries like Denmark, Norway, and the Netherlands, MNOs with significant market power have open access obligations towards MVNOs. In other countries, there are no such regulatory obligations.

The literature on market foreclosure addresses the question of whether a vertically integrated firm can increase its profit by foreclosing the downstream market to rivals.[8] As it is well-known, the monopolist owner of a bottleneck production factor, which is also present in the downstream retail market, may have the incentive and the ability to restrict access to the bottleneck production factor, in order to restrict competition in the downstream retail market.[9] An example of this is a monopolist owner of a public switched telephone network, which may want to restrict access to its local loop in order to restrict competition on the markets of fixed telephony or broadband access to the Internet.

In mobile telephony, because MNOs are not monopolist providers of a network, there are at least three reasons to suspect that MNOs have different incentives than fixed telephony incumbents with respect to giving access to their networks. First, even if an MNO denies an entrant access to its network, there is no guarantee that the entrant will not obtain access elsewhere. Second, an MNO that hosts an MVNO shares the revenue loss caused by an entrant with other MNOs. This mitigates the negative impact that entry may have on the revenues of the host MNO. Third, if entry cannot be blocked, then it is probably better for each MNO to be the one that gives access to the entrant. This allows the host MNO to earn additional wholesale revenues that at least partially compensate the loss in retail revenues caused by the entrant. Altogether, this suggests that

7 Typically, each MVNO buys access from only one MNO, although an MNO may sell access to several MVNOs.

8 The case in which the upstream market is monopolized was reviewed by J. Tirole, The Theory of Industrial Organization 193-4 (MIT Press 1988). The case of oligopolistic vertical integration with an oligopolistic upstream market was analyzed by J. Ordover et al., *Equilibrium Vertical Foreclosure*, 80(1) Am. Econ. Rev. 127-42 (1990).

9 *See, e.g.,* W. Baumol & J. Sidak, *The Pricing of Inputs Sold to Competitors*, 11(1) Yale J. on Reg. 170-202 (1994); G. Biglaiser & P. DeGraba, *Downstream Integration by a Bottleneck Input Supplier whose Regulated Wholesale Prices Are above Costs*, 31(2) RAND J. Econ. 137-150 (2001); N. Economides, *The Incentive for non-Price Discrimination by an Input Monopolist*, 16 Int'l J. of Indus. Org. 271-84 (1998); T. Krattenmaker & S. Salop, *Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power of Price*, 96 Yale L.J. 209-93 (1986); D. Sibley & D. Weisman, *Raising Rival's Costs: The Entry of an Upstream Monopolist into Downstream Markets*, 10 Info. Econ. & Pol'y 451-70 (1998); and D. Weisman, *Access Pricing and Exclusionary Behavior*, 72 Econ. Letters 121-26 (2001). For a dissenting view, *see* R. Bork, *Vertical Integration and the Sherman Act: The Legal History of an Economic Misconception*, 22 U. Chi. L. Rev. 157-201 (1954).

MNOs may face a prisoner's dilemma. They would be better off if entry did not occur. However, each has individual incentives to rush to be the one who gives access to the entrant.[10] This does not mean that such voluntary agreements should necessarily occur. Incumbents may still non-cooperatively foreclose the market or collude to foreclose the market.

B. RAISING RIVAL'S COSTS

Once entry occurs, perhaps due to open access regulation, an upstream monopolist that participates in the downstream market may try to raise the costs of the downstream rivals, for instance by discriminatory quality degradation. By doing so it might induce the downstream rivals to contract their market share, leaving a larger share of downstream oligopoly profits for its downstream subsidiary.[11] In the case of MVNOs, the possibility of quality degradation may be mitigated by the competition between host MNOs. Following a raising rival's cost strategy, the host should consider that the entrant has other alternatives. To the extent that if entry occurs it is better to be the host, the other MNOs will be eager to take its place. Hence, competition between possible hosts should ensure high-quality access service.

COMPETITION BETWEEN POSSIBLE HOSTS SHOULD ENSURE HIGH-QUALITY ACCESS SERVICE.

IV. Impact of Entry on Prices

Entry by an MVNO differs from entry by an MNO. An MVNO is simultaneously a rival and a customer of the host MNO. This affects the host's pricing strategy.

We make the helpful simplifying assumption that the host MNO is paid a constant access price for each of the MVNO's customers. Note that if the access price is set above marginal cost, and if the entrant is otherwise equally efficient, which we assume, then the entrant has higher costs than the incumbents.

Suppose that after the entry of an MVNO, all MNOs set the prices that prevailed before entry. If most of the consumers are already served, entry by an

10 In some cases, the host MNO's may even be better off with entry. The MVNO may attract many new customers to the host network, because it has a comprehensive retail network, such as Virgin, 7-Eleven, or Tesco, or because it has a global brand, such as Disney.

11 Salop and Scheffman (1987) addressed the issue of whether an upstream monopolist participating in the downstream market would raise rivals' costs. Economides (1998) showed that an upstream monopolist that is also present in the downstream market has the incentive to raise costs of its downstream rivals through discriminatory quality degradation, until they are driven out of the market. Vickers (1995) showed that an upstream monopolist present in a downstream oligopolistic market, and regulated under asymmetric information, also has incentives to raise rivals' costs. *See* J. Vickers, *Competition and Regulation in Vertically Related Markets*, 62(1) REV. ECON. STUD. 1-17 (1995); Economides, *supra* note 9; and, S. Salop & D. Scheffman, *Cost-Raising Strategies*, 36(1) J. INDUS. ECON. 19-34 (1987).

MVNO necessarily causes some consumers to switch from the firm they originally patronized to the entrant. If the host MNO loses a significant number of consumers to the entrant, it has an incentive to decrease its price. We call this downward pressure on the prices of the incumbents caused by the entrant stealing customers from them the retail competition effect. There is another effect that is exclusive to the host. If a host decreases its retail price, it gains customers. However, it also decreases the demand of the entrant, and therefore its wholesale revenue. We call this upward pressure on the host's retail price caused by the fact that decreasing its retail price reduces its wholesale revenues the wholesale effect.

The wholesale effect and the retail competition effect have opposing signs. This implies that the impact on prices of the entry of an MVNO is potentially ambiguous. It may cause a price reduction, as one would expect, but it may also lead to higher prices.

The price of the host is more likely to increase when the wholesale effect is large and the retail competition effect is small. This happens when the access price is high, or when the entrant MVNO captures a large fraction of the consumers that switch providers after an increase in the host's price.

If the access price is high, the entrant has higher costs than the incumbents. As a consequence, the entrant may charge a higher price than the prices the incumbents charged prior to entry. Consumers may be, nevertheless, better off due to the increase in variety that the entry brings about. Instead of paying a lower price for a product that they do not have a strong preference for, they pay a bit more for a product that is ideally suited to them and that they get a higher surplus from.

The prices of the non-host incumbents are likely to move in the same direction as the demand for their services. If the wholesale effect is large and the access price is high, the prices of the non-host incumbents are more likely to rise after entry. The reason is that the eventual increase in the host's price and the high price set by the entrant mean that non-host MNOs will have a larger demand, particularly those selling services similar to the host's.

V. Entrants' Product Differentiation Decision

First we discuss the entrant's perspective regarding its product differentiation decision. Suppose that the host MNO and an MVNO have been matched, and that the constant access price per consumer has been set above marginal cost. Before making its product differentiation decision, the entrant should anticipate the implications that this choice will have on price competition.[12] If the entrant chooses to offer a product very similar to the products of any of the incumbent's,

12 We are assuming that the prices decisions can be changed more rapidly than the product characteristics, brand positioning or consumer perception, which are assumed to be fixed for a longer period.

the consumer's choice will be essentially based on price, and the firm with the lowest price will succeed in capturing a substantial number of consumers. The ensuing price competition between the entrant and the incumbent in question will result in lower retail prices, and both firms will end up with lower profits. Consequently, the entrant should try to offer a product as differentiated as possible from those of the other incumbents.

However, the incumbents do not have symmetric incentives in terms of pricing, because one of them is the host. As mentioned in Section IV, among the incumbents, the host has the lowest incentives to cut its price after entry due to the wholesale effect. This means that it would be less damaging for the profit of the entrant to offer a product that consumers view as a closer substitute to the product of the host than to the products of the other MNOs. Thus, anticipating the different price responses of the host and of the other MNOs and holding everything else constant and symmetric among incumbents, the entrant should reduce the level of differentiation of its product compared to the product of the host and increase the level of differentiation of its product compared to the product of the other incumbents. Recall that the higher the access price, the stronger the wholesale effect. Hence, the higher the access price, the closer the entrant should position its product to the host's. However, it is not in the entrant's interest to offer a product identical to the product of the host, because a strong price competition would emerge.

HENCE, THE HIGHER THE ACCESS PRICE, THE CLOSER THE ENTRANT SHOULD POSITION ITS PRODUCT TO THE HOST'S.

We now turn to the host's perspective regarding the entrant's product differentiation decision. The host benefits from entry because it allows it to capture, through the entrant, subscribers that originally patronized the other incumbents. It might seem that the best situation for the host is one in which the entrant offers a product that competes closely with the products of the rival MNOs, but not with the product of the host. However, this is not necessarily true. Given retail prices, if the entrant offers a service that competes closely with the products of the rival MNOs, the wholesale revenues of the host do not cannibalize its retail subscriber base, and hence do not reduce its retail profits. Additionally, with an access price above marginal cost, the wholesale profits will be large. In other words, given retail prices, it is in the host's interest that the entrant offers a product that competes more closely with the products of the other MNOs than with its own product. However, the relative positioning of the product of the entrant affects prices. If the access price is high, the entrant has a severe cost disadvantage compared with the incumbents. Under these circumstances, offering a product that competes closely with those provided by other MNOs and therefore competing essentially on price while having a cost disadvantage, will lead to low revenues for the entrant and, thus, to low wholesale revenues for the host. This is why both the host and the entrant may prefer for the entrant not to offer a product that competes closely with rival MNOs.

VI. Conclusion

Mobile telephony is an oligopolistic market, where the number of competitors was initially limited to the number of licenses assigned by sectorial regulators. MVNOs allow any eventual entry limitations caused by the scarcity of the radio-electric spectrum to be overcome and free-entry equilibrium to be attained.

Additionally, the mobile telephony industry is one of the few in which more than one firm can provide access to a bottleneck input: a license to use the radio-electric spectrum. Incumbents may still foreclose the market. However, competition between them may lead the incumbents to voluntarily concede access to their networks.

Entry by MVNOs may cause prices to decrease. However, entry may also lead to higher prices for the entrant and the host. This is due to the wholesale effect. A host MNO makes both retail and wholesale profits. This gives the host an incentive to raise its retail price. Due to the same effect, and in order to mitigate post-entry price competition, the entrant should seek to position its product such that the host is its closest competitor.

VOLUME 3 | NUMBER 1 | SPRING 2007

Competition Policy International

Contingent Commissions in Insurance: A Legal and Economic Analysis

Richard A. Epstein

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Contingent Commissions in Insurance: A Legal and Economic Analysis

Richard A. Epstein

This paper gives a brief analysis of the role of contingent commissions in insurance markets. These contracts have received a great deal of attention in recent years because they were the focal point of major criminal enforcement actions that New York's then-Attorney General, now Governor, Eliot Spitzer, brought against prominent insurance brokers, including the largest three brokers: Marsh & McLennan, Aon, and Willis. Those prosecutions resulted in fines and other sanctions being lodged against these brokerage houses, as well as continuing criminal prosecution against employees who were engaged in some bid-rigging schemes. On balance, a strong case can be made out for requiring disclosure of contingent commissions and for banning any form of bid-rigging. The adverse consequences of nondisclosures are more difficult to track than those for collusion, given the difficulty of showing in individual cases a connection between the nondisclosure and any pecuniary loss sustained by the insured. The case for banning all contingent commissions in the absence of concealment or bid-rigging, still remains "not proven." It is not easy to come up with a powerful efficiency explanation for the use of contingent commission agreements, but if these agreements continue to be adopted with full disclosure in the absence of collusion, then it seems premature to ban them just because our incomplete knowledge of how brokerage markets work does not supply a compelling efficiency justification for their use.

The author is the James Parker Hall Distinguished Service Professor of Law, University of Chicago, and the Peter and Kirsten Bedford Senior Fellow, The Hoover Institution. In developing his own views, the author had the benefit of an informative and comprehensive paper, *The Economics of Insurance Intermediaries*, by Professors J. David Cummins and Neil A. Down. He is also most appreciative of the excellent research assistance he received from Chad Clamage, Stanford Law School, Class of 2008 and of funding support from the Barbon Institute.

I. Introduction

The purpose of this short paper is to give a brief analysis of the role of contingent commissions in insurance markets. These contracts have received much attention in recent years because they were the focal point of major criminal enforcement actions that New York's former Attorney General, Eliot Spitzer, brought against prominent insurance brokers, including the largest three: Marsh & McLennan, Aon, and Willis. Those prosecutions resulted in fines and other sanctions being lodged against these brokerage houses, as well as continuing criminal prosecutions of employees who were allegedly engaged in some bid-rigging schemes.

The gist of these settlements is captured in the terms that Marsh & McLennan entered into with Spitzer's office: in addition to paying US$850 million over four years into a client compensation fund, Marsh agreed that:

> "the company will adopt dramatic new reforms, including an agreement to limit its insurance brokerage compensation to a single fee or commission at the time of placement, a ban on contingent commissions, and a requirement that all forms of compensation will be disclosed to and approved by Marsh's clients."[1]

The merits of these settlement provisions require a great deal of attention, but before addressing those issues, it is best to begin with an account of what these contracts are, and the role they play in the overall insurance industry.

This paper is organized as follows. The first section relies on the general theory of insurance to give a definition of contingent insurance contracts and to offer a tentative efficiency explanation for their selective use in some insurance market segments. The second section then analyzes the two main objections that have been made against the use of these contracts based on the laws of fiduciary duties regarding disclosure on the one hand, and the antitrust laws on the other. The third section then evaluates whether it is wise to ban or regulate these contracts in the absence of either nondisclosure or collusion, and concludes that, as of yet, the case for any further regulatory initiative is as yet unproved. In this regard the blanket ban found in the New York settlement appears to go beyond the exigencies of the situation.

1 Press Release, New York State Attorney General & New Your State Insurance Department, Insurance Broker Agrees to Sweeping Reforms (Jan. 31, 2005), *available at* http://www.oag.state.ny.us/press/2005/jan/marshsettlement_pr.pdf.

II. Contingent Commissions

The essential function of a contract of insurance is to shift the risk of certain specified losses in whole or in part, from the insured to the insurer. The insurance company receives a premium from an insured that obliges it to compensate the insured for losses that arise on the occurrence of certain designated events. As with all useful contracts, the reassignment of risk cannot be a sterile transaction, from which neither side receives any gain. The transaction costs of finding a suitable insurer and negotiating terms and premiums are positive, so that entering into an insurance contract only makes sense if each side expects to receive some net benefit from the contract. As with all transactions, from the ex ante perspective, each side must think itself better off from the transaction to enter into it.

THESE TWO CONDITIONS SUGGEST THAT THE TRANSACTION WILL GO FORWARD ONLY IF THE SUM OF THE GAINS TO BOTH SIDES EXCEED THEIR COMBINED TRANSACTIONS COSTS. THE QUESTION IS WHAT FACTORS MUST BE TAKEN INTO ACCOUNT TO EXPLAIN WHY THIS OUTCOME WILL RESULT.

More formally, two conditions must be satisfied. On the one side of the market, the gain to the insured must exceed in expectation the sum of the premium paid and the transaction costs incurred in setting up the transaction. On the other side of the market, the premium received from the insured (plus any subsequent investment income from said premium) must exceed the present value of the insurer's future payoffs plus the transaction costs it incurs to put the deal together. Taken together, these two conditions suggest that the transaction will go forward only if the sum of the gains to both sides exceed their combined transactions costs. The question is what factors must be taken into account to explain why this outcome will result.

The first point to note is the source of the gain from the shifting of risk between the two parties. On the insured's side, this gain usually comes from smoothing the flow of income and expenses over time in different states of the world. For individuals, where it is hard to diversify risk, the need for insurance is often quite great. In the corporate setting, the shareholders may well have diversified portfolios, so that insurance becomes a less pressing issue. But even here, firm managers are often not fully diversified, and they may pressure the firm to take out insurance, knowing that an adverse event which produces sharp fluctuations in income could hurt their individual prospects by exposing the firm to a risk of bankruptcy or the loss of working capital. Many businesses therefore take out insurance in order to stabilize their future revenues and, through that, their profit position and the position of their key managers. That practice can be found in businesses both large and small, in both partnerships and corporations.

In addition, in some markets the insurer does more than smooth the insured's loss function. It also takes steps that help the insured organize its business to reduce and manage the risk of loss, and the insurer backs its promise of assistance

by assuming responsibility for these losses in the event that these (reduced-risk) events come to pass. That outcome is quite common with liability insurance, where the twin obligations to provide indemnification and defense are best understood as a way to make credible the commitment to engage in extensive accident prevention activities. The inspection firm that fails in its fundamental obligation now has to face the consequences. But if it has done its job well, then payouts on these losses could easily be a tiny fraction of the total premium dollar.

The insurance company achieves its own protection against loss in several ways. First, it diversifies some kinds of risk by taking on many insureds, always taking care to see that their risks are independent, so that payment on one policy will not correlate with payment on many or all such policies. This is one reason why insurance policies, especially in the property and casualty area, often sensibly exclude coverage of certain catastrophic losses—e.g., flood damages—that tend to occur in large bunches. Second, the company can pass on some portion of the risk through reinsurance contracts with a range of other carriers, so as to further diversify its risks across geographical regions and loss type. Finally, as noted above, the insurer can provide incentives (such as renewals at favorable rates) to insureds to take greater care and to avoid risky behaviors in order to reduce the probability of a claim.

In order to achieve these gains, it is necessary to find some way to pair insurers with insureds at reasonable cost. There is no single business strategy for discharging this critical search function. Many insurers hire in-house agents to sell their products. These agents often work extensively in the personal lines (home, auto, disability) in which the coverages offered are relatively standardized, and the competition in question usually comes down to the premium, the policy deductible, and the limits in light of the history of the insured (e.g., driving record). In other markets, however, the need for more specific or tailored forms of coverage is greater. Enter the independent brokers, who act as matchmakers between the insurers and the insureds. Brokers are typically hired by the insured as their agents, often taking on the task of finding suitable coverage from a full range of insurers with whom they have ongoing business relations.

The logic of the brokerage contract mirrors that of the basic insurance arrangement. The deal will go through only if both parties gain. On the one side, the insured must be satisfied that the broker's services in finding coverage and securing favorable terms cost less than the incremental gains the broker delivers from getting superior coverage, or a lower price, or some combination of the two. The relevant comparison does not ask whether the insured is better off with insurance than without it. Rather, it is whether the incremental costs of hiring a broker produce an insurance policy that is better than the insured could have acquired on its own, taking into account his own costs of search and negotiation. On the other side, the broker's expenses in finding a client suitable coverage must be lower than the expenses it incurs in rendering the services.

There is no one contractual formula for insurance brokerage, just as there is no one way to compensate employees for their labor. The most prevalent contract formula, however, calls for the brokerage commissions to be paid in one lump sum, set as a fixed percentage of the policy premium. The fee is generally high for the initial booking of the contract, usually in the neighborhood of 10 percent, but lower for repeat business, reflecting the benefits of stability in the business relationship. The broker who needs to perform fewer services receives a lower commission for his efforts. In many niche commercial markets, however, the information needed to provide for stable insurance markets is not available, so it is not all that surprising that in general about 4 to 5 percent of brokers' revenues come from contingent commissions.[2] As the name implies, these commissions are contingent on factors such as the profitability of the account to the insurer, or the duration or volume of the business that the broker has placed with the insured. They are typically paid by the insurer, rather than by the insured, as a reward for landing good accounts.

Choosing optimal insurance brokerage contract terms often turns on the complexity of the underlying business transaction. In many cases, particularly in personal insurance lines, the markets are relatively thick and sufficiently abundant, and reliable data on risk is available both for large populations and for the individual insurance applicant. Oftentimes individuals with little knowledge of the overall market turn to brokers who find it relatively easy to bring the two sides together. The effectiveness of this matching system is evidenced by the strong market position held by independent brokers. A.M. Best estimates that independent agents and brokers handled 67 percent of commercial lines property-casualty business and 33 percent of personal lines business in 2003. Estimates of the independent agent trade association put those numbers at 79.8 and 36.6 percent, respectively.[3]

These relatively routine transactions are frequently handled by standardized contracts, which are one way to provide assurance to inexperienced clients that they are not receiving less favorable treatment than other clients. But these transactions hardly tell the whole story. Most large commercial clients have unique risks that are hard to evaluate.[4] The terms of commercial contracts often vary by explicit agreement, as well as by the use of so-called manuscript policies, whereby standard print policies are altered, sometimes by hand, to take into account the specific circumstances of individual cases. The choice of policy limits and deductibles, the purchase of excess layers of coverage from other insurers,

2 J. David Cummins & Neil A. Doherty, *The Economics of Insurance Intermediaries* (May 20, 2005) (working paper), *available at* http://www.huebnergeneva.org/documents/cumminsdohertybrokers%205-20-05d.pdf.

3 *Id.* at 8, n. 4.

4 *Id.* at 7.

and the need to retain risks at certain levels all make it highly unlikely that a single standard form of insurance will work for all first and third party lines.[5] For different kinds of risks, different forms of coverage have to be devised.

One common feature of many of these complex deals is that it may be difficult to estimate the profit or loss that the insurer will receive from the transaction over the life of the policy. The usual public forms of information may be insufficient to allow the insurer to make an accurate estimation of the potential risk. This problem is pervasive in many complex commercial insurance transactions, regardless of whether contingent commissions are used, because the distinctive information about the nature and extent of any given risk is often chiefly, if not exclusively, in the hands of the insured, not the insurer. Accordingly, the law has imposed on the insured a duty to disclose all material circumstances that relate to the anticipated frequency and severity of losses.

The great risk in these cases is that of adverse selection, and it falls on the insurer, not the insured. Consider two parties that appear to present the same risk. The party with private information that his expected losses will be greater than the norm is more likely to purchase the insurance because he gets the standard rate even though he presents the higher risk. Yet any party with private information that his expected losses will be less than the norm is more likely to find that the insurance is not worth the cost. The low-risk customers exit the market, while the high risk customers stay.

THE GREAT RISK IN THESE CASES IS THAT OF ADVERSE SELECTION, AND IT FALLS ON THE INSURER, NOT THE INSURED.

The problem of adverse selection is endemic to all insurance markets. In certain difficult markets, insurers have to go to great lengths and considerable cost in order to counter the risk of adverse selection. One way to do this is to ask the broker, who has a closer relationship with its client, to vouch for the suitability of the insured as a risk. One way for the broker to demonstrate its belief that the insured is an appropriate risk is to bind itself to the transaction in such a way that the profit that it receives from the transaction will be reduced if the insured turns out to be of higher-than-expected risk.

The simplest way to achieve this result is to adopt the profit-based contingent commission, whereby some fraction of the profit that the broker hopes to receive from the transaction is held back and is conditional on the insurer making a profit out of the transaction. That scheme may have some use, but it is far from perfect as a sorting device because there are still likely to be many cases in which the original risk is, in fact, low, but the loss experience is nevertheless high. Yet the adverse financial outcomes do not mean that the broker has understated the

5 The former cover losses such as property damage or business interruption insurance. The latter cover various risks of liability to third persons.

relevant risks. The poor outcome (for the broker) could stem from a random roll of the dice.

The profit-based contingent type of payment is not unique to the insurance industry, and it can occur in any cases where two conditions are satisfied:

(1) there is a high variance in the potential payoffs to one party to the contract, and

(2) it is difficult for the party at risk to observe the underlying effort or risk associated with its trading partner.

Contingent payment systems in common use in other areas also reflect these dual concerns by pegging commissions not to completed transactions, as with ordinary brokerage fees, but to the profits generated by the deal. The most familiar version of this is the lawyer's contingency fee, which ties the service payment to the level of the recovery in the underlying case. Although this may look, in form, like the fee that a broker collects on selling a home, the underlying risk is surely much greater, given the possibility that the defendant prevails in a case so that the lawyer receives nothing at all. This is one reason why contingency fee lawyers work to obtain settlements which reduce the variance for both their clients and themselves. Viewed in this light, a contingent commission that is closely tied to the profitability of the transaction is likely to make sense in cases where the potential insureds have little or no previous track record.

It is possible to adopt similar fee arrangements in other markets, but their inherent complexity may well lead to a competitive disadvantage. Such appears to be the case with mortgage brokers who work, in the first instance, for consumers. These brokers are also compensated by a premium rate when they supply lenders loans that yield an interest rate in excess of par. This is called yield-spread premium. In order for brokers to collect that yield-spread premium, they should, in principle, have to reduce proportionally their upfront charges to clients. But the complexity of the market may prevent smooth adjustments, so that this payment scheme could harm consumers by giving mortgage brokers an added incentive to offer above-par loans to increase their compensation, without reducing their upfront charges.

This market failure may well have happened with mortgage brokers. In 2004, the U.S. Federal Trade Commission published its study on the effect of the disclosure of brokers' compensation agreements on consumer's choice. The study found that

---

"[i]f consumers notice and read the compensation disclosure, the resulting consumer confusion and mistaken loan choices will lead a significant pro-

> portion of borrowers to pay more for their loans than they would otherwise. The bias against mortgage brokers will put brokers at a competitive disadvantage relative to direct lenders and possibly lead to less competition and higher costs for all mortgage customers."[6]

That study was directed toward consumer markets where these risks are likely to be greater, even in cases of disclosure, which supports the conclusion that customers will shy away from products that they do not fully understand. But the persistence of the contingent commission in the commercial insurance context suggests that repeat players are more likely to surmount these information obstacles. So while it is sensible to predict the demise of these contracts in one market, it hardly follows that they will necessarily fall into disuse in other markets.

In those cases where some contingent commission survives, however, it should not be supposed that its use eliminates all conflict of interest between the parties. In both the brokerage and the lawyer situation, one risk that remains is that the agent will quit work too soon because the agent has to bear all the cost of additional work to land the contract or to recover a verdict, even if the agent only receives a fraction of the additional gain. Nonetheless, these conflicts are endured as a cost of doing business for at least two reasons. First, the parties who get paid under these arrangements tend, as repeat players, to develop strong reputations in their markets, and hence can be counted on to put out some extra effort today in order to improve their odds of getting additional business from the insurers tomorrow. Second, the alternative compensation systems could be worse because, in removing one set of conflicts, they create a second set that is more acute: the use of hourly fees could easily result in brokers and lawyers running up bills while doing little or no labor of any value. Additional factors may be operative in various individual cases.

In the end, therefore, contingent commissions in insurance, like other forms of contingent payments, may prove to be the best solution in certain critical segments of the market. The persistence of their use among commercial parties over long periods of time should be treated as some evidence of their economic value, especially when they take place between sophisticated parties who have the ready option to return to fixed commissions payable in full when the transaction is completed.

The arguments above help explain the use of commissions that are contingent on the level of profit the insurer achieves from the account. But there are still some unresolved issues. In some substantial fraction of cases, the contingent fees

6 James M. Lacko & Janis K. Pappalardo, *The Effect of Mortgage Broker Compensation Disclosures on Consumers and Competition: A Controlled Experiment*, Fed. Trade Comm'n Bureau of Econ. Staff Report (Feb. 2004), at ES-1, *available at* http://www.ftc.gov/os/2004/01/030123mortgagefullrpt.pdf.

are tied not to profit, but to volume or to renewals from a particular client. The use of volume measures is somewhat puzzling because it is not clear just what contingency is represented in volume transactions; that is, it is not clear why any subsequently acquired information is needed to determine the payout to the broker. A simple response might be to use, instead, a standard form of volume discount that just lowers the fixed commission at the front end, which may well be done in some cases.[7] To be sure, any institutional practice would have to account for both the advantages and disadvantages of placing large-volume accounts with a single insurer. On the plus side of the ledger, there are lower transactions costs to service the account, which could justify the higher payment based on the total amount of business generated by an individual client. But on the other side, writing extensive coverage for a single firm could expose the insurer to certain forms of correlated risk that are difficult in the abstract to calculate. If volume is achieved through multiple clients as opposed to a few large ones, then the insurer achieves greater diversification of his portfolio. Moreover, if insurers offer volume-based commissions, they must believe that it is profitable to insure large companies, all costs and benefits considered. If so, then contingent commissions based on volume might operate as a surrogate for contingent commissions on profits. And if they do not, then we should not expect their use to survive over time.

By the same token, the use of contingent commissions based on future renewals seems less difficult to understand. The renewal decision of the insurer represents its judgment that the account continues to be worth holding. The payment of a commission at this time is simply a statement that the initial account was more profitable than the insurer could have obtained by using only its own underwriting skills, and thus resembles a contingent commission based on profits. It is also worth noting that the size of that commission could be effectively limited because the insured, which knows its own payout history, could also insist on a reduction in the premiums that it pays.

There are some real business questions as to why and how these commissions are used. But whatever the uncertainties as to their effectiveness, it seems inappropriate to conclude that these commissions are simply and solely illicit covert devices to pay off brokers for steering business in a certain direction.

In sum, there are some real business questions as to why and how these commissions are used. But whatever the uncertainties as to their effectiveness, it seems inappropriate to conclude, as did Eliot Spitzer, that these commissions are simply and solely illicit covert devices to pay off brokers for steering business in a certain direction, which is thought to justify their ban even in the absence of collusion or nondisclosure. Any secret payment

7 For more discussion of the economic rationale of volume-based commissions, see Cummins & Doherty, *supra* note 2, at 17.

could have that effect, even if no contingencies are involved. More concretely, the fraud risk with contingent commissions looks to be no greater than that associated with ordinary brokerage commissions. There is trouble any time brokers receive secret payments for steering clients to higher-priced insurers for comparable coverage.

III. Twin Pitfalls of Contingent Commissions: Nondisclosure and Bid-Rigging

In light of the above arguments, it is not surprising that the recent litigation over contingent commissions generally does not rest on the assumption that these contracts were improper in any and all cases. Instead, the perceived risks are nondisclosure and bid-rigging. Each requires a few more words.

A. NONDISCLOSURE

The duty of disclosure is a pervasive norm in many commercial contexts as a source of protection to the uninformed party against conflicts of interest. Even if it is accepted that strangers deal with each other at arm's length, it is widely agreed that agents owe a fiduciary duty to their clients. As a matter of basic legal principle, contingent commissions should be subject to the standard duty to disclose. Normally, the agent is paid only by his principal. Yet now, in the absence of such disclosure, the broker would also receive a secret payment from the insurer with whom he is doing business. The obvious fear is that the agent's loyalty will follow the secret commission, thereby saddling the principal with an inferior insurance contract from which the agent makes a larger profit.

The case for requiring disclosure with contingent commissions is, to give one useful comparison, even stronger than it is in securities cases. Securities regimes require disclosure to the general market, and the information involved could concern all the risks and potentials of the proposed venture, which could easily prove valuable to the competitors of the firm. Moreover, deciding which disclosures are material and which are not is a delicate task which often results in massive litigation over what are often only trivial omissions in the disclosure process. But in the case of contingent insurance commissions, any disclosure is private and is made only to a single party. There is little to no risk of communicating vital information to the competitors of the firm.

Furthermore, it is possible to put sensible limits on what should be disclosed. In this regard, it would be unwise to insist that the entire contingent commission arrangement should be disclosed by the broker to the insured. A simple disclosure that the broker has received some contingent commission from the insurer should trigger the interest of any commercial insured, who can then ask for further information if that is desired for its own protection. According to Cummins and Doherty in their 2005 paper, the typical basic commission today covers part

of the fee to the broker, who still receives about 2 percent contingent commission directly from the seller.[8] That rate is an obvious subject of negotiation. At the same time, it must be remembered that disclosure of the contingent commission does not preclude the agent from insisting on the original deal.

It is also worth remembering that no disclosure obligation prevents a broker from seeking at any time to modify or terminate an agreement that no longer works to its advantage. To be sure, in the initial position, the duty of loyalty obligates the agent to take steps to improve the position of the client, even if these obligations work to his own disadvantage. His sole benefit comes from the compensation that the agreement supplies against these contingencies. If carrying out these duties generates losses to the agent that exceed any contract gains to the principal, it may make sense for both sides to terminate the relationship going forward because it reduces the net worth of the pair. Because no readjustment in fees can generate a net profit, the parties are better off without any arrangement at all. The precise distribution of the loss will turn on the specific contractual provisions and the relative bargaining skills of the parties. Where conflicts arise that are less acute, to the extent that the relationship is worth preserving, the two parties could agree to modify the agreement so as to keep it alive. The principal could ask the agent to alter his compensation schedule, to look for new trading or additional partners, or to explore a different set of contractual terms. Given the disclosure, the performance, termination, or renegotiation of any contingent commission contract follows ordinary contractual principles. Whether the terms of the policy changes after disclosure, however, is a business and not a legal concern.

The creation of any general disclosure obligation also must be put in context. That obligation is not the only source of protection available to potential insureds. In light of the general industry knowledge surrounding their use, any firm concerned with these lurking commissions is entitled, and surely prudent, to announce requests for proposals (RFPs) for competitive bids that raise the question front and center. These proposals routinely "request the complete disclosure of all compensation to be earned on the account. That compensation package will be expressed in terms of direct commission and/or fee, reinsurance, wholesale commission, contingent commission, etc."[9] Once a client asks point blank whether the broker has received contingent commissions from insurers, any refusal to answer that direct question honestly is a garden-variety version of fraud. An ambiguity from the common law disclosure obligation is effectively removed.

---

8 *Id.* at 20.

9 William J. Kelly, *Whom Do You Trust? The Selection, Evaluation and Compensation of Insurance Brokers*, Risk Mgmt. Mag. (Apr. 2006), *available at* http://www.rmmagazine.com/Magazine/PrintTemplate.cfm?AID=3077. Kelly is a risk manager and Chairman of the International Federation of Risk and Insurance Management Associations, and he attests to such practices.

Empirically, there is some evidence that buyers have started submitting RFPs in recent years.[10] But it is difficult to know for sure how much benefit these disclosure requests generate for insureds. One study conducted by Advisen, Ltd., in May of 2004 concludes that "less than 20 percent of the buyers at the 330 surveyed companies felt the level of disclosure they received from their insurance brokers about contingent commissions was entirely adequate."[11] But the response rate to this survey was not mentioned, nor do we know the fraction of premium volume covered by the fully satisfied clients. Nor, for that matter, do we know the baseline disclosure rates for other sorts of brokerage payment systems. A second study in November of 2004 also found, with a low response rate of 16 percent, that "57 percent of the 684 respondents believe their brokerage firms do not fully disclose all sources of income related to insurance transactions."[12] The same survey also found that "nearly two-thirds of the respondents said they were not yet considering changing brokerage firms," which could be evidence that the perceived shortfalls in disclosure are less harmful than might be supposed. The acid test on this matter is how, with risk exposure held constant, insurance premiums in transactions with disclosure stack up in dollar and cent terms against identical transactions in which either no disclosures or inadequate disclosures have been made. Is there in fact a price differential that hurts the insured? How often, and in what cases? On this point there is, to my knowledge, little or no systematic research.

The unsettled market situation clearly is capable of improvement especially in light of the recent litigation. One possibility is for major players in the brokerage industry to issue general statements of policy as to whether they do or do not accept contingent commissions from insurers. Given that the economic case for using contingent commissions is uncertain, many firms might choose to clear the air by announcing that they will not resort to them at all. Just that result was undertaken, for example, by Willis in the aftermath of the New York investigation. The *Willis Client Bill of Rights* states categorically that "Willis will not accept contingency compensation from insurers."[13]

But what about those cases in which this explicit disclaimer has not been made. Suppose, for example, that for some reason an insured remains ignorant

10 *Id.*

11 *See* Press Release, Advisen, Majority of Commercial Insurance Buyers Say Contingent Commission Practice Is Conflict of Interest (May 24, 2004), *available at* https://www.advisen.com/HTTPBroker?action=jsp_request&id=articleDetailsNotLogged&resource_id=28386431.

12 Press Release, Advisen, Advisen Survey Finds Corporate Insurance Buyers Seek Transparency on Broker Compensation, Transaction Terms and Prices (Nov. 16, 2004), *available at* https://www.advisen.com/HTTPBroker?action=jsp_request&id=articleDetailsNotLogged&resource_id=36216473.

13 Willis Client Bill of Rights, *at* http://www.willis.com/The%20Way%20We%20Do%20Business/extras/ClientBillofRights_letter.pdf.

about the use of contingent commission in its contracts. Nonetheless, the extent of its ensuing harm is hard to determine, given the fact that other institutional safeguards also help to protect the uninformed insured. The most obvious such safeguard is competition itself. That competition expresses itself in many ways. Although systematic evidence is scant, some large insureds may choose to work through more than one broker—some national and some regional—for different portions of their insurance portfolio.[14] This strategy of segmentation means that a firm does not turn all of its business over to one large brokerage house, but can instead parcel its accounts by size and complexity to multiple brokers.[15] The constant input from many brokers provides observable bases for price comparisons, as each current broker seeks to expand its fraction of the overall business from an established client. Nor is potential competition limited to the stable of established brokers. Other brokers, anxious to gain new business, are also able to review the prices the insured pays, and would be able to let a prospective client know, if such is the case, how poorly it is being treated.

The ability to shift accounts between insurers could thus take place even if the insured has no knowledge of an undisclosed secret commission. In equilibrium, these competitive forces are not likely to lead to perfect pricing, for the costs of search are positive, even for experienced businesses. But except in the highly unlikely circumstance that every competitive firm (remember collusion has been put to one side for the moment) engages in the same practice of nondisclosure, the actions of these competitors still remain an important protection against abuse.

---

14 Quoting Kelly:

> I have, on rare occasions, moved discreet pieces of business to niche brokers that have developed a particular specialty. For example, a previous employer had a relocation subsidiary, a firm that facilitates executive moves for third party corporations. As part of the business, we had a portfolio of approximately 9,000 residential homes throughout the country. A representative of a very small Connecticut brokerage offered to bid on the portfolio. As the property and liability coverages were placed by a top brokerage firm, I did not expect the niche company to be successful. However, he returned with a three year, non cancelable program, from a top rated insurer with absolutely compelling cost savings.

See William J. Kelly, Everything I Ever Wanted to Say to an Insurance Broker, Address to Willis Exceptional Producers' Meeting (Apr. 14, 2000), *available at* http://www.ifrima.org/DOWNLOAD/WILLISINSURERBROKERAGE.PDF.

15 Quoting Kelly, *supra* note 9:

> . . . Larger corporations that have significant insurance needs in each of the major coverage areas of property, casualty, and management liability may, and often do, elect to utilize the services of multiple brokers. As these insurance programs are usually discreet from each other and led by different specialist insurance companies, they can be separately managed through different insurance brokerage firms.

This approach allows the insured to remain both a client and a prospect, with each broker continuing to vie for that portion of the risk they do not have and with no one provider becoming overly comfortable in the relationship.

It is also important to ask about the importance of the disclosure option in the ordinary course of business. It is surely of great moment when the insured, as principal, pays commissions to the broker at the normal rate, for then he has no reason to suspect that this broker has received a commission from the party on the other side of the transaction. But if the insured sees an unusually low stated commission, then, based on past experience, he might be able to infer that the broker has received some compensation from the insurer, for otherwise the transaction does not offer enough gain for the broker to accept it.

The size of the direct commission could prove relevant in the event of litigation for damages once an insured learns of a previous nondisclosure. A disappointed insured could sue, for example, the broker, to turn over the contingent commission, or perhaps to obtain a reduction in rates to the level that they might have been if the full disclosure had been made, so that the client could have tested the market with other brokers. As in all such cases, the disclosure serves as the basis of a successful claim only if the insured can prove that the nondisclosure caused some economic loss. The broker could, therefore, be free to argue that an unusually low commission provided sufficient information of the contingent commission to constitute effective notice to the principal, thereby implying tacit acquiescence. In some cases, it seems at least an arguable question of fact whether sophisticated purchasers would believe that a highly complex and delicate brokerage transaction would generate only a below-normal payoff to the successful broker. These uncertainties about causation, however, are something that both sides would do best to avoid. The strong case for routine disclosure of contingent commissions makes sense precisely in that it eliminates the need to resolve the messy problems of proof that inevitably arise in the event of nondisclosure.

THE STRONG CASE FOR ROUTINE DISCLOSURE OF CONTINGENT COMMISSIONS MAKES SENSE PRECISELY IN THAT IT ELIMINATES THE NEED TO RESOLVE THE MESSY PROBLEMS OF PROOF THAT INEVITABLY ARISE IN THE EVENT OF NONDISCLOSURE.

B. ANTITRUST RISK

The second risk associated with the use of contingent commissions involves collusion or bid-rigging, of which there was incontrovertible evidence in the New York cases against the leading brokerage houses. Here the illegality of the practice is unquestioned under the antitrust law, which imposes strong sanctions against these forms of collusion. But in this setting, the objection to outright collusion also rests on principles of ordinary contract law. No insured would ever consent to a transaction whereby a broker presents it with phony high bids from nominal competitors just to create the illusion of competitive bidding. At the very least the industry collusion is aggravated by fraud. To be sure, even if the market for using contingent commissions is complex, the antitrust issues are not: these cases are simple instances of price fixing and market division. They do not

offer any difficult attack on the standard vertical arrangement between a broker and an insurer. As a first approximation, the horizontal restraint of trade looks every bit as illegal in these two-sided insurance markets as they do anywhere else.

But once the illegality is established, other questions still remain. What kind of remedy, either civil or criminal should be imposed in these cases? In this situation, it is useful to distinguish between imposing sanctions against the individuals who knowingly engaged in the wrongful transactions, and imposing sanctions against the brokerage house or insurer at which they worked. The former question is straightforward because the actual participants to the scheme do not appear to have any substantive defense against either civil or criminal sanctions, although it is always wise to examine the full record to be sure.

The liability of the brokerage houses is more complicated. On the civil side, the actions in question were surely within the scope of employment, so damage awards or other civil sanctions are surely appropriate. But the criminal side is much more difficult. If the bid-rigging were authorized by persons higher up in the firm, the criminal sanctions would properly reach up through the firm hierarchy. But, even if that were the case, the question of criminal responsibility of the firm, as an entity, for the actions of its employees is a separate matter. It is highly debatable whether any firm—which necessarily means the innocent shareholders in public corporations—should be asked to pay the price for wrongs in which they did not participate.

Even under current law, which uses broad definitions of vicarious liability to rope in corporate defendants, the question of prosecutorial discretion looms large. In principle, any decision to launch a criminal investigation against the firm is likely to depend in large measure on the frequency and pattern of the bid-rigging incidents, which is, for example, the situation in the New York cases. If these incidents were confined to a small number of key people on only a few occasions, the corporate criminal sanction (which could lead to a firm dissolution Arthur Andersen-style) seems to be massive overkill. It is far better to stick with the individual sanctions that do not pose that risk. But if the bid-rigging practices were endemic, the balance starts to shift. Exactly where the balance should tip in any case is hard to say.

The evidence on the frequency and distribution of wrongs within the brokerage houses is, however, important for an additional reason. It gives some guidance as to the level of appropriate fines. The remedy of choice in the New York settlements was restitution of the revenues received by the firms in all their contingent fee transactions.[16] The argument made in favor of dollar for dollar resti-

16 Press Release, New York State Attorney General & New Your State Insurance Department, Insurance Broker Agrees to Sweeping Reforms (Jan. 31, 2005), *available at* http://www.oag.state.ny.us/press/2005/jan/marshsettlement_pr.pdf.

tution of all contingent commissions paid was that these were "almost pure profit" derived from wholly corrupt transactions, which were used solely to steer business to the insurance company that paid the largest contingent commissions.[17]

It is unlikely that this system of rough justice hit on the right remedy, because the proper calculations are more difficult to make than this simple restitution formula suggests. The first step is to figure out the extent to which the bid-rigging increased the cost of premiums to the insured or, in the alternative, lowered the level of coverage for any given level of premium. Clearly, there should be no restitution for contingent commissions paid without taint of bid-rigging, at least in cases of full disclosure. Even in those cases where the bids were rigged, the proper measure of damages is not the amount paid under the contract. Rather, it is solely the price increment from the conspiracy in restraint of trade that should be trebled, not the full amount of the commissions paid. That calculation could prove difficult if there were some partial offset in the direct premiums or commissions paid by the insurer. It is possible that these supracompetitive profits, once trebled, were large enough to wipe out the revenues from these transactions. But any grand assertion that the entire contingent premiums counted as "almost pure profit"[18] could be correct only if there would have been no reduction in the base premiums paid on this policy in the absence of the contingent commission. Yet that contention seems highly questionable. Cummins and Doherty report that "[p]remium-based commissions account typically for about 10-11% of premiums, compared with an average of 1-2% of premiums for contingent commissions."[19] The elimination of contingent commissions in these contexts is likely to produce at least some adaptive response from brokers, who in all likelihood would charge at least the same flat rate as before, and perhaps more. In fact, our knowledge of these various practices does nothing to rule out the possibility that eliminating contingent commissions in competitive markets could lead to higher brokerage fees for businesses, if there are any losses of efficiency advantages. How this plays out, given the available state of knowledge, is uncertain.

Any grand assertion that the entire contingent premiums counted as "almost pure profit" could be correct only if there would have been no reduction in the base premiums paid on this policy in the absence of the contingent commission. Yet that contention seems highly questionable.

17 *Hearing on Insurance Brokerage Practices Before the Subcomm. on Financial Management, the Budget and International Security of the S. Comm. on Governmental Affairs, 108th Cong. 7* (2004) (statement of Eliot Spitzer, Attorney General, New York State), *available at* http://www.oag.state.ny.us/press/statements/insurance_investigation_testimony.pdf.

18 *Id.*

19 Cummins & Doherty, *supra* note 2, at 2.

In light of these complexities, the correspondence between the wrong and the remedy should be proved, and not presumed. The risk here is that the threat of criminal prosecution leads to the imposition of remedies beyond those needed to promote market efficiency. The subject of prosecutorial discretion is beyond the scope of this essay, but the dangers of overdeterrence should never be overlooked, especially in the prosecutor's hour of triumph. The major risk is that the consequences of any decision to prosecute are necessarily amplified because prosecution triggers a broad range of collateral regulatory responses. Insurance commissioners in every state have to investigate whether to impose additional sanctions—loss of licenses and tighter reporting requirements, for example—once the indictment has been filed. In some jurisdictions the licenses could be pulled immediately. These sanctions impose severe penalties even if the charges are dismissed as unfounded down the road. The irony is that a defendant has stronger protections against the conviction than against the indictment, even though the indictment poses far greater risk. Given this giant lever, private brokers could easily make settlements that overstate the extent of any social loss (even if trebled) attributable to its bidding practices. The social losses from over-enforcement, moreover, cannot be lightly ignored if it leads a brokerage firm to avoid business practices that might have a high expected social value. How to control prosecutorial discretion is beyond the scope of this paper, but the problem will not quietly disappear in the near future. Its systematic risks extend far beyond the risks in contingent commission cases.

IV. Legislative Reform

As noted, legislative reform on the matter of contingent commissions warrants careful attention. As so often happens, the impulse for legal reform often takes place even when the existing laws have imposed heavy sanctions on the parties. And all too often the inquiry is not whether any shortfall in current enforcement should be fixed by the more effective use of existing institutions and sanctions against wrongdoers. Instead the usual public reaction is to ask what new sanctions could be added to the arsenal to nip various forms of misconduct in the bud—without asking, however, whether tougher sanctions will stifle beneficial conduct as well. As befits this situation, the pressure is placed on both the disclosure and the antitrust fronts, and each requires somewhat different treatment.

On the question of disclosure, it is unclear how often contingent commissions have been disclosed. For these purposes, suppose that no disclosures have been made, but that no bid-rigging has taken place as well. In these cases, the magnitude of the problem is uncertain given that competitive forces have remained operative. It is always hard to know whether any consistent lack of disclosure should be treated as strong evidence of a long-term problem, or whether it just means that the level of market distortion is relatively small. Indeed one reason to require the disclosures is that once they are made, it removes any need to spec-

ulate over this difficult counterfactual. Nor need any broker wait for outside parties to impose a duty to disclose. Their first line of defense could always be voluntary disclosures that make the legislative or administrative intervention largely unnecessary. In this regard, note that the settlements with New York preclude the use of undisclosed commission by the signatories. If non-signatories follow suit, then the problem has taken care of itself. The only possible efficiency loss here arises if these undisclosed commissions have positive economic value, at which point the legislative ban results in unnecessary efficiency losses. On the antitrust side, there is no need for any change in the appropriate legal rule because the bid-rigging was already illegal under tough laws in effect at the time it was practiced.

On the antitrust side, there is no need for any change in the appropriate legal rule because the bid-rigging was already illegal under tough laws in effect at the time it was practiced.

The hard question that remains is whether Congress or the states should ban the use of all contingent commissions, even when the broker has complied with all disclosure and antitrust regulations. That objective has been touted in New York, so much so that the ban on contingent commissions in all contexts is now regarded as a major legislative objective. The case for the legislation is not made out by the demonstration of either nondisclosure or bid-rigging, because the use of contingent commissions requires neither. Surely, we would not ban standard commissions in their entirety because of nondisclosure or bid-rigging, so why do it in the case of contingent commissions? The preferable strategy looks therefore to avoid such an overbroad prohibition. To that position there are possible objections. The first of these is that the risk of overbreadth is minor because contingent commissions turn out to play little role in a competitive market with full disclosure. At this point the ease of enforcement might in principle justify the broader restraint on conduct. Why not ban these commissions if their only use is to distort insurance markets by the illicit steering of business? Yet the result represents some measure of regulatory excess if any efficiencies do follow from the use of contingent commissions in their familiar historical niches. It is a taller order to explain why routine business practices should be banned across the board than it is to require their disclosure to clients. Unfortunately, the New York initiative does not discuss why these contingent commission contracts might prove valuable in some contexts. If there is any evidence that the practice long predates the recent abuses in New York, then the best that can be said is that the case for the total ban is "not proven". ◀

Volume 3 | Number 1 | Spring 2007

Competition Policy International

The *Impala* Judgment: Does EC Merger Control Need to Be Fixed or Fine-Tuned?

Rachel Brandenburger and Thomas Janssens

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

The *Impala* Judgment: Does EC Merger Control Need to Be Fixed or Fine-Tuned?

Rachel Brandenburger and Thomas Janssens

In its *Impala* judgment last year, the Court of First Instance annulled a European Commission unconditional merger clearance decision for the first time. As a result, the Commission is having to carry out a new investigation into a transaction that closed over two years ago. In this judgment, the Court applied the three-limbed test for collective dominance from *Airtours* judgment. But this time it assessed strengthening, as opposed to creation, of collective dominance. Importantly, the Court made it clear that the Commission must base a clearance on equally solid grounds as a prohibition.

We examine a number of the fundamental issues that the *Impala* judgment has raised. These have significance beyond the factual context of the case itself, both for the way the Commission must conduct its investigations and for the role of judicial review by the EC courts. We conclude by suggesting some changes in Court and Commission practices that would, we believe, strengthen the effectiveness of EC merger control.

The authors are partners in Freshfields Bruckhaus Deringer.

I. Introduction

On July 13, 2006, the Court of First Instance of the European Communities (CFI)[1] annulled the clearance[2] by the European Commission of the joint venture between Sony's and Bertelsmann's global recorded music businesses.[3] This was the first (and so far only) time the CFI had overturned an unconditional Commission clearance decision under the EC Merger Regulation.[4] As a result of the judgment, the Commission is having to carry out a new investigation into the SonyBMG joint venture, which has been in operation since August 2004. The CFI's judgment, which Sony and Bertelsmann have appealed to the European Court of Justice (ECJ),[5] raises a number of fundamental issues for EC merger control that have significance beyond the factual context of the case itself and the way in which the Commission conducted that particular investigation.

II. The Commission's U-Turn

The *Sony/BMG* Decision appears to have been a remarkably reluctant clearance. Rather than explaining why the SonyBMG joint venture should be approved, the Decision concluded that the evidence was not sufficient to support a prohibition. The CFI held that the Decision had departed from the Commission's Statement of Objections (SO) without giving sufficient reasons for this change of mind, notwithstanding the arguments the Commission offered in support of its clearance during the court proceedings.

In its SO of May 24, 2004, the Commission had reached the preliminary view, based on strongly worded adverse findings of facts, that the SonyBMG joint venture would strengthen an existing position of collective dominance (coordinated effects) in both the physical and the online recorded music markets. Following the parties' response to the SO and an oral hearing that took place on June 14 and 15, 2004, the Commission reversed its position, described subsequently by the CFI as a "fundamental U-turn",[6] and cleared the SonyBMG joint venture on July 19, 2004 without, however, fully explaining the reasons for the

1 Case T-464/04, Independent Music Publishers and Labels Association (Impala) v. Commission [hereinafter *Impala* or *Impala* judgment] (not yet reported) (2006).

2 Commission Decision 2005/188/EC [hereinafter the Decision], Case COMP/M.3333, *Sony/BMG* [hereinafter *Sony/BMG*], 2005 O.J. (L 62) 33.

3 Sony's activities in Japan were not contributed to the joint venture.

4 In 2001, the CFI annulled the unconditional clearance of a merger under the European Coal and Steel Community Treaty, in RJB Mining v. Commission, Case T-156/98, 2001 E.C.R. II-337.

5 Appeal brought on October 10, 2006, Case C-413/06 P, 2006 O.J. (C 326) 25.

6 *Impala*, *supra* note 1, at 283.

U-turn in its Decision. Rather, the Commission concluded that its "detailed analysis [...] showed some indications of coordinated behaviour which were as such, however, not sufficient to establish existing collective dominance"[7] and approved the transaction on that basis.

On December 3, 2004, Impala, an association of independent music companies, lodged an application for annulment of the Decision, requesting that the CFI adjudicate the case under the expedited (or "fast-track") procedure for merger appeals.[8]

III. The CFI'S Key Criticisms

As a court of review rather than appeal,[9] the CFI's task was not to rehear the facts of the case nor to establish whether the conditions of collective dominance in the recorded music industry were fulfilled, but to review how the Commission had conducted its investigation and reached its conclusions.

In its judgment, the CFI criticized the way the Commission had conducted its investigation and defended the Decision in court. In particular, the CFI held that the Commission's finding that the transaction would not strengthen existing collective dominance was inadequately reasoned, and the CFI pointed out numerous inconsistencies between the Decision, the Commission's SO and its submissions before the CFI. Although lack of reasoning would have been a sufficient ground, in itself, for annulment of the Decision, the CFI also ruled that the Decision was vitiated by a manifest error of assessment in so far as "the elements forming the basis of the Decision did not constitute all the relevant data that must be taken into consideration and were not sufficient to support the conclusions drawn from them."[10]

THE CFI POINTED OUT NUMEROUS INCONSISTENCIES BETWEEN THE DECISION, THE COMMISSION'S SO AND ITS SUBMISSIONS BEFORE THE CFI.

The CFI considered the Commission had ignored the elements of existing collective dominance previously postulated in its SO, and had based its clearance on insufficiently solid evidence—an error it could not rectify in the CFI proceedings. In particularly harsh terms, the CFI noted that the Commission "cannot

7 *Impala*, supra note 1, at 109.

8 Court of First Instance, Rules of Procedure, 2000 O.J. (C 34) 39, at art. 76a.

9 EC Treaty, at art. 230 (4).

10 *Impala*, *supra* note 1, at 542.

suppress certain relevant elements on the sole ground that they might not be consistent with its new assessment"[11] and that:

> "explanations proffered during the proceedings before the Court or, a fortiori, checks relating to an essential aspect of the Decision cannot compensate for a lack of investigation at the time of the adoption of the Decision and eliminate the manifest error of assessment by which the Decision is thus vitiated, even if that error had no effect on the outcome of the assessment."[12]

The CFI also criticized the fact that the analysis in the Decision concerning the possible creation (as opposed to strengthening) of collective dominance was "extremely succinct"[13] and noted that the Commission's "few observations, which are so superficial, indeed purely formal, cannot satisfy the Commission's obligation to carry out a prospective analysis."[14]

IV. *Airtours* Expanded?

Throughout its judgment, the CFI referred to the three-limbed test for the assessment of collective dominance, established in *Airtours*.[15] Noting that the *Airtours* case law was originally developed in relation to the assessment of the risk of the creation of collective dominance (which entails an entirely prospective analysis), the CFI applied the *Airtours* criteria in the Impala judgment also to the strengthening of existing collective dominance. This, according to the CFI, requires "a concrete analysis of the situation existing at the time of the adoption of the Decision" and thus "must be supported by a series of elements of established facts, past or present, which show that there is a significant impediment of competition on the market."[16] In this respect, the CFI suggested, in an obiter dictum, that the existence of collective dominance (based on the three conditions of *Airtours*) could be established indirectly on the basis of "what may be a very mixed series of

11 *Id.* at 300.

12 *Id.* at 458.

13 *Id.* at 525.

14 *Id.* at 528.

15 Case T-342/99, Airtours v. Commission [hereinafter *Airtours*], 2002 E.C.R. II-2585.

16 *Impala*, *supra* note 1, at 250.

indicia and items of evidence relating to the signs, manifestations and phenomena inherent in the presence of a collective dominant position."[17] According to the CFI, price parallelism might be an indicator of collective dominance in some cases. In the absence of an alternative reasonable explanation:

> "close alignment of prices over a long period, especially if they are above a competitive level, together with other factors typical of a collective dominant position, might [...] suffice to demonstrate the existence of a collective dominant position, even where there is no firm direct evidence of strong market transparency, as such transparency may be presumed in such circumstances."[18]

In reviewing the Decision, the CFI focused on the first two limbs of the *Airtours* test: the degree of market transparency and the possibility of retaliation. The existence of countervailing factors (the third limb of the test) was not examined, as they were not covered in the Decision and, therefore, not part of Impala's appeal.

The *Impala* judgment confirms that the *Airtours* test can be applied to strengthening of existing collective dominance as well as to the prospective analysis of creation of collective dominance.

IV. The SonyBMG Re-Examination: Old Rules Applied in a New Context

The CFI did not require the parties to dissolve their joint venture.[19] Instead, the Commission is having to conduct a re-examination of the SonyBMG joint venture.

Sony and Bertelsmann parties re-notified their joint venture to the Commission on January 31, 2006—some six months after the *Impala* judgment.[20] As the original notification was made prior to May 1, 2004, when the revised EC Merger Regulation entered into force, the re-examination of the joint venture is

17 *Id.* at 251.

18 *Id.* at 252.

19 As a court of review, the CFI does not have powers to order this. The Commission may order the dissolution of an implemented merger if the merger has been declared incompatible with the common market (*see* Article 8(4) of the EC Merger Regulation).

20 2007 O.J. (C 29) 12.

governed by the procedural timetable and substantive "dominance" test of the previous EC Merger Regulation, but must take account of current market conditions. Interestingly and uniquely, this enables the Commission to assess the impact that the joint venture has had on competition over the past two and a half years since it started operating—the ultimate natural experiment!

The Commission's re-examination is taking place in parallel with Sony and Bertelsmann's appeal to the ECJ to overturn the *Impala* judgment. The re-examination is not suspended by the appeal (see Figure 1).

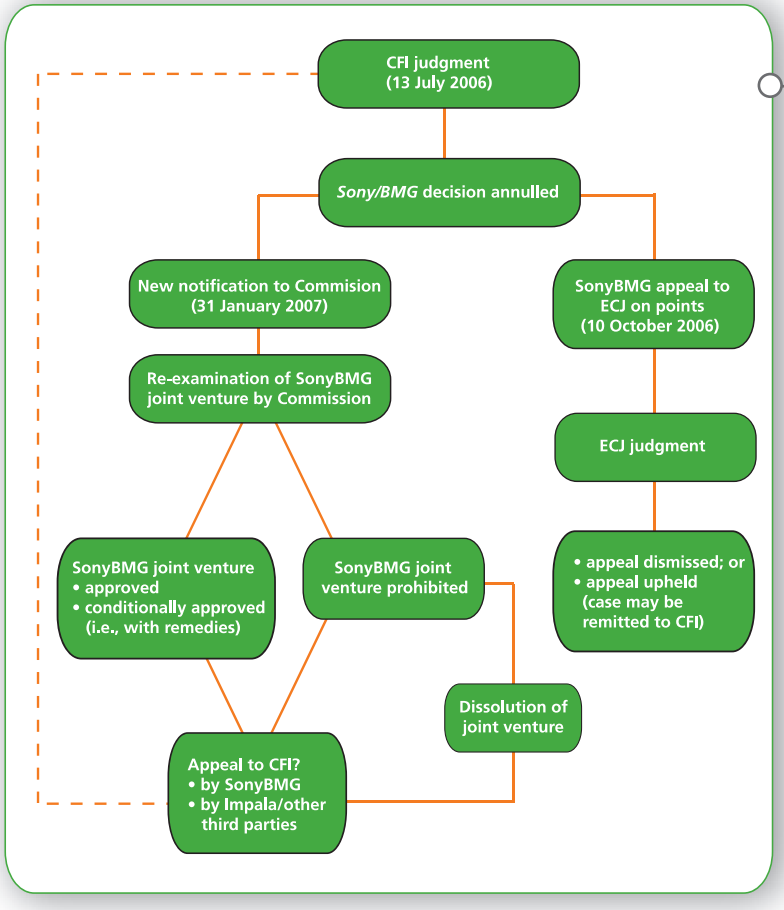


Figure 1

SonyBMG re-examination

V. Raising the Bar for EC Merger Approvals?

The *Impala* judgment sent shockwaves through the EC merger control regime, similar to those that followed the "trilogy" of CFI annulments of Commission prohibition decisions in 2002.[21] While the *Impala* judgment is very fact-specific, it raises a number of questions that are of broader relevance to the way in which the Commission conducts its investigations.

First, does the Commission's SO in effect constitute a benchmark for its final decision? The *Impala* judgment does not necessarily mean that the Commission's preliminary findings in an SO on the facts and on their legal significance are set in stone. Indeed, the CFI recognized that the Commission "is not obliged to explain any differences by comparison with the statement of objections, since that is a preparatory document containing assessments which are purely provisional in nature."[22] But—somewhat in contrast—the judgment does suggest it is incumbent on the Commission to justify any material departure from its initial objections, by refuting them on the basis of evidence that is "at the very least [...] particularly reliable, objective, relevant and cogent".[23] Thus, if the Commission expresses its SO in strongly adversarial terms, as it did in *Sony/BMG*, subsequent reversal of its position in the final decision may become more complicated and time-consuming than in the past. Alternatively, the Commission may refrain from adversarial SOs in the future.

If the Commission expresses its SO in strongly adversarial terms, as it did in *Sony/BMG*, subsequent reversal of its position in the final decision may become more complicated and time-consuming than in the past.

Second, is the Commission now required to conduct a new market investigation following the merging parties' response to the SO? The *Impala* judgment makes it clear that, to support a "U-turn", the Commission cannot rely on information provided only by the merging parties without at the same time seeking views from third parties as that, in the CFI's view, would amount to delegating "without supervision, responsibility for conducting certain parts of the investigation to the parties to the concentration."[24] But, the standard EC Merger Regulation timetable does not allow for a meaningful further market investigation at such a late stage in the proceedings. Are we, therefore, going to see extended investigations in such circumstances?

---

21 *Airtours*, *supra* note 15; Case T-310/01, Schneider Electric v. Commission [hereinafter *Schneider*], 2002 E.C.R. II-4071; Case T-5/02, Tetra Laval v. Commission [hereinafter *Tetra (CFI)*], 2002 E.C.R. II-4381.

22 *Impala*, *supra* note 1, at 285.

23 *Id.* at 414.

24 *Id.* at 415.

Finally, has the CFI raised the standard of proof for merger clearance decisions? In the *Sony/BMG* Decision, the Commission concluded that it had not found sufficient evidence of competitive harm, and it therefore approved the transaction. But the CFI considered this was not enough. This raises an important question: is there a presumption in EC law that mergers are compatible with the common market? Advocate-General Tizzano in *Tetra Laval* indicated that there was when he said: "in the case of uncertainty as to whether or not the transaction is compatible with the common market, the interest of the undertakings seeking to make the merger must prevail."[25]

The CFI, in its *Impala* judgment, has not departed from this or reversed the presumption, thus requiring merging parties to demonstrate why their transaction should be approved, as some have claimed. But it has confirmed that the Commission must carry out its analysis with great care. This implies not only a requirement to base its analysis on "sound economics" and "hard evidence", as the CFI famously stated in *Airtours*,[26] but also the need to conduct, and, as importantly, be seen to conduct, its investigations in a robust and unbiased way.

There are already signs that, as a matter of practice, the Commission may be changing its approach in light of the CFI's *Impala* judgment. In particular, the Commission's information requests in merger cases are becoming more lengthy and its merger analysis increasingly document- and data-intensive, increasing the burdens on both merging parties and third parties.

In *Impala*, the CFI did not address what the Commission should do if, notwithstanding a thorough investigation, the evidence does not clearly point one way or the other. This situation could arise increasingly as good counseling reduces the number of obvious prohibition cases that see the light of day.

VI. The Role of Judicial Review

The *Impala* judgment also confirms, once again, that, nowadays, judicial review is an integral part of EC merger review. An increasing number of high-profile merger decisions are challenged, whether by third parties or the merging parties, and the EC courts have been generous in accepting the admissibility of appeals against merger decisions.

25 AG Opinion (Tizzano) of May 25, 2004, Case C-12/03P, Commission v. Tetra Laval [hereinafter *Tetra (ECJ)*], 2005 E.C.R. I-987, at 79. According to Advocate-General Tizzano, by stipulating that, if the Commission does not make a decision in good time (*see* Article 10(6) of the EC Merger Regulation), then a concentration must be deemed to be authorized, the EC legislature demonstrates as a matter of fact that it considers that there is such a presumption. The ECJ did not, however, address this question in its judgment.

26 *See Airtours*, *supra* note 15; *see also Schneider*, *supra* note 21 and *Tetra (CFI)*, *supra* note 21.

Nevertheless, the judicial control exercised by the CFI and ECJ does not amount to a full appeal. It is limited to a review of the Commission's decisions based on limited grounds of annulment.[27] In their appeal to the ECJ, Sony and Bertelsmann have argued that the CFI exceeded the scope of judicial review by substituting its own assessment for that of the Commission. The ECJ has previously recognized that the provisions of the EC Merger Regulation:

> "confer on the Commission a certain discretion, especially with respect to assessments of an economic nature, and that, consequently, review by the Community Courts of the exercise of that discretion, which is essential for defining the rules on concentrations, must take account of the margin of discretion implicit in the provisions of an economic nature which form part of the rules on concentrations."[28]

But, as the CFI pointed out in its *Impala* judgment,[29] the ECJ has also confirmed the importance of judicial review, stating that the Commission's margin of discretion "does not mean that the Community Courts must refrain from reviewing the Commission's interpretation of information of an economic nature."[30]

A key concern for the parties to a merger remains the ability to obtain judgment within a short time period. Although the expedited procedure was followed in *Impala*, it took 24 months from the Commission's decision on July 19, 2004 to the CFI's judgment of the CFI on July 13, 2006.[31]

The *Impala* judgment has reignited the debate about the need for a specialized EC competition court,[32] or even the introduction of US-style merger litigation allowing for a full appeal, rather than limited judicial review, of Commission decisions.

27 EC Treaty, at art. 230 (4).

28 *Tetra (ECJ)*, *supra* note 25, at 38.

29 *Impala*, *supra* note 1, at 328.

30 *Tetra (ECJ)*, *supra* note 25, at 39.

31 Unusually, the CFI made Impala bear 75 percent of its own costs of the proceedings, as its behavior was found to be inconsistent with an expedited procedure.

32 Article 225a of the EC Treaty enables new tribunals to be established as courts of first instance for specific areas.

VII. Conclusion

The *Impala* judgment is a further chapter in the line of CFI cases that began with *Airtours*, confirming the Commission's duty to conduct its merger investigations thoroughly and to base its decisions on solid grounds backed by complete and accurate information. The judgment also confirms that merging parties are increasingly having to take account of the risk of litigation, and that third parties can play a significant role both during the Commission's investigation and before the EC courts. For the Commission, the challenge now will be to take the CFI's criticisms into account while still respecting the rights of all the parties involved in its investigations. Conducting U.S.-style merger investigations within the straightjacket of the EC Merger Regulation's timetable and subject to a requirement to write fully-reasoned decisions[33] may be asking the impossible of the Commission. In turn, this may lead to an increased willingness on the part of both the Commission and merging parties to settle difficult cases, potentially resulting in over-enforcement.

RATHER THAN REQUIRING A COMPLETE OVERHAUL OF THE CURRENT PROCEDURES, SOME RELATIVELY SMALL CHANGES TO THE WAY THE COMMISSION CONDUCTS ITS MERGER INVESTIGATIONS MAY CONTRIBUTE TO THE IMPROVED EFFECTIVENESS OF EC MERGER CONTROL. BY ADOPTING AN INVESTIGATIVE APPROACH, RATHER THAN AN ADVERSARIAL ONE, THE COMMISSION MAY AVOID KEELING OVER WHEN TAKING A "U-TURN".

Rather than requiring a complete overhaul of the current procedures, however, some relatively small changes to the way the Commission conducts its merger investigations may contribute to the improved effectiveness of EC merger control. For example, by adopting an investigative approach, rather than an adversarial one (especially with regards to its SO), the Commission may avoid keeling over when taking a "U-turn". Similarly, some adaptations to the CFI's procedures, such as those relating to its working languages and its ability to enforce the accelerated timetable in merger cases, may go a long way to addressing the concerns that have been voiced following the *Impala* judgment. ◀

33 This is not a requirement in the United States, although the U.S. antitrust agencies sometimes issue a brief statement in relation to a merger clearance in important cases.

Volume 3 | Number 1 | Spring 2007

Competition Policy International

Review of Michael Whinston's *Lectures on Antitrust Economics* (MIT Press, 2006)

Massimo Motta

Reprinted in *Competition Policy International* (print ISSN 1554-0189, online ISSN 1554-6853), Spring 2007, Vol. 3, No. 1. *Competition Policy International* is a free publication. To order or download additional copies, visit eSapience.org.

Review of Michael Whinston, *Lectures on Antitrust Economics* (MIT Press, 2006)

Massimo Motta

Michael Whinston is one of the economists who have contributed most to the understanding of antitrust issues. His works, alone or with co-authors (especially Douglas Bernheim and Ilya Segal), have shed light on such issues as exclusive contracts, tying, and multi-market collusion among others. For this reason, the publication of his book *Lectures on Antitrust Economics* is an event many people have looked forward to.[1] They will not be disappointed.

The book is not intended to be comprehensive, as it limits itself to three particular topics, namely price-fixing, horizontal mergers, and exclusionary vertical contracts. However, the insights given, the new perspectives offered when surveying both theoretical and empirical work, and the depth with which the arguments chosen are treated, make the book well worth its price and the time devoted to read it.

Apart from economists who have a research interest in antitrust issues, the main audience for the book should be graduate students who have already a background in industrial organization. (The book takes for granted that the reader knows the basics of industrial economics and, to a lesser extent, of antitrust law: there is a brief introduction on U.S. law.) Indeed, the treatment is at too high-level for undergraduate students and for lawyers.

1 M. Whinston, Lectures on Antitrust Economics (MIT Press 2006).

The author is professor in the Department of Economics at European University Institute, Florence, Italy.
He is very grateful to Chiara Fumagalli, Joe Harrington, and John Vickers for comments on a previous draft.

Graduate teachers may also use the book for a selected topics course in a Ph.D. program, if properly complemented with other readings. An alternative title for the book may have been "Invitation to Antitrust Economics" as graduate students and economists fluent in modern microeconomics but unfamiliar with antitrust might use this book for a first approach to the field. Hopefully, Whinston's selection of topics and his thoughtful remarks will push some readers to know more of antitrust, and do research work on it.

The book is composed of a short introductory chapter and three chapters that I now succinctly describe and comment on. Since we economists suffer from the referee's bias syndrome, I will focus more on those (rare) matters on which I have some critical remarks. But these minor remarks do not modify my overall conclusion that this is an excellent and thought-provoking book which is highly recommended. A consequence of this bias is also that I will mainly deal with Chapter 2, which I feel warrants more discussion, whereas I will say very little about Chapters 3 and 4, which are outstanding and very accomplished in my view.

Chapter 2 deals with price-fixing (i.e., agreements among competitors to restrict output or raise prices—synonyms include the terms "cartel" and "explicit collusion"), and it starts in a provocative way, by underlining that price-fixers may sometimes have pro-competitive justifications for their cartels. It also cites Mankiw and Whinston's result from their 1986 paper that free entry may lead to too few (or too many) firms at equilibrium from the point of view of welfare: relaxing competition may therefore lead to higher welfare.[2] Only after a few pages does Whinston explain how the possible benefits from cartels are not likely enough to justify a rule of reason: given the exceptionality of welfare-improving cartels, it would be too costly for the courts to depart from a per se rule of prohibition of price-fixing (that is, there is no justification which can be invoked to allow a cartel).[3,4]

---

2 *See* G. Mankiw & M.D. Whinston, *Free entry and social inefficiency*, 17 RAND J. Econ. 48-58 (1986).

d'Aspremont and Motta (2000) analyze the trade-off between concentration and competition. They show that—as Whinston argues—very fierce price competition may lead to too-concentrated an industry, but also that joint-profit maximization (the solution that a cartel would choose) is never optimal from the point of view of welfare. *See* C. d'Aspremont, C. & M. Motta, *Tougher Competition or Lower Concentration: A Trade-Off for Antitrust Authorities?*, *in* Market Structure & Competition Policy: Game-Theoretic Approaches (G. Norman & J. Thisse eds., 2000).

3 Incidentally, throughout the book Whinston repeatedly uses the theoretical result of Mankiw & Whinston (1986), to qualify welfare results obtained for exclusionary vertical restraints: for instance, he suggests that entry-deterrence by a monopolist might not be bad if the market led to too much entry (*see* Mankiw & Whinston, *supra* note 1, at 151, 166, and 188). However valid this argument from a theoretical standpoint (but how would one apply it in practice?), from the policy point of view it should be dismissed, for the same reasons Whinston uses to explain why the per se rule of cartel prohibition is appropriate: it would be too costly for courts to consider a monopolist's claim that absent its predatory or exclusionary practices the market would have led to too much entry. Further, how many markets do we know where there are "too many" firms?

4 I would have not started a chapter on cartels by mentioning their possible pro-competitive effects, but I guess this was made intentionally, to arouse interest.

Faithful to his declared objective "to unsettle the discourse a bit" in the most settled area of antitrust,[5] Whinston offers a stimulating perspective in Chapter 2. Rather than dealing with what economics has achieved in explaining collusion,[6] the chapter stresses where economics has been less successful in dealing with collusion. In particular, the main theme of the chapter deals with the difference between firms talking and not talking to each other; that is, the difference between tacit and explicit collusion. Indeed, economics has so far been unable to model this difference: the standard supergames literature applies to tacit collusion as much as to explicit cartels, and does not capture (at least not directly) the effect of competitors talking to each other (i.e., if they engage in price-fixing).[7]

Starting from this basic consideration, Whinston also surveys the empirical literature, trying to answer the question: does it really matter if firms talk to each other? He surveys works which have tried to estimate either the impact of conspiracies (to what extent have they led to higher market prices?), or the impact of antitrust interventions (have they led to a decrease in prices?), and concludes that overall "the published evidence on the effect of price-fixing conspiracies is somewhat mixed."[8] He also appeals to more scientific work in this area: while there is a whole branch of forensic economics that is busy in estimating damages in price-fixing cases, it is rare that this type of work appears in refereed publications.

I SHARE WHINSTON'S CONCERN THAT THERE SHOULD BE MORE ECONOMETRIC WORK ON THE EFFECTS OF CARTELS, BUT I AM A LITTLE MORE SKEPTICAL ABOUT SOME OF THE STUDIES MENTIONED HERE, IN PARTICULAR THOSE THAT INDICATE SCARCE EFFECTS OF ANTITRUST INTERVENTIONS

I share Whinston's concern that there should be more econometric work on the effects of cartels, but I am a little more skeptical about some of the studies mentioned here, in particular those that indicate scarce effects of antitrust inter-

5 WHINSTON, *supra* note 1, at 3.

6 Modern theory on collusion is based on supergames. Through simple models, we are able to understand the problems of firms' incentives to collude and of firms' coordination. We also know a lot about the factors that facilitate collusion, which is crucial for the design of policies against collusion. For a discussion of facilitating practices, not dealt with in this book, *see, e.g.*, M. MOTTA, COMPETITION POLICY: THEORY & PRACTICE (Cambridge University Press 2004).

7 As Whinston observes:

> Of course, most economists are not bothered by this [failure to explain formally the role of talking to each other], perhaps because they believe (as I do) that direct communication (and especially face-to-face communication) often will matter for achieving cooperation, and that pro-competitive benefits of collusion are both rare and difficult to document. Nonetheless, it would be good if economists understood better the economics behind this belief.

WHINSTON, *supra* note 1, at 26.

8 *Id.* at 38.

ventions (one way to see the impact of price-fixing is to see what happens to market prices when there is a cartel indictment).[9,10] Some of the cited papers contain price data that are insufficiently disaggregated, others refer to old cartels, and it is therefore possible that the laws did not provide sufficient deterrence from collusion (Whinston himself underlines that cartel penalties have been increased to serious levels only recently). And finally in some cases a past (overt) agreement might provide focal points to the firms, which could continue to coordinate on high prices even without talking to each other: it is only with time, when demand and supply shocks change the industry conditions, that the impossibility to talk to each other will show its effects.[11]

Speaking of changing industry conditions over time, let me mention what is, in my opinion, one of the most important challenges facing economists in the field of collusion, namely understanding how renegotiation affects collusion. The existing models of collusion are not satisfactory in this respect (and may even arrive at the paradoxical conclusion that the possibility to talk jeopardizes collusion by undermining the credibility of the punishment which should take place after a deviation from collusion) and yet this is probably where—together with helping solve coordination problems—talking to each other helps most. In the real and ever-changing world, firms cannot write complete contracts specifying what to do in any possible occurrence, and they need to talk to each other to fill the gap in their incomplete cartel contract (and to avoid misinterpreting as deviations actions which are instead undertaken because of a changing environ-

9 Connor (2005) reviews hundreds of studies and identifies 674 observations of cartel overcharges, in all times and countries. The median overcharge for all cartels is 25 percent, the mean is 49 percent. Estimating cartel overcharges is not an easy task, since it involves estimating the difference between actual price and a counterfactual, and it is unclear to me how many of the studies cited by Connor would satisfy current economic journals' standards. However, since the data are so numerous and are computed using so many different methods, and yet tend to give similar results, it is tempting to find some truth in them. *See* J. Connor, Price Fixing Overcharges: Legal and Economic Evidence 4-17 (Purdue University, Staff Paper No. 04-16, 2005).

10 In very recent work, Langus and Motta (2006) look at the effects that dawn raids (the first publicly available information that a cartel is being investigated) and European Commission's decisions to fine firms for cartel activities have on the share prices of the infringing firms, by using EC antitrust data and event-study techniques. They find that on average the former decrease firms' valuation by 2.4 percent and the latter by around 1.5 percent. Most of the drop is not caused by the fines (which account for only roughly one percent), so it must be due to the likely cessation of the profitable cartel activity. In turn, this should imply that investors expect investigated and fined firms not to be able to sustain high prices any longer (or to a lower extent). Indirectly, this suggests that antitrust activity does have an effect on market prices. *See* G. Langus & M. Motta, The Effect of EU Antitrust Investigations and Fines on the Firm's Valuation (European University Institute, Working Paper, 2006).

11 Furthermore, in some cases an explicit agreement may entail market-sharing clauses with each firm selling in a separate geographic market. When this is the case, the end of an explicit agreement may not change things that much. A firm will think twice before entering its rivals' markets, anticipating that they would react by entering its own market. Moreover, to the extent that shocks are local, such a collusive situation may survive the existence of shocks.

ment), as described in Genesove and Mullin's beautiful account of the U.S. sugar cartel in their 2001 paper.[12]

The chapter concludes with some discussions on how the law should treat tacit vs. explicit collusion and asks a crucial question: given that firms may be able to reach collusive outcomes even without talking, would not a policy which prohibits explicit—but not tacit—collusion (which is the current policy in the United States and the European Community) be clearly insufficient? Here Whinston contrasts two opposite views. On the one hand, Turner's view that tacit collusion should not be seen as an infringement of antitrust law, and that instead one should intervene by adopting industrial restructuring policies (i.e., forced divestitures) that would lower industrial concentration and therefore reduce the possibility that tacit collusion be sustained (concentration is one of the structural conditions which favor collusion). On the other hand, there is Judge Posner's provocative view that economic and econometric evidence could be used to prove the existence of tacit collusion and thus be used by agencies to impose financial (but not criminal) penalties on firms.

Whinston correctly criticizes both views: because nobody would think today of massive de-concentration programs, among other things because we are much more aware of efficiency arguments; and because there is no court of law which would enter into a guessing game of whether a given firm's action is legitimate because of certain market conditions or illegal because it is undertaken with the objective of tacitly colluding.

The chapter ends here, with the recognition that these are difficult issues, and there should be more public debate on these issues. Yet, this is an area in which more could be said. First of all, modern industrial economics has identified a number of factors, beyond concentration, that facilitate collusion. Therefore, one could intervene (in the spirit of Turner) on the environment in which firms act by making it less likely that they could sustain collusion. Prohibiting firms from exchanging disaggregate information (which helps them monitor each other's actions), or preventing them from using certain price clauses or from coordinating on practices (such as resale price maintenance) that favor transparency on the sellers' side of the market, are some examples. (Incidentally, merger control has the same effects as industrial restructuring programs, except it is a preventive action: it prevents sectors from reaching the conditions that lead to tacit collusion.) Further, advances in the study of auctions illustrate how auctions could be designed to avoid bid-rigging.

Furthermore, it is far from clear that tacit collusion can be sustained over time without competitors talking to each other (see the points made above on the necessity for price-fixers to talk in order to deal with changing market condi-

12 D. Genesove & W. Mullin, *Rules, Communication, and Collusion: Narrative Evidence from the Sugar Institute Case*, 91 Am. Econ. Rev. 379-98 (2001).

tions).[13] After all, firms have known for a long time that they can sustain collusion without express agreements and yet agencies keep on uncovering documental evidence of meetings and communication among firms' managers. This observation somehow reduces the importance of the question of how to treat tacit collusion, and refocuses our attention on the issue of how to break and deter cartels (i.e., explicit collusion).

This is also an area in which there have been important developments, both from the theoretical and the policy point of view. First, the introduction of leniency programs (first in the United States, then in the European Community and in most OECD countries) has shown how firms (and their managers) can be induced to report evidence that allows agencies to successfully prosecute cartels, and to break price-fixing. Second, there has recently been a lot of debate on how to deter cartels, leading legislators around the world to increase financial penalties, introduce (e.g., in the United Kingdom) or increase (e.g., in the United States) criminal penalties, promote private actions for damages, discuss how to introduce compliance programs and codes of conduct for firms, and so on. Finally (and this is a point that Whinston also makes in this chapter), there is more attention on how to detect the existence of collusion, so as to allow agencies to direct their investigative efforts to those markets that may hide cartels.[14]

Chapter 3 deals with horizontal mergers and blends theoretical and empirical aspects in an outstanding way. The first part of the chapter starts with Williamson's trade-off between market power and efficiency saving (which is still the cornerstone of the analysis of mergers), proceeds with an insightful description of Farrell and Shapiro's model of mergers (which provides some useful clues for the practice of merger control), and closes with a detailed analysis of the U.S. merger guidelines. The second part of the chapter—the most interesting in my opinion—surveys the different empirical methods which can be used in the analysis of mergers, both in identifying the relevant antitrust markets (the first step in a merger analysis), and in predicting the likely effects of the mergers. Whinston also surveys (ex post) empirical evidence on the effects of actual mergers, something that is probably of less direct utilization for the practice of merger control, but which gives useful insights as to the reliability and limits of the different econometric methods that antitrust agencies could use.

One might wish to receive a little more practical guidance from the author—for instance which methods to use under which circumstances—but admittedly this is an area where the most promising techniques are of very recent development, and it is therefore difficult to compare their validity and fully understand their limits and advantages. Chapter 3 is really an excellent introduction to the

13 Talking to each other might also be necessary to agree on a market allocation.

14 *See, e.g.,* J. Harrington, *Behavioral Screening and the Detection of Cartels*, *in* European Competition Law Annual 2006: Enforcement of Prohibition of Cartels (C.-D. Ehlermann & I. Atanasiu eds., 2006).

econometrics of mergers, and is highly recommended to all those graduate students who want to apply econometric techniques to the analysis of mergers.

Finally, I also like the fact that Whinston devotes some attention to the long-run consequences of mergers, in particular the impact that they could have on research and development. This is an area in which we know very little and more research is needed.[15]

Chapter 4 is a masterly piece. It focuses on one particular class of vertical restraints, namely exclusivity clauses in vertical relationships.[16] Here Whinston manages to provide a unifying conceptual framework to present all the different models which have appeared in the literature to deal with such issues. The central insight is that it is the existence of contracting externalities (either on parties which are not included in the contracting process or among parties which are included in the contracting process, but arising because contracts are bilateral) that allows understanding of when exclusive clauses will be signed, and what effects they will have on welfare. This idea was already present in Bernheim and Whinston's 1998 article,[17] but here it is not only explained more simply, but is also extended to explain a number of contributions not discussed in their paper. For instance, the presentation of Hart and Tirole's model (in which exclusive territorial clauses are used by a manufacturer in order to restore the monopoly power it would lose due to a commitment problem)—with the contrast between the case in which retailers are independent local monopolists and the other extreme case where they sell undifferentiated products—is very illuminating.[18]

THE CENTRAL INSIGHT IS THAT IT IS THE EXISTENCE OF CONTRACTING EXTERNALITIES THAT ALLOWS UNDERSTANDING OF WHEN EXCLUSIVE CLAUSES WILL BE SIGNED, AND WHAT EFFECTS THEY WILL HAVE ON WELFARE.

15 Mergers might also lead to restructuring of capital, which may have important consequences on prices and efficiency in the medium- and long-run. *See, e.g.*, J. Chen, The effects of mergers with dynamic capacity accumulation (2006) (mimeo, U. California at Irvine) (on file with author).

16 Exclusivity clauses take the form of exclusive dealing when a retailer agrees to buy from one particular manufacturer only and not from other manufacturers (Whinston calls them "exclusives to reduce competition in input markets"), and of exclusive territorial protection when a manufacturer commits to sell to one retailer only and not to others ("exclusives to reduce retail competition").

17 B. Bernheim & M. Whinston, *Exclusive Dealing*, 106 J. POL. ECON. 64-103 (1998).

18 However, treating exclusive dealing and exclusive territories as if they were the same phenomenon might be slightly misleading from the point of view of competition policy practice. The former affects inter-brand competition and the latter intra-brand competition, and most economists would agree that competition agencies should concentrate their efforts to vertical restraints that affect inter-brand competition, whereas a number of efficiency reasons may be invoked to justify clauses that restrict competition among retailers offering the same brand. Further, if I had to name a reason why exclusive territories may harm welfare I would mention Rey and Stiglitz (1995)'s argument that they relax competition among retailers and therefore lead to higher prices. *See* P. Rey & J. Stiglitz, *The Role of Exclusive Territories in Producers' Competition*, 26(3) RAND J. ECON. 431-51 (1995).

The analysis in most of this chapter (not only in Section 4.4, which Whinston himself recognizes as more difficult) is necessarily more advanced than in the other chapters, but the readers who are already familiar with the original papers (some of which are not easy to digest themselves) will find a lot of value in the presentation, which draws together different branches of the literature in a very insightful way. Further, this chapter is highly recommended to all those readers who are not familiar with the literature and want to approach one of the most exciting—and still developing—areas of antitrust.

Whinston also indicates some possible policy implications that can be drawn from the literature on exclusive clauses, and closes the chapter with a discussion of possible pro-competitive effects of exclusive contracts and a brief survey of the (very few) empirical works on the issues.

In sum, this is a nice book that I highly recommend. Hopefully, it will encourage discussions and economic research on a number of important topics. ◀