Concurrences

Revue des droits de la concurrence

III Lisbon Conference
on competition law and economics

Colloque | Concurrences N° 2-2010 | www.concurrences.com

Vítor Bento | President, SIBS Forward Payment Solutions

Michael Katz | Haas School of Business, University of California, Berkeley

David Evans | Lecturer, University of Chicago | Executive Director, Jevons Institute for Competition and Economics | Visiting Professor, University College London

Christopher Bellamy | Senior Consultant, Linklaters, LLP

Peter Freeman | Chairman | UK Competition Commission

Mariana Tavares de Araujo | Secretary of Economic Law of Brazil

James Bushnell | University of California Energy Institute, Berkeley

Frank Wolak | Department of Economics, Stanford University

Ricardo Cardoso | Antitrust: Energy and Environment Unit | European Commission - DG Competition

Luís Pais Antunes | Partner, PLMJ – A. M. Pereira, Sáragga Leal, Oliveira Martins, Júdice e Associados, Sociedade de Advogados, RL

Alberto Heimler | Chair, Working Party 2 | OECD

Carl Baudenbacher | President of the EFTA Court

Philip Lowe | Director-General, DG Competition, | European Commission

Luís Cabral | IESE Business School

Frank Lichtenberg | Graduate School of Business, Columbia University

Richard Gilbert | Department of Economics | University of California, Berkeley

Julia Holtz | Senior Competition Counsel - EMEA, China, India | Google UK Ltd.

Jean Yves Art | Associate General Counsel, Microsoft

David R. Schmidt | Assistant Director, Department of Economics | US Federal Trade Commission

Bo Vesterdorf | Senior Consultant | Plesner, Copenhagen and Herbert Smith, LLP, London

Damien Neven | Chief Economist | European Commission - DG Competition

Thomas O. Barnett | Partner | Covington & Burling LLP

Institut de droit de la concurrence

e-Competitions

@ Colloque

III Lisbon Conference on competition law and economics

thursday **14th January**

Panel I - Two-Sided Markets: A challenge for competition policy and regulation?

MODERATOR **Vítor Bento** | President, SIBS Forward Payment Solutions

Professor Michael Katz | Haas School of Business, University of California, Berkeley
Dr David Evans | Lecturer, University of Chicago | Executive Director, Jevons Institute for Competition and Economics | Visiting Professor, University College London
Sir Christopher Bellamy | Senior Consultant, Linklaters, LLP

GUEST SPEAKER **Peter Freeman** | Chairman | UK Competition Commission

Panel II - Energy markets: To what extent can competition, security of supply and environmental protection be reconciled?

MODERATOR **Mariana Tavares de Araujo** | Secretary of Economic Law of Brazil

Professor James Bushnell | University of California Energy Institute, Berkeley
Professor Frank Wolak | Department of Economics, Stanford University
Ricardo Cardoso | Antitrust: Energy and Environment Unit | European Commission - DG Competition

Panel III - Competition Policy in times of crisis: Which enforcement practices best fit the principles?

MODERATOR **Luís Pais Antunes** | Partner, PLMJ – A. M. Pereira, Sáragga Leal, Oliveira Martins, Júdice e Associados, Sociedade de Advogados, RL

Alberto Heimler | Chair, Working Party 2 | OECD
Prof. Dr. Carl Baudenbacher | President of the EFTA Court

friday **15th January**

Philip Lowe | Director-General, DG Competition, | European Commission

Panel IV - Intellectual property and competition: Complementary policies?

MODERATOR **Luís Cabral** | IESE Business School

Professor Frank Lichtenberg | Graduate School of Business, Columbia University
Professor Richard Gilbert | Department of Economics | University of California, Berkeley
Julia Holtz | Senior Competition Counsel - EMEA, China, India | Google UK Ltd.
Jean Yves Art | Associate General Counsel, Microsoft
David R. Schmidt | Assistant Director, Department of Economics | US Federal Trade Commission

Panel V - Competition Policy and Single Firm Conduct: Recent developments in the EU and the US. What consequences in terms of enforcement actions?

MODERATOR **Luís Cabral** | IESE Business School

Bo Vesterdorf | Senior Consultant | Plesner, Copenhagen and Herbert Smith, LLP, London
Damien Neven | Chief Economist | European Commission - DG Competition
Thomas O. Barnett | Partner | Covington & Burling LLP

Autoridade da Concorrência

14th & 15th January 2010

8.00 - 19.00

Calouste Gulbenkian Foundation | Lisbon
Avenida de Berna, 45A

AUTORIDADEDACONCORRÊNCIA

Papers of this conference are published in the electronic version of the review.
Les actes de ce colloque sont publiés dans la version électronique de la revue.

Michael L. Katz
katz@haas.berkeley.edu
Haas School of Business, University of California, Berkeley

Two-Sided Markets: A challenge for competition policy and regulation?

Two-sided markets: A new challenge for competition policy and regulation?

Abstract

Two-sided markets raise difficult issues but these issues not as new or unique as some commentators assert. Importantly, these issues are not a reason to abandon public policy intervention aimed at protecting competition. Nor are they a reason to abandon competition and use pervasive regulation as a substitute for competition.

La difficulté des questions posées par les marchés bi-face ne doivent pas être sur-évaluées dans la mesure où ces difficultés ne sont ni nouvelles ni sans équivalents. Ces difficultés ne peuvent ni justifier l'absence d'intervention publique destinée à protéger la concurrence, ni le recours à une régulation envahissante comme substitut à la concurrence.

1. In preparing my presentation, I almost simply adopted the title of the session – Two-Sided Markets: A Challenge for Competition Policy and Regulation? – as the title. However, I added the word "new" because I think the answer to the question posed by the session title obviously is "yes". A deeper question is whether two-sided markets are a *new* challenge.

2. Before I attempt to answer that question, let me lay a foundation so it is clear what we are talking about in this session. When we are talking about two-sided markets, we typically are talking about a platform that brings together different groups of users. That is, the platform offers some sort of service that facilitates the users interaction with one another. For example, a payment platform allows a consumer holding a payment card to interact with merchants (*i.e.,* to use the card to make purchases from them). In the case of advertiser-supported media, such as many magazines or broadcast television in some countries, the magazine or broadcast serves as a platform that bring viewers or readers together with advertisers that want to reach them. One can also think of video-game consoles as platforms that bring together people who play video games with enterprises that manufacture video games. My last example is a favorite among economists, possibly because they are socially inept and like to study other people's lives: a singles bar or online dating service serves as a platform for bringing together individuals who would like to meet.

3. Two-sided markets can also have a much more complicated structure. For example, a platform's users may themselves be platforms. Consider an example in which a household connects to an Internet service provider (ISP), which may peer with another ISP, which is then connected to an application website. The ISPs are platforms. The application website also might be a platform, bringing households together with either advertisers or online retailers, for example. In the remainder of my discussion today, I will ignore such complexities because—although they make some of the particulars much more complex—fundamentally such markets raise the same issues that arise in more simple two-sided structures.

4. People who hear the term two-sided markets often ask whether *all* markets two-sided? After all, the meaning of a market is that it brings together two sides, buyers and sellers. By some definitions, a grocery store could be considered a platform that facilitates commercial exchanges between food producers and consumers. I don't believe that such expansive definitions are useful. It certainly is hard to see how two-sided markets could raise any new issues if every market is two-sided. Fortunately, there are other definitions that identify a narrower set of markets as being two-sided.

5. At present, there is no single, universally accepted definition of two-sided markets, and I today will not offer a specific definition of my own. Instead, I will identify some of the characteristics that I believe should be present for a market to be considered two-sided. One is that cross-group network effects be present across the two sides of the platform. A cross-group network effect arises when a user on one side of

the platform cares how many users are on the other side of the platform. Consider a payment card network, for example. If you're a consumer holding a credit or debit card, then you want there to be many merchants who accept that brand of card. Similarly, if you're a merchant, then accepting a payment card is more valuable to you the greater the number of consumers who use the card.

6. Grocery stores are not subject to cross-group network effects in the same way. A consumer purchasing bananas at a particular grocery store generally doesn't care if that store is supplied by one banana vendor or twenty. What matters to the consumer is the product quality and the retail price that the grocery store charges. In the other direction, there is a sense in which a banana wholesaler would like the grocery store to have a lot of customers so that it will order a lot of bananas. But there is also a sense in which the wholesaler doesn't care. Specifically, once the grocery has taken ownership of the bananas, the wholesaler is indifferent as to whether the grocery store sells the bananas to consumers or is forced to dispose of them after being unable to find buyers. Hence, a definition that requires the presence of network effects will eliminate at least some types of market from consideration as two-sided markets.

7. Another factor that it is useful to include in the definition is that the platform can treat users on its two sides differently in ways that matter. A common practice in the United States, for example, has been for bars to have promotions in which they charge lower prices to women than men: so-called "ladies nights." I note in passing that suing people is also common practice in the United States, and there has been litigation alleging that such pricing constitutes illegal discrimination. In any event, the economic rationale for ladies nights is that men will pay more money to be at bar with women, and more women will go to a bar if it charges them lower prices. Hence, the lower prices to users on one side of the platform (*i.e.,* women) can be viewed as an investment in making the platform (*i.e.,* the bar) more attractive to users on the other side (*i.e.,* men). Bars can pursue such policies because – litigation aside – they have the ability to distinguish between men and women. Similarly, a credit card network can distinguish between merchants and card holders, and the network can treat the two groups of users differently from one another, say by charging them different prices for using the network. More generally, when it has the ability to take actions that differentially impact users on the two sides of its platform, a platform owner can try to influence the size of the network on each side of its platform in ways that are beneficial to the owner. The set of potential actions is not limited to the choice of prices. For example, the owner of an online dating site might seek ways to attract more women to the site. Because women are often sensitive to security and safety issues, one possibility would be to have stricter policies for vetting site members, say by running extensive background checks on men wishing to join. Such measures would be a non-price means of making the platform more attractive to users on one side.

8. With the foundation in place, let's turn to the issues for competition policy and regulation in markets with cross-group network effects where platform owners have the ability to take actions that differentially affect the two sides of the market. Several of these issues center on concerns whether two-sided markets perform well. One such concern arises from the presence of network effects when different platforms do not share common networks of users. This is the case, for example, with credit and charge card networks: a merchant that accepts American Express cards may not accept Visa cards. Thus, card holders consider the merchant networks of American Express and Visa to be distinct. When platforms have proprietary user groups, there can be positive feedback effects that create competitive advantage for the leading network. Specifically, as the number of users of a given platform grows, that platform becomes increasingly attractive to other users. As those users join the platform, it becomes still more attractive. This sort of positive feedback can lead to "tipping," which is the tendency of one platform to pull away from its rivals in popularity once it has gained an initial edge. It is possible that a market will tip to the wrong platform in the sense that users would all be better off if they could move in a coordinated way to another network, but they cannot.

9. Notice that this effect does not arise when platforms have non-proprietary user networks. Consider, for example, mobile telephone networks. These networks all share a common base of users in that users on one wireless network can communicate with users on another as the result of network interconnection. Consequently, the fact that a wireless carrier has a large customer base does not, in itself, make its network more attractive to consumers. (Here, I am ignoring attempts that have been made by some wireless service providers to create proprietary networks by charging differential rates for calls made to users on other networks.)

10. One might respond to the problems associated with proprietary user groups by using competition policy or regulation to force platforms to be interoperable with one another. However, such policies can restrict platform competition because, absent such policies, a winner-take-all structure can lead to vigorous competition, at least until there is a winner that takes all. Such policies can also limit product variety or make it difficult to innovate. Suppose, for example, that there are two platforms which are subject to mandatory interconnection. How would public policy enforcers respond if one of the platform owners announced that it wanted to introduce a new technology that offered large consumer benefits but was incompatible with the other platform? In an industry with the potential for rapid innovation, a policy of forcing every platform to remain compatible with every other platform could stifle significant innovation.

11. Another issue that has gotten significant attention in some circles is that economic theory indicates that competition between platforms often will not yield an efficient or socially optimal outcome. Indeed, in theory, the competitive equilibrium can be worse than the monopoly one. To see why, consider a platform owner that is thinking about what prices to charge to the two sides of the market (*e.g.,* an online dating service choosing membership fees for men and women). In part, the owner's decision is going to depend on each side's elasticity of demand. Loosely speaking, if one group has less elastic demand (*i.e.,* is less sensitive to the price it is charged),

then the platform owner will want to charge a higher price to that side and a lower price to the other side. Users who are willing to pay a lot without pulling off the platform are the ones to whom the platform owner wants to charge high prices. In the face of competition, a platform owner doesn't think about users' overall willingness to pay for the service (what are known as market elasticities). Instead, each owner is concerned with users' willingness pay for its specific platform (what is known as the firm-specific elasticity). It's quite possible to have the following situation. Users on one side of the platform, say men, have a high overall willingness to pay to be on a platform (*i.e.,* have low market elasticities) yet are very willing to switch among platforms (*i.e.,* have high firm-specific elasticities). Women might be the reverse in that they are less willing to pay overall but have strong preferences for specific platforms. In such a setting, socially optimal pricing would entail high charges to men in order to encourage overall participation on the collection of sites. If there were a single, monopoly owner of all the sites, it would tend to adopt such a pricing structure. However, competing platforms would set relatively low prices to men in order to keep them from choosing rival platforms. In this respect, the competitive equilibrium could be further from the optimum than would the monopoly outcome.

12. It is important to recognize that there are circumstances in which competition leads to socially undesirable outcomes even in markets that are not two-sided. For example, a similar problem of responding to firm-specific rather than market elasticities arises in markets in which multiproduct firms engage in a form of Ramsey pricing to cover their fixed costs and earn profits. And, as a matter of theory, too many firms may enter a market, resulting in inefficiently small firms that are unable to realize economies of scale fully. It is also theoretically possible to have too many firms from the point of view of optimally promoting R&D investment. The fact that, even in the absence of two-sided market effects, competition may lead to less-than-ideal outcomes has not stopped the majority of economists working on these issues from concluding that public policy should generally seek to promote, or at least protect, competition. Consider mergers that potentially reduce competition in the provision of advertising. Economic theory tells us that competitive markets can have excessive advertising. Yet I am unaware of any competition policy enforcer's having approved an anticompetitive merger on the grounds that the resulting restriction in the supply of advertising would be beneficial. Similarly, the fact that competitive two-sided markets don't necessarily attain the best possible outcome is not – in my view, at least – a reason to abandon the merger policy in two-sided markets.

14. There are other reasons why equilbirum outcomes in two-sided markets may be inefficient even when platforms to not possess the degree of market power that typically is required to trigger competition-policy concerns. One of the other things that can go wrong is that there can be what economists call a Spence distortion. Suppose you are an owner and you are evaluating whether to take steps to increase the value of your platform to consumers. The particular consumers you care about are your *marginal* customers, the ones who will cease purchasing your product if you raise the price even a little bit. If you can figure out how to make your marginal customers willing to pay more, then you can raise your prices without losing any customers. Your firm will also have so-called *inframarginal* customers, who buy the good and are willing to pay much more than the current price. Suppose you implement an improvement that makes your platform more attractive to inframarginal customers but not to marginal ones. Then the improvement may do little to raise your profits: your inframarginal customers would enjoy an even greater surplus of benefits over price, but you would not be able to raise your price without losing your marginal customers. If the improvement appeals to marginal consumers, however, then it will allow your platform to raise its prices without losing customers. Hence, a platform owner has have incentives to respond to the preferences of marginal customers. This fact can lead to inefficient outcomes when marginal customers have different preferences with respect to platform features than do inframarginal customers and, thus, the preferences of marginal customers do not represent the preferences of consumers overall.

15. Although there are multiple reasons why equilibrium prices in a two-sided market can be inefficient, most of those reasons are present in other markets as well. The Spence distortion, for example, can be present in a market whether or not it is two-sided. There is no reason to expect that actual markets—two-sided or otherwise—attain the theoretical optimum. Competition policy is based on the premise, not that competitive markets yield perfect outcomes, but rather that they lead to better outcomes than could the alternatives of uncompetitive markets or highly regulated markets. Although competition in two-sided markets does not necessarily lead to the theoretical optimum of market performance, to date we do not have strong reasons to believe that policy makers could successfully identify and implement policies that abandon reliance on competition and instead use regulatory fiat to improve consumer welfare and economic efficiency.

16. That said, the fact that equilbirium outcomes in two-sided markets may be inefficient has led some people to suggest that pricing in two-sided markets should be regulated. Interestingly, some of the suggestions call for regulation that addresses only the pricing *structure*. By structure, I mean the *relative* levels of the prices charged to users on two sides of a platform rather than the overall levels of those prices. For example, various public agencies around the world have expressed concern that credit card networks have rules that lead to prices charged to merchants that are too high and prices charged to cardholders that are too low. An important question which regulators have had difficulty answering is what the regulated prices should be in the light of the fact that economic theory indicates that the answer is highly sensitive to characteristics of demand about which regulators have little information.

17. In the United States and some other countries, there is also movement toward so-called network neutrality regulation. Network neutrality means different things to different people, but a common element is a call for a ban on two-sided pricing strategies under which a household's ISP would charge both the household and application providers

communicating with the household for the household's Internet access service. Everybody agrees that it can be appropriate for an ISP to charge its household customers. The debate concerns whether the ISP should also be allowed to charge application providers for the ability to communicate with the ISP's household customers. Those arguing that the answer should be "no" often overlook the effects of such a policy on households. Namely, that by denying an ISP the ability to charge application providers, such a public policy could raise the prices paid by households. Think about the issue from an ISP's perspective. If you are an ISP and you can make a lot of money charging applications provider for the right to reach your household customers, then you have incentives to lower your prices to households in order to attract more of them and increase your ability to earn profits from application providers. Indeed, you might be willing to subsidize households in order to get more of them. That's the reason why Google subsidizes search. If there were a public policy that prevented Google from charging advertisers for access to searchers, then Google very likely would cease offering free search. When people say they're in favor of not letting ISPs charge application providers for reaching the ISPs' household customers, those people are implicitly saying that they're in favor of charging higher prices to households. It is important to remember that, if public policy pushes prices down on one side of the market, it may be raising them on the other.

18. I would like to close this discussion of regulating price structures by briefly mentioning another argument that has been made in favor of a regulatory ban on strategies that charge positive prices to both sides of the platform. The argument is based on the claim that it is somehow immoral, unethical, or just plain unfair to charge both sides of the market simultaneously for a single good or service, such as Internet access that connects the two parties with one another. If you think about this claim for even a minute, you will see that it is nonsense. Consider an ISP that facilitates interaction between a household subscribing to the ISP and an application provider. The claim made by some opponents of two-sided pricing is that charging the household one Euro and charging the application provider nothing is ethically superior to charging each side 50 cents. You might try to justify this argument by asserting that the application provider is somehow deserving of better treatment than is the household. However, those arguing against two-sided pricing on the grounds that it is unethical should also assert that charging the application provider one Euro and the household zero would be superior to charging each 50 cents. In my view, the claim that two-sided pricing is unethical "double dipping" is illogical and unfounded. Unfortunately, it is an argument that resonates with some people.

19. Thus far, I have been discussing issues that arise from concerns regarding the nature of equilibrium outcomes in two-sided markets. Some of the other issues that can arise with two-sided markets have to do with the analysis of such markets rather than the nature of the market outcomes. Several of these issues arise from the fact that it can be misleading to look at one side of a two-sided market without also considering the other side. My earlier discussion of singles bars illustrates this point. Someone who saw a bar selling half-price drinks to women might calculate that these drinks were being sold below cost and must somehow be predatory. However, once one takes into account the increased ability to earn profits from selling drinks to men that results from attracting women to the bar, one sees that the strategy may well be competitive, not predatory.

20. Of course, even taking a broader view of the market, it might be difficult to tell whether the pricing was competitive or predatory. In thinking about the implication of this fact for competition policy, it is important to recognize that there can be competitive reasons for suppliers to set prices below costs even in markets that are not two-sided. For example, a firm may engage in penetration pricing designed to attract consumers to sample a product with which they have no previous experience. In other words, diagnosing predatory pricing is always a difficult problem.

21. In the United States, there is a standard test applied by the courts to the analysis of pricing behavior in non-two-sided markets which asks if price is below average variable cost. Although the test is imperfect, one could adapt it to two-sided markets by asking whether the *combined* margins of prices minus variable costs on the two sides of the market are positive or negative. There would be no need to know anything about the underlying demand conditions or what the platform owner was trying to do to balance the interests of users on the two sides of its platform. One would simply ask if the platform passed the overall profitability test. The price-versus-average-variable-cost test has problems; among economists, there is no uniformly accepted definition of—or test for—predatory pricing. But those problems have little or nothing to do with whether a market is two-sided. Hence, I don't see a reason to consider two-sided markets as being so special that they have to be treated in a different matter.

22. It is also important to consider both sides of the market to make sure that one is not misdiagnosing the presence of market power. For instance, someone observing a singles bar charging high prices to men might conclude that the bar must have market power and the ability to earn excess profits. However, if the subsidies for women's drinks are necessary to create the ability to charge high prices to men, then those subsidies should be considered a cost of serving men. In this light, the bar might well be seen as breaking even or even losing money overall. In order to have a sense of the strength of competition or the degree of market power, it can be important to look at both sides of the market at once. Ladies nights may seem like a trivial example, at least to those of us who are married, but similar forces are at work in other markets as well.

23. Subsidizing one side of the market to increase the value of the platform to the other side is an example of a strategy intended to create network effects that raise the value of the platform to users and, thus, raise the potential for the platform owner to earn profits. As a general matter, there is no economic principle stating that a platform has to earn equal amounts of money from the two sides of the market. In fact, it will frequently not be the optimal thing to do. Thus, differential treatment of the two sides of the

market is neither evidence of the exercise of market power nor sufficient grounds for concluding that regulation would improve market performance.

24. In conclusion, I believe the answer to the question posed by the title of the session is "yes, two-sided markets do pose challenges to competition policy and regulation." However, the answer to the question raised by the title of my presentation is "yes, some of these issues are new, but they aren't as new and unique as some people seem to think." These are difficult issues, but many of them are issues with which public policy already grapples. I reject the argument that the challenges posed by these issues are reasons for public-policy enforcers to take a hands-off approach to two-sided markets out of concern that, because policy enforcers won't know what they're doing, they will more likely to harm competition and efficiency than promote them. I believe that it is appropriate to apply competition policy to protect competition in two-sided markets. However, I also reject the argument that two-sided markets need more pervasive intervention than can be undertaken within a competition policy framework. I agree that the performance of two-sided markets may, in some cases, be far from ideal. But it does not follow that these problems are so strong that we should give up on competition and try to regulate our way to the right outcome. There are specific circumstances in which regulation may be appropriate. But I have yet to hear a sound argument that, as a matter of course, regulation is likely to lead to better outcomes in two-sided markets than would competition subject to competition policy oversight. So where does this leave us regarding the role of competition policy and regulation in two-sided markets? I think it leaves us squarely in the middle. There is a broad role for competition policy to protect competition; there is also a potential role for regulation, but only in limited circumstances where there are specific reasons to believe that regulation is likely to lead to an improvement in the market outcome. ■

David Evans
david.evans@ucl.ac.uk

Lecturer, University of Chicago
Executive Director, Jevons Institute for Competition and Economics
Visiting Professor, University College London

Two-Sided markets: A challenge for competition policy and regulation?

The web economy, two-sided markets, and competition policy

Abstract

The web-economy has grown rapidly in the last decade. Online businesses have several key features that are important for understanding the pro-competitive and anti-competitive strategies they may engage in. The two-sided market literature helps elucidate many of these strategies. It also provides guidance for the antitrust analysis of market definition and exclusionary practices for web-based businesses.

L'économie numérique s'est développée très rapidement cette dernière décennie. Les entreprises de ce secteur présentent plusieurs caractéristiques importantes pour la bonne compréhension des stratégies pro- et anticoncurrentielles. La doctrine existante en matière de marchés bifaces permet de comprendre nombre de ces stratégies d'entreprises. Elle fournit également une grille d'analyse utile pour la définition des marchés et des pratiques d'exclusion spécifiques à l'économie numérique.

1. Ten years ago a tweet was something a bird did. We generally did not poke our friends. When we sent an email about buying a car we would have jumped out of our chairs if an advertisement for BMW all of a sudden popped up on our screens. And our mobile phones did not have application stores. Change has occurred rapidly following the development of the commercial internet but it has a long way to go. It takes years for entrepreneurs to come up with ideas, for businesses to start, and for industries to evolve and sort themselves out. It will take a decade or two, perhaps, for things to settle down. Competition policy will find itself increasingly grappling with mergers, exclusionary practices, and collusion as the web economy becomes more prominent, as it matures, and as it goes through a period of significant turmoil.

A decade ago we would have finished the phrase "*two-sided*" with "*coin*" and not markets and we would have thought that a person who engaged in multihoming had a place in the country. Like the web economy the study of two-sided platforms has grown rapidly since its birth at the turn of the century. The literature has flourished with many theoretical papers and much empirical research. Most major competition authorities around the world have are using the two-sided market approach in a broad range of cases from credit cards to shopping malls.

Competition policy matters involving web-based businesses will provide fertile ground for using two-sided analysis. In this paper I want to use the lens of two-sided markets to describe some key features of the analysis of market definition, market power, and exclusionary practices particularly as they relate to entry.

A tour of the web economy

2. The web economy is constantly evolving. Today businesses fit into one or more of the following categories. (1) E-commerce includes massive shopping malls like eBay, Amazon, and Baidu. It also includes many retailers that have set up shop online such as walmart.com. (2) There is online media which includes everything from portals like MSN to online video like YouTube, to newspapers like The Guardian.com, to all of you that have blogs. (3) Social networking is the new kid on the block on the internet. Many of you have profiles on Facebook or Bebo or some kind of site like that. (4) Online gaming has become enormous. It ranges from social networking games like Farmville, which you can find on Facebook, to Xbox Live.

Many of these web-based businesses make money from attracting eyeballs and selling access to those eyeballs to advertisers. This is where on-line advertising comes in. Many web sites run advertising sales themselves just like traditional newspapers and magazines have. That includes Yahoo and reuters.com. But then there are many businesses on the web that act as advertising intermediaries. They operate networks of advertisers and media properties and they pool inventory. Advertising is important for another reason on the web. The e-commerce and the media properties

are advertisers themselves. What could be better for online advertising to drive clicks to them ? AOL has become a very big player.

We now take a brief bus tour of some of the more interesting properties on the web.

3. Google, of course, has to be the first stop. We hear so much about this ten-year-old company that we need to put it in some perspective. Google makes virtually all of its money from selling text-based ads on search-results pages and a bit more with contextual ads on its publisher network. It is not making much money from many of the other businesses that it has gotten into and it has largely abandoned its efforts to make money from selling traditional ads for radio and television. It is an important player, nevertheless, for three key reasons. First, it's the dominant player in helping people find things on the web and that is terribly important. Second, it gets billions of dollars of revenue a year from search and contextual ads and this money helps it fund grander ambitions. Third, it does not like other firms controlling parts of the web because that gets in the way, or can get in the way, of it selling more advertising.

4. The iPhone is the next stop. What makes the iPhone so important is that it is a platform for developing applications. There are more than a 100,000 applications that people can download for their phones. More of these are being written every day and most are are web-based applications. These applications are helping to transform other industries –below we will see the example of Square which may disrupt payments.

5. Facebook is a six-year-old company that is run by a twenty-five-year-old. It has taken over the social networking space but it is actually a late entrant in that segment. There was a flurry of social networking sites that started in the late 1990s such as Six Degrees of Separation. Then Friendster was the king of the segment in the early 2000s. MySpace toppled Friendster and then Facebook has leapfrogged MySpace. Like YouTube, Twitter and other sites that attract eyeballs, Facebook really has not figured out how to make huge dollars from the huge traffic that it generates around the world. The idea is advertising. The problem is that when you go visit your friends you do not necessarily want to have a stranger trying to sell you male enhancement drugs. If you are an advertiser like Procter and Gamble you may not want your ads appearing next to pictures of drunk and half-naked people.

6. How do these web celebrities get along in the neighborhood? Google doesn't like Facebook because it cannot get its « spiders » (which pull content for its search engine) onto this large and growing part of the web. Facebook likes the iPhone because it is a great device for people to go visit their friends. Google and Apple used to get along so well that Google had a couple of representatives on the Apple board. But then Google could not stand having Apple control a large part of the mobile space for inserting ads. As a result, Google unfriended Apple and they are now engaged in the mobile platform wars.

Key features of the web economy

7. The web economy has many interesting economic features. I would like to highlight several that are important from a two-sided perspective.

CRITICAL MASS. We have seen a lot of entry of platforms over the last 15 years. For every one that succeeds, and that you have heard about, far many more have failed. There were more than 40 video sites that secured enough viewers to be counted around the time that YouTube began back in 2005. Almost all of them are gone now. More generally, most platforms fail to take off. The growth of these platforms is driven by network effects and they face a critical mass problem. The problem is analogous to an exchange which can survive only if it gets enough liquidity. Much of the work that web entrepreneurs like the founders do when they start is trying to figure out ways to get enough customers on board the platform to take off. YouTube as an example of tackling the critical mass problem. It had to figure out how to get enough people to upload videos and how to get enough people to view those videos and how to get both of those groups onboard the platform in enough numbers to ignite the platform and to get rapid growth. Businesses that get enough liquidity (such as Hulu and YouTube) ignite while those that don't (such Revver, another video sharing site) impode. It is as if there is an invisible wall. Once the entrepreneur pushes the platform through that wall the platform can take off.

We have known for a long time in a two-sided literature that exclusivities are a way to solve the chicken and egg problem. The reverse is true as well. One way for an incumbent firm to prevent a new firm from taking off is to make it hard for it to get enough critical mass by entering into exclusives with enough of at least one major customer group.

Free. Many of the web-based platforms discussed above are free to at least one group of customers. That is a well studied and documented phenomena in the two-sided literature. The platform generates value by getting one or more customer groups together. It can be profitable to charge one group of customers nothing just to get them on board so that the platform can charge another group of customers for access to them. We often see this offline. There are many free newspapers, for example. We often see « free » online because the marginal cost of serving another user is zero. Thus, you do not have to pay for your Facebook page. That social network makes money by selling your eyeball to advertisers and selling complementary goods like virtual gifts.

There is a tendency in competition policy cases to ignore the customers that are getting things for free. That happens for two reasons. One is that analysts tend to equate the business with the money side. We have seen this with Windows. An important group of customers for Microsoft are software developers that use the Windows APIs. Microsoft has chosen to provide most of its services to these customers for free. But those customers get a lot of value from Microsoft and should be considered in any welfare calculation. Another reason that analysts ignore the free side is that the traditional methods of market definition focus attention on a single group of customers even though the members of the two groups have welfare that is inextricably intertwined.

INVISIBLE ENGINES. If you looked for the heart and soul of the web business what would you find? It is not the server farm in South Dakota with all the lights flashing. It is the software. If you decided to start a new web business like Pandora or Skype or Twitter, you would mainly spend your time writing software code. All of these businesses on the web are based on thousands of lines of code. This software can be locked down so no one can get access to it. It can also be opened up so that others can use the features of it ; in that case it becomes a software platform that can support the development of complementary applications. Many of the web-based businesses have started software platforms by opening up their code. You might wonder how Firefox managed to cut Internet Explorer down to size. Much of it had to do with encouraging developers to write applications that increased the value of the Firefox browser. Facebook has done the same thing and there is an active developer community on Facebook with more than 500,000 applications written so far. Farmville is an example of a Facebook application. Google Maps has become such a powerful product because Google made the APIs available so developers could build applications that integrate mapping.

The software platform model is transforming the web. Almost every major web property has followed that strategy. It is propelling rapid growth and innovation.

From a competition policy perspective these applications that have been built on top of these platforms cut two ways. For one, they are the source of great value. The developers benefit from the platform directly because it makes it possible for them to engage in innovation, to write applications for your iPhone or Facebook or for Google for example, at a low cost. The consumers of the applications also benefit. But these applications also pose a possible barrier to entry. It is the old chicken and egg story. It takes a lot of effort to compete with the incumbent platforms that already have both sides onboard. As a result, if you want to compete with the iPhone, you have to cope with the fact that the iPhone has a 100,000 and counting applications.

MASHUPS AND MORPHING. We see a lot of « mash-ups » on the web. That means creating new services by combining things. Square is an example. That is a new payment system that was created by Jack Dorsey, who is one of the founders of Twitter. He has a software application that works with the iPhone. A small merchant can add a square attachment to your iPhone and can accept cards (after singing up for a processing agreement). Consumers who swipe their cards in that square device enter their emails and become part of the network. This new platform provides an alternative payment system.

9. It is also relatively easy for web businesses to morph rapidly in ways that few might have expected. You might think that LinkedIn is like Facebook. It is not. LinkedIn is a job board and recruiting tool. It makes money basically by acting as a recruiting tool, selling job postings and so forth.

Mash-ups and morphing are important for analyzing market definition and market power. A few years ago TomTom, which was a leading supplier of handheld navigational devices, bought TeleAtlas which was one of the few major suppliers of maps used for navigation.The Commission approved the merger but it twas controversial. Google has just completely disrupted that business since then by mashing up its Android phones and Google Maps. The Android phone can be an incredibly powerful navigational tool that people can use in their cars and pretty much everywhere at zero marginal cost. The stock price of TomTom and Garmin which is another navigational device maker plummeted last October after the Android phone came out with this navigational device.

IT'S ONLY JUST BEGUN. There is a tendency to think that we are at the end of history. The latest new thing is the last thing. That is what people thought when Friendster created a successful social network. And that is what they thought about MySpace when it killed Friendster. Is Facebook really the final thing ? At some point we will be at the end of history for some of these categories. Some company will nail it and will dominate for a long period of time until something completely new comes along. But it takes a long time in really only in retrospect to know when that happens.

Even then, the pace of innovation in this area should give one pause on how secure any dominant position is on the web. No one predicted in the early 2000s that Google would be playing rope-a-dope with Microsoft or that Microsoft would be complaining about someone else leveraging their monopoly power.

Competition policy for the web economy

There is no reason that competition policy cannot deal with all the issues that are going to emerge in the web economy over the next few years. The web is hardly alone in being more complicated and different than Adam Smith's pin factory. Moreover, many of the two-sided issues that arise for the web occur in traditional industries, such as physical exchanges, payments and shopping malls.

11. If there is one place where the analysis could go wrong it is market definition. Done correctly market definition ought to be a tool for understanding competitive constraints and helping to evaluate unilateral and coordinated effects. The problem arises when mechanical methods are used to draw hard market boundaries and when the perspective of this market provides a distorted view of the competition that is actually taking place in the real world.

There is a growing consensus among economists that the current tools for market definition face numerous problems especially when there is product differentiation which there almost always is. The hypothetical monopolist test is quite difficult to implement reliably. The results are highly dependent upon whatever assumption the economist is making about the shape of the demand schedule and the sequence of the products considered among other things. All those problems become much more difficult when the markets are two-sided. Market definition, for example, does not deal very well with the complementary products that

characterize two-sided markets. Many of the problems that critical loss has in one-sided markets become an order of magnitude harder in two-sided markets.

For the web economy competition policy will do better to rely on methods that are less formulaic and based more on qualitative research into the nature of competition in the business ecosystem.

As with any new kind of business the web businesses will provide creative ways of monopolizing unlawfully that we have not thought of. There will also be pro-competitive explanations that we have not imagined for practices that look suspicious. I mentioned critical mass earlier. Firms use many methods to achieve critical mass and obtain platform ignition. Firms can use these same sorts of methods to prevent their rivals from getting critical mass. It may not take much to prevent a rival from achieving ignition. That is something for competition authorities to be concerned about.

12. Google will likely become the testing ground for the next decade of entrust analysis of the web economy. The economics in this company is really pretty simple. It makes money from advertising. That means it wants spaces to put ads and it wants eyeballs to look at those ads. The more space and the more eyeballs they have the better. My suspicion is that Google enters other parts of the web primarily to ensure that nothing comes between them, the advertising space, and the eyeballs looking at that space. We are seeing this now in mobile. Advertising and eyeballs are going to mobile devices. Google has started the Android operating system and launched it own line of phones to help ensure access to mobile advertising inventory. As Google keeps entering other parts of the web ecosystem, we are going to return to the an debate. Is Google a dominant firm, leveraging its way into other businesses. Or is this a double marginalization story where a dominant firm in one market can make consumers better off by either making a related market competitive or extending its own dominance into that market. Thus, is it an anti-competitive story or an efficiency one.

13. The web economy is intellectually interesting, the issues are complicated, and there is much room for debating whether practices are good or bad. All of this will make the practice of competition policy most intriguing and perhaps profitable, for good or not, for many years to come. ■

Sir Christopher Bellamy
christopher.bellamy@linklaters.com
Senior Consultant

Two-Sided Markets: A challenge for competition policy and regulation?

Two-sided markets: A challenge for competition policy and regulation?

Abstract

The author considers competition law issues arising in the "two-sided markets" for credit and debit cards. He considers four main issues. 1. What is the legal basis for competition law intervention in the various fees charged to the various players in credit/debit card payment systems? 2. Even if, quod non, competition law is an appropriate tool to apply, what is the correct analytical approach to assessing the "fairness" or "efficiency" of the charges borne by the various parties? 3. These questions should also be considered in a banking and policy context, in which the need to reduce importance of cash and to create the right incentives to improve payment systems are very important. 4. At present, there is insufficient raw data to enable reliable competition assessments of "value" to be made, even if that is conceptually appropriate. In consequence, competition authorities should be very cautious in applying competition law in this area.

La présente contribution analyse quatre questions du marché double face des cartes de crédit et de débit : 1) Quelle est la base juridique pour l'application du droit de la concurrence à l'égard des commissions imposées aux différents intervenants des systèmes de cartes de paiement ? 2) Dans l'hypothèse où le droit de la concurrence serait un outil approprié, comment évaluer le caractère juste et efficace des différentes commissions ? 3) En tout état de cause, il est par ailleurs nécessaire de prendre en compte des considérations de politique bancaire qui considèrent comme importants la réduction de la circulation d'argent liquide et la création d'incitations à améliorer les systèmes de paiement. 4) En l'état actuel, il n'existe pas suffisamment de données disponibles pour évaluer de manière fiable le surplus concurrentiel. En définitive, les autorités de concurrence ne devraient intervenir dans ce secteur qu'avec la plus grande prudence.

1. What a great pleasure it is to be back in Lisbon among so many friends. The warmth of the welcome is matched by the warmth of the weather, at least as far as those that come from Northern Europe are concerned. I am going to try to talk a little bit more specifically about two-sided markets in the context of credit and debit card payment systems, and in particular, the levels of the various fees that pass between the various parties, which have been the subject of two decisions at the European level.

2. These were the Visa decision of 1997, which granted an exemption that expired at the end of 2007, and the subsequent MasterCard decision, which is currently under the appeal to the CFI on these issues. There are four points that I want to make, and I am going to just quickly summarize them. Firstly, in terms of European competition law, the legal basis for any intervention at all in the credit card and debit card systems is not yet clear. Secondly, assuming that it is a matter for competition law, the correct analytical approach is not yet clear and the value judgements are extremely difficult. Thirdly, I would suggest the overall policy issues are very complex and not yet fully understood. And fourthly the raw data is not yet adequate or complete to enable a sound judgement to be arrived at.

3. So, let us quickly take those four points and start by trying to identify what it is that we are talking about. In a credit card payment system, or a debit card payment system, there are at least four, possibly five, main players. We'll just quickly identify them. I, and probably everybody in this room is a card holder of some kind or another with a card issued by the bank. So the first player is the card holder, and the second player is the bank who issues the card. The card is only useful, of course, if it can be accepted at a retailer. So the third player is the retailer who accepts the card. When the retailer accepts the card, the retailer is reimbursed for the price of the transaction less a fee by what's called the acquiring bank, who is the fourth player. The acquiring bank arranges for the retailer to be put in funds as a result of the agreement between the retailer and the acquiring bank. The acquiring bank is then reimbursed by the issuing bank, debits the card holder's account and that's basically the way the transaction works. So far, then, we've got the card holder, the issuing bank, the retailer and the acquiring bank. There is, of course, the fifth player: the platform that is providing the overall service, which in international terms are essentially the Visa and MasterCard systems, which are there, really, to benefit all the players: the various banks, the card holder and the retailer.

4. There are various flows of fees. The two chief ones are that the retailer pays a fee to the acquiring bank which is called the merchant service charge (MSC); and the fee which the acquiring bank pays to the issuing bank which is called the multilateral interchange fee or the MIF. The essential complaint of the retailers is that the merchant service charge is too high, and that they are being effectively required to pay the MIFs, which go to the issuing bank.

5. The overall argument in this area is about how the cake, as it were, should be split up. In such a "two sided" market, where no one element can function without the other, it is always difficult to know whether one side or the other bears too much cost. So, it's worth identifying the services and commercial transactions that are going on in this system. As between the two banks concerned there is the question of authorisation, of processing and of transmitting the payments to each other. As between the issuing bank and the card holder, the issuing bank is providing a service to the card holder that may well be, and very often is, part of a general banking service which includes a current account, probably an overdraft facility, money transmission services and so forth, in which the credit or debit card is only one feature.

6. There may be or may not be a separate fee for the credit card paid by the card holder, but the credit card is essentially one part of the banking relationship between the issuing bank and the customer.

7. As regards the retailer/acquiring bank relationship, when the acquiring bank reimburses the retailer, the retailer gets immediate cleared funds, he gets protection against fraud and he is able to make a sale to a customer who is buying on credit without himself having to supply that credit. So he doesn't have to do a credit check on the customer, and he doesn't have to incur his own cost of credit. He can, as it were, make the sale without himself advancing the credit. Typically, the retailer will pay for those services in the MSC. Let's say, for argument's sake, that he pays the acquiring bank one per cent of the transaction value. The acquiring bank then pays the interchange fee or MIF across to the issuing bank and, let's say, that's half a per cent. This half a per cent MIF is reflected in the MSC charged to the retailer. What the retailers effectively say is that the MSCs are too high, because the fact that the MIF flows from the acquiring bank to the issuing bank means that the merchants, are effectively paying for this transaction between the acquiring bank and the issuing bank, and that this is a distortion in the market place.

8. On the other hand, it is strongly argued that the fact that the MIFs flow to the issuer, incentivize the issuing bank to issue cards, to expand the system, to encourage more people to hold cards, to innovate to create new payment systems, to improve fraud control and so forth. All those benefits actually go back indirectly or directly to the retailer, because the more card holders there are, the more useful it is to the retailer to belong to the system, so that all the relationships in this complicated system are inextricably interrelated.

9. It said that, for example, if there was no interchange fee moving from the acquiring bank to the issuing bank, then there might have to be a separate charge to the card holder or the existing charges to card holders would have to be raised, which would discourage card holders from having cards. That would lead to a disadvantage to the retailer, because he would have less people presenting cards, so he would have less sales etc. We have all these interrelationships, and the challenge for competition lawyers is how on earth we are going to analyse all this. Instinctively, I suppose, the analysis focuses on the fact that the interchange fees between the banks is set by the banks themselves, who are members of the system collectively deciding what the interchange fees are going to be, as distinct from negotiating by bi-lateral arrangements between all the various participating banks. So it looks like, on a very superficial level, a sort of price fixing, making people a bit suspicious about what is going on.

10. How do we analyse all this in competition terms? Let us now just briefly see what has been said in the cases so far. What has been said essentially in both the Visa and MasterCard decisions is that the multilateral interchange fees or MIFs restrict competition between the acquiring banks.

11. The argument is that because this is a common cost that all the acquiring banks effectively have to bear which they pass onto the retailers, this in some way affects the competition between the acquirers and reduces their ability to negotiate lower charges to the merchants. That was effectively the line that was taken in the Visa decision for the application of article 81 (1) or article 101, as it is now. That analysis, that first front line analysis of whether competition law applies at all, was not challenged by Visa in the Visa decision, because they were given an exemption, in the days when you could get an exemption.

12. In the Visa exemption decision there was basically an agreement as to what the level of the MIFs should be. And what was said in that decision was that the MIFs can be calculated by reference to the costs of issuers. So, you can take into account the cost of processing, the cost of fraud and fraud prevention measures; and you can take account of the cost of the free funding period, on the basis that that is a cost and it's also something that indirectly benefits the merchants, because the merchants benefit from the fact that they can sell goods on credit without having to provide the credit themselves. In the Visa decision on a weighted average basis the MIFs for international transactions were limited, to 0.7% of the transaction value. The point to bear in mind is the case is essentially about the amount of the MIFs. (There have been various other cases at national level, with a pretty patchy record of success I have to say and I am not going to talk about those).

13. The second case building on from that is the MasterCard case in which there was some change of approach by the Commission. The Commission basically said words to the effect "we are not convinced that the MIFs are necessary for the system to operate at all. You could do without MIFs altogether and it's up to you to convince us under article 101 (3) and show us why MIFs are either necessary, useful, or efficient, or produce benefits. You, MasterCard, have failed to produce any evidence or sufficient evidence, to that effect".

14. Then in the Decision, the Commission sat back and said "*No, no you haven't proved a thing, under Article 101(3) therefore we ban your MIFs altogether.*" That case is now under appeal to the Court of First Instance; the primary argument being, as far as one can tell, that article 101 does not apply at all to these arrangements.

15. The Mastercard Decision was followed by a settlement about a year later, an interim agreement, I should say, between MasterCard and the European Commission in which various

public statements were made by the Commission to say "*No, no, no. We are not against MIFs as such, they may well produce good effects, etc. We are entirely in favour of the* right *MIFs*." And the "right" MIFs according to the oracle as pronounced in the press release, are 0.2% for a debit card transaction and 0.3% for a credit card transaction.

16. No doubt that was to some extent a negotiated deal, but one of the conceptual reasons put forward for it which was not particularly elaborated, is a new approach to the setting of MIFs, which is called the merchant indifference test. According to this view the MIFs should be so adjusted, or the system should be so balanced that, from the merchant's point of view, when a card is presented, he shouldn't mind whether it is a debit card, a credit card or cash, because the cost to him or the overall cost, taking into account all his costs, the opportunity cost and the rest of it, will be the same. So he shouldn't care which payment method he accepts. And that's apparently the basis for the MasterCard settlement. Further proceedings have now been opened against Visa. Their exemption was not renewed when it expired at the end of 2007 and their card system is now the object of statements of objections.

17. So we've got at least two conceptual methods for trying to decide what is a "fair" value as between the two sides of these two-sided markets' (card holders and retailers). We've got the *issuers' cost method* in the first Visa decision and we've got *merchant indifference* in the MasterCard decision.

18. Now to come to the very first point that I made: what is the legal basis for any intervention at all under competition law, under article 101 (1) of the Treaty. As a general matter, competition law finds it extremely difficult to grapple with arrangements that intrinsically require cooperation between a number of parties. The obvious example is sport. Sport can't happen, football can't happen, unless you've got the rules of football that everybody signs up to - that you've got the rules for running a football league, that you've got two teams that will compete against each other, and the whole thing is a structured system. Nobody suggests that the offside rule in football is a restriction on competition and the reason is that you couldn't have the sport in the first place unless you had a set of rules to enable it to happen.

19. The first question that rises conceptually is: could you have a credit card system on the scale of MasterCard and Visa, dealing with international transactions without any kind of rule about the direction in which the various fees and charges in the system should flow? Could you do it with everybody negotiating with everybody else as to what charges they would pay each other for acquiring and so forth?

20. Originally it was said by competition authorities that you could manage these matters on a bilateral basis, and every bank should negotiate with every other bank, and that would be a competitive solution. There are about 5,000 banks that belong to each of these systems, so I forget exactly what the arithmetic is, but it means about 3 million separate bilateral transactions at the extreme. In international systems such as these, that is a hopelessly impractical and inefficient way of doing it. What you need to have, and this is effectively what the MIFs are, is a default rule as to what is to happen if there is no bilateral agreement between the two banks concerned. That is essentially what the MIF is, it is a default rule.

21. So the first question is, could the system operate without a default rule? Even if the default rule is that there should be no MIFs, that is still a rule. And arguably, even a rule that said there should be no MIFs would itself be on the correct analysis a restriction of competition. So if you accept, for argument's sake, which seems to be a credible point of view, that you need a default rule, that helps us analytically, because it is then clear you are not arguing about the rule as such, you are arguing about the amount. You are effectively arguing about how the cost should be shared, either in fairness, if you believe that's an approach, or in terms of economic efficiency, if you believe in that approach, between the two sides of this two-sided market. And that is an extremely difficult question for competition law to answer, certainly in the European context, to decide what is fair, what is efficient between these two sides is extremely difficult, using competition law.

22. It should also be said, I think, that the argument, which is essentially the Commission's argument, that the common cost puts a floor in the acquiring market and therefore restricts competition between acquirers, is itself pushing the idea of restriction of competition under Article 101 (1) a pretty long way. It wouldn't normally be said that the fact that suppliers of diesel fuel, for example, face a common cost of oil or a common excise tax restricts the way they compete with each other. The way that suppliers compete with each other, the way that acquirers compete with each other, depends on the service they can offer, the price they are offering, the quality of the goods and so forth and so on.

23. It can fairly be said that it is a pretty strained construction of the idea of "restriction in competition" under the Treaty to go down this route. It may also be that the argument is not actually, as a matter of fact, correct, because the acquiring bank who is reimbursing the retailer may well have a banking relationship with that retailer, such as the retailer's bank account, the lending to the retailer, the cost of handling his cheques, the cost of delivering cash, all those things.

24. In those circumstances, it does not necessarily follow that the MSC is going to be separately negotiated, in a way that is outside the ambit of the whole banking relationship. The MSC may be just one element in a wider banking relationship and it is conceptually possible to find that there are MSCs that don't actually cover the MIFs. That's the question or fact.

25. So the first point is that there is still a lot of debate to be had as to the conceptual basis for competition law to intervene at all. But assuming, quod non, that there is a basis to intervene: what is the correct analytical approach? We have the approach of cost methodology, we've got this merchant indifference methodology, but have we got anywhere near a complete theory to help us with this difficult question? Competition law, at least in Europe, and probably worldwide, is very bad at deciding what the "fair value" of something is, whether something is being unfair or inefficient between one lot of players and another lot of players. For example, it

is very difficult to work out what exactly are the elasticities: whether, and to what extent, a rise in cost to cardholders would lead to a reduction in credit card usage and therefore be to the detriment of the merchants and to what extent retail prices would fall if you reduce the MSCs. All these are very difficult issues.

26. One of the most difficult issues, certainly if you try to look at it from the efficiency point of view, is that credit and debit cards are only two of several payment methods. There is also, of course, cash, as well as cheques and other sorts of money transmission. We are in a world in which in my view at least, we already face massive distortion from the fact that cash is not priced at all. We often assume that cash doesn't cost anything. You don't pay anything when you go to your bank to get some cash, you get it at par, you pay the retailer in cash at the par value of the goods, but the cost of cash is enormous. It's not just the inconvenience and time spent by the consumer going to the bank getting cash or going to the ATM and getting cash and carrying it around and losing it, getting it stolen and all the rest of it. But also at the retailer's endthe handling of cash is very expensive, very time consuming, sometimes very risky and so forth and so on. Cash has an enormous cost to the economy; but it is not priced at all. There are similar arguments on cheques as well. So it's rather difficult to just look at credit and debit cards without a very good understanding of what the cost of cash is and what the cost of alternative payment systems are, if you are trying to get the pricing signals right. So that's one very important element.

27. And secondly, as I have said, it's quite difficult in banking terms to isolate these particular transactions, because from the card holder's point of view, the card is only one aspect of his banking relationship, which includes his current accounts and his borrowing facilities, money transmission, internet banking, and all the rest of it. And on the retailer's side the acquiring activities are also very often part of a banking relationship. Not to mention the fact that although we talk about four party systems with card holder, issuing banks, retailers and acquiring banks, the issuing bank and the acquiring bank might very often be the same bank, because banks are both issuers and acquirers in many countries – not in all, because there are some very, perhaps, undeveloped acquiring systems in some member states – but the so called "*on-us transactions*" where the acquirer and the issuer is the same bank are by no means an unimportant aspect of the system.

28. Then you have the fact, and this brings us back to the two sided nature of the market, that the merchant is undoubtedly both benefiting from the service provided to the cardholder and the service provided to him, the retailer, because, as I say, the retailer doesn't have to worry any longer about giving credit, credit checks, about being paid or about charging interest. It's only a comparatively recent idea that retailers are not involved or shouldn't be involved in providing credit and should not have to pay for being relieved of that burden. Until banking became more universal it was very common and still is quite common for a retailer to be involved in the provision of credit. That still, of course, carries on with store cards.

29. So it is very difficult to know conceptually how to do it if you are going to make some kind of balance in the system either from the point of view of efficiency or from the point of view of fairness between the two sides. On top of all that, there are very wide policy issues, particularly the incentives that are needed to invest in new technology, and where those incentives are to come from - the incentives to set up a proper European payment system through the various SEPA suggestions. I mean it is outrageous, isn't it, that in Europe we still can't do a cross border direct debit and such like. But how a properly integrated European payment system is going to work is part of the overall conceptual issue. And this is already a part of the banking context.

30. And lastly, have we got enough information, enough reliable information, enough data, to even begin to start making these judgements? The available information about cost: cost of cash, cost of cheques, the information: about investment, about incentives, about the demand relationship between the two sides of these markets, is still pretty rudimentary. There seems to be, in my view, a big question mark as to whether we are anywhere near having the raw data and the intellectual basis for a proper analysis of these matters.

31. My own view is that we are some way away from taking sound decisions on these issues. And I would encourage competition authorities, whether at European or domestic level, to think very hard and go quite slowly in trying to solve these problems, because they are very difficult. Thank you very much indeed. ■

Frank A. Wolak
wolak@zia.stanford.edu
Department of Economics, Stanford University

Managing demand-side economic and political constraints on electricity industry re-structuring processes

Abstract

This paper identifies the major political and economic constraints that impact the demandside of electricity industry re-structuring processes. It then describes how these constraints have been addressed and how this has harmed market efficiency and system reliability. Finally, the paper proposes demand-side regulatory interventions to manage these constraints in a manner that limits the harm to wholesale market efficiency.

1. Economic and political factors constraint electricity industry re-structuring processes. Powerful entities that existed before re-structuring continue to exercise this political clout in the new regime. Former state-owned or privately-owned vertically-integrated geographic monopolies maintain their dominant position in the new regime. Existing regulatory agencies continue to exercise control over market participant behavior even if these actions adversely impact wholesale market efficiency. New market participants find their demands ignored in favor of those by the more politically power incumbent firms. Conflicts between regulatory agencies arise because of the uncertain boundaries of the authority between these agencies brought about by the restructuring process.

2. The primary economic factor constraining most re-structuring processes is a physical infrastructure poorly suited to the wholesale market environment. A transmission network with insufficient transfer capacity between generation unit owner locations makes it extremely difficult for competition between suppliers to discipline wholesale electricity prices at all locations in the transmission network. The lack of hourly meters on the premises of final electricity consumers prevents retailers from setting default retail prices that vary with hourly wholesale prices. These political and economic constraints on the re-structuring process exert the greatest influence on the demand-side of the market. For example, the motivation often offered for bid caps and other market power mitigation mechanisms is to protect final consumers from an inadequate physical infrastructure to support competitive market outcomes.

3. This paper identifies the major political and economic constraints that impact the demandside of electricity industry re-structuring processes. It then describes how these constraints have been addressed in previous re-structuring processes and how this has harmed market efficiency and system reliability. Finally, the paper proposes demand-side regulatory interventions to manage these constraints in a manner that limits the harm to wholesale market efficiency. Each of the next three sections of the paper is devoted to one of these tasks.

I. Major political and economic demand-side constraints

4. Because electricity is a necessary input to so many economic activities, there are significant political obstacles to charging business and residential users retail prices that reflect the hourly wholesale price of electricity. A long history of retail electricity prices that do not vary with realtime system conditions makes this task even more difficult. Finally, the lack of hourly meters on the customer's premises makes it impossible to determine precisely how much energy each customer withdraws in a given hour. These factors combine to make it virtually impossible to allow retail prices to allocate the available supply to final consumers willing to pay the market-clearing price as is the case for other energy sources such as oil and natural gas.

1. The political economy of electricity prices

5. Under the vertically-integrated geographic monopoly regime in the United States (US), retail electricity prices are set by state public utilities commissions (PUCs). Although these geographic monopolies are usually privately-owned firms, they are also among the largest employers in the state, so the PUC must balance the interests of ratepayers and employees of the company in the price-setting process. The usual regulatory bargain in the US is that the vertically-integrated monopoly utility must serve all demand at the prices set by the PUC, and the PUC must set retail prices that allow the utility an opportunity to recover all prudently incurred costs incurred to serve that demand.

6. This regulatory history has established a public precedent for retail electricity prices that only recover total production costs, or prices equal to the long-run average cost of supplying electricity. However, prices set through market mechanisms can often be vastly in excess of or substantially below the average total cost of supplying the product. This is particularly true for wholesale electricity because of a number of features of the technology of electricity supply discussed in Wolak (2004) that make these markets extremely susceptible to the exercise of unilateral market power by generation unit owners.

7. Setting retail prices that pass through hourly wholesale prices is even more difficult in the US because there are explicit regulatory prohibitions against consumers paying wholesale prices that reflect the exercise of unilateral market power. As discussed in Wolak (2003b), the Federal Power Act of 1930 requires that the Federal Energy Regulatory Commission (FERC), the United States wholesale market regulator, to ensure that consumers do not pay unjust and unreasonable wholesale prices. FERC has determined that market prices that reflect the exercise of unilateral market power by suppliers are one example of unjust and unreasonable prices.

8. This problem is further compounded by the fact that most state PUCs have prohibitions against passing on unjust and unreasonable wholesale prices in the retail prices they set. For example, if the FERC determines that certain wholesale prices are unjust and unreasonable because they reflect the exercise of unilateral market power, then it is illegal for the state PUC to set retail prices that recover these costs. Unjust and unreasonable wholesale prices are imprudently incurred costs and therefore the PUC has no obligation to set a retail price that recovers these costs.

9. This regulatory structure creates an almost impossible situation for introducing an active demand-side into the wholesale market. Requiring consumers to manage hourly wholesale pricerisk will create the necessary price-responsive final demand that limits the opportunities for suppliers to exercise unilateral market power in the short-term market. This retail pricing policy will also very likely lead to wholesale prices that are vastly in excess of the average cost of supplying electricity during a number of hours of the year, which may mean that consumers are being charged illegal prices for wholesale electricity during these hours. However, without the active demand-side participation enabled by charging final consumers prices that reflect hourly wholesale prices, electricity suppliers will face a final demand that is virtually inelastic with respect to the hourly wholesale price and implies significant opportunities for suppliers to exercise unilateral market power.

10. Although this explicit regulatory conflict between retail prices that pass through hourly wholesale prices and the legality charging these prices to consumers does not exist in other countries, the same precedent exists for setting retail electricity prices equal to the average total cost of supply. In the former state-owned monopoly regime a government agency or regulatory body was charged with setting retail electricity prices to allow the firm to recover its production costs. Outside of the US, there was also a strong aversion to retail electricity price volatility. Substantial input cost increases were slowly phased into retail electricity prices.

11. In developing countries, there is even greater pressure to keep nominal electricity prices as low as possible because of the crucial role electricity is thought to play in the development process. These concerns have often led to retail prices that only recover the variable costs of supplying electricity. In some of these countries, electricity prices are used to pursue political goals. For example, since 1977 politicians in various regions of India have run on a platform of subsidized or even free electricity for farmers (Mukherjee, 2007).

12. These political constraints emphasize why it is so difficult for the political process to require final consumers to purchase wholesale electricity at prices that reflect hourly wholesale prices. The fact that few consumers have meters on their premises that measure their consumption on an hourly basis ensures that this situation will not change without significant regulatory intervention.

2. The economics of hourly metering

13. Virtually all electricity meters that exist in the United States and other industrialized countries record the total amount of electricity consumed on a continuous basis. A customer's electricity consumption over any time interval is the difference between the value on the meter at the end of the time period and value at the beginning of the time period. In the US, meters are typically read manually on a monthly or bi-monthly basis. A meter reader must show up at the customer's premises and record the value on the meter. If a meter reader is unable to make it to the customer's premises there are rules for determining the customer's consumption during that billing period.

14. Another feature of electricity retailing in the US is that customers receive their bill for last month's consumption during the current month. With bi-monthly metering and billing, the delay between consumption and invoicing can be more than one month. If the only information a customer receives about the cost of their consumption during the previous billing cycle is provided at the end of this billing

cycle plus a processing delay, it is unclear how hourly retail prices that vary with hourly wholesale prices can be used to cause final consumers to alter their demand in real-time. Some signal about the value of the hourly wholesale price must be provided to final consumers to cause them to alter their real-time demand.

15. Hourly metering technology can both record consumption each hour of the month and provide information to the customer on the value of hourly prices. There are a variety of technologies available to accomplish this, but all of them share similar cost structures. There are significant upfront costs in terms of infrastructure to install the meters and the technology necessary to read and record the output from the meters. In addition, the average cost of installing meters is much less if they are installed in volume in over a small geographic area. Once installed in volume, the monthly average cost of operating the system is very low, less than $0.50 per meter-month.

16. Consequently, the tradeoff for an investment in interval metering is whether the cost saving in terms of the reduced labor costs associated with monthly manual meter reading and wholesale energy purchase costs to serve final consumers are sufficient to recover the up-front costs of installing the meters plus the monthly cost of operating the automated meter reading network.

17. Figure 1 provides a graphical illustration of an automated meter reading network. The meter must automatically communicate either by wire or by wireless technology to the data center each hour of the day to send consumption information back to the data center. From this data center the information is sent to the retailer, who can also share it with the final consumer. For example, virtually all automated meter reading networks have the capability for consumers to download information on their consumption of electricity as soon as it is recorded at the data center.

18. The major drivers of the economics of installing of an automated meter reading networks are labor costs and the level and volatility in wholesale electricity prices. In regions where labor costs are higher, the cost savings from eliminating manual meter reading are larger. In areas with higher and more volatile electricity prices, the cost savings in wholesale electricity purchase costs from being able to use price signals to shift demand throughout the day, week, or month are much greater than in a region with low and/or stable wholesale electricity prices. Consequently, a fossil fuel-based system which has substantial price fluctuations within the day, has a much greater potential to realize significant cost savings from an automated meter reading network than a hydro-based system which typically has fairly constant wholesale prices throughout the day.

19. Wolak (1999) compares the time series behavior of prices in re-structured electricity markets in Australia, New Zealand, England and Wales and Norway and Sweden. Australia and England and Wales are fossil fuel-based systems with substantial amounts of price variation within the day, whereas New Zealand and Norway and Sweden are hydro-based industries with small price fluctuations within the day-ahead, although there can be substantial price differences across seasons of the year, depending on hydrological conditions. By the above logic, the wholesale electricity purchase savings from an automated meter reading network in Australia and England and Wales should be significantly higher than those in New Zealand or Norway and Sweden.

20. It is important to emphasize that this statement does not imply that New Zealand and Norway and Sweden would not benefit from retail prices that pass through wholesale prices. Because the major source of price variation in these markets is across seasons of the year or across years, virtually all of the savings from passing through wholesale prices in retail prices could be accomplished with monthly meter reading.

21. Customers with larger annual electricity bills can expect to realize greater total benefits from hourly meters than smaller customers. Any given percentage saving in wholesale electricity purchase costs from having an hourly meter will translate into a larger total annual dollar savings, which increases the likelihood that the annual total benefits of an hourly meter for that customer will exceed the annual cost. For example, a five percent savings in wholesale electricity purchase costs applied to an annual electricity bill of $1000 only yields $50 in saving. Applying this same percentage to an annual electricity bill of $10,000, yields $500 in saving, which more than covers the annual cost (including fixed costs) of installing and operating an hourly meter.

22. Another factor determining the benefits a customer might realize from hourly metering is the amount that customer can reduce its demand in response to price signals. The magnitude of a customer's demand responsiveness depends on the mechanism used by the retailer to deliver the price signal. There has been a substantial amount of recent research on ways to deliver wholesale price signals to final consumers to maximize the cost savings realized from providing these price signals.

23. Simply passing through the hourly wholesale price in an hourly retail price may not provide the greatest aggregate reduction in wholesale purchase cost savings by customers with hourly meters. By coordinating the demand reduction efforts of consumers with hourly meters it is possible to realize an additional source of benefits from price-responsive final demand besides those that result from customers shifting consumption to lower-priced hours of the day from the high-priced hours of the day. Coordinated actions in the same time interval to reduce demand by all consumers with hourly meters can reduce total system demand which can then lead to lower wholesale prices. These coordinated actions increase the total benefits realized from hourly meters because customers without hourly meters benefit from lower wholesale electricity purchase costs enabled by a lower wholesale electricity price. Section 3 discusses alternate hourly pricing mechanisms that attempt to capture both the load shifting benefits and the wholesale price-reducing benefits of hourly metering of final consumers.

24. The costs of hourly metering and the magnitude of the typical annual household electricity bill in most developed countries make it difficult for the expected benefits from

universal hourly metering not to exceed the costs, particularly given the significant economies to scale and geographic economies to scope in the installation and operation of hourly meters. This economic logic is consistent with recent regulatory decisions made in a number of jurisdictions. The state of Victoria in Australia, the province of Ontario in Canada and the state of California in the US have all decided to implement universal hourly metering for all consumers. In Victoria, the plan is to install approximately 2.5 million hourly meters by 2013. In Ontario, 5 million meters are to be installed by 2010. In California, no firm deadlines have been set, but the plan is to install hourly meters for all customers of the three large investor-owned utilities. In all these jurisdictions, the cost of the hourly metering technology will be included in the regulated cost of local distribution services.

25. For developing countries, the economic case for universal hourly metering is much less favorable because the labor costs associated with manual meter reading are much lower and annual residential and business electricity bills tend to be much lower. However, in the areas where affluent households live there are residential customers with sufficiently large annual electricity bills to pass the cost/benefit test for hourly metering. There are also likely to be many industrial and commercial customers that are viable candidates for hourly meters. Consequently, determining precisely where to draw the line between customers that pass the net benefit test for hourly meters and those that do not is much more difficult in developing countries. This logic also suggests an alternative approach to organizing the retailing segment of the industry in developing versus industrialized countries discussed in Section 4.

3. The political economy case against real-time pricing

26. The need for explicit regulatory intervention to install the metering infrastructure necessary for the widespread implementation retail pricing plans that reflect real-time wholesale prices has significantly slowed the pace of adoption of these technologies. Particularly in the US, the regulatory framework governing the electricity industry re-structuring process has further conspired against the adoption of interval meters and real-time pricing plans. Because the cost of conventional manual metering reading is sufficiently high that in places like California, the case for the adoption of automated meter reading technology can be largely made using the costs savings associated with the elimination of manual meter reading.

27. Despite the adoption of automated meter reading technology, most state PUCs have been extremely reluctant to implement retail pricing plans that reflect hourly wholesale market conditions because of the apparent contradiction with their regulatory goals. Mandating that all customers face an hourly retail price that passes through the hourly wholesale price would create strong incentives for final consumers to respond to hourly wholesale prices. A real-time demand for electricity that responds to hourly wholesale electricity prices is crucial to limiting the opportunities for suppliers to exercise unilateral market power in the wholesale market and ensuring competitive wholesale market outcomes. However, such a requirement would also expose final customers to prices that reflect the exercise of significant unilateral market power during a number of hours of the year. For this reason, a state PUC might argue that setting a pass through of the hourly wholesale price as the default retail price is inconsistent with its regulatory mandate to protect consumers from unjust and unreasonable retail prices.

28. The fact that this very straightforward solution to the lack of a price elastic wholesale electricity demand has been rejected by all US state PUCs, suggests these entities view a default retail price that passes through hourly wholesale prices in hourly retail prices as explicitly or implicitly inconsistent with their regulatory mandate. As we discuss in the next section, the response of state and federal regulators in the US to the risk of very high hourly wholesale prices has been to implement regulatory interventions that limit price volatility but very likely increase average wholesale prices and reduce system reliability.

29. The situation in wholesale markets in other industrialized countries is not much better. There is very little penetration of hourly metering technology in most of these markets because of the reluctance of the regulators to mandate its adoption. As noted above, recently this trend has begun to reverse, but it remains to be seen if once these meters are in place default prices that pass through hourly wholesale prices will be adopted.

II. Mechanisms adopted to deal with constraints on demand-side participation

30. The desire of policymakers to shield final consumers from wholesale price risk has led to a number of regulatory interventions that significantly degrade the efficiency of the re-structuring industry in the short-term and long-term. The first is the implementation of the bid caps and other market power mitigation mechanisms in the short-term market. This has led retailers and final consumers to engage in an inadequate amount of hedging of short-term price risk and claims by generation unit owners that the existence of bid caps and other market power mitigation mechanisms prevent them from full revenue recovery. These claims have led to a number of regulatory interventions that provide additional revenue to generation unit owners. However, these revenue flows also raise total wholesale energy costs to consumers and decrease the likelihood they will receive any benefits from industry re-structuring.

The one bright spot in these regulatory adaptions to political and economic constraints on demand-side participant is experimental evidence on real-time pricing programs that final electricity consumers find politically palatable. As discussed in Section 3.4, these programs share the real-time price risk between electricity retailers and final consumers in a manner that can share the benefits between the two parties.

1. Offer caps and market power mitigation mechanisms

31. Virtually all short-term wholesale markets operating in the world have offer caps that limit highest price offer a supplier can submit or a price cap that limits the value the market clearing price can take on. In the US, FERC has set the maximum offer cap in the eastern US markets at $1,000/MWh. In California, the current offer cap is $400/MWh, but it is scheduled to increase to $1,000/MWh to be consistent with the offer caps in the eastern US markets. All US markets also have local market power mitigation mechanisms that limit the maximum bid a generation unit owner can submit when it is determined to possess local market power.

32. Wholesale markets in other parts of the world also have offer caps. For example, the Australian market currently has an offer cap of 10,000 Australian dollars per MWh. The Alberta electricity market currently has a price cap equal to 1,000 Canadian dollars per MWh. In the Nord Pool, the market operator sets a maximum and minimum offer price each day as part of the operation of the day-ahead energy market.

33. These bid caps and price caps and local market power mitigation mechanisms are proposed to address the fact that the wholesale demand for electricity is completely price inelastic because of the lack of hourly meters and default retail prices that pass through hourly wholesale prices. One of the standard tests to determine whether a supplier possesses local market power worthy of mitigation in US wholesale markets is whether that supplier or a small number of suppliers is pivotal to meet demand in a congested portion of the transmission network. A supplier or group of suppliers is pivotal if removing their supply implies that demand could not be met by the remaining suppliers alone. For example, with five firms each owning 100 MW of capacity, if the demand for electricity is above 400 MW, then each of the five suppliers is pivotal. If the demand is not completely price inelastic then no supplier could be pivotal, because there would always be a price at which the demand would equal the available supply.

34. There would be significantly less need for bid caps and local market power mitigation mechanisms if final consumers were required to manage short-term wholesale price risk. If all final consumers had hourly meters and were required to pay the hourly wholesale price as part of their default hourly retail price, consumers would sign fixed-price forward contracts for their essential demand so that they could consume this quantity of electricity each hour regardless of the hourly price. These consumers could then to alter their hourly demand around this contracted essential demand in response to hourly price signals. In this way, the need for bid caps and other market power mitigation mechanisms would be significantly reduced.

2. Inadequate hedging of short-term prices and the reliability externality

35. These offer caps and market power mitigation mechanisms create incentives for market participant behavior that can significantly degrade market efficiency and system reliability. Offer caps limit the potential downside to electricity retailers and large consumers (able to purchase from the short-term market) delaying their purchases of electricity until real time operation. These offer caps also create the possibility that real-time system conditions can occur where the amount of demanded at or below the offer cap is less than the amount suppliers are willing to offer at or below the offer cap. This outcome implies that the system operator must be forced to either abandon the market mechanism or curtail load until the available supply offered at or below the offer cap equals the reduced level of demand. Because random curtailments are used to make demand equal to the available supply at or below the bid cap, this mechanism creates a reliability externality that further increases the incentive of retailers to rely short-term market purchases.

36. Particularly for markets with very low offer caps, retailers have little incentive to engage in sufficient fixed-price forward contracts with generation unit owners to ensure a reliable supply of electricity for all possible realizations of demand. For example, a 200 MW generation unit owner that expects to run 100 hours during the year with a variable cost of $80/MWh would be willing to sign a fixed-price forward contract to provide up to 200 MWh of energy for up to 100 hours of the year to a retailer. Because this generation unit owner is essentially selling its expected annual output to the retailer, it would want a $/MWh price that at least exceeds its average total cost of supplying energy during that year. This price can be significantly above the average price in the short-term wholesale market during the hours that this generation unit operates because of the offer cap on the short-term market and other market power mitigation mechanisms. This fact implies that the retailer would find it expected profit-maximizing not to sign the forward contract that allows the generation unit owner full cost recovery but instead wait until the short-term market to purchase the necessary energy at prices that are the result of offer caps and the market power mitigation mechanism.

37. Although this incentive for retailers to rely on a mitigated the short-term market is most likely to impact generation units that run infrequently, if the level of demand relative to the amount of available supply is sufficiently large, it can even impact intermediate and baseload units. Because of the expectation of very low prices in the short-term market and the limited prospect of very high prices because of offer caps or market power mitigation mechanisms, retailers may decide not to sign fixed-price forward contracts with these generation unit owners and purchase their energy in the short-term market. By this logic, a mitigated short-term energy market always creates an incentive for retailers to delay purchasing some of their energy needs until real-time, when the market power mitigation mechanisms on the short-term market can be used to obtain this energy at a lower price than the supplier would willingly sell it in the forward market.

38. The lower the offer cap and the more stringent the market power mitigation mechanisms are, the greater is the likelihood that the retailer will delay their electricity purchases to the short-term market. Delaying more purchases to the short-term market increases the likelihood of the event that insufficient supply will bid into the short-term market at or below the offer cap to meet demand. Because of the lack of hourly metering, there is no way to determine precisely how much electricity each customer is consuming during these time periods. For this reason, system operators manage these shortfalls by curtailing sufficient load to allow the available supply meet the remaining demand. If retailers know this is how supply shortfalls in the short-term market will be met, this creates an additional incentive for them to rely on the short-term market.

39. If a retailer knows that part of the cost of its failure to purchase sufficient fixed-price forward contracts will be borne by other retailers and large consumers, then it has an incentive to engage in less fixed-price forward contracts than it would in a world where all customers had hourly meters and all customers could be charged hourly prices high enough to cause them to reduce their demand to equal the amount of supply available at that price. As discussed in Wolak (2003), all of the wholesale markets in Latin American recognize this incentive to purchase too much energy in the short-term wholesale market when it is subject to offer caps or other market power mitigation mechanisms. These countries address this incentive to under contract by mandating forward contract coverage ratios for retailers and large consumers that have the option to purchase from the short-term market. For example, in the Brazilian market all retailers and large consumers are required to have 100% of their final demand covered in a fixed-price forward contract.

40. Without these forward contracting requirements on retailers and large consumers, a wholesale market with offer caps and stringent market power mitigation mechanisms and final consumers without hourly meters face significant reliability challenges in both the short-term and long-term. In the short-term market the lower the bid caps and more stringent the market power mitigation mechanism the greater the likelihood of that there will be insufficient supply offered into the shortterm market at or below the offer cap to meet demand. Because of the mitigated short-term market and inadequate fixed-price forward contracting by retailers, there is a likelihood that new generation entrants will be unable to earn sufficient revenues from the selling in the short-term market and therefore unwilling to construct new generation units to serve load growth, which increases the likelihood of future supply shortfalls.

3. Capacity markets and other "cures"

41. A number of "remedies" have been proposed for bid caps and market power mitigation mechanisms necessitated by the lack of hourly metering and the pass-through of hourly wholesale prices in the default retail prices. Capacity payment mechanisms are the most common. The major rationale for capacity markets in the US appears to be a holdover from the vertically-integrated regulated regime when capacity payments compensated generation units for their capital costs, because the regulatory process typically reimbursed unit owners for their variable operating costs.

42. It is important to emphasize that in a wholesale market regime all generation unit owners have the opportunity to earn the market-clearing price which is typically above a generation unit's average variable cost when the unit is operating. In this way, the generation unit earns a return to capital during each hour it produces electricity. This paradigm for earning a return on capital from the difference between the market price and the firm's average variable cost of production has managed to provide the appropriate incentives for investment in new productive capacity all workably competitive industries. There is little reason to expect that it could not work in the wholesale electricity industry with an active demand side.

43. Capacity payments typically involve a dollar per kilowatt year ($/kW-year) payment to individual generation units based on some measure of the average amount of their capacity available to produce electricity within the year. For example, a baseload coal-fired unit would have a capacity value very close to its nameplate capacity, whereas wind generation facility would have a capacity value significantly below its nameplate capacity.

44. Capacity payment mechanisms differ along a number of dimensions. In some regions, the payment is made to all generation unit owners regardless of how much total generation capacity is needed to operate the system. In other regions, the independent system operator (ISO) specifies a system-wide demand for capacity equal to peak system demand plus some planning reserve, typically between 15 to 20 percent, and only makes capacity payments to enough generation units to meet this demand.

45. There have been attempts to use market mechanisms to set the value of the $/kW-year payment to the generation units needed to meet the total demand for capacity. However, these market mechanisms have been largely unsuccessful because they are extremely susceptible to the exercise of unilateral market power, because of the pivotal supplier problem created by the inelastic demand for capacity. In the eastern US markets, there have been numerous instances of the exercise of the enormous market power in these capacity markets. During the off-peak months of the year when no single supplier is pivotal in the capacity market, the price of paid for capacity is very close to zero, which is the marginal cost of a supplier providing an additional megawatt (MW) of available capacity from existing generation capacity. During the peak and shoulder months when one or more suppliers are pivotal in the capacity market, there is no limit on the price a supplier can charge. For example, suppose a market has 10 suppliers each of which owns 1200 MW and the peak demand for the system during the peak month is 10,000 MW. Under these circumstances all suppliers know that the aggregate available capacity requirement of say 11,500 MW (=1.15 x 10,000 MW) cannot be met without some of their capacity. As consequence in all of the Eastern US markets, very stringent market power mitigation measures have had to be put in place. Consequently, capacity prices typically

fluctuate from very close to zero to the regulatory price cap. It is difficult to see how these very volatile prices provide very useful signals about the need for new investment in generation capacity.

46. This market power problem leaves open the question of how to determine the value of the $/kW-year capacity payment. In most regions, the value of the capacity payment is based on the regulator's estimate of annual $/kW fixed cost of a peaking generation unit. This is backed by the logic that because of the offer cap on the short-term market and other market power mitigation mechanisms this peaking unit could only set a price slightly higher than its variable operating costs. Because this generation unit and all other generation units are missing the hours when the market price would rise above its variable operating costs because a price-responsive final demand would set the market price, the annual $/kW cost of the peaking unit is needed to compensate all generation units for the revenues they do not receive because of the offer cap and market power mitigation mechanisms. This logic for the value of $/kW-yr capacity payment explicitly assumes that the realtime demand for electricity is completely price inelastic and that suppliers are unable to exercise significant amounts of unilateral market power in the short-term market. Both of these assumptions are clearly false.

47. In addition, it is unclear why electricity is so fundamentally different from other products that it requires paying suppliers for their generation units to exist. Consumers want cars, not automobile assembly plants; point-to-point air travel, not airplanes; and a loaf of bread, not a bakery. In these markets producers do not receive capacity payments for owning the facilities needed to provide these products. All of these industries are also high fixed cost, relatively low marginal cost production processes, yet all of these firms earn their return on capital invested by selling the good that consumers want at a price above the variable cost of producing it. Cars, air travel, and bread are in many way essential commodities, yet capacity payments are not needed to ensure that there is sufficient productive capacity for these products to meet our needs.

48. Capacity payment mechanisms virtually guarantee that consumers will pay more for electricity their annual electricity consumption than they would in a world with active demand-side participation in the wholesale market. Recall that the capacity payment is made to either all generation units in the system or all generation units needed to meet the ISO's demand for capacity. On top of this, all suppliers receive a market-clearing price set by the highest generation offer needed to meet system demand. Thus to the extent that suppliers are able to exercise unilateral market power in the short-term market, they can raise energy prices significantly above the variable cost of the highest cost unit operating within the hour for all hours of the year.

49. For a number of reasons, a wholesale market with a capacity payment mechanism makes it more likely that suppliers will be able to exercise unilateral market power in the short-term wholesale market relative to a market with active demand-side participation and no capacity payment mechanism. Capacity payment mechanisms are typically accompanied by offer caps and market power mitigation mechanisms that significantly limit the incentive for final consumers to become active participants in the short-term wholesale market. For example, if the maximum wholesale price in an hour in $400/MWh because of an offer cap at this level, then a 1 KWh reduction in demand for a residential customer (a very large demand reduction) during an hour only saves the customer 40 cents, which seems unlikely to be sufficiently attractive to cause that consumer to reduce its demand. This lack of an active demand-side of the wholesale market impacts how generation unit owners offer their generation units into the wholesale market. Active participation by final demand substantially increases the competitiveness of the short-term wholesale market because all suppliers know that higher offer prices will result in less of their generation capacity being called upon to produce because the offers of final consumers to reduce their demand are accepted instead. Without an active demand-side of the wholesale market suppliers know that they can submit offers that are farther above their variable cost of supplying electricity and not have these offers rejected. Therefore, a market with a capacity payment mechanism charges consumers for the $/kW-year fixed cost of a peaker unit for their entire capacity needs and then give suppliers greater opportunities to exercise unilateral market power in the short-term market.

50. Another argument given for capacity payments is that they reduce the likelihood of long-term capacity inadequacy problems because of the promise of a capacity payment provides incentives for new generation units to enter the market. However, until very recently capacity payments in most markets around the world were only promised for at most a single year and only paid to existing generation units. Both these features substantially dulled the incentive for new generation units to enter the market, because the unit that entered often had no guarantee of receiving the capacity payment for one year and no guarantee that if it received it the first year it would continue to receive it. This has led the eastern US ISOs to focus on the development a long-term capacity product that is sold two to three years in advance of delivery to provide the lead time for new generation units to participate. As we discuss in Section 4, this solution unlikely to lead to a lower cost solution for consumers than the long-term contract adequacy approach described in that section.

4. Politically palatable real-time pricing

51. One benefit of the political and economic constraints associated with implementing an active demand-side in wholesale markets in the US is that there have been a number of experiments to determine the real-time price-responsiveness of retail electricity consumers. These experiments typically install hourly metering on a sample of customers and require a fraction of these customers to pay retail prices that vary with hourly system conditions and the remainder to pay according the standard retail price schedule.

52. These experiments have been run in a number of jurisdictions and found statistically significant evidence that retail customers are able to substantially alter their

consumption of electricity in response to hourly retail prices. Although these results are not surprising, the more surprising conclusion from this research is that how real-time wholesale price signals are provided to final consumers can impact the magnitude of the price response.

53. Regulators and many final consumers often argue that responding to hourly price signals would be too complex and time-consuming for most retail customers. Customers would have to continually monitor the price of electricity each hour of the day to determine whether it makes economic sense to alter their consumption. In addition, hourly real-time electricity prices can be extremely volatile and customers are likely to find it difficult to determine how long price spikes are likely to last and whether it is worth taking actions to reduce their consumption in response to a very high price during a single hour. For example, an electricity intensive industrial customer, may only be able to reduce their consumption significantly by shutting down an entire 8-hour production shift. This customer would need to have an intimate understanding of the time series behavior of real-time electricity prices to be able to make an informed decision about whether it makes economic sense to shut down production for an entire 8-hour shift in response to high prices in a single hour. Patrick and Wolak (1997) study responsiveness of large industrial and commercial customers in England and Wales to retail prices that pass-through half-hourly wholesale prices and find significant diversity in the magnitude and pattern of the demand responses with the day. All of these customers have extremely large monthly wholesale electricity bills, in the thousands of dollars, so they have a strong financial incentive to invest in the expertise needed to respond to half-hourly wholesale prices.

54. For smaller customers with the flexibility to reduce their consumption during a few hours of the day, it may not make sense to do so in response to every hour with a high real-time price. There are 744 hours (= 24 hour/day x 31 days) in a month. If there is a fixed-cost to taking action to reduce demand in an hour or number of hours of the day, the price increase with a single hour of the month must be very high to cause the customer to take action. Attaching some numbers to this calculation is helpful. Suppose for simplicity that the household's average hourly consumption is 2.5 KWh (roughly equal to the average hourly consumption of a California household) in all 744 hours of the month and the average wholesale price is $0.05/KWh. Suppose the household is able to reduce its consumption by 0.5 KWh by taking the actions that cost $5 in psychic or actual costs. The wholesale price in this hour would have to rise to at least $100/KWh = ($5 cost of taking action)/(0.5 KWh saving) for the customer to find the cost of taking action for that single hour to exceed the benefit. This $100/KWh price translates into $100,000/MWh, which is vastly in excess of the bid cap on any market currently operating in the world and is 100 times higher than the offer cap on the wholesale market in the eastern US.

This calculation illustrates a very important point about nature of hourly price spikes required for small electricity consumers with relatively small fixed-costs of taking actions to reduce their consumption in response to a price spike in a single hour of the day. If this fixed cost of taking action is $1, then the required wholesale price to take action falls to $5,000/MWh, which is still five times the offer cap on all of the eastern US markets.

55. Another factor that can reduce the magnitude of the wholesale price spike required to cause customers to take action is the duration of the price spike. In our simple example, the longer the customer expects the price spike to last the greater the likelihood the customer will take action, because once the customer pays the cost to act it can reduce its consumption by 0.5 KWh for as many hours as it would like. Therefore, at the $5 cost to take action, a two-hour price spike would only need the price to average $50,000/MWh or more.

56. This cost of taking action to reduce hourly electricity demand expressed as a fraction of the customer's typical monthly electricity bill is likely to be smaller the larger is the customer's monthly bill. For example, a large industrial user with an hourly consumption of 1 MWh is likely to have many ways to reduce its hourly consumption by 5 percent that cost significantly less than $50 to implement per event. This 0.05 MWh reduction in consumption only requires a wholesale price of $1000/MWh or more to yield energy savings sufficient to justify the $50 cost of taking actions to reduce demand within the hour. This logic implies that larger customers need a smaller price spike to find it profitable to take actions to reduce their hourly demand for electricity.

57. If there are offer caps and other market power mitigation mechanisms that limit the level of wholesale electricity prices, other mechanisms for passing through real-time price signals must be devised to reduce the cost of customers responding to real-time prices or increase the benefits they receive from responding. This logic has led to the design of critical peak pricing (CPP) programs that share the risk of responding to real-time prices between retailers and final consumers in order to both reduce the customer's cost of responding and the benefits it expects to receive from responding.

58. Under this sort of real-time pricing program customers pay according to a single fixed-price or an increasing block tariff during the month with a fixed price for each block of the household's monthly consumption. The retailer is then allowed to call a certain number of critical peak days within a given time interval. Typically, this is done the day before by a telephone or e-mail, but the program could be modified to notify the customer closer to the time of CPP event. During an agreedupon peak period of a CPP day the customer must pay a substantially higher price. For example, if the customer normally pays 8 cents/KWh for energy, during the peak period of a CPP day it would pay 35 cents/KWh. This mechanism does not require the final consumer to follow the hourly wholesale price or know anything about wholesale market conditions. The retailer declares CPP events on the days that it would like customers to reduce their consumption. Another benefit of the CPP program is that the peak period of the day during which a CPP customer pays the higher retail price is typically between four to six hours long. This implies a longer period over which a CPP customer has to accrue benefits by reducing its consumption.

59. If the retailer has enough customers on the CPP pricing program, then the structure of the program causes all CPP customers to focus their demand-reduction efforts during the same time period, which increases the likelihood that declaring a CPP event will result in lower wholesale prices during the CPP period because of the reduced system-wide demand for electricity. This further increases the benefits realized from implementing real-time pricing because it reduces the cost to the retailer of serving its remaining customers.

60. One variation on the standard critical peak-pricing program that is very popular with customers involves a rebate for consumption reductions relative to a reference level on critical peak days. Under this scheme the customer is paid a $/KWh rebate for every KWh of consumption less than some reference level during critical peak periods instead paying for all consumption at the higher price. For example, if a customer's peak period reference level is 8 KWh and the customer consumes 6 KWh, then it is paid the $/KWh rebate for 2 KWh. If the customer does not reduce its consumption below this reference level then it does not receive any rebate and its does not have to make any payment. Mathematically, the payment received by the customer during CPP days is p_{rebate}*max(0, q_{ref} - q_{actual}), where p_{rebate} is the $/KWh rebate, q_{ref} is the reference level for rebates, and q_{actual} is the customer's actual consumption during the peak period.

This mechanism implies greater risk for the retailer because it could pay out more in rebates than it saves in wholesale energy purchase costs. This real-time pricing program is more attractive to customers than the conventional CPP program because the customer cannot lose from participating the program. At worst, the customer does not receive any rebate payments.

61. Wolak (2006a) analyzes household-level price responsiveness under a CPP program with a rebate for the City of Anaheim in southern California. This program paid customers a $0.35/KWh rebate for reductions in consumption relative to their reference level during peak period of CPP days. During all other hours, the customer pays a price of 6.75 cents/KWh for monthly consumption less than 240 KWh and 11.02 cents/KWh for monthly consumption above 240 KWh. The peak period of the day for the purposes of the Anaheim CPP mechanism with a rebate is noon to 6 wpm. Wolak (2006a) found that during CPP days the mean difference in the difference in consumption between the CPP customers and the control group of customers is a reduction of approximately 13 percent. If this mean consumption reduction associated with a CPP event could be scaled to all residential consumers in California, approximately one-third of the consumption in California, this would imply slightly more than a 4 percent reduction in system demand as a result of a CPP event. Applying this to a peak demand in California of 50,000 MWh implies a 2,000 MWh reduction in demand, which means that California can avoid building and paying for almost 2,000 MW of new generation capacity as a result of this demand response capability.

62. The magnitude of the response to a critical peak day estimated in Wolak (2006a) is likely to underestimate the potential demand reduction possible, because of a number of new technologies to monitor and control electricity consumption automatically. There are a number of standards for allowing advanced meters to communicate with appliances throughout a geographic area using both wireless and wireline technologies. For example, a household could program a personal computer to alter electricity use based on wholesale prices or other signals provided by the retailer. The ZigBee Alliance (www.zigbee.org) is perhaps the most popular of these standards. It is a wireless network designed to monitor and control appliances and was organized as a nonprofit corporation in 2002. A number of companies are offering appliance control networks that are compliant with the ZigBee standard. Homeplug Powerline Alliance is powerline-based open standard for communications (www.homeplug.org) aimed at providing, among other services, monitoring and control of appliances. These technologies are likely to reduce overall electricity consumption as well as reduce the cost of responding to real-time price signals and the magnitude of the demand response.

63. The development of politically attractive real-time pricing plans and technologies that reduce the cost and increase the magnitude of demand response strongly argues in favor of introducing mechanisms that require final consumers to manage real-time price risk. The nontrivial cost of hourly meters and the technologies to reduce the cost of demand response favor a phased-in approach that focuses on customers realizing the greatest net benefits from these technologies and respects the political constraints facing regulators and policymakers in allowing active demand-side participation in wholesale electricity markets.

III. Managing demand-side economic and political constraints

64. This section proposes a retail market regulatory structure that addresses the economic and political constraints described in Section 2 with minimal harm to wholesale market efficiency and system reliability. This retail market structure emphasizes the necessity of hedging short-term wholesale price risk either through fixed-price forward contracts or active demand-side participation to ensure a reliable supply of electricity and the long-term financial viability of the industry. Another guiding principle is symmetric treatment of generation unit owners and final consumers in the sense that both sets of market participants face a default price that reflects all real-time price risk. Finally, this regulatory structure recognizes that hourly meters may not make economic sense for all retail customers at the present time, but these circumstances may change in the future as the price of electricity rises and the cost of hourly meters falls.

1. Hedging short-term wholesale price risk

65. There are two types of wholesale price risk that can harm electricity consumers. The first is prices persistently above competitive levels. This pattern of wholesale prices is typically the result of suppliers exercising unilateral market power in the short-term market by withholding output. The second is a short duration of very high prices usually accompanied by

stressed system conditions because of a generation unit or transmission line outage or an extreme unexpected weather event. Each form of wholesale price risk is best dealt with using a different set of actions by final consumers.

66. The risk of short-term prices persistently above competitive levels is best managed with fixed-price forward contracts between generation unit owners and retailers or large consumers able to purchase directly from the wholesale short-term market. As discussed in detail in Wolak (2000), fixed-price forward contract commitments by generation unit owners reduce their incentive to exercise unilateral market power in the short-term energy market because the supplier only earns the short-term price on any energy it sells in excess of its forward contract commitment.

67. To understand this logic, let pc equal the forward contract price at which the supplier agrees to sell energy to an electricity retailer and qc equal to the quantity of energy sold. This contract is negotiated in advance of the date that the generation unit owner will supply the energy, so that the value of pc and qc are pre-determined from the perspective of the supplier's behavior in a short-term wholesale market. The quantity of fixed-price forward contract obligations held by the supplier impact what price the firm finds profit-maximizing given its marginal cost of producing energy, the supply offers of its competitors, and the level of aggregate demand. Incorporating the payment stream a generation unit owner receives from its forward contract obligations, its variable profit function for a given hour of the day is:

$$\pi(p) = (pc - c)qc + (qs - qc)(ps - c), \quad (4.1)$$

68. where qs is the quantity of energy produced by the generation unit owner, ps is the price of energy sold in the short-term market and c is the supplier's marginal cost of producing electricity. The first term in (4.1) is the variable profit from the forward contract sales and the second term is the additional profit or loss from selling more or less energy in the short-term market than the supplier's forward contract quantity. Because the forward contract price and quantity are negotiated in advance of the delivery date, the first term is a fixed profit stream to the supplier from the perspective of its participation in the day-ahead market. The second term depends on the price in the short-term market, but in a way that can significantly limit the incentive for the supplier to raise prices in the short-term market.

69. For example, if the supplier is too aggressive in its attempts to raise prices by withholding output, it could end up selling less in the short-term market and than its forward contract quantity, and if the resulting market-clearing price is greater than the firm's marginal cost, c, the second term in the firm's variable profit function will be negative. Consequently, only in the case that the supplier is confident it will produce more than its forward contract quantity in the short-term market does it have an incentive to withhold output in order to raise short-term prices.

70. The quantity of forward contract obligations held by a firm's competitors also limits incentive of that supplier to exercise unilateral market power in the short-term market. If a supplier knows that all of its competitors have substantial fixed-price forward contract obligations, then this supplier knows these firms will be bidding very aggressively to sell their output in the short-term wholesale market. Therefore, attempts by this supplier to raise prices in the short-term market by withholding output are likely to be unsuccessful because of the aggressiveness of the offers into the short-term market by its competitors with substantial fixed-price forward contract obligations.

71. Short periods of extremely high prices are best managed through active demand-side participation in the wholesale market, because many of these price spikes are driven by unexpected events that occur too quickly for the supply side of the market to respond to. The outage of a large generation unit can often be managed by the generation units providing operating reserves increasing their output. However, the outages are sometimes severe enough that the only way to manage them is to reduce the demand electricity.

72. Although it is possible to manage the risk of the exercise of unilateral market power in the short-term market with demand response alone, this could impose significant hardship on consumers. For example, in a hydro-dominated system where water comes primarily in the form of winter snowpack, if the amount of water available to produce electricity is much less than normal, then the fossilfuel suppliers will have a greater opportunity to exercise unilateral market power until the following year. As discussed in Wolak (2003b), this describes the initial conditions in the western US immediately before the start of the summer of 2000. To limit the ability of suppliers to exercise unilateral market power under these system conditions, consumers would likely have to reduce their demands for long periods of time period until the next year's snowfall melted, which could impose significant hardship on electricity consumers. Consequently, a strategy that involves a lower downside to consumers would be to hedge their expected demand for electricity each period in fixedprice long-term contracts. That way if low hydro conditions arise the fossil fuel suppliers will have less of an incentive to exercise unilateral market power in the short-term wholesale market because of their substantial fixed-price forward contract obligations.

73. Fixed-price long-term contracts can be used to protect consumers against short-term price spikes, but this is likely to be more expensive for consumers than managing this risk with active demand-side participation in the wholesale market. To hedge against the risk of price spikes, consumers or their retailers would have to purchase fixed-price forward contract coverage for 100% of their demand requirements. Because the realized demand for electricity is unknown at the time a retailer signs the fixed-price forward contracts, it would have to purchase more forward contracts for more than 100% of its expected demand. This implies that during many hours, the retailer would be selling back energy purchased in the forward contract at a low spot market price because its actual demand is less than the amount it purchased in the forward contract. This further increases the effective price consumers pay for the electricity. A numerical example helps to illustrate this point. Suppose the distribution of the retailer's demand has a mean of100 MWh and a standard deviation of

20 MWh. For this reason, the customer purchases 130 MWh in a fixed price forward contract at a price of $50/MWh, to guard against the risk of paying very high spot prices if its demand is unexpectedly high. If a retailer's realized demand is 100 MWh and the real-time price is $20/MWh, then the retailer makes a loss of $900 by selling the 30 MWh it bought for $50/MWh at a price of $20/MWh. This implies an effective price for the 100 MWh consumed of $59/Mwh = ($50/MWh*100 MWh + $900)/100 MWh, almost a 20% price increase. A lower cost strategy for the retailer is simply to purchase the expected demand of 100 MWh in the forward market and manage the remaining short-term price risk by altering the demand of its customers in response to real-time prices.

2. Contract adequacy in wholesale electricity markets

74. Adequate fixed-price forward contracting by electricity retailers and large customers able to purchase from the short-term wholesale market is a necessary condition for both competitive shortterm market outcomes and adequate generation capacity to meet future demand. These fixed-price forward contracts must be negotiated far enough in advance of delivery for all possible sources of supply to compete. Signing a fixed-price forward contract a day, month, or even a year ahead of delivery can limit the number of suppliers and modes of supply that are able to provide this energy. For example, a contract negotiated one day in advance limits the sources of supply to existing generation unit owners able to produce energy the next day. Even a year in advance limits the sources that can compete to existing generation unit owners, because it takes longer than a year to site and build a substantial new generation unit. To obtain the most competitive prices, at a minimum, the vast majority of the fixed-price forward contracts should be negotiated far enough in advance of delivery to allow new entrants to compete with existing suppliers.

75. Regulators should focus on ensuring contract adequacy, not on generation adequacy. Specifically, retailers and large consumers should have adequate fixed-price forward contract coverage for their expected future demand signed far enough in advance of delivery to obtain the most competitive prices. By purchasing a hedge against the spot price risk at the locations in the network where the retailer or large consumer withdraws energy, the buyer can rely on the financial incentives that the seller faces to provide the contracted for energy at least cost.

76. A major mistake made by the California Department of Water Resources (CDWR) in negotiating the forward contracts signed by the State of California during the winter and spring of 2001 is that it focused on purchasing power plants instead of hedges against the spot price of energy at the locations where the three large electricity retailers withdrew energy from the transmission network. This procurement strategy created a number of market inefficiencies that significantly increased the cost of these forward contracts and prices in the wholesale market, because they often called for more expensive generation units to operate (than those required for a least-cost dispatch of California's generation resources) in order for the seller's contractual obligations to met.

77. By focusing on contract adequacy rather than building generation facilities, California would have had a portfolio of forward contracts that provided incentives for least cost production of electricity in the short and long term. Firms that sold these forward financial contracts would have strong incentives to ensure that the spot prices at the locations in the California ISO control area where these contracts clear are as low as possible. That is because as equation (4.1) demonstrates, once a supplier has signed a fixed-price a forward contract that clears against the spot price at a given location in the network, the supplier's revenue stream is fixed for this quantity of energy, so it has the strongest possible incentive to ensure that the cost of meeting this forward contract obligation in real-time is as low as possible. Most of the contracts signed by the State of California had durations of eight years and longer. If these contracts were hedges against short-term wholesale prices at locations where the major California retailers withdraw electricity, the sellers of these forward contracts would want to construct any new generation units needed to meet these obligations to limit the magnitude of transmission congestion the new generation units face.

78. An active forward market has other hedging instruments besides swap contracts where a supplier and a retailer agree to a fixed price at a location in the transmission network for a fixed quantity of energy. Cap contracts are also very effective instruments for guarding against price spikes in the short-term market and for funding the appropriate amount of peak generation capacity. For example, a supplier might sell a retailer a cap contract that says that if the price at a specific location exceeds the cap's exercise price the seller of the contract pays the buyer of the contract the difference between the spot price and the cap exercise price times the number of MWh of the cap contract sold. For example, suppose the cap exercise price is $300/MWh and market price is $400/MWh, then the payoff from the cap contract is $100/MWh = $400/MWh - $300/MWh times the number of MWh sold. If the spot price is less than $300/MWh, then the buyer of the cap contract does not receive a payment.

79. Because the seller of a cap contract is providing insurance against price spikes, it must make payments when the price exceeds the cap exercise price. This price spike insurance obligation implies that the buyer must make a fixed up-front payment to the seller in order for the seller to be willing to take on this obligation. This payment can then be used by the seller of the cap contract to fund a generation unit that provides a physical hedge against price spikes at this location, such as a peaking generation unit. The Australian electricity market has an active financial forward market where these types of cap contracts are traded. These contracts have been used to fund peaking generation capacity to provide the seller of the cap contract with a physical hedge against this insurance obligation.

80. One question often asked about the contract adequacy approach is whether sufficient generation resources will be built to meet demand if consumers only buy forward

financial hedges against spot price risks at their location in the network. In this regard it is important to bear in mind that the incentives faced by the seller of the forward financial contract once this contract has been sold. The supplier has an obligation to insure that the forward contract quantity of energy can be purchased at the agreed upon location in the spot market (or whatever market the forward contract clears against) at the agreed upon forward price or less. The seller bears all of the risk associated with higher spot prices at that location. In order to prudently hedge this risk, the seller has a very strong incentive to construct sufficient generation capacity to ensure that the risk associated with guaranteeing the price in the short-term market at that location in the network is minimal for the quantity of energy sold in the fixed-price forward contract.

81. This logic implies that if a supplier signs a forward contract guaranteeing the price for 500 MWh of energy for 24 hours a day and 7 days per week at a specific location in the network, it will construct or contract for more than 500 MWh of generation capacity to hedge this spot price risk. Building only a 500 MW facility to hedge this risk would be extremely imprudent and expose the supplier to significant risk, because if this 500 MW facility is unavailable to provide electricity, the supplier must effectively purchase the energy from the spot market at the price that prevails at the time. If this generation unit is unavailable, it is very likely that the spot price will be extremely high.

82. Different from the case of a capacity market, the contract adequacy approach does not require the regulator to specify the amount of total generation capacity needed to meet demand. Instead the regulator ensures that retailers and large customers have adequate fixed-price forward contract coverage of final demand and then relies on the incentives that the suppliers of these contracts face to provide sufficient generation capacity to meet these forward contract obligations.

83. Implementing the contract adequacy approach in a world with offer caps and market power mitigation mechanisms is complicated by the fact that retailers and large consumers have an incentive to rely on the short-term market as discussed in Section 3. To address the incentives caused by these regulatory distortions, the regulator must mandate certain levels of fixed price forward contract coverage at various horizons to delivery.

84. For example, the regulator could require that a large fraction of the retailer's year ahead and two-year ahead demand forecasts be covered by fixed-price forward contract obligations. How large this fraction needs to be depends on a number of factors. First, the larger the fraction of final demand paying a retail price that passes through the hourly wholesale price, the smaller this fraction needs to be. Second, the greater the share of electricity coming from hydroelectric sources, the greater this fraction needs to be because hydroelectric energy has an additional supply shortfall risk not relevant for fossil fuel-based sources: insufficient water behind the turbine to meet the unit's forward contract obligations. Higher electricity prices will not cause more water to show up behind the turbine, but it is very likely to increase the amount of fuel that can be profitably sold to a fossil fuel-fired generation unit owner. As Wolak (2003a) emphasizes, the vast majority of Latin American markets mandate minimum fractions of fixed-price forward contract coverage of the retailer's or large consumer's demand at various horizons to delivery as way to deal with the incentive of retailers to rely on the short-term wholesale market.

85. It is important to emphasize that mandating these contracting levels should not impose a financial hardship on retailers that lose customers in a competitive wholesale market regime. If a retailer purchased more fixed-price forward contract coverage than it ultimately needs because it lost customers to a competitor, it can trade this obligation in the secondary market. Unless the market demand in the future is unexpectedly low, this retailer is just as likely to make a profit on this sale as it is to make a loss, because one of the retailers that gained customers is going to need a forward contract to meet its regulatory requirements for coverage of its final demand. Only in the very unlikely case that the aggregate amount of forward contracts purchased is greater than the realized demand, will there be a potential for stranded forward contracts held by retailers than lose load.

3. Symmetric treatment of load and generation

86. As noted in Section 2, the economic and political constraints on demand-side participation in wholesale electricity markets in the US have led state PUCs to set fixed default retail prices that have a significant risk of failing to cover the retailer's wholesale energy purchase costs. In addition, many states allow customers taking service from a competing retailer to switch back to the regulated retail price whenever they would like. This further increases the regulated supplier's wholesale energy price risk, because customers are most likely to switch back to the regulated retail price when it benefits them to do so and these benefits are greatest when the wholesale price of electricity is extremely high. This ability to switch back at will leaves the regulated retailer with an enormous unhedged risk against movements in the short-term price of wholesale electricity.

87. The best way to solve this problem is to make the default retail price pass-though the hourly real-time price of electricity. Any attempt to set a fixed retail price that consumers can switch to at their own discretion is an invitation to create a "California Problem," in the sense that there is a risk that the implicit fixed wholesale price in the regulated retail price is less that the wholesale price of electricity. Treating all final consumers like generation unit owners in the sense that their default price is equal to the hourly real-time price of electricity solves this problem. This is the same default rate faced by all electricity generation unit owners. Unless owners of generation units enter into forward market agreements, they will receive the hourly spot price for all electricity they deliver in real-time. Similarly, all final customers, including residential and small business customers should have to purchase all of their consumption a retail price that reflects the hourly real-time wholesale price plus the relevant transmission and distribution

charges. However, all customers should also be able to enter into forward contracts and other forward market hedging agreements with competitive retailers, if they desire, just as generators are permitted to do. No final consumer must purchase any of its energy at the real-time price if it is willing to pay for spot price risk management services.

88. It is important to emphasize that this mechanism would not require any customer to purchase even a fraction of their consumption at the hourly real-time price, only that this is the default price that the customer pays for wholesale electricity if he does not enter into a hedging arrangement. This requirement is no different from what occurs in other markets, such as air travel where the customer always has the option to purchase the ticket at the airport at the time they would like to fly. Customers rarely do this because of a desire to hedge the spot price associated with this real-time purchasing strategy.

89. An important necessary condition for providing valid economic signals for customers to manage real-time price risk is to set a default rate that requires customers to manage this risk and sets the price of insurance against short-term wholesale price volatility appropriately. Figure 2 assumes that final customers have a expected utility functions, $U(E(P_r), (P_r))$, that are decreasing functions of the expected hourly retail price, $E(P_r)$, and standard deviation of the hourly retail price, (P_r) for the retail pricing plans offered. Indifference curves for consumer 0 and consumer 1 are plotted in the figure. Consumer 0 is less risk-averse than consumer 1. This figure also plots the set of feasible pairs $(E(P_r), (P_r))$ that the retailer can offer in their retail pricing plans without facing a significant risk of going bankrupt. The "Feasible Expected Price and Price Risk Frontier" implies that the retailer must increase the value (P_r) in order to offer a pricing plan with a lower value of $E(P_r)$. Finding the point of tangency between each customer's indifference curve and this frontier yields that customer's optimal pricing plan choice. For customer 0 this process yields the point $((E(P_r)_0, (P_r)_0)$ and for customer 1 the point $((E(P_r)_1, (P_r)_1)$. It is important to emphasize that the reason each customer chose a plan that required it to take on some hourly price risk is because it faces the default retail rate that is a pass through of the hourly wholesale price.

90. Figure 3 illustrates the choices of consumer 0 and 1 if a low regulated retail price is set that completely eliminates all retail price risk, as is currently the case in all US wholesale markets. The original indifference curve for consumers 0 and consumer 1 are drawn as U_{01} and U_{11}. Two indifference curves with a higher level of utility for each consumer are drawn as U_{02} and U_{12}. These represent the utility levels that consumers 0 and 1 would achieve if a default fixed retail price, $E(P_r)_d$, was set that eliminated all price risk faced by these two consumers. Because $U_{01} < U_{02}$ and $U_{11} < U_{12}$, both consumers would achieve a higher level of expected utility by choosing $E(P_r)_d$ instead of any point along the Expected Price and Price Risk Frontier. This diagram illustrates the necessity of setting a default retail price that is a pass through of the hourly wholesale price or setting a fixed default price that contains a substantial risk premium so that it does not interfere with the choices the customers make along the Expected Price and Price Risk Frontier. This suggested fixed default price is given by the vertical line on the far right of the graph.

91. It is important to emphasize that requiring the default retail price to at least pass through the hourly real-time wholesale price is only making explicit something that must be true on a long-term basis: All wholesale electricity costs paid by the retailer must be recovered from retail rates. If this is not the case, then the retailer cannot remain in business over the long-term because it will be charging a price that is less than the amount it pays for wholesale electricity.

92. Therefore, a prohibition on hourly meters and real-time pricing in the name of protecting consumers from real-time wholesale price volatility does not mean that consumers do not have to pay these volatile wholesale prices. On an annual basis they must or the retailer supplying them will go bankrupt. The regulatory prohibition on hourly meters and a default retail price that passes through the real-time wholesale price only prevents consumers from obtaining a lower annual electricity bill by altering their consumption in response to hourly wholesale prices. A default fixed retail price requires the consumers to pay the same wholesale price for electricity every hour of the year regardless of the wholesale price.

93. A final point to emphasize with respect to the question of symmetric treatment of load and generation is that all retail customers must face the real-time hourly price as their default price unless they find an entity willing to provide a hedge against this risk. The same logic applies to electricity generation unit owners. Unless they are able to find an entity willing to provide a hedge against short-term wholesale price risk, they will sell all output they produce at the hourly real-time price.

94. Symmetric treatment of load and generation creates the following sequence of market efficiency-enhancing incentives. First, final consumers must sign long-term contracts to obtain a fixed-price hedge against their wholesale market spot price risk. Retailers then would attempt to hedge their wholesale market risk associated with selling this fixed-price retail contract to the final consumer. This creates a demand for fixed-price forward contracts sold by generation unit owners. Therefore, by requiring both generation unit owners to receive and final consumer to pay the hourly real-time price by default, each has a strong incentive to do their part to manage this real-time price risk.

4. A core/non-core approach to retail market operation

95. This section proposes a core/non-core customer approach to organizing the retail segment of the industry that recognizes the economic and political constraints on active demand-side participation in wholesale electricity markets described in Section 2. This approach recognizes the need for adequate fixed-price forward contracting by electricity retailers and large customers and the fact that with offer caps and market power mitigation mechanisms there is less of an incentive for these agents to sign the necessary quantity of

fixed-price forward contracts. It also recognizes that there are very few regions with hourly meters in place at the start of restructuring so it is necessary to determine which customers will receive these meters and what prices these customers will face once they have hourly meters.

96. The core/non-core distinction refers to the fact that core customers remain with the regulated retailer and are not required to have hourly meters and the non-core customers are required to have hourly meters and purchase directly from the wholesale market or from a competitive retailer. All non-core customers face a default retail price that passes though the hourly wholesale price. The regulated retailer is required to take a non-core customer back at this retail price if the competitive retailer serving that customer goes bankrupt or terminates service with that customer. The regulated retailer does not have an obligation to offer this customer any other retail price that provides some short-term risk management services.

97. In order to switch from the core segment to the non-core segment, a customer must have an hourly meter installed on their premises. As discussed in Section 2, it seems likely that hourly metering will soon replace conventional meters for most jurisdictions in the industrialized world and that metering services will be provided as a regulated distribution service. However, the process of installing these meters will take time so it is important to emphasize that a customer cannot switch to the non-core segment without an hourly meter. This is necessary because of the requirement that the default retail rate for all non-core customers is a pass through of the hourly wholesale price and without an hourly meter it is impossible to measure the customer's consumption during each hour of the day.

98. Customers in the core segment would not be required to have hourly meters, but those with hourly meters could remain in the core segment. A major challenge faced by the regulatory process is to set tariffs that define the Feasible Expected Price and Price Risk Frontier presented in Figures 2 and 3. The regulator must guard against setting a fixed retail price at an unrealistically low level to drive out any incentive by core customers to manage wholesale price risk as described in the previous section. This is the most important factor to consider in setting the default price for core customers, because if this price is set too low, the sequence of events outlined in Figure 3 will occur and the risk of bankruptcy for the regulated retailer will be significantly higher. The regulator must set a fixed retail price for a year that guarantees that the retailer will have sufficient revenue to meet its core customer wholesale energy costs for the following year.

99. The regulator must be confident that even if it is fixed for a year, this retail price will provide the retailer with sufficient revenue to cover its wholesale energy costs. The expectation is that this retail price will be adjusted only once a year. The regulator should also mandate 100% forward contract coverage of the expected hourly demand of its core customers signed one year in advance of delivery. Following the process of validating adequate forward contract coverage, the regulator can set the fixed retail price for the year taking the total forward contracting costs divided by the retailers annual load forecast as the average wholesale price in the retail rate.

100. Under this scheme, the regulated retailer then faces only the quantity risk associated with serving an uncertain retail load. It is free to manage the remaining revenue risk through real-time pricing programs offered to its customers. For example, the retailer can offer its core customers a CPP rate or CCP rate with a rebate to ensure that its total demand during certain hours of the year is consistent with its forward contracting purchases made one year in advance.

101. This core customer retail pricing scheme encourages active demand-side participation in the wholesale market because it sets the fixed retail price sufficiently high to leave room for customers to choose expected price and standard deviation of price combinations that provide higher levels of expected utility for final consumers either from the regulated retailer or its competitors. Consistent with the economic and political constraints on active demand side participation in the wholesale market, all market participants will take on this wholesale price risk voluntarily. The retailer serving core customers must offer programs that core customers find beneficial relative to the fixed-price retail rate to manage the quantity risk associated with this fixed wholesale price.

102. As discussed in Section 3, offer caps and market power mitigation mechanisms create the possibility that the wholesale market price cannot rise to a level where amount supplied at this price equals the amount demanded. For this reason, it is important to specify what will happen when there are supply shortfalls in the short-term market. As noted earlier, the usual approach to solving this problem involves random curtailment. This outcome is unavoidable because the technology to switch off certain customers is not universally available. However, to limit the risk of this outcome, all customers are required to pay a penalty rate for their consumption during hours of system emergency. This penalty rate is designed to provide both core and non-core customers with the strongest possible incentive to reduce their demand during these periods and to take preventive actions to ensure that supply shortfalls do not occur. For example, if the offer cap on the ISO's realtime market is $1,000/MWh, the penalty rate for consumption during these periods should be sufficient to ensure that non-core consumers will make the greatest possible efforts to reduce their consumption. For example, a penalty price of $5,000/MWh would provide strong incentives for noncore customers to reduce their demand during system emergency periods so that random curtailment of load is not necessary to manage a temporary supply shortfall.

103. It is important to emphasize that this penalty rate need never be active. It is only imposed to ensure the credibility of the offer cap in the wholesale market. Specifically, in order to avoid paying the penalty rate, both non-core customers and retailers serving core customers could be expected to bid demand response into the ISO's real-time market at or below the offer cap to ensure that economic curtailment (less demand clears the day-ahead and real-time market) takes places before it is necessary to invoke random curtailment. If insufficient demand is offered into the day-ahead and

real-time markets at or below the offer cap to prevent system emergencies, this should be taken as strong evidence that the offer cap is set to low or the penalty price is too low.

104. Large retailers can even use their customers with hourly meters to reduce the wholesale prices they pay to serve all of their customers. This requires that retailers charge real-time pricing customers a different wholesale price in a given hour than the retailer is actually paying for power in that hour. Both the CPP and CPP with a rebate pricing mechanisms are simple examples of this sort of program. Because all real-time pricing programs offered in this core/non-core scheme are voluntary, the regulator does not need to set these real-time pricing rates. For core customers, the retailer must offer the fixed retail rate set by the regulator and the retailer is free to offer any other retail pricing contracts it would like. For the non-core segment, the retailers are free to offer whatever plan customers would like, the only requirement is that the core customer's default rate on return to the regulated retailer is an hourly pass through of the wholesale price.

105. Retailers can reduce their total wholesale purchase costs for a given number of total MWh by reducing their total demand during hours when the aggregate bid supply curve is very steep and increases its demand in hours when the aggregate bid supply curve is flat. Consider the following two-period example of a single retailer exercising its unilateral market power as a buyer. Suppose this is a core retailer is serving customers on a fixed price retail rate and paying a real-time pricing rate.

106. Let PW_i equal the wholesale price in period I and PR_i the price charged to retail customers on the real-time pricing program in period 1. Let $D_i(p)$ equal the demand of real-time pricing customers at price p in period 1. Suppose that the retailer commits to guaranteeing that demand served on the real-time pricing contract will provide no marginal contribution to retailer's profits. This imposes the following constraints on the expected profit-maximizing values of PR_i for i=1,2:

$$PR_1 (D_1(PR_1) + PR_2 (D_2(PR_2) = PW_1(D_1(PR_1) + PW_2 (D_2(PR_2), \quad (4.2)$$

107. The total payments by customers facing real-time prices, PR, equals the total payments the retailer makes to the wholesale market to purchase this energy, because PW is wholesale price in that hour that the retailer pays for all its wholesale market purchases.

108. Suppose the retailer maximizes the profits associated with serving customers on fixed retail rates. Let PF equal the fixed retail rate and QF_i the demand for customers facing price the PF in period i. Let $S_i(p)$ equal the aggregative bid supply curve in period i. The profit function for the firm assuming the constraint (4.2) is:

$$\Pi(PR_1,PR_2) = PF(QF_1 + QF_2) - PW_1 QF_1 - PW_2 QF_2$$

The wholesale price for each period, PW is the solution to $S_i(PW_i) = D_i(PR_i) + QF_i$. This equation implies that PW_i can be expressed as:

$$PW_i = S_{i\,-1}(D_i(PR_i) + QF_i),$$

which implies that PW_i is a function of PR_i.

109. The simple two-period model of choosing PR_i to maximize the retailers expected profits can be illustrated graphically. Figure 4 makes the simplifying assumption that D(p) and S(p) are the same for periods 1 and 2. The only difference is the amount of fixed-price load the retailer must serve in each period. I assume that $Q_1 < Q_2$. I define P_i as the value of the wholesale price in period i if the retailer passively bids in the real-time demand function D(p) in each period. In this figure, PW_i is the wholesale price in period i assuming that the retailer chooses PR_i, the price charged to real-time pricing customers, to maximize daily profits. The large difference in PR_2 and PW_2 shows the tremendous benefit in high demand periods from the retailer exercising its market power. In order to satisfy the constraint that the retailer makes less than or equal to a zero profit from serving realtime pricing load, the retailer must set PR_1 below PW_1. The two lighter shaded areas in the Period 1 and 2 diagrams are equal, illustrating that the constraint (4.2) given above is satisfied. The large difference between P_2 and PW_2 versus the relatively small difference PW_1 and P_1 illustrates the large reduction in daily average wholesale prices from the retailer using its real-time pricing customers to exercise market power versus simply using their demand curves non-strategically. The darker shaded rectangles in the Period 1 and Period 2 figures, shows the profit increase achieved by the retailer as a result of exercising its buying power. Some of the difference between the large dark rectangle in Period 2 and the small dark rectangle in period 1 can be given to the real-time consumers as payment for their price response efforts.

This strategy for retailers to exercise market power extends in a straightforward manner to multiple time periods within the day, week or month. It represents a major source of potential benefits from a price responsive final demand in the retail segment.

110. A final aspect of this core/non-core model for electricity retailing is a change in the mission of the industry regulator. Although the regulator's primary role in the former vertically integrated regime was setting retail prices, there is less need for this role in the core/non-core model, particularly if there is universal interval metering. In fact, if the regulator sets the fixed-retail price too high this will only encourage more customers to manage real-time wholesale price risk with a competing retailer or the core retailer. For this reason, the regulator should focus its attention on providing information to retail customer to help them better manage their real-time price risk. For example, the regulator might manage a web-site that has all of the plans offered and illustrates the price risk and standard deviation of price tradeoff inherent in each plan.

111. If there are a significant number of core customers without hourly meters the regulator's job becomes more difficult because a moral hazard problem in electricity

retailing arises that is similar to the one that exists in retail banking. The fear in retail banking is that the bank will take customer deposits and invest them in extremely risky assets in an effort to deliver a very favorable return to the investor and the bank's shareholders. However, engaging in this risk-taking behavior may lead to outcomes that render the bank unable to meet certain future obligations to its depositors. An analogous chain of events can happen in the electricity retailing industry. The retailer has a strong incentive to under-invest in forward contracts to cover their future load obligations when it sells a fixed-price commitment to a customer for one or two-year period. It may be able to earn a higher expected return by taking risks that increase the probability of bankruptcy but also have the prospect of very high positive profit levels due to low wholesale prices.

112. Consequently, similar to the retail banking sector regulation, state PUCs must change its focus from retail rate setting to monitoring the forward contract procurement process and ensuring forward contract coverage requirements of all retailers relative to their forecasted retail market commitments. Clearly, if firms are always required to hold the 100% of their forecast demand in fixed-price forward contracts one year in advance, then these firms will find it profit-maximizing to honor their retail market commitments.

This market monitoring process should require all retailers to submit to their state PUC on a monthly basis a list their retail market commitments by duration and price and their wholesale market coverage by quantity and price. The role of the PUC would be to verify that the retailer met these risk management prudency standards and assess penalties and sanctions for violations.

113. Consider the following example of how this might work. The second and third column of Table 1 contains a list of the quantity-weighted average wholesale price implicit in the fixed retail price retail and quantity obligations that the retailer has agreed to supply for various delivery months in the future. The fourth and fifth columns of Table 1 contain the quantity-weighted average fixed wholesale price and quantity commitments the retailer has signed with wholesale energy suppliers. The sixth columns contain the desired percentage of the total monthly quantity of fixed-price wholesale quantity commitments that the state PUC deems that it is prudent for the retailer to hold as a hedge against its fixed price retail commitments for each future delivery date. The last column contains the product of the percentage in the sixth column and the fixed price retail obligation quantity given in the second column.

In this example there are several delivery horizons where the desired hedge quantity is greater than the amount given in the fourth column. In these instances there are several actions that the state could take. First, it could assess a substantial penalty per MWh on the positive part of difference between desired quantity in the seventh column and the actual quantity in the fourth column. The PUC could also prohibit this retailer from selling more fixed-price retail obligations at this time horizon or shorter until the retailer submits a monthly report that is not out of violation for all months longer than this delivery horizon.

114. For the case given in Table 1, the first month the retailer is out of compliance is month 4. This means that retailer is prohibited from signing fixed price commitments for deliveries longer than 3 months in the future during the next month unless it submits proof of compliance in the next month for all delivery horizons up to 3 months. There are other prudency standards that state PUCs could impose on hedging behavior of retailers that uses risk measures based on the prices of retail obligation versus the price of wholesale commitments that cover them. Fortunately, these hedging standards do not need to be set using very sophisticated methods in order provide a reasonable level of assurance that all retailers will be able to meet their fixed price retail obligations with a high degree of certainty.

115. The other role of the state PUC in a competitive retail market is to ensure that all retailers have equal access to the billing and metering services provided by the regulated monopoly local distribution company. The PUC must establish rules that prevent the local distribution company from favoring its competitive retailing affiliate.

5. Developing country issues

116. Developing countries complicate several features of this core/non-core model. First, in many developing countries a significant fraction of customers lack of any sort of meter on their premises. Second, substantial fractions of customers in a number of countries do not pay their bills. Third, a significant fraction of the population do not have access to electricity. Although crafting a satisfactory solution to all of these problems is beyond the scope of this paper, a few promising directions to consider are suggested.

117. Electricity networks are well-suited to implementing group payment programs for electricity bills because all customers in a given geographic area typically take their energy from the same location in the high voltage transmission network. The lower voltage distribution network that serves a given geographic area typically interconnect at this location and the system operator is able to meter total withdrawals from these locations in real-time. This fact suggests allocating the liability for the cost of all wholesale energy withdrawn at the lowest voltage location in the transmission network that the system operator is able to meter withdrawals to all customers taking service from this location in the transmission network.

118. The wholesale market operator could be made responsible for terminating service for all customers at this location after a certain period of nonpayment. Because it is impossible to determine how much electricity was consumed by each customer in a given time period because of the lack of meters or the lack of hourly meters, assigning payment liability to each customer in the geographic region and collecting payment from them is an extremely complex task. This problem should be easier to solve by asking other customers in the same area to ensure that all other customers in the area pay their bills and do not steal electricity. Allowing the wholesale market operator to curtail power at lowest level in the network at which it has this capability

provides credibility to threat that nonpayment will result in curtailment. Credible demonstration of this threat by the system operator will make it easier for electricity retailers to address the problem of nonpayment, because a substantial fraction of nonpayments in many countries is due to theft.

119. The use of social pressure to ensure prompt payment has been successfully used most notably in the area of providing microfinance. Johnson and Rogaly (1997) describe the successful use of group liability in the provision of microfinance. Borrowers are formed into groups by the microfinance banks and these groups assume joint liability for repayment of each member's loan. By the same logic, the set of electricity consumers connected to the transmission network at a given location must assume joint liability for payment for the total amount of electricity withdrawn at that location, or jointly face the risk of no electricity for all consumers in the geographic area until the liability is paid.

120. Although this may seem like a drastic measure to ensure payment, as Wolak (2006b) emphasizes for the case of India, without a change in the attitude of consumers toward paying for electricity, it is unlikely that India will ever be able to attract private investment in the electricity sector. Even spending government money on this sector seems misguided if final consumers do not pay for the electricity that is produced. Determining the magnitude of the total amount of KWh consumed and assigning it to all customers in that geographic area and alerting these customers to the joint liability nature of their electricity supply costs should help to improve payment rates.

121. The second issue concerns the need to built out the transmission and distribution network to serve more customers in many developing countries. These customers should be treated as core customers and their retail prices determined as describe above for non-core customers. As these regions grow, it may make sense to install hourly meters and convert some of these customers to the non-core segment.

IV. Concluding comments

122. All existing electricity markets in the US and virtually all markets that exist in other jurisdictions have failed to introduce the necessary demand-side incentives for setting the lowest possible prices for wholesale electricity consistent with the long-term financial viability of the industry. In the name of protecting financial consumers, state PUCs in the US have denied consumers the ability to benefit from being active participants in the spot market. We have argued in this paper, by handicapping the demand side of the market the PUCs are only increasing the likelihood that wholesale suppliers will be able to raise prices through their own unilateral bidding and scheduling behavior.

123. Final consumers must bear the full cost of high wholesale prices and have the ability to realize the full benefits from taking actions in the forward and spot markets to respond to these high prices. Investments in hedging instruments and demand-responsiveness technology will then lead to a more competitive wholesale market that will, in turn, lead to lower average prices than the former vertically integrated monopoly regime when final demand was a passive participant in the wholesale market.

124. The well-known dictum of "there's no such thing as a free lunch" applies to the case of introducing competition into a formerly regulated industry. Unless competition changes the behavior of some market participants, it cannot benefit consumers relative to the former monopoly regime. For example, if generation unit owners continue to produce the same amount of electricity in the same manner as they did under the former monopoly regime and all input costs for all companies remain the same, then total production costs will not change. Similarly if consumers continue to demand the same amount of electricity in each hour of the year their annual electricity bills cannot decrease.

125. Only by providing incentives for more efficient operation of generating facilities and more efficient consumption signals can a market result in lower annual average prices than under the former monopoly regime. The retail market infrastructure presented in this paper provides the strongest possible incentives for consumers to alter their behavior to reduce the cost of producing wholesale electricity and making most efficient use of the generating capacity that currently exists. ■

References

Johnson, Susan, and Rogaly, Ben (1997) Microfinance and Poverty Reduction, Oxfam.

Mukherjee, Andy (2007) "Why you need to run a power plant at home," Livemint, May 25, 2007 at http://www.livemint.com/2007/05/25001738/Why-you-need-to-run-a-power-pl.html.

Patrick, Robert H. and Wolak, Frank A. (1997) "Estimating the Customer-Level Demand for Electricity Under Real-Time Market Prices," August 1997, (available at http://www.stanford.edu/~wolak/).

Wolak, Frank A., Nordhaus, Robert, and Shapiro, Carl, "Preliminary Report on the Operation of the Ancillary Services Markets of the California Independent System Operator (ISO)," August 1998 (available at http://www.caiso.com/docs/2000/09/14/200009141610025714.html).

Wolak, Frank A., Nordhaus, Robert, and Shapiro, Carl, "Report on the Redesign of the Markets for Ancillary Services and Real-Time Energy," March 1999 (available at http://www.caiso.com/docs/2000/09/14/200009141610025714.html).

Wolak, Frank A. (1999) "Market Design and Price Behavior in Restructured Electricity Markets: An International Comparison," in Competition Policy in the Asia Pacific Region, EASE Volume 8, Takatoshi Ito and Anne Krueger (editors) University of Chicago Press, 79-134.

Wolak, Frank A. (2000) "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market," International Economic Journal, Summer 2000, 1-40.

Wolak, Frank A. (2003a) "Designing Competitive Wholesale Markets for Latin American Countries," available at http://www.stanford.edu/~wolak.

Wolak, Frank A. (2003b) "Diagnosing the California Electricity Crisis," The Electricity Journal, August/September, 11-37.

Wolak, Frank A. (2003c) "Regulating Wholesale Electricity Markets in the Aftermath of the California Crisis," The Electricity Journal, August/September, 50-55.

Wolak, Frank A., (2004) "Managing Unilateral Market Power in Wholesale Electricity," in The Pros and Cons of Antitrust in Deregulated Markets, edited by Mats Bergman, Swedish Competition Authority.

Wolak, Frank A. (2006a) "Residential Customer Response to Real-Time Pricing: The Anaheim Critical-Peak Pricing Experiment, available at http://www.stanford.edu/~wolak.

Wolak, Frank A. (2006b) "Reforming the Indian Electricity Supply Industry," available at http:/www.stanford.edu/~wolak.

Table 1: Sample Monthly Forward Contract Filing						
	Retail Obligations		Forecast Wholesale Purchases		Compliance Levels	
Future Delivery Date for Energy (months)	Total Quantity (MWH)	Average Implicit Wholesale Price ($/MWH)	Total Quantity (MWH)	Average Purchase Price ($/MWH)	Hedge Factor (%)	Desired Hedge Quantity (MWH)
1	10000	44.56	10000	40.12	100	10000
2	10000	45.60	10000	45.00	100	10000
3	10000	42.00	11000	40.21	100	10000
4	12000	50.00	11000	49.00	100	12000
5	13000	54.00	12000	52.00	100	13000
6	11000	51.00	9000	50.12	100	11000
12	10000	48.00	10000	45.29	100	10000
18	10000	44.23	9000	39.56	85	8500
24	12000	44.00	10000	42.03	80	9600

Figure 1

Advanced Metering Communication Networks

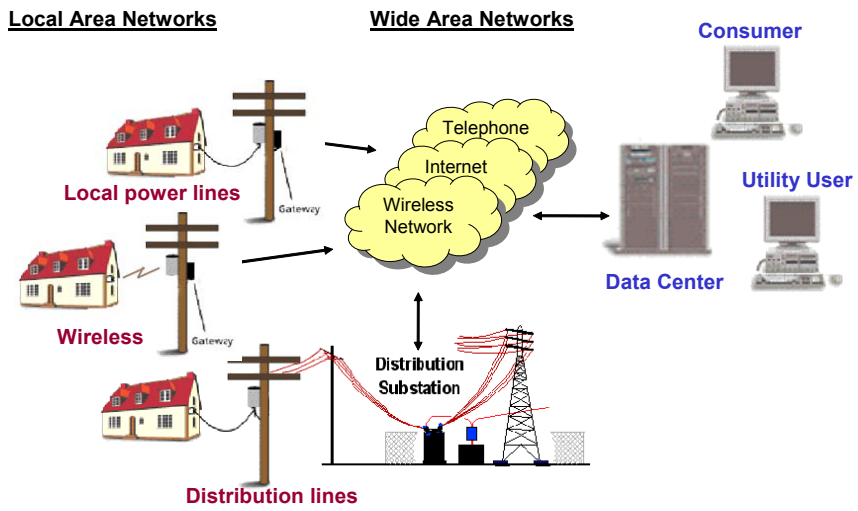


Figure 2

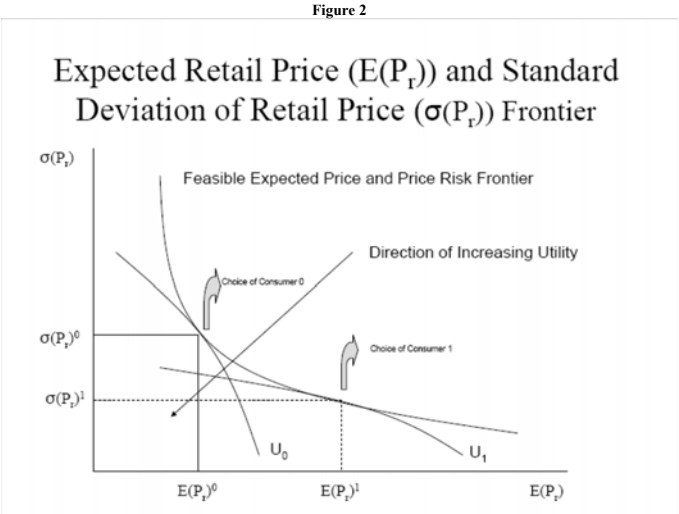
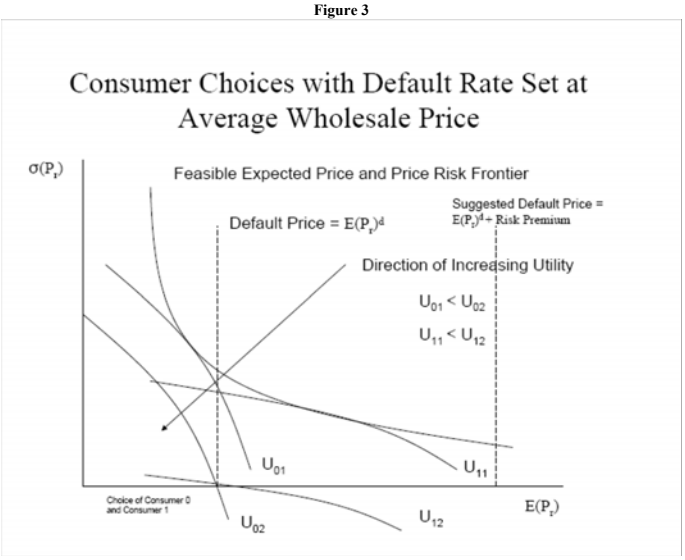
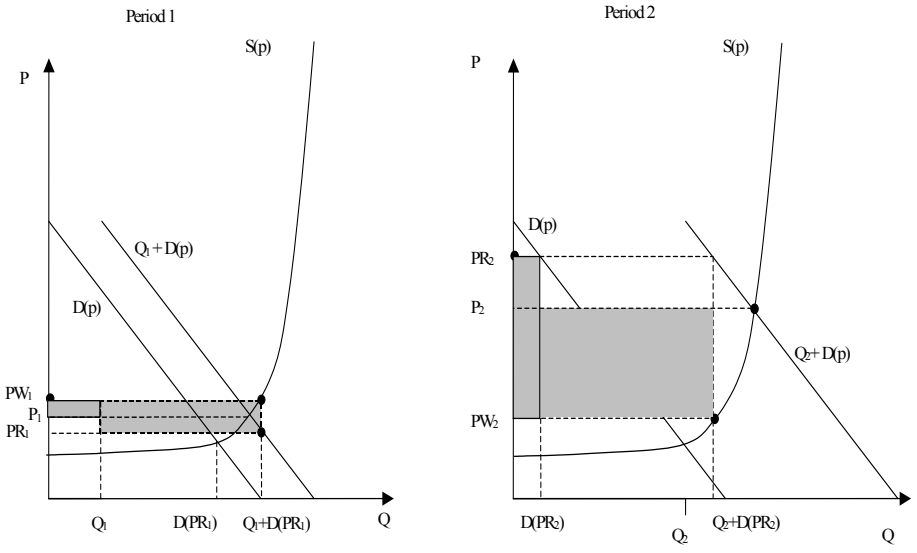


Figure 3



57

Figure 4



58

III Lisbon conference on competition law and economics
PANNEL II
14-15 January 2010

Ricardo Cardoso de Andrade
ricardo.cardoso@ec.europa.eu
Case Handler
DG Competition, European Commission

Energy markets: To what extent can competition, security of supply and environmental protection be reconciled?

Competition, security of supply and environment: Enemies or allies?

Abstract

This presentation argued that the European Commission's objectives in the energy sector of competitiveness, security of supply and environmental goals are mutually enhancing and not in tension with each other. In particular, the presentation lists the positive effects competition cases run by the European Commission have had in improving the functioning of the European energy markets.

Cette présentation soutient que les objectifs de la Commission européenne dans le domaine de l'énergie en termes de compétitivité, sécurité d'approvisionnement et d'environnement, non seulement ne sont pas contradictoires mais se renforcent mutuellement. En particulier, la présentation énumère les effets positifs dans l'amélioration du fonctionnement des marchés européens de l'énergie induits par les affaires de concurrence menées par la Commission européenne.

1. I would like to start by thanking the organizers of the conference for inviting me and giving me the opportunity to speak to you about a topic which is as relevant and important as the issue of the relationship between competition, security of supply and the environment. Now, I'm at some advantage to my co-panellists because I'm playing at home here in Lisbon. But not only that; we are all here in the Gulbenkian Foundation and for those who may not be familiar with him, Calouste Gulbenkian made all of his fortune in the energy sector in the first half of the century before retiring here to Portugal and funding this wonderful museum.

2. Coming back to the theme of this panel, there's no doubt in my mind that the European Union remains committed to creating competitive energy markets. They are of fundamental importance to the competitiveness of the European economy and I think this is particularly relevant in times of crisis. We want our future European energy market to be an open and welcoming market that presents enormous opportunities to players on equal terms. The Commission's actions via the third liberalisation package and a strict and diligent enforcement of competition law constitute the stepping stones that will allow us to reach our objectives which are sustainable, secure and fairly priced energy. In recent years the Commission has probably dealt with more cases in the energy markets than in any other sector. I think this shows how much we find that this is a priority for us. The Commission has acted using all the tools at our disposal to achieve the liberalisation objectives which have been set for European energy markets.

3. The title of the panel as stated here, reads, and I quote, "*Can competition, security of supply and environmental protection be reconciled?*" However, this makes it sound simply like "*Can they be reconciled?*" In my view the question should be asked not like this, but as "*How can competition contribute to achieving security of supply and environmental protection?*" The purpose of my presentation today will be to explain to you and defend the thesis that competition is not in contradiction to the other two objectives but rather it is a means to achieve them. The possibility of conflict between competition and security of supply or environmental goals is an allegation often brought forward by energy incumbents to justify foreclosure of domestic markets. My presentation will focus mostly on the issue of security of supply and I will also touch briefly on the matter of environmental protection which I think has already been dealt with rather extensively by the two speakers preceding me.

4. We are constantly reminded of the importance and the vulnerability of our energy supply. There have been numerous crises, such as the supply crisis with Ukraine, the issue with Georgia, the issue with volatile fuel prices throughout 2008 and then again in 2009. And, of course, you can also read in many newspapers long articles with titles such as, "*Security of supply is decreasing. European citizens expect the Union*

to act". But what does security of supply mean? Well, I would define it as continued energy supply at reasonable and foreseeable price levels. And within this definition I would then add that: firstly, interconnection of networks is crucial for ensuring this aim; as is, in times of greater demand, what has been called "solidarity" between shippers. Secondly, it is important to have a sufficient and diversified number of supply sources. Diversification of supply routes provides more stability and less dependency on a sole supplier in the case of gas, for example. Finally, prices must reflect real market outcomes and create pressure on suppliers and there should be trust in price formation mechanisms.

5. In the gas sector, security of supply takes the form of diversified supply routes to consumption regions, as well as sufficient capacities intra EU and intra Member State to change supply directions, if required. In the power sector the key is a diversified mix of fuel for generation and sufficient investment in new generation to ensure an adequate equilibrium between supply and demand. In both cases the key is encouraging sufficient investment to support our security of supply objectives.

6. So, how has security of supply developed over the past years? Well, traditionally security of supply was seen as a national concern. The issue was looked at by each member state and generally the sector was put in the hands of a single incumbent under direct or indirect control of the State. This situation has been developing, particularly in light of the liberalisation which really began in the 90s in the European Union. It's become apparent that energy security doesn't stop at the border. Its international dimension makes it a matter of EU concern and also makes it more efficient: the geographic location of different Member States allow a better diversification of supplies, for example, Spain or France have access to Algerian LNG or the UK has access to Norwegian gas. The benefit that we can derive from this is that, as markets become more interconnected, different parts of Europe can get access to sources of supply which didn't use to be available to them. Finally, as the sector becomes progressively more privatised the importance of regulation and supervision has increased. This is because, without the proper framework shifting companies' incentives, maximization of profit will not always necessarily lead to security of supply.

7. Let me now discuss what competition brings to the energy sector, to security of supply and to our environmental goals. And let me begin by asking a rather provocative question: why do we bother with competition in the energy sector? What is wrong with the old system of, as some might put it, letting incumbents provide steady supplies of energy to Europe? I mean, why are we disrupting the status quo? Now, of course, you would expect me to defend the need for competition and in this, you would be absolutely right. In my view there are a multitude of benefits to bring in competition to European energy markets. First, freedom of choice. Europe's citizens have very different expectations on the energy market. Some focus solely on prices. Others want to choose green electricity. Still others may wish to generate their own electricity and feed it into the grid. Whatever the expectations, consumers should have the choice. Second, energy prices. Competitive markets are playing an important role in curbing energy price increases and protecting consumers against unjustified price increases. However, the Commission cannot guarantee that, with liberalisation, energy prices will go down. The problem is that liberalised energy markets are, naturally, heavily influenced by world energy prices and several of the exogenous factors which I discussed earlier. Finally, the issue of increased investments of infrastructure. A competitive market with correct price signals, provides the indicator for companies to increase investment in new infrastructure. And, as we have witnessed in some Member States following liberalisation of the gas markets, market conditions make investors choose the most cost-effective units, provided the price signals are right. Competition between suppliers then ensures that the lowest production and service costs are achieved. I think those are some of the benefits that competition brings to the energy sector which leads us to the crux of this presentation. As I said earlier, competition is a means to helping us achieve the real objectives of energy policy, that is: sustainability, competitiveness and security of supply. We in the Commission believe that these three are mutually enhancing.

8. In what ways can the application of this tool, competition law enforcement, improve market functioning to the benefit of the Commission's triumvirate of policy objectives? I would like to describe the positive impact of our more recent cases on European energy markets.

9. First, our cases are leading to better interconnection between gas networks and to less shackled flows between Member States. In particular, I would point to cases challenging the undersizing of interconnectors, for example in the ENI case, where we accused the company of failing to increase capacity in major import pipelines in order to protect its dominant position on supply markets. The Commission also carried out cases ensuring fair access to capacity, such as the *RWE* case, where we accused the company of failing to release unused capacity. Similarly, in another case we accused EON and GDF of blocking access to markets where they were dominant through a combination of long term upstream supply contracts and matching long term capacity reservations. Note that both these cases challenge historic long-term capacity reservations. Finally, we have cases removing artificial barriers to trade, for example the alleged collusion between incumbents in the form of the market sharing between EON and GDF. In July of last year the Commission issued a decision fining each company 553 000 000 euros for having colluded to share the German and French gas markets over a very long period of time. All of these cases promote increased gas exchanges across the EU both through encouraging or permitting companies to sell across borders and through the demand this then creates for cross-border pipeline capacity. This fosters investment. Naturally, the greater the ability of shippers to sell gas anywhere in Europe, the higher the security of supply obtained for consumers.

10. Second, our cases are allowing true signals for investment in electricity grids and in power generation. I would like to point out our *SVK* case (SVENSKA KRAFTNÄT, the Swedish transmission system operator). We published,

in November 2009, the market test of the commitments offered by SVK to solve our concerns that it could be artificially limiting electricity exports from Sweden to neighbouring states, thus infringing article 102 of the TFEU (formerly article 82 of the EC Treaty). The company is proposing to subdivide Sweden into several price zones, thus removing the anticompetitive export system which is currently in place. This case brings real price signals as to the value of interconnection and allows operators to correctly assess the necessity of investments in network infrastructure.

11. Third, our cases also create conditions for market entry or facilitate market entry. Some cases remove foreclosure of customer markets, for example, in the EDF long-term power contracts case, we recently finished carrying out the market test for the commitments proposed by the company to resolve the identified issues. The proposed commitments include guaranteeing that sufficient electricity comes back to the market on a yearly basis, removing resale restrictions from downstream sales contracts and limiting their length to a maximum of five years. Regarding this case, I would further like to point out that requests from customers to continue having long-term visibility on pricing are not incompatible with this objective – for example, the Exeltium agreement falls within it. I would also point to, in the gas sector, our cases removing territorial restrictions from supply contracts – such as the ones which used to be found in supply contracts with Algerian or Russian sellers. These cases increase the attractiveness of European energy markets to companies and their ability to operate throughout the EU. They also prevent incumbents from locking in customers and denying competitors the possibility of selling on their market.

12. Fourth, the cases we have carried out help prevent the manipulation of market prices – for example, the recent E.ON case dealing with reduction of production, so called "withdrawal of available generation capacity". In November 2008, in the electricity sector, the Commission adopted a commitment decision in the E.ON case by which it accepted substantial remedies (5000 MW divesture of generation plants) that structurally change the German electricity market to the benefit of consumers. In a parallel case the Commission investigated whether E.ON had abusively raised network costs to the benefit of its generation affiliates. To resolve these concerns E.ON agreed to sell its ultra-high voltage network. Although I haven't directly mentioned them yet, merger cases are also one of the crucial elements of competition law enforcement. Within this fourth group of cases, I would also cite the EDF/British Energy merger, where we identified issues relating to the possibility of the merged entity manipulating market prices. At the end of 2008 the Commission approved the merger, subject to conditions. The package of remedies we obtained from the parties wholly addresses these concerns. These cases, both antitrust and merger, show the Commission's determination to ensure that market prices reflect real economic fundamentals and allow companies to make accurate investment decisions. As you will remember, one of the key elements of security of supply was that price formation mechanisms must be truthful.

13. Fifth, our cases open possibilities for third parties to invest in power generation. As well as the aforementioned EDF/British Energy cases, I would also mention the EDF/SPE case, approved in November 2009, where, to remedy competition concerns the Commission had in relation to the reduced incentives of EDF to continue its plans to build additional electricity generation capacity in Belgium after the proposed acquisition, EDF has committed to immediately divest the assets of one of its companies in charge of the development of one of EDF's planned power station projects and, should EDF decide not to invest in a second planned power station by a specific date, to divest the assets of the company in charge of that project as well. Clearly, giving as many companies as possible the opportunity to invest in generation assets across the EU will increase the security of supply of the European power market.

14. Very briefly, since it's been discussed in greater detail by previous speakers, I would just like to touch upon the issue of competition enforcement and achieving our environmental goals. I would, in general, simply argue that competitive markets mean more cost-reflective prices and that application of competition law increases the cost-reflectiveness of prices, as I indicated regarding some of the cases described earlier. These costs can also include environmental costs when supported by the appropriate regulatory framework. Competitive, transparent and integrated markets provide the necessary price signals to allow environmental schemes such as the EU ETS to function optimally. They also stimulate the development of new environmentally friendly technologies. To me, it's not an issue whether competition can co-exist with achieving ambitious environmental goals, but how to design a structure which puts in place the required incentives.

15. Finally, I think I would like to share with you a quote by Niels Bohr, the Danish Nobel prize physicist who said that "*prediction is very difficult, especially about the future*." Now, whilst I couldn't agree more with the statement I believe that the Commission's competition enforcement actions in the past have already begun to lay down the groundwork for a future integrated and diversified energy market. This is in my view the key to achieving our objectives of sustainable, secure and fairly priced energy. Such a market will provide a solid base enabling us to reduce our vulnerability and establish clear relations with energy suppliers as well as ensuring a higher level of cross border arbitrage, which is important in times of shortage (such as last year's gas crisis). Further, a well integrated and functioning EU market will also be attractive to foreign investors and producers, offering enormous opportunities in a single market with almost 500 million consumers. Specifically regarding security of supply, I think there is a real problem with the argument which we sometimes hear supporting a one supplier per region policy. In my view, only a diversified market with many suppliers purchasing from different producers can provide the security we are striving for European consumers. The Commission is acting now to have competitive, sustainable and secure energy markets in the future. We are confident that the diligent enforcement of competition policy makes a real difference promoting these objectives. I for one believe that one day customers will view the idea of having an energy supplier forcefully thrust upon them depending on the country they live in, as anachronistic as having only two TV channels or going to the well for water. ■

Alberto Heimler
alberto.heimler@me.com
Scuola superiore della pubblica amministrazione, Rome

The economic crisis: Which jurisdiction is better placed the US or the EU?

Abstract

The US economy is much freer than the European economy. As a result new opportunities are much easier to be exploited. Europe has changed dramatically in the past 20 years under the leadership of the European Commission. The gap with the US is not filled. A great effort still needs to be made in order to eliminate a number of unjustified regulatory restrictions that still block domestic economies. The introduction of competition impact assessment techniques would help and so would the creation of a more vocal advocate for competition, either a minister of competition like in Australia or giving the chairman of the competition Authority ministerial status. However Europe is better placed than the US because of the existence of State aid rules that limit the amount of subsidies national governments are providing to firms, helping Governments resist business requests for bail outs. In the US there is no discipline on amount of aid to be granted. As a result the European economy will exit from the crisis more efficient than it would have otherwise. The same on antitrust. The less ideological approach that with respect to the US the EC has developed in the field of exclusionary abuses will tend to maintain in the market efficient firms that otherwise, with a contracting economy, might leave the market. Active antitrust enforcement in Europe will help mitigate the effect of the crisis, while the hands off of the US may well exacerbate it.

L'économie américaine est plus que l'économie européenne. Cela implique une plus grande facilité à saisir les opportunités du marché. Même si l'Europe, par l'intermédiaire de la Commission européenne, a radicalement évolué au cours des vingt dernières années, l'écart avec les USA n'est toujours pas comblé. Un effort conséquent reste e ncore à faire pour éliminer un certain nombre de restrictions réglementaires injustifiées qui freinent les économies nationales. L'introduction de techniques d'évaluation d'impact de la concurrence serait une première étape ainsi que l'institution d'un « défenseur de la concurrence », que ce soit, comme en Australie, sous la forme d'un ministre de la concurrence ou encore d'un Président d'autorité ayant le statut ministériel. Mais l'Europe est mieux placée que les USA grâce à l'existence des règles sur les aides d'Etat limitant le montant des subventions que les Etats peuvent accorder aux entreprises. Ces dispositions réduisent le risque les gouvernements soient entraînés par des entreprises dans des opérations de sauvetage. Les USA n'ont aucune règlementation limitant le montant des aides publiques. Il en va de même pour l'antitrust. L'approche moins idéologique de la Commission européenne dans le domaine des pratiques d'exclusivité permet de maintenir en vie des entreprises efficaces qui auraient pu se voir contraintes à sortir du marché sous le coup de la crise. En définitive, la mise en œuvre des règles européennes de concurrence devrait donc atténuer les effets de la crise à l'inverse des dispositions américaines qui devraient les aggraver.

1. These are bad times for competition and for competition oriented reforms. The community of antitrust authorities is very worried that the "golden decades" of liberalization are over. It should not be. There is nothing new in this negative attitude against competition. All throughout the years competition was fiercely resisted, except in some countries and for some (short) periods. The reasons are very profound and are independent of the current crisis. Competition operates via two channels: creation and destruction. The sequence is clear: because of greater opportunities new activities are created and, as result, old obsolete activities are no longer competitive and are shut down. New entrants gain from competition. Inefficient players lose. This is a process that takes time, much more so than in trade liberalization where, once tariffs are cut, foreign cheaper products enter the domestic market very quickly, the only constrain being the capacity of production of the exporting manufacturer.

2. With competition oriented reform the process of creation (and therefore also of destruction) is much smoother than with trade liberalization. There is time for every player to adjust to the new environment. However, in the political economy of competition oriented reforms, the timing of the effects of liberalization is irrelevant and all the effects are supposed to take place immediately. As a result, those that suffer from competition tend to emphasize the reductions in jobs that a process of liberalization will bring about, while those that may gain from it, either are not present in the debate (the new entrants) or gain too little individually as to care too much (consumers). Furthermore while competition is presented by its promoters as an instrument for promoting product differentiation, innovation and growth, it is contrasted by its detractors as a wicked tool that would undermine the attainment of some important general interest objectives like, among others, industry competitiveness, market stability, universal service and employment. As a result, advocates for competition are forced to prove that the proposed competition solution would not be so negative with respect to the attainment of these important general interest objectives. A debate where the opponents of competition would have had to argue against it showing that the benefits of competition would not be attained, is transformed into a discussion where competition needs to be defended. Not a win-win solution.

3. Furthermore liberalization does not take place in a vacuum. It is directed to dismantle regulatory protections that may have existed for decades. The protected category has much to lose from liberalization because in the service sector where today's restrictions are particularly present those that could benefit from liberalization are probably new entrants, for example the big supermarkets, the younger lawyers, etc. As a result, the opposition of the protected category towards liberalization cannot be weakened, like in manufacturing, by the prospects of larger markets.

4. This is why European wide liberalization are so important: domestic protectionist coalitions are less likely to be successful. European pro-market instruments are numerous and a free trade regime would not have been effective. The treaty guarantees within the Union the respect of the four fundamental freedoms, i.e. the free movement of goods, services, labor and capital, it introduces antitrust provisions and it impedes anticompetitive subsidies. All these provisions were necessary to address private and government restraints that segmented national markets, thus potentially undermining the common market. No other international organization or even no other sovereign country has a similar portfolio of instruments aimed at achieving an integrated market.

5. While all these instruments are clearly beneficial in an expanding economy, are they still beneficial in a crisis? If the answer is yes, then it means that Europe is better placed today than other jurisdictions that lack such a rich and articulated portfolio of instruments.

I. State aid[1]

6. Europe is the only jurisdiction in the world that has introduced a rigorous system of State aid control. In the US where there is no State aid control many have argued that state aid control is not necessary since most subsidies (tax breaks) are meant to induce new firms to locate in the subsidizing State. The argument goes as follows: since States compete for companies to be localized in their territory, they should be allowed to offer to them all sorts of services and competitive advantages: good infrastructure, good schools, good health services, etc.. All these are not considered State aid, even if they are provided for free. On the other hand State aid is prohibited, even if it might help under resourced countries overcome their comparative disadvantage. Without State aid control, competition for locations will be won by the region offering the most advantages (real and financial) to companies willing to locate there.

This argument cannot be dismissed prima facie and has some value. For example it could be argued that in the United States where there is no State aid policy, individual States operate under a strict balanced budget constraint and are mostly responsible for their own finances. In such instances, even if States grant aid to companies, the voting mechanism and the reduction in the tax base originating from people leaving a bankrupt State can well discipline it, even in the absence of State aid control at the Federal level.

7. However the argument is based on the assumption that the length of the political and the economic cycle is the same. Should it be so, policy makers of a fiscal disciplined State would not find it attractive to grant ineffective State aid for fear of them not being re-elected. However, the political cycle is much shorter than the economic cycle and policy makers maintain a positive incentive to grant an excessive amount of State aid (in comparison to real advantages) even under a rigorous fiscal discipline. As a consequence, even in those jurisdictions like the US where individual States operate under a strict balanced budget constraint, the introduction of State aid control could therefore nonetheless be necessary to impede an excessive amount of aid from being granted.

8. This is the more so in Europe. Member States do not operate under a strict balanced budget constraint and especially now with the common currency any fiscal largesse (the Maastricht treaty budget parameters impose effectively a soft budget constraint) is transferred to other sovereign bodies, for example as a result of a weaker currency. The same is true for most local governments that also do not operate under a strict budget constraint and, in case of need, are bailed out by their national governments. Free riding can therefore lead to an excessive amount of State aid in Europe, but also in the US. The control of State aid is therefore necessary, even for locational aid.

9. The more so in the economic crisis when, without State aid control, there is no limit on the subsidies to be given for restructuring too big to fail firms. Here some control on State aid is necessary in order to avoid moral hazard, that is disciplining companies from taking excessive risks knowing that in any case they will be bailed out. That the question of moral hazard is so crucial can be seen by the discussion we are having these days in the US about bonuses to bank managers in subsidized banks. What about the car industry? Should we introduce also there a cap on managers income? But even if we do so, do we really avoid moral hazard by introducing income caps? I do not think so. The only way to avoid moral hazard is to limit State aid and indeed the legal provision in the Community impede the Commission from being too generous. Article 87 provides a strong discipline for member States. In the present circumstances aid that remedies a serious disturbance in the economy can yes be exempted but only temporarily and under a strict definition of what is a serious disturbance. The Court in Europe can play an active role even by disciplining the Commission. Nothing of this sort can even be imagined in the US.

10. In Europe legal provisions on State aid make sure that the economy is not overly subsidized and therefore also in the crisis markets will tend to remain more competitive than without them.

II. Antitrust

11. The past 10 years have been the years of cartel detection. There is nor risk of false positives with cartels. The crisis does not eliminate the need to fight against cartels. To the contrary. In order to manage the reduction of capacity, crisis cartels may be created. More importantly however the crisis tends to make cartels less stable because a shrinking market provides strong incentives for cheating on a cartel agreement. As a result it might well be that cartels will become less important.

12. As for unilateral conduct, the crisis will accelerate the exit from the market of the weakest firms, both in market in which there is dominant company and also in more competitive markets. The confidence that an expanding economy will bring relief to those excluded (both firms and people) will therefore weaken. In such circumstances the fear, almost the panic, developed in the US against false positives can be criticized not only on its merit, but also because it will exacerbate the crisis.

13. Today the EC and the US are not two worlds apart as they were until the 1990's. With a 20 years delay Europe has bridged the gap with respect to the use of economic analysis as a tool for identifying a violation of the law.

1 For a more comprehensive analysis of State aid policy, see Heimler, Alberto (2009), "European State Aid Policy in Search of a Standard: what is the Role of Economic Analysis", presented at the Fordham antitrust law institute conference in September 2009.

14. The greater importance of economic analysis in EC antitrust has been prompted by the introduction of the merger regulation in 1989. The emphasis on economic analysis that it brought with it started to move the Commission away from form-based to effects-based enforcement, first in merger control and after also in antitrust (restrictive agreements and abuse of dominance). The communication on the relevant market was issued in 1997; the new block exemption on vertical restraints in 1999. But only very recently with the introduction of Regulation n. 1/2003 and the elimination of the notification system for exemptible agreements, the effect based approach has made it in European antitrust. After 40 years, substance is what matters in European antitrust enforcement.

15. The U.S. influence was very important in this respect. The definition of the relevant market, the treatment of vertical restraints and the way to analyze mergers have a clear U.S. origin. The Chicago School made it clear that form-based antitrust enforcement is completely ineffective, with the exception of hard-core cartels which are always prohibited. Economic analysis, and its insistence on efficiency, has provided the glasses through which to interpret antitrust enforcement provisions.

16. Just as an example of this achieved convergence on principles, the OECD Competition Committee held a roundtable discussion on competition on the merits in 2005. It was clear from the Commission submission that also on abuse of dominance the Commission is moving towards an effects-based approach. Writes the Commission: "The protection of the competitive process is not protection of competitors. When analyzing the effects of behavior by dominant companies competition authorities should not disturb competition by protecting competitors that are inefficient […] On the same notion in cases regarding refusal to deal competition authorities should not disturb the competitive process by intervening in order to grant access to the market to competitors who, as an efficient operator, should be able to create their own access to the market."

17. However the past is not completely overcome and the Commission concludes: "Dominant companies should be able to successfully defend themselves against challenges of abuse by demonstrating that there is an objective justification for their behavior".

18. While in the US such objective justification for the behavior of a dominant company would not be required, the Commission statements show that, contrary to the past, economic analysis has become an essential part of the Commission approach on abuse of dominance. The difference between the US and the EC antitrust enforcement is no longer an ideological one. Convergence on principles is achieved. The difference, and it remains big, is in the details of enforcement and, in particular on the burden of proof.

19. First of all in the US two Supreme Court judgements, Trinko and especially Linkline, have made it clear that a regulatory duty to deal does not imply an antitrust duty to deal. In the EC a regulatory duty to deal is the same as an antitrust duty to deal. In Europe exclusionary conduct by regulated companies is pursued much more aggressively than in the US also outside the regulated area.

20. As for exclusionary conduct by non regulated firms, the EC has developed the as efficient competitor standard and actual exclusion plays a very minor role. In the US identifying a violation without the excluded company leaving the market is extremely difficult. Both in Le Page, in Concord Boat and in Virgin-British Airways, the question was whether competitors left or did not leave the market. In the EC, the Britsh Airways case can be criticized because the Commission did not show convincingly that as efficient competitors would have to price below cost in order to match BA discounts, but there is no question that the Commission was trying to do that, without any conclusion being drawn by the fact that competitors did not have to leave the market as a result of those discounts.

21. In the face of the economic crisis and the lower ability of weaker firms to find market alternatives, the European approach, by not giving up efficiency but being more open to exclusionary claims, is more able to keep firms in the market and therefore making the crisis less severe.

III. Liberalization

22. On liberalization the general attitude of the US is for open markets and has been so for decades. There is no question that anticompetitive government restraints are much less important in the US than in the EU. The European Community has been a follower in this respect, but its action has been essential for bringing larger and more open markets in Europe. The problem is that every effort by the Community has been resisted by member States, luckily not always successfully. However on most circumstances the opposition by member States to competition oriented reforms watered down the original Commission proposal and at best delayed the reforms. Here there is ample room for melioration especially at the domestic level. A few examples may help[2].

1. Telecommunications

23. In the late 1980s, the telecommunications sector was characterised by legal monopolies in most member States. A Commission directive issued in 1988 on the basis of Article 86 of the Treaty introduced competition in the market of telecommunications terminal equipment. An interesting phenomenon, which shows the type of pressures that originate from pro-competitive reforms and the ways in which the Commission's powers and related institutional machinery have helped to address these powers, is that Member States participated fully in the discussions that led to the directive. However, when the directive entered into force, five Member States (France, Italy, Belgium, Germany, and Greece) challenged it before the Court of Justice. The Court ruled

2 These examples are taken from Heimler, Alberto (2009), "Regulatory Reform and Competition: How to Push the Agenda Forward. A European Perspective", *Comparative Economic Studies* 51: 540-557.

conclusively in favour of the Commission. After this decision, the liberalisation process gathered steam. In 1990, the Commission issued directive No. 388 which liberalised value-added services and data transmission.[3] Only voice telephony was left as a monopoly because a number of countries opposed its liberalization, even though voice telephony was characterizes by high inefficiency in many Member States. After France and Germany offered support for full liberalisation, all Member States finally agreed on a timetable for the comprehensive liberalisation of telecommunications infrastructure (Council resolution of 22 December 1994).[4] Starting on 1 January 1998, the telecommunications sector was opened up to full competition.

24. Of course, liberalisation did not create competitive markets overnight. It takes time for new entry to become established. As had already happened in the past, some Member State governments were reluctant to introduce a pro-competitive regulatory structure in a timely way. Further action by the Commission was therefore necessary. In 2002, a package of six directives was approved: the common regulatory framework directive;[5] the universal service directive;[6] the data protection and privacy directive;[7] the directive on access and interconnection;[8] and the authorisation directive for electronic communications.[9] The main thrust of these instruments was to promote the use of competition-based regulatory concepts, going a long way to create a level playing field in the European telecom sector.

25. Today, more than twenty years after the process of liberalization started, market discipline in telecommunications is well accepted by member States and, although the incumbent operator continues to remain dominant in all jurisdictions especially in fixed-line services, it is nonetheless subject to a strong rivalry by competitors allowed to access the unbundled components of the telecommunications network. As for mobile services, although regulation continues to be important for some part of the business, rivalry is quite strong.

The opposition to liberalization was clearly short sighted. Technological progress and the example of the US were the major reason why TLC markets were opened up.

3 Directive 90/388/EEC of 28 June 1990 on competition in the markets for telecommunications services, *Official Journal L 192*, 24 July 1990, pp. 10-16.

4 Council Resolution 94/C 379/03 of 22 December 1994 on the principles and timetable for the liberalization of telecommunications infrastructures, *Official Journal C 379*, 31 December 94, pp. 4-5.

5 Directive 2002/21/EC of 7 March 2002 on a common regulatory framework for electronic communications networks and services (Framework Directive), *Official Journal L 108*, 24 April 2002, pp. 33-50.

6 Directive 2002/22/EC of 7 March 2002 on universal service and users' rights relating to electronic communications networks and services (Universal Service Directive), *Official Journal L 108*, 24 April 2002, pp. 51-77.

7 Directive 2002/58/EC of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), *Official Journal L 201*, 31 July 2002, pp. 37-47.

8 Directive 2002/19/EC of 7 March 2002 on access to, and interconnection of, electronic communications networks and associated facilities (Access Directive), *Official Journal L 108*, 24 April 2002, pp. 7-20.

9 Directive 2002/20/EC of 7 March 2002 on the authorisation of electronic communications networks and services (Authorisation Directive), *Official Journal L 108*, 24 April 2002, pp. 21-32.

2. Electricity

26. The degree of opposition to reform the electricity sector exceeded that which occurred in relation to telecommunications reforms. Furthermore legally, while the Commission had the final word in the liberalization of telecommunications, the Council was in charge of electricity and the Commission had only an initiative role. As a result, from December 1990 to June 1996, the initial Commission position to allow direct transactions between as many producers and consumers as possible was blocked by the opposition of Member States believing that a vertically integrated structure for the industry with no possibility for direct transactions by consumers with generators was preferable. In June 1996, after a long debate, the Council of Ministers agreed on a directive concerning common rules for the industry.[10]

27. The 1996 directive began the process of introducing competition while leaving important choices to the discretion of individual Member States. For instance, the directive allowed Member States either to provide for free entry in electricity generation or to introduce a tendering procedure in order to identify the actual generator that entered the market, maintaining central control on the technology to be used and the timing of entry. Furthermore, a grid operator could be made responsible for power transmission and for guaranteeing the physical equilibrium of the system and Member States could designate a single buyer with responsibility for ensuring supply to non-eligible customers.

28. Indeed, leaving to much discretion open, the directive was quite ineffective in changing the market and regulatory structures of Member States. As a result, in order to create a level playing field among suppliers, further important measures were introduced by Directive 2003/54/EC[11] and Regulation (EC) No 1228/2003 on "Cross border Electricity Trading"[12]. Directive 2003/54/EC aimed at complete market opening, requiring that all non-household electricity customers become eligible by 1 July 2004 and all household customers by 1 July 2007. However, in sectors such as electricity where entry requires substantial investments and involves a lengthy authorisation process, simple market opening could not automatically lead to the introduction of vigorous competition. Structural measures such as divestiture would have been be necessary. The directive was silent on this issue, reflecting different beliefs among Member States on the benefits of stronger competition. On their own initiative, some Member States imposed capacity divestitures on the former legal monopolist sometimes coupled with temporary measures to increase competition such as market share caps.[13]

10 Directive 96/92/EC of 19 December 1996 concerning common rules for the internal market in electricity, *Official Journal L 027* , 30 January 1997 pp. 20-29.

11 Directive 2003/54/EC of 26 June 2003 concerning common rules for the internal market in electricity and repealing Directive 96/92/EC – Statements made with regard to decommissioning and waste management activities, *Official Journal L 176,* 15 July 2003, pp. 37-56.

12 Regulation (EC) No 1228/2003 of 26 June 2003 on conditions for access to the network for cross-border exchanges in electricity, *Official Journal L 176,* 15 July 2003, pp. 1-10.

13 In the UK and in Italy, the existing state owned monopolists were split up into competing undertakings in order to create competitive markets, a move which in Italy has nonetheless maintained an incumbent operator with a significant market power.

29. Directive 2003/54/EC also obliged Member States to introduce a regulated third party access regime, removing the possibility of negotiated third party access which had been permitted under the 1996 directive. Furthermore, the directive mandated the appointment of an independent national regulator. As for transmission and distribution, the directive required legal unbundling – stopping short of proprietary unbundling that had been proposed in the OECD as the most effective solution for aligning the incentives of the infrastructure owner with the general interests of society (OECD 2001 and 2006).

30. As the foregoing account implies, pro-competitive reform in the electricity sector has not gone as far as it has in telecommunications. In many cases, markets remain concentrated and national in character. According to the Commission sector enquiry on gas and electricity published in January 2007[14], the incumbent operator is vertically integrated in almost all Member States and the degree of cross-border competition is weak, due in part to a lack of inter-connection capacities. Nonetheless, other evidence indicates that, where effective competition has been introduced, important benefits have been generated for consumers. As the International Energy Agency[15] reports, the benefits of competition are quite strong: in the UK there is a clear falling trend in the price of electricity which is attributable to increased competition promoted by vertical and horizontal separations. In Portugal, vertical unbundling of electricity markets resulted in a 45-80% decline in access prices and a tripling of real investments in transmission facilities in the period 1999-2006[16].

31. Most countries started off in the early 1990's with a vertically integrated electricity company. The experience of the UK, Italy, Portugal, Argentina and many other countries shows that vertically separating the grid infrastructure is greatly beneficial for consumers. The benefit of vertical separation can be best understood by considering that the size of the market depends on the capacity of the transmission grid. If there are bottlenecks in the grid, regions are isolated, the market is not as wide as it could be and market power is raised within each region. A vertically integrated company will not have the incentive to invest in the grid so as to enlarge the geographic market. Consumers would greatly benefit from vertical separation. And indeed Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity suggests that "ownership unbundling, is clearly an effective and stable way to solve the inherent conflict of interests and to ensure security of supply"

32. Furthermore, market power depends on the level of demand. For example at off-peak times, only the "base load" generation will be active in the market. It can price up to the cost of the "peaking" generation without any risk that the peaking generation will increase output. Therefore, at least at off-peak times, the relevant market definition should focus on just the "base load" generation, again significantly increasing the apparent concentration. In addition, the inelasticity of the demand curve for electricity means that, in any case, at peak times, once demand exceeds a certain percentage of total capacity, market power may increase in an unlimited way. As a result, horizontal reorganization should lead to the creation of companies which would include both base load and peaking generation in a single generation portfolio, ideally to form at least four balanced companies in each region, a result that can be easily achieved with integrated markets.

33. This is exactly the problem that we face in Europe where national electricity markets, although connected, are not fully integrated, with the partial exception of the Nordic countries. Greater integration is achievable by imposing vertical separation in every country, creating a structure where the incentives to invest in transmission are independent from the market power in generation. The primary advantage of structural separation is that it eliminates the incentive on the owner of the transmission line to deny access to rivals, which of course would also include cross border generators (and the best strategy is to build as scarce a transmission capacity as possible). Since larger markets (that extend cross borders) imply electricity flows both ways, in order for the incentive structure to achieve these results, vertical separation has to be imposed everywhere. This is what the recent EU Council Directive is trying to achieve. Unfortunately proprietary unbundling is only a choice and countries are allowed to maintain a vertically integrated company provided that they create really independent transmission system operators (functional separation). We will see in the next years whether this reform is sufficient to modify the incentives to strengthen the transmission lines and indeed achieve larger markets.

3. Private services[17]

34. Like with telecommunications and electricity, the path to regulatory convergence and greater competition in private services has been full of resistances. However in private services it has been easier for the protectionist coalitions to block the Commission's competition-oriented reforms. In fact many service providers are individuals, not multinationals. As a result, public opinion perceived liberalization of these services not so much as beneficial for consumers (or a way to punish big and inefficient business such as the telecommunications or the electricity national monopolies), but rather as a cost, almost a punishment, for stakeholders.

35. Articles 43 and 49 of the European Treaty establish the freedoms of establishment and to supply services, prohibiting, according to European case law, not only discriminations based on nationality, but also all national measures that may hinder the exercise of such freedoms. Restrictions are allowed only if they are strictly necessary for achieving a public interest objective. This requirement of strict necessity makes it is very difficult for the Courts to intervene unless

14 See EC Commission (2007b), *DG Competition report on energy sector inquiry {SEC(2006)1724}*, Brussels, 10 January 2007.

15 *International Energy Agency (2005)*, Lessons from Liberalised Electricity Markets, *Paris.*

16 *Geraldes, Pedro (2007), Address to the Roundtable on Competition at the National and International Levels: Energy, Intergovernmental Group of Experts on Competition Law and Policy, 8th Session, UNCTAD, Geneva, July.*

17 For a more complete analysis of the Service directive, see Heimler, Alberto (2006) "La direttiva Bolkestein", *Mercato, Concorrenza, Regole*, 8, 1: pp. 95-109.

such restrictions are clearly not proportionate or unjustified, which is very rarely the case. As a consequence leaving the removal of regulatory restrictions to the direct application of articles 43 and 49 of the EC Treaty was quite ineffective and in the course of the years very few regulations have been successfully challenged.

This is the reason why the Commission in 2004 proposed the adoption of a Council Directive on services, the so called Bolkenstein Directive, based on a horizontal approach aimed at fully achieving the freedom of establishment and the free movements of services.

36. As regards the free movement of services, the draft Directive contained a widely criticized "country of origin" principle, according to which service provision in each member State would be regulated by the country of origin of a given establishment, without the need of any formal recognition by the "importing" country. The criticism of the country of origin principle was that it would induce social dumping. This came as surprise to the Commission (and it should not!). Neither in the presentation of the draft Directive nor in the discussions that followed, reference was made to the posted workers Directive that ensures that workers on temporary service from another Member State are paid at the conditions established in the host country. In other words the posted workers Directive made sure that the country of origin principle would not lead to cross border dumping on labour costs. The criticism was therefore unfounded, but, as a result of lack of proper information. public opinion remained convinced that the draft Directive would strongly reduce standards of living in member States (especially public opinion in the old 15 member States).

37. The opposition to globalization that characterized (and still does) public opinion in Europe was completely ignored by the Commission. In the official communication statements accompanying the draft Directive the Commission insisted that its aim was to introduce the country of origin principle, mentioning only slightly the impact the directive would have on regulatory reform, leading for the debate to go astray. First of all, the fear that the country of origin principle would reduce the quality of services in many member States led to a long list of sectoral exclusions from the obligations the Directive was imposing. Furthermore newspapers were full of articles that described the Directive as being responsible for "Polish plumbers" to move to richer countries, competing with domestic plumbers at Polish pay, a right that is enshrined in the Treaty (free movement of people) and that the new Directive could not touch. Public opinion became convinced that the Directive was the "mother of all evils" with respect to globalization and strongly reacted against it, with widely attended demonstrations in all EU capitals.

38. The Directive was finally adopted in 2006 (but it entered into force only two weeks ago, on January 1 2010) but, as a result of this strong opposition, is now much weaker and less effective than it could have been. The country of origin principle is gone and the Directive is back to the principle of mutual recognition, which means that each member State has a positive duty to authorize service providers if they are authorized to operate in other member States (denying an authorization is possible only under exceptional circumstances strictly defined in the Directive), delaying the entry of more efficient foreign competitors into domestic markets. The list of sectoral exclusions is also long: finance, communications, transport, temporary work agencies, healthcare, broadcasting, gambling, social services, private security services and notaries. Finally on liberalization, many of the provisions contained in the draft Directive that would have made illegal unjustified domestic regulatory restrictions have been eliminated. For example the draft directive contained a ban on all prohibitions of advertising, now that is gone. A better communication strategy by the Commission aimed at showing that competition would not have disrupted the "European social order" might have avoided all this.

IV. Domestic competition advocacy

39. Most liberalizations in Europe have originated from the Community. This is a problem in so far as liberalizations are mostly needed in the non-tradable sector. With respect to private services the Community is generally impeded from intervening because of the absence of a legal base (effect on intra-community trade is a prerequisite for Community action). The service directive is an important exception in this respect but is unable to eliminate most anticompetitive regulations. Its major task is to eliminate all sort of discriminations based on nationality considerations.

40. Indeed to fill the gap between tradables and non tradables, the Italian antitrust law, like in many other jurisdictions, assigns to the Authority the power to advocate in favor of competition oriented reforms. The law gives to the Authority the power to intervene in the legislative process with advices and reports, but does not introduce any obligation on the part of the legislative or executive body to listen! And indeed, being competition not very popular, only a few of the almost 450 reports the Authority issued since its establishment have been followed.

41. Competition advocacy by the competition authority was certainly very important in Italy to bring competition in the political debate, but it has not shown to be effective.

42. Domestic reforms still have a big role to play. And in Italy last year, having recognized the political relative ineffectiveness of competition advocacy (the Authority's advices are seldom discussed by Parliament), the Italian Parliament adopted a law establishing a duty on the Government to present every year to Parliament a proposal containing a number of competition oriented reforms. The law prescribes that Government, sixty days after having received the annual report by the Italian Authority, presents a law to Parliament based on (but not limited by) the liberalization measures proposed by the antitrust Authority in the course of the year, making it therefore mandatory for Parliament to discuss at least some of the Authority's competition advocacy proposals. The law is very recent and will be applied in 2010 for the first time. We all look forward to it.

43. A second very effective reform originates from Australia that a few years ago established a minister for competition in charge of promoting competition in government policy making. Their reasoning was that all other general interest policies, from the environment, to health and justice, have a representative in the government. Only competition does not. The weakness of the Australian approach is that the Minister for competition is a junior Minister and does not provide a great influence in Government decision making.

44. A third approach is that of South Korea where the Chairman of the KFTC has ministerial status and sits in the Council of ministers as a non voting member.

45. Finally introducing specific provisions mandating competition impact assessment of all rules having an economic consequence may prove very useful because it would help identifying the less restrictive regulatory solution able to achieve the general interest objectives pursued by legislation and everybody would be informed of the details of the more competitive solution before any decision is taken. An OECD recommendation suggesting that member Governments adopt competition impact assessment mechanisms has been recently adopted by the OECD council of ministers.

46. However competition impact assessment is a technical issue and having an advocate for competition in the Council of ministers makes it more likely that the most protectionist approach is not followed!

V. The importance of stake holders

47. But what is really important is to create stake holders that gain by competition and are willing to speak up for it.

European Commission directives are appealed in Court by the same member States that did not oppose the Commission initiative in the various advisory committees. Council directives are resisted or implemented with delay. Domestic liberalization measures are strongly opposed. And this much before the economic crisis. In Italy in 2006-2007 the Prodi government (under the initiative of minister Bersani) undertook an extensive liberalization effort in private services by issuing two Government decrees. As soon as the proposals were made public (before they were approved by Parliament) all categories affected reacted very strongly against them, mostly using general consideration arguments, such as trust in the professional-client relationship, universal service in pharmacies, stability consideration in banking and insurance, competitiveness of the industry in the case of airlines, etc. No affected category representative declared publicly that greater competition would yes reduce the income of some individuals, but that it would increase productivity and reduce prices. They tried to gain the sympathy of the public by claiming that these liberalization measures would negatively affect general interest objectives, a claim that could be easily dismissed, but it had a strong appeal.

48. What is interesting is that the decree liberalized as much as it could. For example in the case of pharmacies, the decree anticipated the protests of pharmacy owners (a minority of all pharmacists of the country), allowing the commercialization of non-prescription drugs only at the presence of a qualified pharmacist. In the two years since the decree entered into force, contrary to what had been expected, supermarkets did not generally open a non-prescription drugs section, but 1500-2000 small shops of a new type have been established, so called "para pharmacies", almost all run by qualified professional pharmacists. These pharmacists do not own a regular pharmacy and are becoming an important lobby for the liberalization of the commercialization of all drugs. "We are qualified professional pharmacists and already have a shop where we sell health products, why not let us also sell prescription drugs?", is what they have already started to claim.

49. A reform that was defensive in nature (imposing an unnecessary burden on an activity) has proved to be quite successful in terms of possible future developments (the yet to come full liberalization of the pharmacy sector).

Indeed, much more that any legal provision opening up markets, creating vested interest that gain from further liberalization is proving the most effective way of pursuing a competition oriented reform.

50. The government of which Mr Bersani was part resigned in June 2008. Since then there have been a number of proposals aimed at abolishing his reforms, especially with respect to the professions, including of course the pharmacy sector. The existence of vested interests, and in particular new entrants, much more than consumers and the press, have been very effective in preventing backlash (at least until now).

VI. Conclusions

51. The crisis did not led people to change their opinion on competition, it just confirmed their original skepticism. In this sense what is going on in most countries, with Parliaments and Governments trying to undue many of the liberalizations of the last decades is not surprising.

52. The problem is that competition exercises its beneficial effects in the medium/long run because it takes some time for firms to take advantage of the new opportunities, very often associated with the setting up of new organizations and with the necessity of new investment. It takes time for the opportunities of competition to be exploited also when liberalization is related to the conduct of firms, like it was the case with professional services, because of inertia and the lack of an immediate need to change by existing firms.

53. As a result, many people do not link the advantages they receive from the market to the liberalization decision, often taken two or three years before and competition continues not to be appreciated by the general public. Furthermore, special interests always find very clever ways to justify the restrictive regulations that protect them by showing that they

are very important for pursuing general interest objectives that are to the advantages of everybody. Now the crisis is an additional argument against competition, but it is not the only one, nor the most important. In this sense, what we are witnessing in recent months is "business as usual".

54. Advocacy report by the antitrust Authority and communication of the benefits of competition by consumer associations or by the press are certainly useful, since they promote the culture of competition with the general public and, especially, with decision makers, but certainly they are not sufficient.

55. What is needed is the creation of vested interests that may gain from further liberalization and that would loose if liberalization was undone. Restraining competition is fine in the abstract when an unknown new entrant is blocked, but it is politically impossible when it leads to unemployment and to the closure of firms, as would be the case with the Italian parapharmacies if the Bersani reforms would be undone. The protected categories know this very well and this is why they fight every minimal liberalization as it would create a disaster. They know that by letting a new entrant in the market, the liberalization process can only continue further. Once liberalization starts the clock cannot be set back. Not even in an economic crisis.

56. In this respect open markets are much more a characteristic of the US than the European economy. In the past 20 years great progresses have been achieved in Europe through the leadership of the European Commission. As a result domestic protectionist coalitions were much less successful in blocking reforms as they would have been had these proposals been originated at the domestic level. However a great effort still needs to be made in order to eliminate a number of unjustified regulatory restrictions that still block domestic economies. The greater importance that competition impact assessment techniques have in government decision making makes sure that competition is actually one of the general interests to be considered in assessing the necessity of a new regulation. However a more vocal advocate for competition may be needed at the domestic level, either a minister of competition like in Australia or giving the chairman of the competition Authority ministerial status. Italy has moved one step in that direction by making it mandatory for government to present every year to parliament a liberalization law based on, but not limited by, the advocacy reports of the competition Authority.

57. As for directly applicable legal provisions, European State aid rules limit the amount of subsidies national governments are providing to firms, helping Governments resist business requests for bail outs. As a result the European economy, contrary to the US where State aid provision do not exist, will exit from the crisis more efficient than it would have otherwise. The same on antitrust. The less ideological approach that with respect to the US the EC has developed in the field of exclusionary abuses will tend to maintain in the market efficient firms that otherwise, with a contracting economy, might leave the market. Active antitrust enforcement in Europe will help mitigate the effect of the crisis, while the hands off of the US may well exacerbate it. ■

III Lisbon conference on competition law and economics
PANNEL III
14-15 January 2010

Prof. Dr. Carl Baudenbacher
carl.baudenbacher@unisg.ch
President of the EFTA Court
Director of the Institute of European and International Economic Law at the University of St.Gallen HSG

Ass. iur. Frank Bremer
frank.bremer@unisg.ch
Lecturer and Senior Research Associate at the University of St.Gallen HSG

Competition Policy in times of crisis: Which enforcement practices best fit the principles?

Overcoming the financial crisis in the banking sector – The role of European Competition Policy

Abstract

The current financial crisis has sparked calls for the strengthening of regulatory oversight of financial institutions. However, financial regulation is ill-suited to deal with the immediate consequences of the crisis. Restoring the balance of the financial services market may require that the actors which were responsible for the emergence of the crisis be rescued. It is public intervention mainly in the form of State aid and nationalisations which then constitutes the last resort for financial institutions. However, States' measures which provide bridging finance to distressed financial institutions raise genuine competition law concerns in regard to compliance with State aid and merger rules.

La récente crise financière a suscité de nombreux appels pour le renforcement de la régulation des institutions financières. La régulation financière n'est toutefois pas adaptée pour répondre aux conséquences les plus immédiates de cette crise. Rétablir l'équilibre du marché des services financiers peut rendre nécessaire des mesures de sauvetage des acteurs à l'origine même de la crise. L'intervention publique sous la forme d'aides d'Etat et de nationalisations constituent le dernier recours pour les entreprises de ce secteur. Cependant, les relais de financement fournis par les Etats soulèvent des questions de concurrence au regard des règles en matière d'aides d'Etat et de concentrations.

I. Introductory remarks

1. The current financial crisis has sparked calls for the strengthening of regulatory oversight of financial institutions. However, financial regulation is ill-suited to deal with the immediate consequences of the crisis. This has also been acknowledged by the European Commission (the Commission) according to which "*under generalized financial crises, such as the recent crisis, there may be no alternative other than to use public funds to support the banking sector*".[1] Restoring the balance of the financial services market may require that the actors which were responsible for the emergence of the crisis be rescued. It is public intervention mainly in the form of State aid and nationalisations which then constitutes the last resort for financial institutions. However, States' measures which provide bridging finance to distressed financial institutions raise genuine competition law concerns in regard to compliance with State aid and merger rules.[2]

II. State Aid control

1. Legal framework prior to the crisis

2. Despite the fact that the Treaty on the Functioning of the European Union considers State aid in general to be an evil, some aid is more evil than other. It is the responsibility of the Commission to mitigate the evil of State aid by ensuring as far as possible a level playing field among the banks, in particular for those which did not receive any public funding. Otherwise, yesterday's rescues would encourage tomorrow's risk-taking,[3] or, to put it more bluntly, they would promote the survival not of the fittest but of the fattest banks.

3. Prior to the financial crisis, exemption of State aid for financial institutions from the general prohibition of Article 107(1) TFEU (ex Article 87(1) EC) was assessed under Article 107(3)(c) TFEU (ex Article 87(3)(c)). According to that provision, a derogation is allowed for "*aid to facilitate the development of certain economic activities or of certain economic areas, where such aid does not adversely affect trading conditions to an extent contrary to the common interest*". On that basis the Commission

1 Commission Staff Working Document, Impact Assessment Accompanying the Communication from the Commission on an EU Framework for Cross-Border Crisis Management in the Banking Sector, COM(2009)561 final, 20.10. 2009, Brussels, p. 8.

2 For the sake of completeness, it may be added that also the fundamental freedoms may be affected. See the letter of the EFTA Surveillance Authority of 4 December 2009 to Norton Rose LLP, London, <http://www.forsaetisraduneyti.is/media/frettir/Bradabirgdanidurstada_ESA.pdf> (last visited on 9.11.2010)

3 The Economist, Penance for their sins, 8.10.2009.
See <http://www.economist.com/businessfinance/displaystory.cfm?story_id=14587609> (last visited on 22.11.2009).

in 1994 adopted the Community Guidelines on State aid for rescuing and restructuring firms in difficulty which were lastly revised in 2004 (R&R Guidelines).[4] Subject to strict conditions, the R&R Guidelines allowed a firm in difficulty[5] to receive rescue aid, designed to keep it afloat for the time needed to set up a restructuring or a liquidation plan, on the one hand, and restructuring aid, designed to restore the firm's long term viability, on the other.[6] With regard to the substantive conditions under which aid may be granted, the rescue aid must be in the form of loan guarantees or loans granted at an interest rate comparable to those for loans to healthy firms, must be reimbursed within a period of not more than six months after disbursement of the first instalment, be warranted on the grounds of serious social difficulties, have no unduly adverse spill-over effects on other Member States, be accompanied, on notification, by an undertaking given by the Member State concerned to communicate to the Commission within six months a restructuring plan, a liquidation plan or proof that the loan has been reimbursed in full and/or that the guarantee has been terminated, be restricted to the amount needed to keep the firm in business for the period during which the aid is authorised, and respect the "one time, last time" principle.[7] As a general rule, state aid exceeding six months may only be authorised as restructuring aid. In order to be approved, a restructuring plan must be implemented that restores the firm's long-term viability within a reasonable time; compensatory measures must be taken to prevent or to minimise the risks of distortion of competition (divestiture of assets, a reduction in capacity or market presence or a reduction of entry barriers); the aid must be limited to the strict minimum; the beneficiaries must contribute to the restructuring plan from their own resources; and the Commission must be put in a position to make sure that the restructuring plan is being implemented properly, through regular reports communicated by the Member State concerned.[8]

2. Application in the R&R guidelines in the first phase of the crisis

4. In the first phase of the crisis, the Commission relied on the R&R Guidelines when assessing individual cases of state aid for financial institutions. This stage started in mid-September 2007 with the "subprime crisis" and the bank run on Northern Rock and lasted until the bankruptcy filing of Lehman Brothers in September 2008.[9]

4 Communication from the Commission - Community Guidelines on state aid for rescuing and restructuring firms in difficulty (OJ C 244, 1.10.2004, p. 2).

5 See R&R Guidelines, para. 9: «*[…] the Commission regards a firm as being in difficulty where it is unable, whether through its own resources or with the funds it is able to obtain from its owner/shareholders or creditors, to stem losses which, without outside intervention by the public authorities, will almost certainly condemn it to going out of business in the short or medium term*".

6 See R&R Guidelines, paras. 15 and 17.

7 See R&R Guidelines, paras. 25 et seq.

8 See R&R Guidelines, paras. 31 et seq.

9 Commission Decision of 5 December 2007, Case NN 70/2007, *United Kingdom Rescue Aid to Northern Rock*; Commission Decision of 30 April 2008, Case NN 25/2008, *WestLB risk shield, Germany*; Commission Decision of 4 June 2008, Case C 9/2008, *Sachsen LB, Germany*; Commission Decision of 31 July 2008, Case NN 36/2008, *Denmark – Roskilde Bank A.S.;* Commission Decision of 1 October 2008, Case NN 41/2008, *United Kingdom Rescue aid to Bradford & Bingley*; Commission Decision of 2 October 2008, Case NN 44/2008, *Germany – Hypo Real Estate Holding AG.*

3. Insufficiency of the R&R guidelines due to the systemic nature of the crisis

5. However, following the general collapse of confidence after the Lehman bankruptcy, doubts arose as to whether the R&R Guidelines still provided an appropriate framework. Now, even fundamentally sound financial institutions were forced to ask for State aid. The "one time last time" principle, too, became an issue of concern.[10] Moreover, the R&R guidelines provide that rescue aid in the banking sector granted in a form other than loan guarantees or loans cannot consist in structural financial measures related to the bank's own funds.[11] However, shielding financial institutions from the effects of volatile markets and asset values in the wake of the financial crisis often required measures of essentially structural character.[12] In order to instil confidence in the financial sector and to reassure depositors that they will not suffer losses, it became necessary to approve rescue guarantee schemes for a sufficient period of time.[13] Finally, decisions of the Commission on aid measures normally take several months. 6 months on average are needed for decisions based on a preliminary investigation of notified measures, and 20 months if the Commission opens a formal investigation.[14] Considering that many financial institutions came of bankruptcy within days or even hours, there was a need to speed up State aid control procedures. Lastly, and more generally, the R&R guidelines address the problem of how to deal with a single failing firm. The systemic nature of the crisis required a broader approach.

4. The new crisis framework in a nutshell

6. In view of the deepening of the crisis and the insufficient framework provided by the R&R Guidelines, the Commission recognized in the second phase the necessity to apply the special provision of Article 107(3)(b) TFEU (ex Article 87(3)(b) EC) which allows aid to be granted to address "*a serious disturbance in the economy of a Member State*".[15] Based on that provision, the Commission adopted four sets of guidelines between October 2008 and July 2009, namely the Banking Communication,[16] the

10 See R&R Guidelines, paras. 72-77.

11 See R&R Guidelines, para. 25(a), fn. 3. The problem became evident in *United Kingdom Rescue aid to Bradford & Bingley*, loc. cit. In this case, rescue aid was approved despite the fact that the aid measures consisted, inter alia, in the winding-up of the company and the selling of its retail deposit book. Unsurprisingly, the Commission avoided dealing with this problem when examining the form of the aid (see paras. 43-46). See also *Sachsen LB, Germany*, loc. cit., para. 65.

12 Under the R&R guidelines, structural emergency measures in the banking sector are therefore only admissible as «restructuring aid». However, it was almost impossible for States to accompany these measures with a restructuring plan meeting the conditions of paras. 32-51 of the R&R Guidelines. In practice, only few structural emergency measures would have met the various conditions for approval as «restructuring aid» by the Commission.

13 See R&R Guidelines, para. 25(a).

14 See MEMO/09/208, «State aid: Commission adopts Best Practices Code and Simplified Procedure to accelerate state aid decisions – frequently asked questions", 29.4.2009, Brussels.

15 Kroes, (6.10.2008),"Dealing with the Current Financial Crisis", Speech/08/498, Brussels.

16 Communication from the Commission - The application of State aid rules to measures taken in relation to financial institutions in the context of the current global financial crisis, 13.10.2008 (OJ C 270, 25.10.2008, p. 8).

Recapitalization Communication,[17] the Impaired Assets Communication,[18] and the Restructuring Communication[19] (the "Communications" or "Crisis Framework"). In addition to the guidelines directly addressing financial institutions, the Commission introduced the Temporary Framework for State aid measures to counteract the increasing difficulty of the *real economy* to obtain credit and other types of financial support.[20] As of 12 November 2009, the Commission has adopted 69 decisions under the new Crisis Framework with eleven cases under consideration.[21] Except for three decisions, where approval of the aid was subject to certain conditions,[22] the Commission decided not to raise objections. Finally, the EFTA Surveillance Authority (ESA) which has competence to apply the State aid rules of the European Economic Area (EEA) Agreement has adopted four sets of guidelines which are largely identical to the Crisis Framework of the Commission.[23] The EFTA States parties to the EEA Agreement are Iceland, Liechtenstein, and Norway. So far, ESA has issued four positive decisions in relation to State aid measures designed to counteract the effects of the financial crisis.[24]

4.1. The basic structure of the crisis framework

7. In the Banking Communication the Commission laid down the basic rules and conditions under which financial institutions could receive State aid.[25] One of the main purposes was to reassure bank depositors that they will not suffer losses, so as to limit the possibility of bank runs and undue negative spill-over effects on healthy firms. Furthermore, the Commission regarded state guarantees covering the liabilities of banks as necessary in order to restore confidence among financial institutions and reactivate interbank lending.[26] The Recapitalization Communication dealt in particular with capital injections into financial institutions designed to provide emergency support and thereby to prevent possible insolvencies.[27] In spite of the public guarantee schemes and the recapitalisation measures, the uncertainty regarding the quality of bank balance sheets remained and kept undermining confidence.[28] In response, the Commission adopted the Impaired Assets Communication in which it considered in particular the separation of impaired assets from good assets by transferring them to so called "bad banks" as separate legal entities, whose losses would ultimately be shared between the "good bank" and the State, and asset insurance schemes according to which banks retain impaired assets on their balance sheets but are indemnified against losses as potential asset relief measures.[29] Lastly, the Restructuring Communication was adopted which sets out the conditions for authorising restructuring aid in particular to minimise the distortive effects on competition and to ensure long-term viability without reliance on State support.[30]

4.2. The crisis framework – Answers to the regulatory need

8. Instead of restricting State aid to companies in difficulty, the Communications merely require Member States to show a "*serious disturbance in the economy of a Member State*" in accordance with Article 107(3)(b) TFEU. Although the Commission obligatorily emphasised the necessity of a restrictive interpretation,[31] Member States enjoy considerable leeway in granting aid to financial institutions. Other than that the entire functioning of financial markets must be jeopardized, the Communications provide no further direction as to what constitutes a "*serious disturbance in the economy*".[32] This condition is intrinsically vague, and whether it is justiciable in the present economic environment may be doubted. Tellingly, the Commission acknowledged that should national authorities declare to the Commission that there is a risk of such a serious disturbance, this shall be of particular relevance for the Commission's assessment.[33] In light of this, it becomes understandable why the Commission ultimately seized the chance to pronounce a fixed expiry date albeit only with regard to the application of the Restructuring Communication.

9. Second, compared with the R&R guidelines, the "one time last time" rule was relaxed. Departure from that rule is permitted under the Banking Notice. More specifically, the Restructuring Communication provides that "*additional*

17 Communication from the Commission - The recapitalisation of financial institutions in the current financial crisis: limitation of aid to the minimum necessary and safeguards against undue distortions of competition, 5.12.2008 (OJ C 10, 15.1.2009, p. 2).

18 Communication from the Commission on the Treatment of Impaired Assets in the Community Banking Sector, 25.2.2009 (OJ C 72, 26.3.2009, p. 1).

19 Communication from the Commission - The return to viability and the assessment of restructuring measures in the financial sector in the current crisis under the State aid rules, 23.07.2009 (OJ C 195, 19.8.2009, p. 9).

20 See Commission Communication - Temporary framework for State aid measures to support access to finance in the current financial and economic crisis, 17.12.2008 as amended on 25.2.2009 (OJ C 16, 22.1.2009, p. 1; consolidated version OJ C 83, 7.4.2009.

21 See MEMO/09/499, "State aid: Overview of national measures adopted as a response to the financial/economic crisis", 12.11. 2009, Brussels.

22 Commission Decision of 21.10.2008_on State aid measure C 10/08 (ex NN 7/08) implemented by Germany for the restructuring of IKB Deutsche Industriebank AG; Commission Decision of 12.5.2009 on the State aid No C 43/2008 (ex N 390/2008) implemented by Germany for the restructuring of WestLB AG; Commission Decision of 28.10.2009, C 14/2008, nyr.

23 Based on Article 61(3)(b) of the EEA Agreement which, like Article 107(3)(b) TFEU, allows aid "*to* remedy a serious disturbance in the economy *of an EC State or an EFTA State*" (emphasis added), the EEA Crisis Framework consists of ESA Decision 28/09/COL of 29.1.2009 - The application of state aid rules to measures taken in relation to financial institutions in the context of the current global financial crisis; ESA Decision 28/09/COL of 29.1.2009 - The recapitalisation of financial institutions in the current financial crisis: limitation of aid to the minimum necessary and safeguards against undue distortions of competition; ESA Decision 191/01/COL of 22.4.2009 - The treatment of impaired assets in the EEA banking sector, nyr.; ESA Decision 472/08/COL of 25.11.2009 - Return to viability and the assessment of restructuring measures in the financial sector in the current crisis under the state aid rules, nyr.

24 See <http://www.eftasurv.int/state-aid/state-aid-register/iceland/nr/1064>; <http://www.eftasurv.int/state-aid/state-aid-register/norway/nr/1080> (last visited on 9.11.2010). A comprehensive report on the impact of the financial crisis on Iceland, Liechtenstein and Norway is provided in the State Aid Scoreboard for 2008 for the European Economic Area EFTA States, Autumn 2009.

25 However, the Banking Communication also covered recapitalisation measures (paras. 34-42) and the controlled winding-up of financial institutions (paras. 43-50).

26 See Banking Communication, paras. 19-21.

27 See Recapitalisation Communication, paras. 4-6.

28 See Impaired Assets Communication, paras. 5-7.

29 See Impaired Assets Communication, Annex II, I.

30 See Restructuring Communication, para. 5.

31 See Banking Communication, para. 8 and the case-law cited.

32 See Banking Communication 11.

33 See Banking Communication, para. 9.

aid during the restructuring period should remain a possibility if justified by reasons of financial stability".[34] However, the Commission has not altogether abandoned the "one time last time" principle. It distinguishes between fundamentally sound financial institutions solely affected by the current restrictions on access to liquidity and beneficiaries that are additionally suffering from more structural solvency problems.[35] With respect to financial institutions whose difficulties were attributable to inefficiencies, poor asset-liability management or risky strategies in the first place, the Commission considers the application of the normal R&R Guidelines to be appropriate.[36]

10. Further, contrary to the R&R Guidelines, the Communications permit rescue aid in the banking sector to consist of structural measures, i.e. aid not temporary but irreversible in character. The Banking Communication expressly provides that the current circumstances may allow, *inter alia*, exceptional measures such as structural emergency interventions.[37] The most important structural measures envisaged are the recapitalisation of financial institutions,[38] aid schemes to relieve banks from their impaired assets,[39] as well as the controlled winding-up of financial institutions and the potential sale of its divisions to other companies.[40] The main simplification brought about is that structural measures may be introduced on short notice without an approved restructuring plan in place. However, the Commission has not relinquished its supervisory function. In view of the need to provide a tool-kit of urgent structural rescue measures while still safeguarding general respect for State aid rules, it merely abandoned its past *ex ante* based review system in favour of a more *ex post* oriented regime. Accordingly, structural emergency measures in support of the financial institution are to be followed up either by a report on the implementation of the measures, or a restructuring plan which will be separately assessed by the Commission.[41]

11. Also with regard to the permissible time limits for emergency rescue aid, a more flexible approach has been adopted. Unlike the R&R Guidelines, the Banking Communication accepts a period of up to two years with the possibility of further extension as long as the financial crisis requires so.[42] The extended temporal scope *mutatis mutandis* also applies to recapitalisation schemes,[43] whereas the Commission's approval for asset relief measures is granted only for a period of six months.[44]

34 See Restructuring Notice, paras. 7, indent 4 and 27.

35 See Banking Communication, paras. 14, 33, 35, indent 5.

36 See Banking Communication, para. 14.

37 See Banking Communication, para. 10.

38 See Banking Communication, paras. 35, indent 4, 42.

39 See Impaired Assets Communication, para. 49.

40 See Banking Communication, para. 43.

41 See Recapitalisation Communication paras. 40 et seq.; Impaired Assets Communication, Annex V.

42 See Banking Communication, para. 25(a).

43 See Banking Communication, para. 35, indent 2.

44 See Impaired Assets Communication, Annex V.

12. Finally, in recognition of the need for *ad hoc* decision making, the Communications introduce a fast track procedure for assessing and approving aid to financial institutions. The Commission pledges to adopt swift decisions upon complete notification of the aid measures, if necessary within 24 hours and over a weekend.[45] Even if one must not overlook that in most cases informal talks will be held before the notification is lodged, this self-commitment sent an important signal to Member States and to the markets that a definitive assessment of the conformity of aid measures with the Treaty will be provided within a short time.[46]

5. Key elements of the new crisis framework

5.1. Temporal scope

13. It is noteworthy that the Commission adopted the Communications only as a temporary framework. The Banking Communication stated that recourse to Article 107(3)(b) TFEU, on which all Communications are based, is possible only as long as the crisis situation justifies its application.[47] Subsequently, the Commission limited the applicability of the Restructuring Notice until the 31 December 2010, after which the normal rules of the R&R Guidelines adopted under Article 107(3)(c) TFEU will again become fully effective.[48]

5.2. Personal scope

14. Only financial institutions are eligible to receive State aid under the Communications. According to the Banking Communication, the use of Article 107 (3)(b) TFEU "*cannot be envisaged as a matter of principle in crisis situations in other individual sectors in the absence of a comparable risk that they have an immediate impact on the economy of a Member State as a whole*".[49] That being said, the Communications remain silent on what precisely constitutes a "financial institution". It would appear that at least credit and financial institutions within the meaning of the Banking Directive 2006/48 are covered.[50] Like the Treaty rules on State aid, the Crisis Framework applies to banks in private and public ownership.

15. The Communications apply to financial institutions irrespective of their size. At first glance, this could be questionable as the Banking Communication states that Article 107(3)(b) TFEU becomes only applicable if in the absence of state aid the entire functioning of financial

45 See Banking Communication, para. 53.

46 The Commission approved State aid measures already under the R&R Guidelines within 24 hours of their notification. See, e.g., Commission Decision of 1.10.2008, Case NN 41/2008, *United Kingdom Rescue aid to Bradford & Bingley*; Commission Decision of 2.10.2008, Case NN 44/2008, *Germany – Hypo Real Estate Holding AG*.

47 See Banking Communication, para. 12.

48 See Restructuring Communication, para. 49; Commission Press Release, IP/09/1180 «State aid: Commission presents guidelines on restructuring aid to banks», 23.6.2009, Brussels.

49 See Banking Communication, para. 11.

50 See, in particular, Article 4 of the Directive.

markets is jeopardized.[51] On a strict understanding, this requirement is satisfied only by big banks with the result that smaller banks would have to continue to rely on Article 107(3) (c) TFEU and the respective R&R Guidelines. However, in a volatile market environment even the exit of smaller banks can have severe destabilizing effects. Systemic importance can therefore also accrue to smaller banks. This is recognised by the Commission which expressly envisages general schemes of State aid for "*several or all financial institutions in a Member State*" to address the systemic risks.[52] A broad scope of application has also a political dimension: to avoid the situation of banks not only too big to fail, but also of banks too small to be rescued. In light of the role played by big banks in the development of the financial crisis, this would hardly be acceptable.

5.3. Substantive scope

16. The Communications are based on the principles underpinning the R&R Guidelines. They provide guidance as to how the Commission will apply its general principles in the specific context of the financial crisis.[53] It follows that the R&R Guidelines are only superseded to the extent that their provisions have been modified by the Communications.[54] As regards the relationship of the different Communications, they largely address different measures designed to stabilise the banking sector and therefore largely complement each other. Where they deal with the same subject matter, the latter communication can generally be considered to refine the previous communication in analogy to the principle of *lex posterior derogat legi priori*.[55]

17. To reiterate, although issued specifically in response to the financial crisis, the Communications are modelled on the general rules laid down in the R&R Guidelines. As the overarching principle, State aid must be in line with the principle of proportionality.[56] Accordingly, State support has to be, first, appropriate to fulfil the objective pursued, i.e., capable of keeping the financial institution afloat (rescue aid) and bringing it to long-term viability (restructuring aid) with the further aim of restoring stability in the financial markets; second, necessary to achieve the objective, i.e., limited to the strict minimum needed in time and amount to allow a financial institution to cope with the financial crisis for the purpose of reducing consequential distortions of competition; and, finally, proportionate *stricto sensu*, i.e., striking the right balance with regard to other important Community interests such as compliance with budgetary discipline and monetary stability.

The rather abstract proportionality principle translates into the following more specific requirements which are inherently interrelated with each other:

a. Burden sharing

18. The principle of burden-sharing is a direct response to the criticism that banks and their employees privatise profits but socialise losses. It stipulates that the costs of aid must be shared between the States, the banks and their capital holders. However, the principle of burden-sharing is not an innovation of the Crisis Framework. Already under the R&R Guidelines beneficiaries were expected to make a significant contribution to the restructuring from their own resources with contributions of at least 50 % in particular for large firms.[57] The Communications reaffirm the principle, but, at first sight somewhat surprisingly, significantly lower the required contribution level from the private sector. In the view of the Commission, it is "*not appropriate to fix thresholds concerning burden-sharing ex ante in the context of the current systemic crisis, having regard to the objective of facilitating access to private capital and a return to normal market conditions*".[58] Given the enormous size of the aid support required, there are good reasons for the more lenient policy of the Commission having in mind that the contribution may be increased to an appropriate level *ex post* if States have made a respective reservation. Thus, where the price for State aid has been initially significantly below the market price, States are requested to ensure that the terms of the financial support are revised in the restructuring plan.[59]

19. The principle of burden-sharing serves a multitude of purposes. It aims to ensure that the aid is limited to the minimum required, that banks carry adequate responsibility for their past behaviour, do not receive unjustified benefits at the taxpayer's expense, and, finally, that the markets believe in the long-term viability of the financial institution concerned. Member States therefore have to take appropriate steps to guarantee a significant contribution from the aid beneficiaries.[60]

b. Avoidance of undue distortions to competition

20. Financial institutions profiting from State aid have a considerable competitive advantage over non-aided banks. In this connection, Christine Lagarde, the French Minister of Finance, used the colourful description of "banks on steroids".[61] In order to maintain a level playing field and to prevent beneficiaries from abusing their preferential situation, undue distortions of competition must be minimized as far as possible.[62] In essence, financial institutions must be prevented from pursuing aggressive market strategies on

51 See Banking Communication para. 11.

52 See Banking Communication para. 9.

53 See MEMO/09/350, «State aid: Commission presents guidelines on restructuring aid to banks - frequently asked questions», 23.7.2009, Brussels with regard to the Restructuring Communication.

54 See Banking Communication, para. 10.

55 See Commission Press Release, IP/08/1901 «State aid: Commission adopts guidance on bank recapitalisation in current financial crisis to boost credit flows to real economy», 8.12.2008, Brussels with regard to the relationship between the Banking Communication and the subsequent Recapitalisation Communication which both deal with recapitalization measures.

56 See Banking Communication, para. 15; Recapitalisation Communication, para. 11; Impaired Assets Communication, para. 16; Restructuring Communication, paras. 30-33.

57 See R&R Guidelines, paras. 43-44.

58 See Restructuring Communication, para. 24.

59 See Restructuring Communication, para. 34.

60 See Banking Communication, paras. 25-26; Recapitalisation Communication, paras. 31-34; Impaired Assets Communication, paras. 21-25; Restructuring Communication, paras. 22-27.

61 See BBC News „'Banks on steroids' worry France"<http://news.bbc.co.uk/2/hi/8351766.stm> (last visited on 11.12.2009).

62 See Banking Communication, paras. 27-29; Recapitalisation Communication, paras. 35-39; Restructuring Communication, paras. 28-45.

the back of State support. Non-aided banks must not be punished because of aid given to other financial institutions. This is to be achieved mainly with the help of compensatory measures. To stay with the metaphor, although banks had to be provided with steroids, they must now be slowed down by attaching weights to their feet, delicately, without making them stumble. Compensatory measures may be either of behavioural or structural character. The more distortive the aid, the more comprehensive the compensatory measures expected from the financial institutions concerned.[63] Behavioural commitments may take the form of a price leadership ban, meaning that aid beneficiaries must not offer terms to customers which cannot be matched by non-aided competitors;[64] a prohibition on the use of State aid for the acquisition of competing businesses;[65] a ban on advertising the receipt of State aid when marketing financial offers;[66] or a restriction on the issuance of stock options for management which would make it more attractive for employees of its competitors.[67] Significantly, the Commission requires Member States to complement such behavioural limitations with provisions allowing enforcement of these behavioural constraints.[68] Likewise, the Communications envisage a variety of options with regard to structural remedies. Aid recipients may be required to divest subsidiaries, customer portfolios or business units. Furthermore, they may be prevented from expanding in certain business or geographical areas,[69] or be restricted in their expansion through a market share ceiling.[70] Also, structural relief may be ensured by placing limits on the size of the institution's balance sheet.[71] In the context of structural adjustments, the Commission is particularly anxious to avoid a retrenchment to national markets, in other words a negative impact on the Single Market resulting from businesses being split along national boundaries.[72] Finally, access to State aid must be guaranteed without discrimination between financial institutions from different Member States.[73] Thus, the Commission approved the Irish guarantee scheme only after the Government announced that it was also open to certain foreign credit institutions "*with a significant and broad based footprint in the domestic economy*".[74] Equally, where State aid is conditional upon meeting certain lending targets to the real economy, the Commission will view such plans more positively if the targets extend beyond the territory of the Member State granting the support.[75]

c. Long-time viability

21. State aid must not be allowed to become the lifeblood of financial institutions. The longer financial institutions rely on State support, the more they risk losing their competitive instincts. Respect for the interest of the markets and consumers becomes substituted for that of politicians and bureaucrats. Furthermore, the longer financial institutions rely on State support, the less non-benefiting competitors will be able to adhere to market behaviour on competitive terms. They will be strongly tempted to also seek government intervention. Finally, a generous provision of State aid may fuel the risk of a subsidy race between Member States. Therefore, Member States are required to ensure the return to long term viability of the aid beneficiary and the timely phasing out of the rescue schemes.[76] In particular, restructuring plans must set out a coherent concept demonstrating how the banks will achieve long-term viability. Long term viability means that the beneficiary is able to cover all its costs including depreciation and financial charges and provide an appropriate return on equity.[77] Rather than proposing one-size-fits-all solutions, the Restructuring Communication favours measures that are tailored to market characteristics. As a new requirement, banks receiving aid must undergo a stress test based on common parameters agreed at Community level assessing the long term viability also under worst-case assumptions.[78]

22. Member States must provide safeguards that encourage banks to end their reliance on State support as quickly as possible. As a general rule, the higher the amount of State aid, the more necessary it becomes for Member States to set out a clear exit mechanism.[79] To this end, Member States have to include in their reports and restructuring plans information on the "*path towards exit from reliance on State capital*".[80] As regards concrete ways ensuring a rapid exit, the Commission recommends in the Recapitalisation Communication that an add-on is generally applied to the entry price for recapitalisation measures as well as other built-in incentives such as step-up and redemption clauses.[81] Moreover, a restrictive dividend policy is suggested to ensure the temporary character of State intervention and to incentivise exit. With respect to guarantee schemes of an overly long duration, the Commission considers deterrent pricing conditions and appropriate quantitative limits on the debt covered to be appropriate.[82]

63 Kroes, (18.11.2009), "Commission outlines conditions for state aid to KBC, ING and Lloyds", Speech/09/541, Brussels.

64 See Restructuring Communication, para. 44.

65 See Restructuring Communication, paras. 40-41. The prohibition should apply for at least three years and derogation may only be made in exceptional circumstances and with prior authorisation from the Commission.

66 See Restructuring Communication, para. 44; See Recapitalisation Communication, para. 36.

67 See Banking Communication, para. 27, sub-indent 3.

68 See Banking Communication, para. 27, indent 2.

69 See Restructuring Communication, paras. 35 and 36.

70 See Banking Communication, para. 27, sub-indent 2. One possible option could be to prohibit large banks to combine high street retail banking with risky investment banking or funding strategies.

71 See Banking Communication, para. 27, sub-indent 3.

72 See Restructuring Communication, para. 36.

73 See Banking Communication, paras. 18, 35, indent 1.

74 Commission Decision of 13.10.2008, Case NN 48/2008, Ireland – Guarantee scheme for banks in Ireland, paras. 41, 47. Originally, the guarantee scheme was only applicable to six specified Irish banks.

75 See MEMO/09/350, "State aid: Commission presents guidelines on restructuring aid to banks - frequently asked questions", 23.7.2009, Brussels.

76 See Restructuring Communication, paras. 9-21; Impaired Assets Communication, paras. 48-59.

77 See Restructuring Communication, para. 7, indent 1.

78 See Restructuring Communication, para. 13.

79 See Recapitalisation Communication, para. 34.

80 See Recapitalisation Communication, para. 40(e); See also Restructuring Communication, Annex - model restructuring plan, 5.8.2.

81 See Recapitalisation Communication, paras. 31-34.

82 See Banking Communication, para. 24.

III. Merger control

23. The Commission has acknowledged that mergers and acquisitions also constitute an important instrument to consolidate the financial markets.[83] In particular, the sale of ailing banks to sound financial institutions represents a convenient avenue to ensure adequate burden-sharing on the part of the private sector. State aid and mergers are not mutually exclusive. Thus, the Commission opines that the sale of a bank may also involve State aid to the buyer and/or to the sold activity.[84] In this respect, the Commission has stressed that the State aid rules and the EC merger control must work in tandem. The relaxing of the merger control rules cannot be considered an alternative to State aid support.[85] As a matter of fact, "rescue mergers" create considerable difficulties of their own. Mergers between large financial institutions, so called mega-mergers, raise the risk of creating entities which are too big to fail and as such contribute to moral hazard. Furthermore, oversized banks are likely to have considerable negative repercussions on competition in the financial markets.

24. In contrast to its approach to State aid, the Commission opted against introducing a special financial crisis merger framework and instead continues to apply the existing merger rules. In the words of Commissioner Kroes, it is "*business as usual in merger control – for all our sakes*".[86] Thus, the Commission considers that, where a sale would *prima facie* result in a significant impediment of effective competition, it should not be allowed unless the distortions to competition are addressed by appropriate remedies accompanying the aid.[87] While being unreservedly committed to the existing merger rules, the Commission has vowed to take full account of the special economic environment caused by the crisis. First, the Commission emphasised that pursuant to Article 7(3) of the EC Merger Regulation, it may grant derogations from the normal standstill provisions pending a definitive outcome of the proceedings.[88] Consequently, provided there is urgency and there are no *a priori* competition concerns, immediate implementation of merger transactions as part of rescue operations is possible. Second, where applicable, the Commission is prepared to approve a merger on the basis of the "failing firm defence".[89] However, the requirements for the failing firm defence have not been relaxed. It is noted that under the EU Merger Regulation the Commission is not allowed to take account of public interest considerations other than competition policy. The legal situation is in marked contrast to that of some of the Member States. For example, in the U.K., a new public interest ground of preserving financial stability was introduced in connection with the Lloyds/HBOS merger.[90] Although the Office of Fair Trading (OFT) found significant *prima facie* competition concerns, the Secretary of State cleared the transaction claiming that "*the merger will result in significant benefits to the public interest as it relates to ensuring the stability of the UK financial system and that these benefits outweigh the potential for the merger to result in the anti-competitive outcomes identified by the OFT*".[91] Finally, nationalisations of financial institutions by Member States have become a distinctive feature of the current financial crisis. In accordance with Article 345 TFEU (ex Article 295 TEC), the Commission does not treat nationalisations fundamentally different from acquisitions of companies by private parties. In general, a nationalisation measure must respect EU competition rules including mandatory notification to the Commission under the Merger Regulation. However, no prior notification is required as long as, after the nationalisation, the financial institutions will make up an economic unit that retains independent power of decision. Where States hold controlling interests in more than one financial institution, it must be ascertained, in order to exclude an obligation to notify, that there is no room for coordination between different state-controlled banks. Finally, the acquired banks must be in a position to formulate their business strategy and carry out their day-to-day business, typically ensured by adopting budget and business plans on an autonomous basis.[92] On the Member State level, it is noteworthy that Germany decided to waive its national merger review procedures with regard to "rescue nationalisations" of financial institutions. The Financial Market and Stabilisation Act of October 2008 established the Financial-Market Stabilisation Fund. This Fund is tasked, *inter alia*, with acquiring financial interests in banks in distress. The Stabilisation Act specifically exempts the rescue fund from the application of the German Act against Restraints of Competition which contains the respective law on merger control.[93] However, the future acquisition of these interests from the fund by third parties would not escape the German merger law.

25. From a competition standpoint, nationalisations are largely seen as preferable to purely private mergers.[94] It is considerably less difficult to reverse nationalisation than to break up large private conglomerates. Moreover, as a result of their direct managerial influence, Member States are able to implement with less difficulty the measures required to restore the long-term viability of financial institutions.[95] However, there are obvious drawbacks to the involvement of Member States. One of the main objections is that the State is by nature a bad entrepreneur.

83 See Recapitalisation Communication, para. 37.

84 See Restructuring Communication, para. 20; State aid is of particular importance if the State arranges a merger between two ailing financial institutions.

85 Competition and Financial Markets 2009, OECD Policy Roundtables, DAF/COMP(2009)11, p. 237.

86 Kroes, (30.3.2009), «Competition, the crisis and the road to recovery», Speech/09/152, Toronto; Kroes, (12.3.2009), «Many achievements, more to do", Speech/09/106, Brussels.

87 See Restructuring Communication, para. 19.

88 Kroes, (6.10.2008), "Dealing with the Current Financial Crisis", Speech/08/498, Brussels.

89 Kroes, (11.9.2009), "Competition law in an economic crisis", Speech/09/385, Fiesole.

90 Enterprise Act 2002 (Specification of Additional Section 58 Consideration) Order 2008 (SI 2008/2645, 10.10.2008).

91 Decision of 31 October 2008 by Lord Mandelson, the Secretary of State for Business, not to refer the merger to the Competition Commission, para. 12. See <http://www.berr.gov.uk/files/file48745.pdf> (last visited on 14.12.2009). The decision was ultimately upheld on appeal by the Competition Appeal Tribunal (CAT): Merger Action Group v Secretary of State for Business, Enterprise and Regulatory Reform [2008] CAT 36, CAT Case 1107/4/10/08.

92 Note by the European Commission - Competition and Financial Markets, Roundtable 2 on Crisis: The Role of Competition Policy in Financial Sector Rescue and Restructuring, DAF/COMP/WD 2009 12/ADDI, paras. 26, 27.

93 According to Article 2 Section 17 of the Act, Parts I-III of the German Act against Restraints of Competition are not applicable. The exemption from German merger control is based on public interest considerations.

94 OECD, Competition and Financial Markets, Key Findings, 2009, p. 9. See < http://www.oecd.org/dataoecd/9/22/43067294.pdf> (last visited on 22.11.2009).

95 OECD, Competition and Financial Markets, ibid., p. 31.

IV. Concluding remarks

26. From the start, the Commission left no doubt that State aid is part of the solution and not of the problem. The approach to the application of the State aid rules was defined as firm on principle, but flexible on procedure.[96] Although it may still be too early to say for certain, it would appear that the Commission has delivered on the second part of this commitment. As regards the first part, the Commission's prime concern was to prevent State aid from becoming a means of protectionism on the part of Member States. Considering that almost all decisions so far taken were positive, one may question whether the Commission has been the staunch guardian of State aid rules it promised to be. In fact, one must ask whether it is possible at all to relax State aid procedures without also relaxing enforcement of its core principles. It appears to be rather a political gesture of good intent to assert that State aid decisions taken on an ad-hoc basis even within 24 hours will be as principled as decisions taken within the normal timescale which currently tends to be several months. However, this is not to say that the Crisis Framework of the Commission was an exercise in futility. First, in the present economic environment, the primary function of the Commission is (like that of the governments) to restore confidence in the banking sector. Second, it can be assumed that Member States are keen only to notify measures which are in line with the substantive State aid rules. Insofar, the Commission's approach has a preventive effect. Third, the Framework envisages constant monitoring of the State aid schemes and individual measures. One may even say that the New Crisis Framework and its application have accelerated a process which has started earlier: the conversion of European State aid control from being a tool of negative integration to becoming an instrument of positive integration[97]. The Commission has succeeded in imposing its own concept of what constitutes good State aid policy. And unlike in other fields of State aid control where a selection bias has been diagnosed[98], the Member States were forced to bring the cases by the crisis itself.

27. As to the substance of the Crisis Framework, the Commission has been very articulate about the need for a fundamental restructuring of the beneficiary financial institutions. Commissioner Kroes observed that while some banks might be too big to fail, they are not too big to be restructured.[99] Accordingly, in the cases of KBC, ING and Lloyds, the Commission approved restructuring plans only after substantial commitments, in particular divestment packages, had been secured with a view to limit distortions of competition.[100] However, despite its strong rhetoric, the Commission has as not gone as far as Mervyn King, the Governor of the Bank of England, who, observing that if banks are to be too big to fail, then they are too big, suggested that the size of banks should be generally limited so that each can be wound up without causing systemic risks.[101]

28. The Crisis Framework necessarily reflects the learning-by-doing approach of the Commission. In the following, three short points deserving further attention shall be put forward. First, the current Crisis Framework consists of four different Communications. Such fragmentation does not facilitate a straightforward application. Considering that the R&R Guidelines retain residual applicability, the purpose of providing the financial markets with transparency and predictability has not been fully met. Second, the relationship between the R&R Guidelines and the Crisis Framework is somewhat ambiguous. Third, it may be questioned whether it is a wise decision on the part of the Commission to limit the applicability of the Crisis Framework to the current financial crisis. Crisis is not only a constituent element of economic processes, but according to the model of competition as a discovery procedure[102] a necessary condition for the improvement of the system. Finally, whether the Merger Regulation contains a sufficiently flexible framework to deal with the challenges posed by the financial crisis is an open question. Many of the mergers in the financial sector did not have a Community dimension.[103]

29. All in all, the Commission must be credited with having taken a measured approach in dealing with what is perhaps the worst economic crisis since the Great Depression. That being said, important tasks still lie ahead. If the Crisis Framework is not to be regarded as a fig leaf for the legality of State action, it must be followed up with a robust *ex post* enforcement in the framework of monitoring the approved State aid measures. In the aftermath of the current financial crisis, in the absence of imminent systemic risks, a more rigorous assessment of State aid measures will be possible. It has been said that the Commission's capability to collect market information under the State aid rules is rather limited.[104] This may be true in general, but in the current crisis situation, it is different. One will not overlook in this context that the Commission has also crucial competentences in the field of business competition law. ■

96 Kroes, (17.2.2009), "The Road to Recovery", Speech/09/63, Paris.

97 See Michael Blauberger, From Negative to Positive Integration? European State Aid Control Through Soft and Hard Law, MPOfG Discussion Paper 08/4, p. 22 et seq. and passim < http://www.mpifg.de/pu/mpifg_dp/dp08-4.pdf > (last visited on 9.11.2010).

98 Hans W. Friederiszik, Lars-Hendrik Röller and Vincent Verouden, European State Aid Control: an economic framework, September 28th, 2006, p. 32. <http://www.esmt.org/fm/312/European_State_Aid_Control.pdf> (last visited on 9.11.2010).

99 Kroes, (11.9.2009), „Competition law in an economic crisis", Speech/09/385, Fiesole.

100 See MEMO/09/507, "State aid: Commission decision on KBC, ING and Lloyds - frequently asked questions", 18.11.2009, Brussels.

101 Mervyn King (17.6.2009), Speech at the Lord Mayor's Banquet for Bankers and Merchants of the City of London at the Mansion House, See <http://www.bankofengland.co.uk/publications/speeches/2009/speech394.pdf> (last visited on 16.12.2009).

102 See Friedrich A. Hayek, Competition as a Discovery Procedure, in New Studies in Philosophy, Politics, Economics, and the History of Ideas (1978, University of Chicago Press), p. 179-190.

103 So far, the Commission had to deal only with one rescue case under the merger review procedure. See Commission Decision of 14.05.2009, Case No COMP/M.5508, *SoFFin/Hypo Real Estate*. The majority of mergers taking place in the financial sector, even if triggered by the financial crisis, do not fall under the category of genuine "rescue mergers". See, e.g., Commission Decision of 17.12.2008, Case No COMP/M.5363, *Santander/Bradford & Bingley Assets*.

104 Hans W. Friederiszik, Lars-Hendrik Röller and Vincent Verouden, European State Aid Control: an economic framework, September 28th, 2006, p. 32. <http://www.esmt.org/fm/312/European_State_Aid_Control.pdf> (last visited on 9.11.2010).

Does patent protection restrict U.S. drug use? The impact of patent expiration on U.S. drug prices, marketing, and utilization

Frank R. Lichtenberg*
frl1@columbia.edu
Columbia University
National Bureau of Economic Research
Faculty Fellow, Lerner Center

Gautier Duflos*
gautier.duflos@univ-paris1.fr
University of Paris I
Paris School of Economics

Abstract

Microeconomic theory implies that the demand for prescription drugs should be inversely related to drug prices and directly related to marketing expenditure. Patent expiration is likely to reduce both the average price of a drug and marketing expenditure, so the effect of patent expiration on total utilization of a drug is theoretically indeterminate. We use longitudinal, molecule-level data on virtually all prescription drugs sold during the period 2000-2004 to analyze the impact of changes in market structure (primarily resulting from patent expiration) on U.S. drug prices, marketing, and utilization. Price and marketing expenditure both decline by about 50-60% in the years immediately following patent expiration, but the number of prescriptions remains essentially constant during those years. The two effects of generic entry on utilization – positive (via price), and negative (via marketing) – almost exactly offset one another, so the net effect of patent expiration on drug utilization is zero.

La théorie microéconomique prédit que la demande de médicaments devrait inversement dépendre du prix et directement des dépenses promotionnelles. L'expiration d'un brevet étant susceptible de réduire tant le prix du médicament associé que les dépenses promotionnels consenties pour le vendre, l'effet final sur le niveau d'utilisation est théoriquement indéterminé. Nous utilisons des données longitudinales, au niveau de la molécule, sur pratiquement l'ensemble des médicaments délivrés sur ordonnance vendus au cours de la période 2000-2004 aux États-Unis pour analyser l'impact des changements de structure de marché (résultant principalement de l'expiration de brevets) sur les prix, le marketing, et le niveau d'utilisation des médicaments. Il ressort que les prix, comme les dépenses promotionnelles, diminuent d'environ 50-60% les années suivant immédiatement l'expiration du brevet, mais le niveau des prescriptions demeure essentiellement constant pendant ces années. Les deux effets résultant de l'entrée des génériques sur le niveau d'utilisation –positif (par l'intermédiaire du prix), et négatif (par l'intermédiaire du marketing)– se compensent donc presque exactement l'un l'autre de sorte que l'effet net de l'expiration du brevet sur le niveau d'utilisation est nul.

* This research was supported by the Manhattan Institute for Policy Research.
Lerner Center for Pharmaceutical Management Studies at Rutgers University, and to its Director, Mahmud Hassan, for providing IMS Health data to us.

I. Background and objectives

1. U.S. patent law is based on Article I, Section 8 of the Constitution, which states that "the Congress shall have power to promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." The framers of the Constitution believed that, unless inventors were granted a monopoly on their discoveries, they would lack the incentive to pursue them. But this monopoly, the framers believed, should last for only a limited time, since inventions that enter the public domain are likely to be produced by more than one supplier, thereby benefiting the public by bringing down their price and increasing their availability.

2. The question of the socially optimal extent of intellectual-property protection has been hotly debated for more than a century. In particular, with the extension of patents to biotechnology products came a powerful reaction against the sweep of intellectual property rights, raising fears of the privatization of genetic inventions and the appropriation of the Southern Hemisphere's genetic resources by corporations based in rich countries. The reaction spread with the information-technology and Internet boom, which pitted supporters of freeware, file-sharing, and open architecture against the owners and defenders of proprietary products.

3. Determining the optimal level of patent protection involves weighing many factors. One of them is the extent to which patent protection restricts access to (utilization of) inventions: other things (e.g. the cost of developing new products) being equal, the lower the extent to which patent protection restricts access to inventions, the greater is the optimal strength of patent protection. This study will examine the impact of drug patent expiration on three variables: U.S. drug prices; the amount of marketing companies are willing to undertake;[1] and the quantity of drugs consumed. It does so by drawing on comprehensive data on virtually all drugs sold. Many studies have examined the effect of patent expiration and the ensuing entry of generics on drug prices, but we are aware of only two studies (Berndt, Kyle, and Ling (2002) and Lakdawalla et al. (2006)) that examined their effect on companies' marketing efforts and consumers' levels of utilization of previously patent-protected drugs. One of those studies examined data on just two drugs (cimetidine and ranitidine).

4. In general, increasing competition in a market, due to expiration of a patent or for other reasons, might be expected to reduce prices and thereby increase demand for and thus production of a good. However, this may not happen if (1) demand for the good is not very sensitive to price, perhaps because insurers or other third parties

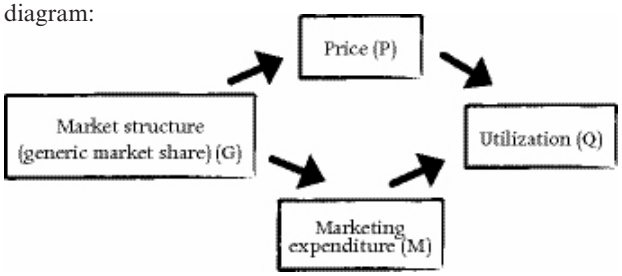
1 Due to data limitations, marketing expenditure will be defined as "cost of professional promotion": the total cost of advertising<but "contacts" aren't advertising> that is directed to the professional audience. It is the sum of three items: the cost of contacts (physician office or hospital calls, service visits, or telephone contacts); dollars spent in medical journals; and the retail value of samples. Direct-to-consumer (DTC) advertising is not included. However, DTC advertising accounts for a very small share of total pharmaceutical marketing expenditure.

usually pay for it, which is true of the U.S. prescription-drug market, where, in 2007, out-of-pocket payments by consumers accounted for only 20 percent of U.S. drug expenditure; and (2) demand is affected by factors other than price, such as marketing efforts on behalf of the product, which are extensive in the United States, and which a patent's expiration would be expected to diminish.[2]

5. In this paper, we will provide evidence on the extent to which patent protection, the loss of it, and various ancillary consequences may restrict or enhance access to, and thus use of, valuable, even life-saving drugs in the United States.

II. Theory

6. The theoretical framework is summarized by the following diagram:



As shown in the right-hand side of the diagram, we hypothesize that utilization of a drug (the total number of prescriptions dispensed) depends on two variables: the average price of the drug and marketing expenditure.[3] In particular, we hypothesize that utilization is inversely related to price and directly related to marketing expenditure.

As shown in the top left of the diagram, we hypothesize that mean price is inversely related to generic market share (generic prescriptions/total prescriptions). Also, as shown in the bottom left of the diagram, we hypothesize that marketing expenditure is inversely related to generic market share.

7. The hypothesis of a negative effect of generic market share on marketing expenditure is based on the following reasoning. Suppose that marketing expenditure has a positive effect on utilization, but that marketing is subject to diminishing marginal returns. We also assume that there are marketing spillovers, whereby the promotion of a branded pharmaceutical by a manufacturer affects the total number of prescriptions written for a range of products containing the underlying molecule and not just the number of prescriptions written for the marketer's own proprietary product.[4] The branded firm will increase marketing up to the point where the marginal private return is equal to the marginal cost of marketing. This implies that an increase in generic market share will reduce marketing expenditure.

8. This conceptual framework has several interesting implications. First, while conventional analysis implies that market structure affects utilization only via its effect on price, this framework implies that market structure affects utilization, conditional on price. Holding price constant, an increase in generics' market share will reduce utilization. Second, since increases in generic market share are hypothesized to reduce both price and marketing expenditure, and these variables are hypothesized to have opposite effects on utilization, the net effect of an increase in generics' market share on utilization is an empirical question.

III. Econometric approach

9. We will use longitudinal, molecule-level data on virtually all prescription drugs sold in the United States to investigate the effect of market structure on price and marketing activity and then the effects of these variables on utilization. We will conduct two types of analyses.

First, we will compute the age profiles of four variables: the fraction of prescriptions for a drug that were for generic products; the average price of these prescriptions; marketing expenditure on the drug, and the number of prescriptions dispensed, where age is defined as the number of years since the drug was first marketed.

Then we will estimate a prescription-drug demand equation, in which the quantity of drugs sold is a function of both price and marketing expenditure, using longitudinal molecule-level data. We will also estimate relationships between each of these variables (drug quantity, price, and marketing expenditure) and generic market share, also using longitudinal molecule-level data.

IV. Data

10. We obtained monthly data for the period 2000-2004 from IMS Health on virtually all prescription drugs sold in the United States. Our dataset contained data on the number of prescriptions, manufacturer-wholesaler revenue, and marketing expenditure (cost of professional promotion), by product and month, for over 19,000 products. In addition, the dataset revealed the following fixed product attributes: product name and manufacturer, active ingredient(s), date the product was first marketed, and product status (branded, generic, branded generic, other). We aggregated the product-level data to the molecule (or combination of molecules) level. We also computed average price (manufacturer-wholesaler revenue per prescription), generic market share, and molecule age, by molecule and month.

11. The dataset contains information on about 1560 molecules or combinations of molecules. A relatively small number of prescription drugs are also available over the counter (OTC),

2 CMS, National Health Expenditure Web Tables, http://www.cms.hhs.gov/NationalHealthExpendData/downloads/tables.pdf

3 Pindyck and Rubinfeld (2009, p. 424) hypothesize that the quantity of a firm's output demanded "depends on both its price and its advertising expenditure in dollars."

4 Marketing has been found to have spillover effects in a variety of industries. Vardanyan and Tremblay (2006) found significant marketing spillovers in the U.S. brewing industry, and Verbeek and Huij (2007) found that mutual funds with high marketing expenses enhance cash inflows to funds in other fund families with low marketing expenses.

i.e., without a doctor's prescription.[5] We determined from the FDA's Orange Book that 3.2 percent (50 out of 1560) of the molecules or combinations were available as OTC products; 7.3% of prescriptions issued from 2000 to 2004 were for drugs that were available over the counter. We do not have any information about utilization of OTC products, so we will exclude molecules that were available over the counter.[6]

Table 1 shows aggregate annual data on the number of prescriptions, manufacturer-wholesaler revenue, marketing expenditure, generics' market share, and average revenue per prescription. The top twenty-five molecules, ranked by total number of prescriptions issued in 2000-2004, are shown in Table 2. Monthly data on the market shares of six major generic drugs with the largest increases in market share during the period 2000-2004 are shown in Figure 1.

V. Empirical analysis

1. Estimation of age profiles of generics' market share, average price, advertising expenditure, and number of prescriptions

12. Estimates of the age profile of generics' market share are shown in Figure 2. Mean generic-market share is essentially zero in years zero (the year the drug was first launched) to six of a molecule's life-cycle. A modest amount of generics enter the market in the next six years; after twelve years, mean generic market share is 10 percent. Generics' market share increases sharply and suddenly after age twelve. By age sixteen, mean generic market share is 54 percent. This finding is quite consistent with the Congressional Budget Office's finding that the average period of marketing under patent protection since enactment of the Hatch-Waxman Act and the Uruguay Round Agreements Act of 1994 is about eleven-and-a half years.[7]

13. Estimates of the age profile of average price (manufacturer-wholesaler revenue per prescription dispensed) are shown in Figure 3. The average price increases about 44 percent (about 3.5 percent per year) from age 0 to age 12. Between age 12 and age 17, the price declines by 61 percent.

14. Estimates of the age profile of total cost of advertising directed to the professional audience are shown in Figure 4. Advertising expenditure rises fairly steadily during years 0-12, and is 2.3 times as high in year 12, when it reaches its peak, as it was in year one. It declines sharply after year 12. It is 20% lower one year after the peak and 60% lower four years after the peak. Berndt et al (2002) found that marketing efforts on four H2-antagonist prescription drugs declined prior to patent expiration. However, these age profiles suggest that the decline in marketing coincides with the increase in generics' market share.

15. Estimates of the age profile of the prescriptions dispensed by pharmacies are shown in Figure 5. The number of prescriptions increases rapidly during the first several years: it is about twice as great five years after launch as it was one year after launch. The number of prescriptions increases by 15% between year eight and year twelve, but remains constant between year 12 and 16, despite the sharp decline in average price shown in Figure 3. Both average price and the number of prescriptions during years 8-16 – the 4 years preceding and the 4 years experiencing the sharpest increase in competition from generics – are shown in Figure 6. These data indicate that increased utilization of prescriptions for generics after patent expiration is almost perfectly offset by reduced utilization of branded prescriptions.

16. The lack of a change in utilization in response to the sharp decline in price contrasts sharply with Lichtenberg and Sun's findings about the impact of Medicare Part D on prescription drug use by the elderly. As shown in Figure 7 (reproduced from their paper), they identified a sharp and immediate increase in prescription-drug use by the elderly when Medicare Part D reduced the cost of medications to them. The absence of any increase in the number of prescriptions during the period of rapidly increasing competition from generics may be due to the sharp decline in advertising shown in Figure 4.[8]

2. Estimation of prescription-drug demand function and other relationships

17. Now we will estimate a prescription-drug demand function, and analyze the impact of changes in competition due to the introduction of generics (which appears to be primarily attributable to patent expiration) on drug prices, marketing, and utilization, using longitudinal molecule-level data. As shown above in Figure 2, the largest increases in

5 In Canada, "the share of non-prescribed drugs in total drug expenditure is expected to have reached 16.7 percent in 2006 and 16.4 percent in 2007." Source: Canadian Institute for Health Information, *Drug Expenditure in Canada, 1985 to 2007* (Ottawa: CIHI, 2008). http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=download_form_e&cw_sku=DRUGEXP8507ENPDF&cw_ctt=1&cw_dform=N

6 A provision of the Waxman-Hatch Act of 1984 granted pioneer manufacturers an additional three years of limited market exclusivity, if they obtained FDA approval for a new presentation and indication for the chemical entity. As noted by Berndt et al (2002), by timing the OTC launch to coincide approximately with the pioneer Rx patent expiration date, a company could potentially benefit from an additional three years of market exclusivity on the OTC version of a drug, thereby offsetting somewhat its loss of Rx sales after the patent has expired. They note that, in theory, "the impact of a brand's OTC introduction on its own Rx sales…could be either positive or negative" (Berndt et al (2002, p. 251)).

7 *How Increased Competition from Generic Drugs Has Affected Prices and Returns in the Pharmaceutical Industry*, July 1998, http://www.cbo.gov/doc.cfm?index=655&type=0&sequence=5. The figure for the post-Hatch-Waxman period is based on the average effective patent term for the 51 drugs approved between 1992 and 1995 that received a Hatch-Waxman extension. The post-Hatch-Waxman figure is based in part on Henry Grabowski and John Vernon, "Longer Patents for Increased Generic Competition in the U.S.: The Hatch-Waxman Act After One Decade," *PharmacoEconomics* (1996).

8 In Figures 3 and 6, price is defined as manufacturer-wholesaler revenue per prescription, whereas Lichtenberg and Sun defined price as the average cost of a prescription to the patient. While the latter is the theoretically preferred measure, as discussed below there is a strong positive correlation across drugs between changes in prices charged by manufacturers and changes in prices paid by patients.

competition from generics usually occur 12-16 years after a drug is first introduced.[9] Therefore drugs introduced during the period 1984-1992 were likely to experience the largest increases in generic competition during the period covered by our IMS data, which was 2000-2004.

18. We estimated four regression equations. The first is a standard demand model, according to which quantity demanded depends on both the price of the good and marketing expenditure. We expect the effect of price on utilization to be negative and the effect of marketing expenditure on utilization to be positive. The model controls for any time-invariant, molecule-specific determinants of demand, and for time-varying factors that influence demand and do not vary across molecules. If the coefficient on price is negative, then the drugs whose prices increased faster than average during the period in question had slower than average growth in utilization, conditional on growth in marketing.

19. The second equation allows us to estimate the effect of changes in competition from generics on the average price (manufacturer-wholesaler revenue per prescription). The third equation allows us to estimate the effect of changes in competition from generics on marketing expenditure, and the fourth equation allows us to estimate the effect of changes in such competition on utilization. We hypothesize that this form of competition affects utilization primarily via its effects on price and marketing.

20. Estimates of the effects of price and marketing on utilization were consistent with our expectations: the price effect is negative and highly significant, and the advertising effect is positive and highly significant.[10] There was also a strong inverse correlation between changes in generics' market share and changes in average manufacturer-wholesaler revenue. The magnitude of this estimate is quite consistent with the age profiles of generics' market share and manufacturer price shown in Figures 2 and 3. Between years 12 and 16, generics' mean market share increases from 8% to 65%. The regression coefficient implies that this should result in a 49% price decline. The actual mean price decline between year 12 and year 16 is 44%.

21. There was also a strong inverse correlation between changes in generics' market share and changes in marketing expenditure. The regression coefficient estimate implies that the increase in generics' mean market share that occurs between years 12 and 16 should result in a 78% decline in marketing expenditure. The actual mean decline in marketing expenditure between year 12 and year 16 is somewhat smaller: 57%.

22. The estimates indicated that changes in generics' market share have no effect on the total number of prescriptions. This is consistent with the age profile of utilization shown in Figures 5 and 6. It is also consistent with the hypothesis that competition from generics does not have any effect on utilization independent of its effects on price and marketing.

VI. Free samples and spillover effects to other drugs within the same class

23. So far we have examined the effect of expiration of a drug's patent(s) on the number of prescriptions for that drug dispensed by pharmacies. But for two reasons, this may not reflect the overall effect of patent expiration on drug utilization. First, some medicines utilized by patients are not obtained from pharmacies: they are free samples obtained from physicians. Second, expiration of a drug's patent may have spillover effects, i.e., it may cause the amount of utilization of other drugs in the same therapeutic class to change ("therapeutic substitution"). Below we will attempt to assess how these two phenomena – free samples and therapeutic substitution – might cause the effect of expiration of a drug's patent on utilization of other drugs that treat the previously patent-protected drug's indication(s) to differ from its effect on the number of prescriptions for that drug dispensed by pharmacies.

1. Free samples

24. About 75% of professional promotional expenditure goes toward providing free samples (Narayanan and Manchanda (2006)). As shown above, on average professional promotion expenditure declines by 60% between years 12 and 16 – when competition from generics rises rapidly – and there is a strong negative correlation across molecules between changes in generics' market share and changes in professional promotion expenditure. This strongly suggests that patent expiration sharply reduces utilization of free samples obtained from physicians.

25. More direct evidence about the effect of patent expiration and competition from generics on utilization of free samples can be obtained from the 1996-2006 MEPS Prescribed Medicines files. MEPS household respondents were asked in each round whether they received any free samples of each reported prescribed medicine during the round. A MEPS variable indicates whether or not a respondent reported having received a free sample of the prescription medicine in the round.[11] We used these data to obtain estimates of the number of people who received free samples of each molecule in each year.

9 The difference between the patent life (usually 20 years) and the effective duration of a drug's market exclusivity is due to the time it takes to complete the clinical trials needed to obtain the FDA approval.

10 Although these parameters have the expected signs, the (absolute and relative) magnitudes of these coefficients are surprising in certain respects. In particular, the magnitude of the price coefficient is smaller than expected. This may be due, to an important extent, to "mismeasurement" of the price of drugs. Patients' demand for drugs presumably depends on the average price that they pay, not on average revenue received by manufacturers and wholesalers. Using data from the Medical Expenditure Panel Survey, we examined the correlation between the average cost of a prescription to patients and the total amount paid for a prescription. We found that there is a strong link between prices paid by patients and revenues received by manufacturers-wholesalers: drugs with above-average reductions in revenues received by manufacturers (e.g., due to patent expiration) tended to have above-average reductions in prices paid by patients.

11 However, respondents were not asked to report the *number* of free samples received, nor was it made clear that free samples were included in the count of the number of times that the respondent reported purchasing or otherwise obtaining the prescribed medicine during the round. Therefore, SAMPLE is not a count variable of free samples; SAMPLE =1 for all acquisitions of a prescribed medicine that a respondent reported getting a free sample of during the round. **http://www.meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h102a/h102adoc.shtml#2725TheSample**

26. Estimates of the number of people receiving free samples in years 0-20 relative to the number of people receiving free samples in year 12 are graphed in Figure 8. The figure also shows the molecule-age profile of professional promotion expenditure, reproduced from Figure 4.

27. The MEPS data indicate that the number of people receiving free samples of a drug increases fairly steadily from year 0 to year 10, when it reaches a peak. Between years 10 and 15, the number of people receiving free samples declines by 50%. The number of people receiving free samples appears to peak about two years before professional promotion expenditure does. However, the age profiles of the two variables are broadly consistent. Both decline sharply during the period in which generics' market share rapidly increases. The effect of patent expiration on the total number of prescriptions for a drug (prescriptions dispensed by pharmacies plus free samples) is therefore lower (more negative) than its effect on the number of prescriptions dispensed by pharmacies. We estimate that, overall, the ratio of the market value of free samples to the sum of the market values of free samples and pharmacy prescriptions in 2003 was 7%. If patent expiration had no effect on the number of pharmacy prescriptions (as suggested by Figures 5 and 6), and reduced the number of people receiving free samples by 50%, it would reduce the total number of prescriptions by 3.5% (= 7% * 50%).

2. Between-drug spillover effects

28. Expiration of a drug's patent may have spillover effects, i.e., it may cause the extent of utilization of other drugs in the same therapeutic class to change. The estimates described above do not account for these potential spillovers. In this section, we will first argue that these spillover effects can go in both directions. Therefore failure to account for spillovers could result in either understatement or overstatement of the effect of patent expiration on drug utilization. Then we will present estimates of a model that accounts for potential spillovers.

29. Positive spillovers. Monopolists may have little incentive to research and develop new products that will compete directly with their currently marketed products. Consequently, "generic entry can […] have a small positive effect on the incentive to innovate."[12] Graham and Higgins (2006) find that "pharmaceutical firms act strategically, targeting the three-year window around the loss of exclusivity to introduce new products." Schering-Plough launched the antihistamine Clarinex shortly before the patent on its older drug Claritin (loratadine) expired (Rubin (2002)). The change in total utilization of antihistamines is presumably much larger than the change in loratadine sales.

30. Negative spillovers. Merck's cholesterol-lowering drug Zocor (simvastatin) lost its U.S. patent protection in June 2006, becoming the largest-selling drug yet to be opened to competition from cheap generics. That change cost Merck billions of dollars a year. But it may have been nearly as damaging to Pfizer, whose rival cholesterol drug, Lipitor, was the world's most popular, with global sales last year of $12 billion. After the patent expired, insurers hoped to convince patients and doctors that cheap clones of Zocor made full-priced Lipitor an unnecessary luxury (Berenson (2006)). The change in total utilization of cholesterol-lowering drugs is presumably much smaller than the change in simvastatin sales.

31. To examine the effect of changes in a drug's market structure on utilization of all drugs in the same therapeutic class (i.e., accounting for potential spillovers), we estimated the relationship between utilization and generics' market share at the level of the therapeutic class as opposed to the molecule level.

32. We used the Anatomical Therapeutic Chemical (ATC) Classification System[13] to aggregate molecules into therapeutic classes. The ATC system is controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology and was first published in 1976. The system divides drugs into different groups according to the organ or system on which they act and/or their therapeutic and chemical characteristics. In this system, drugs are classified into groups at five different levels. There are 14 main groups. The first level of the code indicates the anatomical main group and consists of one letter (Example: C Cardiovascular system). The second level of the code indicates the therapeutic main group and consists of two digits (Example: C03 Diuretics). The third level of the code indicates the therapeutic/pharmacological subgroup and consists of one letter. (Example: C03C High-ceiling diuretics). The fourth level of the code indicates the chemical/therapeutic/pharmacological subgroup and consists of one letter (Example: C03CA Sulfonamides). The fifth level of the code indicates the chemical substance and consists of two digits (Example: C03CA01 Furosemide).

33. We estimated the relationship between utilization and generic market share at both the fourth and third ATC levels. Molecules in the same fourth-level class are likely to be better substitutes than molecules that are in the same third-level class but not the fourth.

34. There was not a significant relationship between changes in utilization and changes in generics' market share at either the fourth ATC level or the third level. This finding indicates that the increases in generics' market penetration do not affect drug utilization, whether or not potential spillovers to other drugs in the same therapeutic class are taken into account.

VII. Summary

35. In general, increasing competition in a market, due to expiration of a patent or for other reasons, might be expected to reduce price and thus to increase demand for a good and thus its total production and consumption.

12 Tirole (1988, p. 392), quoting Kenneth J. Arrow, and Congressional Budget Office (1998, Appendix D).

13 http://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System

However, this need not be the case if the demand for the good is sensitive to factors other than price (e.g., marketing), and if patent expiration has an important impact on these other factors. This study examined the impact on U.S. drug prices, marketing, and utilization of changes in market structure (changes in generic drugs' market share) primarily resulting from patent expiration, using comprehensive data on virtually all prescription drugs sold during the period 2000-2004. We excluded a small number of molecules that were available over the counter because we do not have any information about utilization of such products.

36. We hypothesized that utilization is inversely related to price and directly related to marketing expenditure. Due to marketing spillovers, whereby the promotion of a drug by a manufacturer increases the total number of prescriptions for that drug and not just those of the marketer, the advent of price competition from generics following patent expiration reduces the incentive to maintain marketing expenditures at their former levels. Because a decline in marketing expenditure produces a decline in demand, just as a decline in price increases demand, the net effect of increased competition from generics on utilization is indeterminate, a priori.

37. We conducted two types of analyses. First, we computed the age profiles of generic market share, average price, marketing expenditure, and number of prescriptions, where age was defined as the number of years since the drug was first marketed. We found that there is little competition from generics in the first 12 years of the product life cycle, but that generic market share increases sharply and suddenly in the next four years. This is quite consistent with previous evidence that the average period of marketing under patent protection after enactment of the Hatch-Waxman Act and the Uruguay Round Agreements Act of 1994 is about 11.5 years. Price and marketing expenditure both decline by about 50-60% during years 12-16, but the number of prescriptions remains essentially constant during those years. This finding implies that the effect on utilization of declining price is approximately offset by the effect of declining marketing, and that increased utilization of generic prescriptions after patent expiration is approximately offset by reduced utilization of branded prescriptions.

38. We also obtained estimates of a prescription-drug demand function – the relationship between changes in utilization and changes in average price and marketing – and of models of the effect of generics' market share on price, marketing, and utilization, using longitudinal molecule-level data. Consistent with our expectations, the effect of price on demand was negative and highly significant, and the effect of advertising on demand was positive and highly significant. The estimated effect of price appeared low; this may be due, to an important extent, to "mismeasurement" of the price of drugs. Patients' demand for drugs presumably depends on the average price that they pay, not on average revenue received by manufacturers and wholesalers. Using data from the Medical Expenditure Panel Survey, we showed that the change in the average price paid by patients is correlated across drugs with the change in the average proceeds received by manufacturers, but it is not perfectly correlated.

39. We found a strong inverse relationship between changes in generics' market share and changes in average manufacturer-wholesaler revenue. The slope of the estimated relationship was quite consistent with the age profiles of generics' market share and manufacturer price. There is also a strong inverse correlation between changes in generics' market share and changes in marketing expenditure.

40. We found no evidence of a relationship across molecules between changes in the total number of prescriptions and changes in generics' market share. The two hypothesized effects of increased competition from generics – increased utilization due to falling prices, and decreased utilization due to reduced marketing – appear approximately to offset one another. Competition from generics does not appear to have any effect on utiliz ation independent of its effects on price and marketing.

41. Even if expiration of a drug's patent(s) does not affect the number of (branded + generic) prescriptions for that drug dispensed by pharmacies, it could still affect drug utilization, for two reasons. First, it could affect the number of free drug samples patients obtain from physicians. We found that the number of free samples declined sharply after patent expiration, and therefore that the effect of patent expiration on the total number of prescriptions for a drug (prescriptions dispensed by pharmacies plus free samples) is lower (more negative) than its effect on the number of prescriptions dispensed by pharmacies. We estimated that if patent expiration had no effect on the number of pharmacy prescriptions, it would reduce the total number of prescriptions by 3.5%.

42. Second, expiration of a drug's patent may have spillover effects, i.e., it may cause utilization levels of other drugs in the same therapeutic class to change. These spillover effects can go in both directions. We attempted to account for potential spillovers by estimating the relationship between changes in utilization and changes in generics' market share at the level of the therapeutic class rather than the molecule level. We did not find a statistically significant relationship. Increases in generics' market penetration do not appear to affect levels of drug utilization, whether or not potential spillovers to other drugs in the same therapeutic class are taken into account.

43. Improving public health depends on both the creation and use of new medical goods and services, such as new drugs. As we discussed earlier, there is a continuing debate over the optimal length and breadth of patents, including whether patents – particularly in the health-care context – limit utilization of important medical products. Our findings suggest that, at least in the United States, patent expiration (and the consequent large declines in price) does not significantly increase utilization. Although patent expiration causes a large decline in price, high levels of prescription-drug insurance coverage prevent this price decline from stimulating consumer demand as much as a price decline by itself would otherwise. Moreover, patent expiration causes a sharp reduction in marketing activity, which reduces demand.

44. Concerns have been expressed regarding the role of industry marketing to physicians (physician detailing) and direct-to-consumer advertising. While this study does not address any claims as to the medical appropriateness of such activity (which is, at a minimum, regulated by the U.S. Food and Drug Administration), we do note that marketing has a significant impact on utilization.

45. Questions surrounding patents, marketing, and access are at the forefront of policy debates at both the state and federal level. In the past decade, the U.S. Supreme Court has issued several decisions that have weakened patent protection. In 1999, the Court granted states immunity from claims of patent infringement (Chartrand (1999)). In 2007, the Court, in its most important patent ruling in years, raised the bar for obtaining patents on new products that combine elements of pre-existing inventions (Greenhouse (2007)). As a result, judges now have more leeway to dismiss lawsuits for patent infringement without requiring a jury trial, and patent examiners, who generally grant patent applications unless they find prior references to the same invention, now feel freer to deny claims. These decisions have not reduced patent length, but they have reduced the value of patent protection, and in the long run, weaker patent protection, like shorter patent protection, is likely to reduce the amount of medical innovation – the rate at which novel medical goods and services are created.

46. In principle, the adverse effect of less innovation on public health could be offset by greater access to existing products. However, our findings imply that, in practice, weaker (or shorter) patent protection would not increase Americans' access to prescription drugs, all of which have been synthesized and marketed under a regime affording greater patent protection than some are now proposing. Due to broad prescription-drug insurance coverage and the role of marketing in increasing awareness of both the efficacy and availability of pharmaceuticals, weaker patent protection would not increase utilization of prescription drugs. ■

Annexe

Table 1

Summary statistics

Year	*Total number of prescriptions (000s)*	*Manufacturer-wholesaler revenue (000s)*	*Manufacturer-wholesaler revenue per prescription*	*Generic market share*	*Professional promotion expenditure (000s)*
2000	*2 813 203*	*$129 565 642*	*$46,06*	*37%*	*$12 583 737*
2001	*2 981 866*	*$154 087 916*	*$51,67*	*37%*	*$15 085 286*
2002	*3 146 565*	*$176 087 414*	*$55,96*	*39%*	*$17 412 398*
2003	*3 288 211*	*$202 513 267*	*$61,59*	*41%*	*$20 211 506*
2004	*3 380 304*	*$221 994 992*	*$65,67*	*44%*	*$22 955 232*

Table 2

Top 25 molecules, ranked by total number of prescriptions during 2000-2004

Molecule	*Number of prescriptions during 2000-2004 (000s)*	*Year first marketed*
ACETAMINOPHEN/HYDROCODONE	*433 947*	*1978*
LEVOTHYROXINE	*396 930*	*1963*
ATORVASTATIN	*316 240*	*1997*
AMOXICILLIN	*293 264*	*1974*
ALBUTEROL	*238 338*	*1981*
METOPROLOL	*226 809*	*1978*
ATENOLOL	*220 880*	*1981*
FUROSEMIDE	*215 518*	*1966*
LISINOPRIL	*210 945*	*1987*
ESTROGENIC SUB, CONJUGATED	*186 319*	*1942*
AZITHROMYCIN	*180 271*	*1992*
AMLODIPINE	*175 413*	*1992*
HYDROCHLOROTHIAZIDE	*169 460*	*1959*
METFORMIN	*161 739*	*1995*
ALPRAZOLAM	*161 066*	*1981*
SERTRALINE	*150 556*	*1992*
ACETAMINOPHEN/PROPOXYPHENE	*139 941*	*1975*
PAROXETINE	*137 818*	*1993*
WARFARIN	*137 579*	*1954*
SIMVASTATIN	*135 362*	*1992*
LANSOPRAZOLE	*135 349*	*1995*
HYDROCHLOROTHIAZIDE/TRIAMTERENE	*134 846*	*1968*
FLUOXETINE	*127 738*	*1988*
CELECOXIB	*125 514*	*1999*
CEPHALEXIN	*122 546*	*1975*

Figure 1
Generic market shares of six major drugs with the largest increases in generic market share during the period 2000-2004

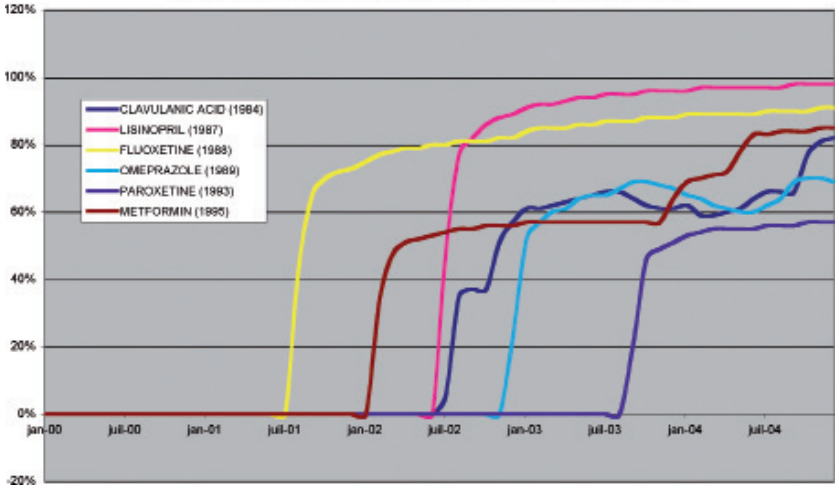


Figure 2
Mean generic market share, by age of

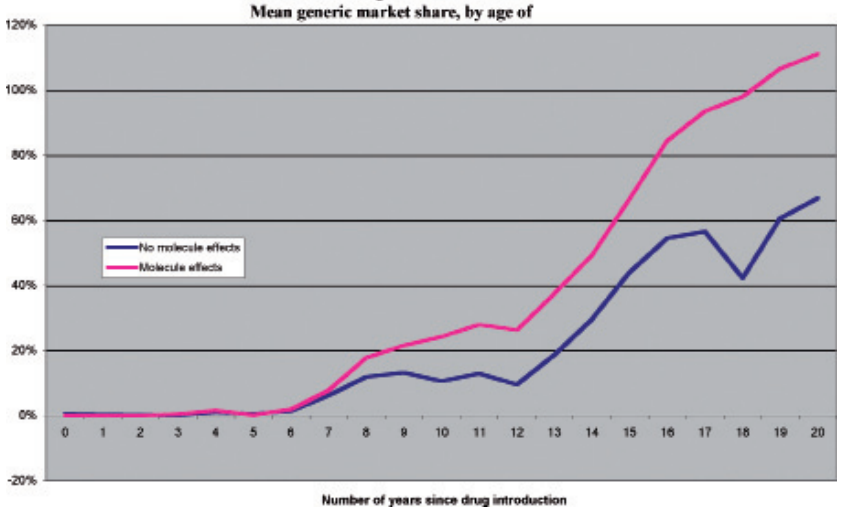
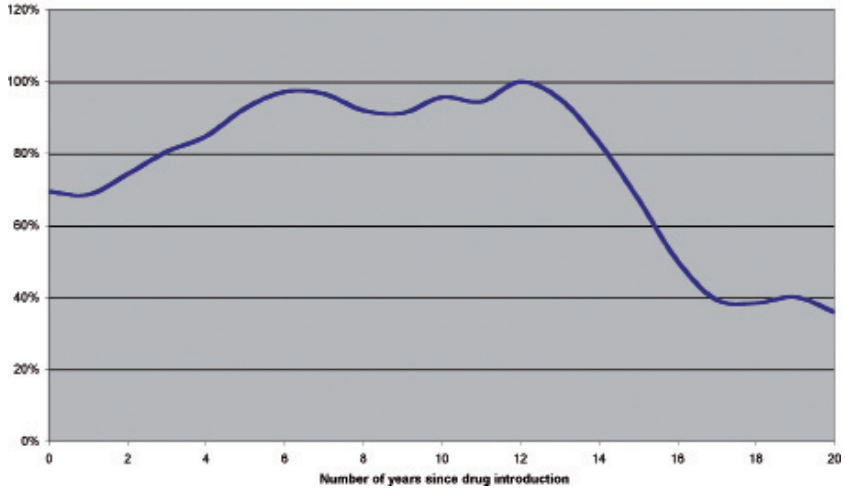


Figure 3
Mean drug price, relative to price in year 12



References

Bagwell, Kyle (2007), "The Economic Analysis of Advertising," Chapter 28 in Handbook of Industrial Organization, vol. 3, ed. by Mark Armstrong and Robert H. Porter, Elsevier.

Berenson, Alex (2006), "Merck Loses Protection for Patent on Zocor," New York Times, June 23, http://www.nytimes.com/2006/06/23/business/23statin.html.

Berndt, Ernst R., Margaret K. Kyle, and Davina C. Ling (2002), "The Long Shadow of Patent Expiration: Generic Entry and Rx-to-OTC Switches," in Feenstra, Robert C. and Matthew D. Shapiro, editors, Scanner Data and Price Indexes (University of Chicago Press for the National Bureau of Economic Research Studies in Income and Wealth), http://www.nber.org/chapters/c9737.pdf.

Chartrand, Sabra (1999), "Patents; A Supreme Court decision leaves many innovators disgusted, angry and outraged," New York Times, June 28, http://www.nytimes.com/1999/06/28/business/patents-supreme-court-decision-leaves-many-innovators-disgusted-angry-outraged.html?scp=4&sq=patents%20supreme%20court&st=cse.

CMS, National Health Expenditure Web Tables, http://www.cms.hhs.gov/NationalHealthExpendData/downloads/tables.pdf.

Congressional Budget Office (1998), "How Increased Competition from Generic Drugs Has Affected Prices and Returns in the Pharmaceutical Industry," July, http://www.cbo.gov/showdoc.cfm?index=655&sequence=0.

Gilbert, Richard, and Carl Shapiro (1990), "Optimal Patent Length and Breadth," RAND Journal of Economics 21(1), Spring, 106-112.

Graham, Stuart J.H., and Matthew J. Higgins (2006), "The Impact of Patenting on New Product Development in the Pharmaceutical Industry," working paper, Georgia Institute of Technology, http://mgt.gatech.edu/news_room/news/2006/reer/files/reer_impact_of_patenting.pdf.

Greenhouse, Linda (2007), "High Court Puts Limits on Patents," New York Times, May 1, http://www.nytimes.com/2007/05/01/business/01bizcourt.html?scp=6&sq=patents%20supreme%20court&st=cse.

Lakdawalla, Darius, Philipson, Thomas and Y. Richard Wang, "Intellectual Property and Marketing", NBER working paper 12557, http://www.nber.org/papers/w12577.

Lévêque, François, and Yann Ménière (2004), The Economics of Patents and Copyright (Berkeley Electronic Press), http://www.bepress.com/cgi/viewcontent.cgi?article=1001&context=leveque.

Lichtenberg, Frank, and Shawn Sun (2007), "The impact of Medicare Part D on prescription drug use by the elderly: evidence from a large retail pharmacy chain," Health Affairs 26(6), November/December 2007, 1735-44.

Narayanan, Sridhar and Puneet Manchanda (2006), "The Role of Free Samples in the Pharmaceutical Industry: An Empirical Analysis," 2006 Informs Marketing Science Conference, University of Pittsburgh. http://www.cci.som.yale.edu/events/documents/Presentation-YCCI-SridharNarayanan.ppt.

National Science Foundation, U.S. Corporate R&D: Volume 1: Top 500 Firms in R&D by Industry Category, http://www.nsf.gov/statistics/nsf00301/expendit.htm.

Pauly, Mark (2004), "Medicare Drug Coverage and Moral Hazard," Health Affairs 23, no. 1: 113-22.

Pindyck, Robert, and Daniel Rubinfeld (2009), Microeconomics, seventh edition (Upper Saddle River, NJ: Pearson Prentice Hall).

Rubin, Rita (2002), "Claritin to go OTC next spring, Clarinex to replace it," USA Today, April 22, http://www.usatoday.com/news/health/drugs/2002-04-23-claritin.htm.

Tirole, Jean (1988), The Theory of Industrial Organization (Cambridge: MIT Press).

Vardanyan, Michael, and Victor J. Tremblay (2006), "The measurement of marketing efficiency in the presence of spillovers: theory and evidence," Managerial and Decision Economics 27 (5), 319-331.

Verbeek, Marno and Joop Huij (2007), "Spillover Effects of Marketing in Mutual Fund Families," ERIM Report Series; EFA 2007 Ljubljana Meetings Paper, February 26. Available at SSRN: http://ssrn.com/abstract=958784.

Wagner, Todd., Heisler, Michele. and Piette, John (2006), "Tiered Co-payments and Cost-Related Medication Underuse?", Paper presented at the annual meeting of the Economics of Population Health: Inaugural Conference of the American Society of Health Economists, TBA, Madison, WI, USA, Jun 04, http://www.allacademic.com/meta/p92398_index.html.

The rising tide of patent damages

Richard GILBERT
gilbert@econ.berkeley.edu

Emeritus Professor of Economics and Professor of the Graduate School at the University of California, Berkeley
Senior Consultant with Compass Lexecon

Abstract

Very large awards and settlements for patent infringement have increased dramatically since the 1980s. A large fraction of these awards have occurred in the computer hardware and software industries. Complex technologies such as computer hardware and software require rights to a very large number of patents. One explanation for the large awards for patent infringement is the bargaining power of a patentee that has a credible injunction threat for a product that requires rights to multiple patents. This can lead to infringement damage awards and settlements that overestimate the patent's contribution to product value.

Les coûts financiers liés à la violation des brevets ont augmenté de façon spectaculaire depuis les années 80. Un grand nombre des dommages et intérêts prononcés dans cette matière concerne le secteur de l'électronique et des logiciels dont la spécificité réside dans dans l'existence d'une pluralité de brevets pour un même produit. L'importance des dommages et intérêts prononcés en matière de violation de brevets tient dans le pouvoir de négociation d'un détenteur d'un brevet disposant d'une menace suffisamment crédible à l'encontre d'un produit couvert par une pluralité de brevets. De telles pratiques peuvent conduire à des montants de dommages et intérêts et de transaction surévaluant la contribution de ce brevet.

1. Debates over the patent system in the United States have often generated extreme positions. Some argue that the patent system is broken beyond repair and must be abandoned. Others say that the patent system is so fundamental to the performance of the economy that any attempt to modify it would undermine technological progress.

Neither position accurately describes the state of the U.S. patent system. The patent system is integral to the economy, but is need of reform, particularly to address the way that patents impact some industry sectors. Signals of the need for reform include a rising trend in very large damage awards and settlements for patent infringement along with evidence that the calculations of infringement damages are prone to error when an infringed patent is only one component of a product's value.

I. Trends in large awards and settlements for patent infringement

2. The number of awards and settlements for infringement of U.S. patents that exceed $100 million in year 2000 dollars has been rising rapidly over the past several decades. Before 1980, awards or settlements for patent infringement rarely exceeded $100 million in inflation adjusted dollars and they were infrequent throughout the decade of the 1980s.[1] The number of large patent damage awards or settlements increased in the 1990s. On average, there were about three awards or settlements each year exceeding $100 million during that decade. Large patent damage awards and settlements exploded after the turn of the century. From 2000 to 2007, infringement awards or damages larger than $100 million averaged about eight per year.[2]

3. The increase in the number of very large awards and settlements for patent infringement suggests that there has been a shift in the monetization of patent rights. This trend alone does not imply that the patent system is broken if the increase in awards and settlements coincides with a more significant role for patent rights in providing incentives for innovation. However, that does not appear to be the case, at least in some industry sectors characterized by products covered by multiple patent rights ("complex technologies").

4. An alternative explanation for the increase in very large awards and settlements for patent infringement is that judges and juries have become more accustomed to awarding very large damages, perhaps for similar reasons that have created an increasing trend in large damage awards in other types of litigation. With regard to patent litigation, many scholarly articles have made the case that the creation of the Court of Appeals for the Federal Circuit in 1982 coincided with an appellate climate that has been much more favorable to patent owners and promoted large damage awards for patent infringement. These factors alone do not suggest that large damage awards and settlements are improper. However, they are troubling if patents are not a significant determination of innovative effort for the economy.

5. Very large patent damage awards and settlements overwhelmingly occur in two broadly defined industry categories: (1) computers, including hardware and software and (2) medical, including pharmaceuticals, biotech and medical equipment. These two industry categories account for more than seventy percent of all awards

* The author is grateful to Michael Katz, Jon Orszag and Carl Shapiro for helpful discussions.

1 All awards and settlement numbers are normalized to the producer price level in 2000.

2 These numbers are calculated from actual awards and settlements collected from publicly available data. While they may include some compensation that is not strictly related to intellectual property, they understate the total to the extent that some awards and settlements are not publicly disclosed.

and settlements for patent infringement in excess of $100 million (in year 2000 dollars). Including the related field of telecommunications increases the share of these very large awards and settlements to more than 75 percent.

6. Awards that go to non-practicing entities (NPEs), defined as patentees that do not practice the technology covered by the patent, figure prominently in two industries – computer hardware and biotechnology (Figure 1). These two industries represent about 30 percent of total large awards for patent infringement, but over 70 percent of large awards to non-practicing entities. Including telecommunications, the corresponding figures are 35 percent of all payments and 80 percent of all payments to NPEs. In the computer hardware industry, NPEs were the recipients of more than half of all payments for patent infringement exceeding $100 million in year 2000 dollars.

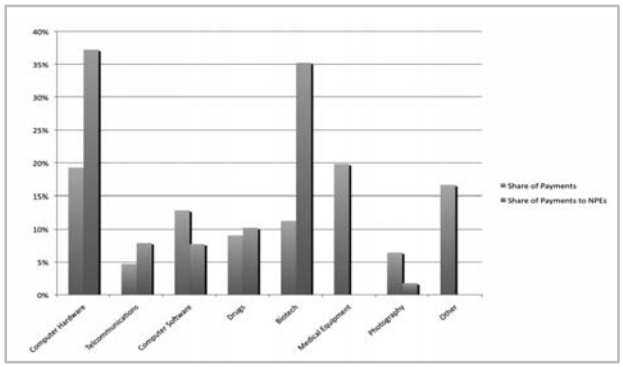


Figure 1. Industry share of all awards/settlements exceeding $100M and industry share of awards/settlements exceeding $100M paid to non-practicing entities.

7. Recent survey data suggest that these figures underestimate the significance of patent infringement actions by non-practicing entities. A survey of nine technology companies reported that in 2008 these companies had a total of 1217 licensing requests and 166 lawsuits pending for patent infringement. Both the number of licensing requests and lawsuits pending show explosive growth from just a few years earlier. In 2004, these companies had 185 licensing requests and 97 pending lawsuits for patent infringement.[3]

8. At these nine companies, more than 80 percent of all patent licensing requests were from NPEs over the period 2004-2008. This is larger than the estimated share of very large awards and settlements for patent infringement paid to NPEs in the computer hardware industry based on publicly available data. However the number likely reflects the increasing role of NPEs in patent infringement cases in this industry. Since 2000, eight of the twelve payments for patent infringement in excess of $100 million in this industry went to NPEs. The website www.patentfreedom.com reports that the number of patent lawsuits filed by non-practicing entities more than doubled from 2004 to 2008.[4]

3 Testimony of Steven R. Appleton, Chairman and Chief Executive Officer, Micron Technology, Inc., Hearing on The Patent Reform Act of 2009 Senate Committee on the Judiciary, March 10, 2009.

4 https://www.patentfreedom.com/research-lot.html accessed March 23, 2009.

9. While the computer hardware and biotechnology industries account for most of the payments to non-practicing entities, there are fundamental differences between NPEs in these two industries and the technical and economic characteristics of their patent claims. Most of the NPEs in biotechnology that received large awards or settlements for patent infringement are small research laboratories or universities. These are entities that specialize in research and their efforts are instrumental to the development of new pharmaceutical products and related technologies.. Furthermore, the technologies covered by the patents generally have a close relationship to a particular product or process. The patent may enable the production of a protein that can be useful for a new biologic drug or the patent may cover a technology for medical testing or drug development. As a result, it is easier to estimate the contribution of a biotechnology patent to the value of a new drug than it is to estimate the contribution of a semiconductor patent to an integrated circuit that also embodies many other patented technologies.

10. The NPEs in the computer hardware industry tend to have different business models compared to NPEs in the biotechnology industry. Most of the NPEs that are the recipients of very large payments for patent infringement in computer hardware are firms that either did not produce a commercial product or are exiting the line of business for which the patent claims are relevant. Furthermore, their patents often address only one or a few features of a complex technology that requires access to numerous other patent rights to make or sell a commercial product. These distinctions are important for the following reasons.

1. Computer hardware requires rights to numerous technologies

11. Unlike many biotech and pharmaceutical patents, the technology covered by patents in computer hardware typically do not define a product or a process to produce a product. Instead, they often cover only *a feature* of a product or a process to produce a product. It can be particularly difficult to value a patent that is one of a great many inputs into a commercially useful product. While this valuation problem is not unique to computer hardware patents, the computer industry is exceptional in that many important products are covered by hundreds or even thousands of patents.

2. Computer hardware patents are often ancillary to R&D efforts

12. Various studies have reached the conclusion that patents have limited value in protecting research programs in the computer and related industries from misappropriation.[5] Trade secrets and complementary investments are more

5 See, e.g., Bessen, James and Michael J. Meurer (2008), *Patent Failure: How Judges, Bureaucrats and Lawyers Put Innovators at Risk*, Princeton University Press; Hall, Bronwyn and Rosemarie Ham-Ziedonis (2001), "The Determinants of Patenting in the U.S. Semiconductor Industry, 1980–1994, *Rand Journal of Economics*, 32 (Spring), p. 101–28; and W.M. Cohen, R.R. Nelson and J.P. Walsh, "Protecting their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)," Working Paper 7552, February, 2000, National Bureau of Economic Research, Cambridge, Mass., revised 2004.

important for competitive advantage in this industry. Trade secrets reflect the fact that manufacturing skills are often more relevant to commercial success than patentable inventions. For an integrated circuit manufacturer, the basic concept of monolithic integrated circuits is a patentable technology, but that does not substitute for the know-how to build circuits with very narrow line widths, which is critical to commercial success.

3. Network effects, switching costs and economies of scale are important sources of value

13. Much of the value in the computer hardware industry is the result of complementary investments made by firms and consumers in the industry. Intel and Microsoft owe their initial success in part to superior technology, but also to the fact that their technologies have become industry standards. Firms and consumers make investments that are specific to these standards and that create value for other users. These networks effects enhance the value of individual investments for the "Wintel" platform and make patent protection a less important determinant of the ability to appropriate returns from investment.

Network effects, switching costs, and economies of scale create value that can be mistakenly attributed to patents. The use a particular patented technology to stack data in a microprocessor can be a source of value, but most of the value comes from investments that support the microprocessor's architecture, create demand for the microprocessor, and add to the cost of switching to an alternative architecture. The threat of an injunction can allow a patent owner to extract a significant fraction of these benefits despite the fact that the patented technology may be of only secondary importance to the value of the product.

4. Failing companies eliminate opportunities to resolve patent disputes

14. Despite the fact that hundreds or even thousands of patents cover computer hardware technologies and other complex products, patent litigation is relatively infrequent. This is because most companies would rather do business with their customers than fight over patent rights in the courtroom. Companies that want the freedom to design and sell products free of infringement litigation have incentives to enter into exensive cross-licensing agreements. Such agreements are common in the computer hardware industry. They are supported by the threat that failure to cross-license can result in the destruction of their businesses from massive patent litigation. Unfortunately, the threat of "mutually assured destruction" is empty when a company is failing or exiting a business and therefore has little to lose from an adverse litigation outcome. Indeed, this is the pattern that emerges from the data on large awards and settlements for patent infringement in computer hardware.

15. Payments for patent infringement to non-practicing entities raise troubling issues when the patents cover a small element of a product or process and when network effects, economies of scale and switching costs are more important than patents as sources of product value. These characteristics are strongly present in markets for computer hardware, software, and information technology. They are somewhat less of a concern in markets for biotechnology and pharmaceuticals. The next section illustrates the potential to over-estimate infringement damages for patents that cover products that benefit from multiple sources of value.

II. Potential to over-estimate damages for complex technologies

16. The Alcatel-Lucent 2007 jury verdict that initially awarded Alcatel-Lucent $1.5 billion for infringement of two MP3 patents provides a clear illustration of the risk that damage awards may greatly exceed a patent's contribution to product value when the product embodies complex technologies. MP3 is a format standard for the storage and transmission of compressed digital audio files on the Internet, personal computers, and portable devices. Lucent-Alcatel alleged that Microsoft's Windows Media Player, which employed MP3 technology as well as other formats for transmitting and storing audio and video files, infringed two of Lucent-Alcatel's patents necessary to implement the MP3 standard. Although the district court judge overruled the jury verdict and an appeals court ruled in favor of the defendant for technical reasons having to do with ownership of the patents, the jury verdict illustrates the potential for very large damage awards for patent infringement despite the fact that the patent represents only a very small part of a product's value.

17. The jury in the Alcatel-Lucent patent case based its damage award for patent infringement on a reasonable royalty of 0.5% per licensed computer. It arrived at the total damage award of $1.5 billion by multiplying the 0.5% royalty times the average price of a personal computer and then applying that figure to the total number of computers sold over the damages period. While not clear from the record, the jury calculation apparently applied the 0.5% royalty to each of the infringed Alcatel-Lucent patents.

18. A key problem with the damages approach accepted by the jury is that it attributed the royalty to the *entire market value* of the computer rather than apportioning the royalty to account for the value contributed by the MP3 patents at issue. The MP3 patents covered technology employed by the Windows Media Player, which Microsoft supplies as a component of its Windows operating systems. While a media player enhances the functionality of the computer, the player is a complement to the operating system software and a prevailing royalty rate reasonably should apply to the software, not to the entire computer. To do otherwise would lead to nonsensical results. For example, a feature-laden computer could cost $2,000. The 0.5% royalty applied to such a computer for each patent would give a value for the two Alcatel-Lucent patents of $20, which is a significant

fraction of the price of the entire operating system. On its face, this result appears to assign too much value to the two MP3 patents at issue given all of the other functionality added to the operating system. Furthermore, Alcatel-Lucent is just one of several entities that together own or license a total of at least 36 MP3 patents.

19. While there is no single correct approach to the calculation of damages that is appropriate for every instance, a reasonable estimate of the economic impact from patent infringement must take into account the contributions from other inputs, including other intellectual property rights. Excessive awards may energize efforts to patent new technologies, but they also increase costs to technology users, which can make it more difficult for those users to develop and commercialize their innovations.

20. A rule that instructed courts to apportion damages for patent infringement would reduce the risk of excessive infringement damage awards such as the jury verdict in the Alcatel-Lucent trial. A statutory apportionment rule is not necessary as evidenced by the corrective action taken by the court in that case. Furthermore, a statutory rule could introduce undesirable rigidities in the calculation of damages for patent infringement. Nonetheless, general guidance is desirable to avoid the most egregious errors that can occur by failing to recognize that an infringed patent is but one of many sources of product value, a fact that is particularly important for complex technologies such as computer software, semiconductors and information technology.

21. Some might argue that real-world negotiations are the only reliable indicators of patent values. For products that require many patents, licensing negotiations depend on the structure of the market in which the negotiations occur as much or more than the technological contribution of the licensed patent. An injunction threat can give a patentee enormous leverage to bargain for a large share of a product's value. If one firm has 100 patents that are essential to make or use a product and another firm has only one, the firm with one patent may use an injunction threat to obtain a large share of the value of the product. But it makes little sense to conclude that one essential patent contributes as much value to a product as 100 equally essential patents.[6] At the same time, it is clearly the case that some patents are much more valuable than others and a patentee should be able to offer evidence to support a claim for a disproportionate share of product value.

22. Another argument is that a patent should earn a "reasonable royalty" and the royalty figure applied by the jury in the Alcatel-Lucent case was "reasonable". The problem with this argument is that the economic underpinnings of a reasonable royalty are weak. At best, a reasonable royalty reflects a likely award assessed by a court for infringement damages. This turns the calculation back onto itself. The court will award damages that reflect a reasonable royalty, and the reasonable royalty is what the court will award. The net result is that neither the court's determination nor the reasonable royalty for actual licensing transactions can be used to justify what is actually reasonable. The amount of the Alcatel-Lucent jury verdict illustrates why a commonly used royalty figure can lead to nonsensical damages for patent infringement, as do examples cited by Lemley and Shapiro in their discussion of royalty stacking.[7] When many patents each earn a "reasonable royalty", the result can be total royalties that are unreasonable by any measure.

23. The apportionment of royalties for patent infringement is not a simple calculation. Such an analysis may require an estimate of the number of patents as well as other intellectual property such as copyrights, know-how, trade secrets and trademarks that cover a technology. Patent owners are sometimes reluctant to divulge information about their patents as it might invite lawsuits to challenge their validity.[8] The calculation may also require an accounting for other inputs that contribute value to a product. But courts should make an effort to elicit damage calculations that reasonably apportion value in patent infringement litigation when many patents cover a technology in addition to the patents being asserted in the case and when intellectual property is only one factor that contributes value to a product. ■

6 See Richard Gilbert and Michael Katz, "Efficient Division of Profits From Complementary Innovations," University of California at Berkeley working paper, 2009. (Derives conditions under which a proportionate sharing rule provides efficient incentives for investment in research and development when many patents are essential to use a technology.)

7 Mark Lemley and Carl Shapiro, "Patent Holdup and Royalty Stacking," *Texas Law Review*, 85(7), p. 1991-2049, 2007.

8 Disclosure might also limit the ability of a patentee to strategically assert its patents against firms that are unaware of the patents' scope. But this strategic flexibility is hardly socially desirable as patent scope is supposed to be in the public domain.

Julia Holtz
jholtz@google.com
Senior Competition Counsel at Google

Why Google's openness makes economic sense

1. This article discusses the economic incentives that Google's openness is based on.[1] The paper is organised as follows: In the first section, Google's business model is discussed, in particular search and advertising. The second section looks a cloud computing and "data liberation" in particular, as well as Google's attitude to open-source software. Section three briefly discusses open and closed systems. Section four concludes.

I. Google's business model

2. In 2008 and 2009, 97% of Google's revenues were derived from advertising. While Google is developing a second pillar of revenue generating services (Google Apps), it is likely that advertising will remain Google's major source of income for the foreseeable future.

1. Product

3. Google did not start life as an advertising company. The search engine was and is at the heart of the company whose mission it is to "*organize the world's information and make it universally accessible and useful*".[2] This goal means not only organising information that is already available online, but also bringing information online that is currently not digitised (e.g., Google Book Search or News Archive).

More engineering time is devoted to search than to any other product at Google, because the company believes that search can always be improved, in particular in terms of relevance and speed.

1.1. Relevance

4. Relevance is improved by constantly trying to perfect the "search algorithm". The software behind the search technology conducts a series of simultaneous calculations requiring only a fraction of a second. While a traditional approach relies heavily on how often a word appears on a web page, Google uses more than 200 signals, including its "PageRank" algorithm, to examine the entire link structure of the web and determine which pages are most important.

Google then conducts hypertext-matching analysis[3] to determine which pages are relevant to the specific search being conducted.[4] The resulting ranking is a combination of overall importance and query-specific relevance. Important pages receive a higher PageRank and are more likely to appear at the top of the search results. PageRank also considers the importance of each page that casts a vote, as votes from some pages are considered to have greater value, thus giving the linked page greater value.

1 See for example, the blog by Jonathan Rosenberg (Senior Vice President, Product Management) on Google's view on openness, http://googleblog.blogspot.com/2009/12/meaning-of-open.html.

2 http://www.google.com/corporate/ .

3 The search engine also analyzes page content. However, instead of simply scanning for page-based text (which can be manipulated by site publishers through meta-tags), the technology analyzes the full content of a page and factors in fonts, subdivisions and the precise location of each word. Google also analyses the content of neighbouring web pages to ensure the results returned are the most relevant to a user's query.

4 The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result. The search results are returned to the user in a fraction of a second.

1.2. Speed

5. Google is interested in returning search results fast. Generally, speed (or reduction of latency) is of utmost importance. This is one of the reasons that Google has invested in a browser (a program with which a user can visit websites), Google Chrome. Google Chrome aims to improve security, speed, and stability. Apart from significant differences in its minimalistic user interface, Chrome's strength is its application performance and JavaScript processing speed, both of which were independently verified by multiple websites to be the swiftest among the major browsers.[5]

1.3. User focus

6. From day one, users have been Google's core asset. The most important corporate motto is therefore: "*Focus on the user and all else will follow*."[6] This means that Google is primarily focused on releasing products that gain wide acceptance and high usage rates, and considers how to monetise these products later on (if they prove to be successful).

This allows Google to experiment with new products. This process is internally known as "*launch and iterate*"[7], *i.e.*, new products are released very early in the development. This means that they may not meet the requirements for a final product for quite a long time (Gmail famously kept its "beta" tag for 5 years[8]), but the feedback the company gets from users help to improve new products quickly. Overall, Google considers that the pace of innovation is greater if products are released early and improved quickly, rather than waiting for perfection before releasing them.

2. Advertising

7. Advertising has proven to be an efficient monetisation strategy for search and other products. This is partly because it has always been Google's goal to offer advertisers measurable, cost-effective and relevant advertising, so that the ads are useful to the people who see them as well as to the advertisers who run them. Advertisers bid in an open and competitive auction to have their ads appear alongside search results for particular keywords. They can specify the geographic location and time of day for their ads to appear. As a result, people see ads that are so useful and relevant and a valuable form of information in their own right.

There is a strict policy to distinguish ads – that someone has paid for – from search results or other content on a page – that nobody pays or could pay for. Ads are labelled as "sponsored links" or "Ads by Google". It is not possible to buy a placement in search results, nor can the ranking be influenced.

Advertisers only pay if a user clicks on the link.

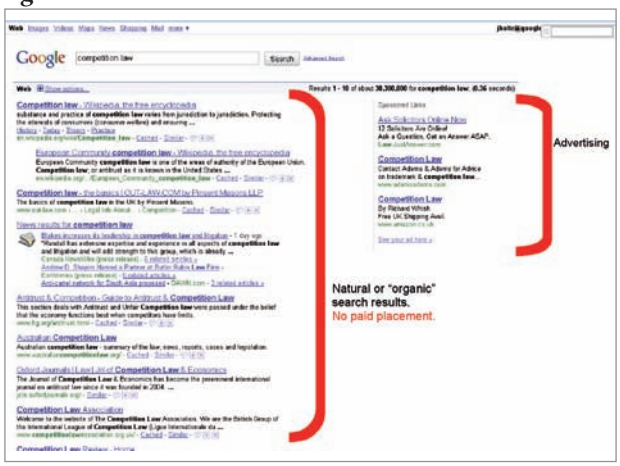
5 See http://en.wikipedia.org/wiki/Google_Chrome with further notes.

6 See http://www.google.com/corporate/tenthings.html.

7 See e.g. http://googleblog.blogspot.com/2008/09/fresh-take-on-browser.html.

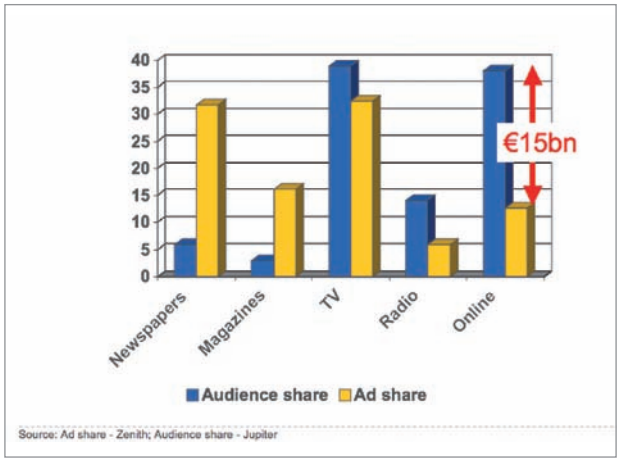
8 http://gmailblog.blogspot.com/2009/07/gmail-leaves-beta-launches-back-to-beta.html

figure 1



8. Advertisers follow eyeballs. These days, the internet accounts for roughly 10% of total advertising revenues. While the percentage is considerably higher in some countries (e.g., the UK) and lower in others (e.g., Portugal), the trend is identical everywhere: it is growing steadily, which is not surprising considering that the time spent online is disproportionately higher than the advertising dollars spent online. In other words, advertising online has significant prospects of growth as the following graph illustrates:

figure 2



9. This means that for Google, the more time is spent online the more it will benefit, albeit indirectly. There is, however, no requirement for users to use Google services; this would in any event be impossible given how the web is structured. There is also no lack of alternatives. However, if Google is successful in attracting users to its products, it will very likely indirectly benefit by generating higher advertising revenues. It is therefore incentivised to innovate at a high pace and to deliver high quality products. In addition, Google is interested in improving the performance of the Internet both in terms of speed and content.

In short, increased online activity is likely to be beneficial to Google's as well as others' bottom line. As further set out below, open systems are supporting the development of the Internet, which receives a steady stream of innovations that attracts users and usage, and grows the entire industry.

II. Cloud computing and data liberation

10. In addition to search and ads, Google's is involved in the supply of web applications ("Apps").[9] Apps were mentioned as an official third strategic pillar in 2007 and have been growing in importance ever since.[10]

Google's best known web applications are Gmail, Google Calendar and Google Docs & Spreadsheets, competing with other webbased productivity applications (*e.g.*, Zimbra) and increasingly desktop productivity applications (*e.g.*, Microsoft Office, Star Office).

11. Web applications are accessible via a browser such as Internet Explorer, Firefox or Chrome. They do not require a piece of software on the desktop, but are accessed remotely. The software and the associated data resides on a server somewhere in the world, or in "the cloud". The user types a URL into the browser, and may need a password to access the service, but it does not matter where he is located.

12. Cloud computing has many advantages over desktop based computing. For example:

→ *Access from anywhere, anytime*: content can be accessed from any computer in the world with an internet connection.

→ *Better collaboration*: because documents live online, it is easy to "share" them with a group of collaborators (*i.e.*, to give access). Everyone in the group can work on the same material at the same time, even if they are working in different buildings, countries or continents. There is no confusion over what is the "latest version" of the document, and comments are immediately visible to the whole group.

→ *Cost savings*: enterprises that rely on cloud computing do not have to invest in their own server infrastructure – this is effectively outsourced. This is particularly useful for SMEs / start-ups as significant cost savings can be achieved.

→ *Less downtime*: the service provider maintains the software on an ongoing basis (many cloud offerings do not have any planned downtime) and the service is not disrupted. Upgrades can be done remotely and often run in the background without the need of downloading the next version of the specific programme.

13. However, cloud computing also faces challenges, inter alia the reliance on a fast internet connection, and legal issues (around data retention). In particular, the user needs to have a high degree of trust in order to agree its data to a third party.

There are two trust aspects that I would like to discuss here: security and choice.[11]

9 http://www.google.com/corporate /. Mobile has been added. Some Google apps are desktop applications: e.g. Chrome (the browser, a programme that accesses the web; or Google Earth), but the vast majority of apps are web-based.

10 http://news.cnet.com/8301-30685_3-10380917-264.html: Eric Schmidt, CEO: Enterprise is the"next billion dollar opportunity".

11 Another aspect is privacy, not further discussed in this document. Google recently released a dashboard showing which data it holds about users. See, http://googleblog.blogspot.com/2009/11/transparency-choice-and-control-now.html.

1. Security

14. Many users and corporations feel that they are better able to secure their data than a third party. However, the following statistics show that this is misconceived:

→ 60% of corporate data resides on unprotected PC desktop and laptops;[12]

→ 1 out of 10 laptops will be stolen within the first year of purchase;[13]

→ 66% of USB thumb drive owners report losing them, over 60% with private corporate data on them.[14]

15. In other words, owning the data on one's own servers or premises does not make the data necessarily more secure. On the contrary, a cloud service is likely to have more sophisticated means to protect their customers' data, as this ability has to be a core competency. A local IT department is less likely to be able to spend similar resources on security.

16. However, the perception is still prevalent that having control over the IT infrastructure means more security. This will likely change. In his book "The Big Switch",[15] Carr makes an interesting observation by comparing the development of energy supply with the IT infrastructure. He points out that around 100 years ago, enterprises used to have their own power plants, locally run and controlled, and that the idea to shift to central utilities was considered too high a risk. The arguments used against centralised energy supply are surprisingly similar to arguments used against cloud computing: that the power infrastructure was a core asset of the company, and that it would be too dangerous to lose control over an indispensable input. However, the shift to central utilities did of course happen, even though the technical hurdles to overcome (such as the switch from DC to AC) were higher than the hurdles we see today with cloud computing.

2. Choice

17. At least as importantly, the user or the enterprise is more likely to trust the cloud supplier if he stays out of choice with the service, rather than by design. Choice means that he has to be able to revoke the trust at any point in time by leaving the service.

18. Google has always operated in a very competitive environment. One cannot think of a product that has lower barriers to switching than search. All a user has to do is type in a different URL into the browser, or, indeed, search for "search engines" to be presented with a wide variety of choice. The service is free, quick to access, there is no need to retrain as the use of the service is very straightforward, and there is no lock-in or bundling with any other service of

12 Source: IDC.

13 Source: Hewlett Packard.

14 Source: Salesforce.

15 See, http://www.nicholasgcarr.com/bigswitch/.

any kind. Online users are also very impatient and not loyal to one service.[16] Finally, search is not characterised by strong network effects.[17]

19. As a result, Google is forced to constantly innovate and improve its search algorithm. If it did not deliver good quality results, users would quickly decide not to come back, which has demonstrably happened to Altavista, Exite[18], Lycos[19] and Yahoo![20] in the past.

20. The fierceness of competition in search forces Google to stay focused on innovation. It considers that this model is the best long-term guarantee for high quality, and this ties in with its focus on the user. It is therefore not surprising that it aspires to make it easy for users to leave its services even though this seems counter-intuitive at first sight. Allowing people to leave increases trust. Since users know they could leave, loyalty is enhanced and they are more likely to stay as a result.

At the same time, the fact that it is easy to leave exposes all products to more intense competition, forcing them to gain users on the basis of product quality – because users switch to the services – and not because it is hard for users to leave.

21. Consider Gmail as an example: email can be a relatively "sticky" product, if it is hard to pull copies of the emails down to a local computer or competing service; and if it is difficult to export all contact information contained in the address book. This means that typically, a user is less likely to switch between email providers than e.g. between search engines or social networks.

22. While Google has always had a policy of not locking users' data in, Google recently started an engineering team to make it even easier for users to export their data.[21] For example, Gmail began offering the internet standard POP protocol early on and later added IMAP which allows the user to connect Gmail with any standard email client and to move all email to a competing service. Similarly, it is possible to download all contact information with a few mouse clicks into a CSV file, which is in turn accepted by most standard email programmes where this information can be uploaded.

23. Today's users have a considerable amount of valuable data stored across a multitude of internet servers, services and companies; from email to photos, from documents to spreadsheets, from instant messages to address books. Previously, the thought of whether or not it is possible to get your data out only occurred at the moment a user or wanted to leave a service, but it seems likely that with increased sophistication, users will start thinking about these issues before they start using a service.

24. According to Google's Data Liberation Front, a user and an enterprise should ask the following questions before entrusting a service with data:

→ Can I get my data out at all?

→ How much is it going to cost to get my data out?

→ How much of my time is it going to take to get my data out?

Ideally, the answers to those questions would be "*yes*", "*nothing more than I'm already paying*", and "*as little as possible*". Many web services still make it difficult to leave their services; users have to pay them for exporting data, or jump through all sorts of technical hoops – for example, exporting your photos one by one, versus en masse. However, locking users in by holding data hostage is no longer an effective way to retain users. Due to negative sentiments, they're unlikely to come back later to try a new product. Since we're constantly innovating, if a users leaves one of our products today, we want them to have a good experience doing that so that they might come back to us tomorrow to try a different product.

25. Consumers as well as enterprises are likely to consider these questions more intensely than they have in the past. By not locking Google's users in, the company forces itself to focus more on innovation as a means of retaining our users, and it would also be in Google's interest to see every web service company follow suit, as this would render cloud computing a more credible alternative overall.

III. Open vs. closed systems

26. However, data liberation is not only an issue when it comes to cloud computing. Data is not necessarily more easily accessible when it is stored locally. There are many reasons why data can be locked in, even though it sits in front of the user on a local machine.

16 For example, on January 31 2009, a coding error corrupted Google search results for about an error, showing an error message under every search result ("warning: this site could harm your computer"). During that hour, the volume of Yahoo searches increased quickly over its normal volume for that time of day. *"From virtually the instant the problem began, queries on Yahoo started shooting up and about an hour later, they reached twice the level they reached the same time the previous Saturday"*. - Prabhakar Raghavan, head of Yahoo Research, Wall Street Journal, 17 March 2009

17 Network effects refer to a phenomenon where the amount that people are willing to pay for a service depends on the number of people that have already adopted a service. The classic example is a fax machine: the amount that a buyer is willing to pay for a fax machine depends on how many of his correspondents already have one. However, a user's decision to use Google is irrelevant to other users. Google benefits from more users insofar as it collects data that it can use to improve the search engine, but given the number of searches conducted, even search engines with a small share get billions of observations, enough to improve their algorithm. Advertising is not characterised by network effects either. It is of course a two-sided market – advertisers want to advertise where users are. However, this does not influence the amount that they are willing to pay on a *per user* basis. The value of a user to an advertiser depends on how likely that person is to buy, not how *many* users there are. A small website about knitting would be relevant to people interesting in this topic, so it may be able to charge higher rates for placing an ad for wool than a website with a large audience. See more detail at http://googleblog.blogspot.com/2008/02/our-secret-sauce.html.

18 "*Almost all of today's search entrepreneurs also say that Google's success lends credibility to their own long-shot quest. When Lawrence Page and Sergey Brin first started tinkering with what would become Google, other search engines like AltaVista and Lycos and Excite were dominant*."(*The New York Times*, 1 January 2007).

19 **"**For a short period in 1999, Lycos became the most popular online destination in the world."(Johan Battelle, *The Search*, 2005)

20 «Yahoo! … the most successful company ever spawned by the World Wide Web. […] This much is clear: Yahoo! has won the search-engine wars and is poised for much bigger things.» (*Fortune*, 2 March 1998)

21 See, http://googlepublicpolicy.blogspot.com/2009/09/introducing-dataliberationorg-liberate.html. It was never impossible to leave a Google service, but sometimes it could be more difficult than necessary. This was never a design decision, but dedicated engineering resources are required to make it easy to leave a service.

→ For example, data on floppy disks today is effectively locked.[22] This data can only be accessed with computers that have the mechanical means to read such a disk, however, no standard computer has a floppy drive these days any more.

→ Data can be lost when a file format is no longer supported (if it is not documented in an open fashion), or where proprietary formats are used that are not supported by competing services. Lock-in can also occur where it is not impossible to convert data, but where conversions cause some data to be lost (e.g. formatting).

27. The web is different. It has its roots in the military world and academia, it was not a commercial venture. It did not, does not and will not belong to any single company.

The web is fundamentally about openness, where more interaction means more innovation and more benefits for users. The web facilitates collaboration and idea sharing across the world. It is an interoperable, ubiquitous and searcheable network where everyone can share information, integrate and innovate, often without having to ask for permission. New fast browsers make the user experience better and more and more similar to the desktop (which for complex applications still beats web apps because they are accessible locally).

28. Google could not have existed without the open web. When Larry Page and Sergey Brin came up with their search technology that they installed at Stanford University[23] did not have to ask permission to crawl the web, or negotiate anything to put up a website. The fact that in many instances, no deals need to be struck and no payments need to be made reduces the barriers for webmasters, developers and also users significantly, spurring innovation.

29. It reduces inefficiencies by allowing developers to concentrate on the problem they want to solve, without the need to start from scratch. To use an "offline" example, a builder will buy lumber and bricks, not grow trees and make bricks on his own. Translated to the online world, if a developer would like to show all coffee shops in Lisbon, he does not have to start creating a map, and after years of efforts, then place coffee shop icons on there. Instead, it is possible to create a mash-up using pre-existing maps APIs, allowing the developer to concentrate on the problem that he wanted to solve.

30. While open systems are prone to spurring competition, it is not necessarily the case that closed systems are inherently uncompetitive. For example, Apple's iTunes (which is a closed system) was a tremendous innovation, which has transformed an entire industry.

However, closed systems lend themselves to complacency. If a company stops innovating and relies on inertia and lock-in to retain their users – which is easier in a closed environment – they are vulnerable to the next company (start-up, corporation, or otherwise) that works harder, innovates more, and just plain makes a better product for their users. And since the cost of starting up and distribution is fast approaching zero, it is easy for new companies to get into the game, and even easier for users to try new products. In a closed system, it will be possible to keep users longer – in the short term.

But as illustrated above, they will finally find a way to leave, which is why Google has decided to bet on a long-term value proposition, ensuring that its products are exposed to competition and thus continued innovation. In an open system, a competitive advantage does not derive from locking in customers, but rather from understanding the environment better than competitors, and using that knowledge to generate superior, more innovative products.

31. Therefore, Google has invested in a number of key open source projects, notably the Google Chrome browser and the Android mobile phone operating system (with the Open Handset Alliance). Recently, it announced a new approach for an operating system for desktop computers, Chrome OS.[24]

All projects required a long-term engineering commitment. Nevertheless, Google decided to open source all three products and make the code available for free.

→ Google Chrome[25]: As explained above, Chrome offers significant innovations in security and speed in particular. While its usage share is still relatively modest (less than 5%), Chrome's immediate impact is more indirectly measurable: since Chrome is open source, other browser vendors are able to look at the code and to use the innovation for their own products. Both Mozilla's Firefox and Apple's Safari immediately included some of Chrome's features.

→ Android[26]: This open source operating system for mobile phones is designed to give users more control over the mobile experience. Before Android, there was no open platform for mobile phones. Hardware manufacturers naturally focused on hardware, and developers had to rewrite their applications to make them work on a large number of phones. Android improved the frameworks to enable more sophisticated development of applications for mobile phones, and the open source / free aspect means a higher adoption rate by manufacturers. As a result, the user experience for the mobile Internet will be vastly increased. Higher internet usage will translate into more revenues for Internet companies, including Google.

→ Chrome OS[27]: Google Chrome OS is a lightweight operating system that will initially be targeted at netbooks (small laptops with less computing power). Speed, simplicity

22 Sony, the last large company to produce them, stopped sales in March this year, see http://www.engadget.com/2010/04/26/sony-shutting-down-japanese-floppy-disk-sales-by-march-2011-kil/

23 See, http://www.google.com/corporate/history.html.

24 In addition, Google contributing over 800 projects that total over 20 million lines of code to open source. An open source project hosting service (http://code.google.com/hosting/) hosts over 250,000 projects, handles basic engineering needs, aids collaboration, and saves time. Google AppEngine allows developers to create apps quickly that scale without needing to set up infrastructure themselves.

25 See, http://googleblog.blogspot.com/2008/09/fresh-take-on-browser.html.

26 See, http://googleblog.blogspot.com/2007/11/wheres-my-gphone.html.

27 See, http://googleblog.blogspot.com/2009/07/introducing-google-chrome-os.html .

and security are the key aspects of Google Chrome OS, which will be designed for the web. The goal is to be able to start up the computer and allow the user to access the web in a few seconds. Most of the user experience will take place on the web.

IV. Conclusion

32. Google has taken some surprising steps, such as open sourcing core products that require high levels of R&D, and making it easy for users to leave its services. However, its commitment to open systems is not altruistic. Rather it represents good business, which is why it is unlikely to change its strategy in the future. ■

Jean-Yves Art*
jeanart@microsoft.com
Associate General Counsel

Intellectual property and competition

Are patents conducive to the supply of innovative products at lower prices? The case of the software industry

Abstract

Some developers in the software industry rely on patents and other intellectual property rights in order to protect their inventions and make their products available to consumers at affordable prices. Others rely on different development and monetization models. This article shows that creative solutions have been developed in order to enable consumers to choose from among a broad array of software products capable of meeting their needs, irrespective of the vendor's preferred development and distribution model

Dans l'industrie du logiciel, certains développeurs se fondent sur les brevets et autres droits de propriété intellectuelle afin de protéger leurs inventions et pouvoir mettre leurs produits à la disposition des consommateurs à des prix raisonnables. D'autres développeurs s'appuient sur des modèles de distribution et de rémunération différents. L'article montre que des solutions créatives ont été développées afin de permettre aux consommateurs de choisir librement au sein d'un vaste choix de logiciels susceptibles de répondre à leur besoins, indépendamment du modèle de développement et de distribution choisi par le concepteur.

1. For many years, government policymakers, scholars and commentators have largely agreed that intellectual property is a key driver of innovation, that innovation in turn spurs competition (and vice-versa), and that both innovation and competition benefit to consumers who obtain new or better products at lower prices.[1] Recently, however, this relationship between intellectual property, innovation, competition and consumer welfare has given rise to some debate, in particular as a result of developments in the pharmaceutical and the software industries. In the latter industry, the debate has been associated with the emergence of the "open source software". This is because the GNU General Public License ("GPL"), one of the licenses used by open source software developers, contains provisions that are generally incompatible with the conditions of enforcement of certain intellectual property rights, such as trade secrets, and automatically involves the grant of a non-exclusive, worldwide, royalty-free patent license under certain of the developer's patent claims to recipients of the software products down the chain.[2]

2. The development of open source software and its co-existence with proprietary software raises complex questions at the juncture of competition and intellectual property law. Indeed, competition in the software industry is spurred by "interoperability." Interoperability is defined as "*the ability to exchange information and mutually to use the information which has been exchanged.*"[3] In layman terms, interoperability describes the ability of two software products or a piece of hardware and a software product to work together, that is, to connect to and with each other and to deliver the functionality which they are intended to deliver. For instance, interoperability describes the ability for Firefox to work on Windows, enabling Windows PC users to use Firefox – not only Internet Explorer – to surf the Web, access and view web pages properly, and download files from the web onto their PC. Thanks to interoperability, users of Apple iPhones and Nokia devices (among others) can synchronize their mail, contacts and other data with Exchange servers. Interoperability also enables users of an mp3 device developed, for instance, by Sony to access and download music from online music stores such as Napster. Aside from the (very crucial) fact that consumers often expect IT products – whether software or hardware – to interoperate with other software and hardware products, interoperability also plays a critical role in competition law. Indeed, in an industry such as the IT industry characterized by strong network effects, regulators have often focused on interoperability as a means to avoid lock-in into a single vendor's offerings. That is because mandating interoperability is viewed as reducing the barriers to entry and expansion in the market, including those which arise from possible network externalities. The ultimate goal is to enable the offer of innovative products to consumers.

*The views expressed in this paper are strictly personal and are not necessarily those of the Company.

1 For a discussion of recent theoretical and empirical studies on the relationship between patents, innovation and competition, see OECD, *Competition, Patents and Innovation*, DAF/COMP(2007)40.

2 See in particular, GPL, Version 3, section 11, at http://opensource.org/licenses/gpl-3.0.html. Note that GPL is only one of the many open source licenses. For other open source licenses approved by the Open Source Initiative, please see at http://www.opensource.org/licenses/category. In addition, several licenses are used for the distribution of open source software although they do not meet the approval criteria set by the Open Source Initiative.

3 Council Directive of 14 May 19991 on the legal protection of computer programs, OJ 1991 L 122, at p. 42 (see Preamble).

3. In practice, interoperability can be achieved in various ways. Under the most straightforward – and often preferred – approach, some software code is included in the two products concerned in order to ensure that the messages sent by one of them are properly received and understood and can be processed by the other. This software code may be based on standards adopted by standard organizations – such as "HTTP" the standard protocol for the formatting and transmission of messages on the Web and for interaction with Web servers and Web browsers in PCs and other computing devices, or "mp3" the ISO/IEC digital audio encoding format standard file format supported by most digital audio devices. Alternatively, the software code that enables interoperability can be exposed through specifications privately developed by companies or individuals but made available broadly. This applies for instance to a large number of communication protocols[4] and application programming interfaces[5] integrated in the Windows operating system, which Microsoft has developed in order to ensure proper delivery of the services which are the competitive differentiator between Windows and other operating systems such as Apple's Mac OS X or Ubuntu. Sometimes, such communications protocols are so widely used that they can be regarded as *de facto* standards – for instance, TCP/IP, the suite of communications protocols used to connect computers to the Internet, has become a *de facto* standard for the transport of data over networks. In other words, there is a broad spectrum of solutions to achieve interoperability.

4. In all cases, the specifications may be covered by patents and software developers need a patent license in order to implement them in their products. This applies to some of the interoperability technologies which are developed by individuals and companies. It also applies to standard technologies. Indeed, many standards (including the very popular mp3 ISO/IEC standard) read on patents held by companies which may (or may not) contribute to the development of the standard. Some standard-setting organizations such as W3C and OASIS have defined an IP policy requiring, for instance, that participants in the standardization process license the essential claims related to that standard free of any royalty or under Fair, Reasonable and Non-Discriminatory terms ("FRAND"). Other organizations do not adopt that approach, or have more flexible policies. The vast majority of standard-setting organizations do not require participants to waive such IP claims altogether.

5. There is clearly some tension between the various FRAND terms that these standards setting organizations allow in a patent license or the traditional terms in patent licenses pursuant to which individuals or companies license proprietary technologies on the one hand, and the GPL family of licenses, on the other.[6] The tension for software licensed under GPLv2 (a frequently used variant of GPL licenses) arises from Section 7, which only allows distributors to take a third party patent license if they are able to pass all the license rights they receive and rely upon to downstream recipients of the code. This means that a would-be GPL implementer cannot take a patent license unless the patent holder is willing to give up the possibility of earning royalties on all the sublicensed copies of the implementer's software – which could be nearly all copies, under the open source model. GPLv3 is arguably even more restrictive in that it expressly restricts the ability of an implementer to pay any royalty to a third party patent holder. Both versions of the GPL were intentionally crafted to be in conflict with the traditional per unit royalty term associated with patent licenses (and other terms associated with FRAND patent licensing).[7] Obviously, these provisions in the two GPL versions may create challenges for the implementation of interoperability technologies which are protected by patents.

6. The first point to be noted in that respect is that the same tension does not apply to open source licenses other than GPL. Indeed, under such other licenses, distributors are generally free to accept any and all terms from third party patent holders.[8] The second, important point is that, in the mixed (open and proprietary) source world, very few commercial vendors only use GPL. Indeed, many of them manage to keep code separated in order to maintain a proprietary layer without violating the GPL and it may be possible to implement certain royalty-bearing standards under this architecture. This appears to be the case, for instance, for the GSM and mp3 standards.[9]

7. This being said, some developers may decide to distribute their products only under GPL. For those vendors, the use of patented technology in software code distributed under the GPL may give rise to IP infringement risks. In 2006, Microsoft and Novell – a leading developer of open source products – entered into an agreement which delivered a solution to this difficulty, in the form of a covenant given by Microsoft not to assert its patent rights against customers who use certain Novell products (Novell gave a similar covenant to users of Windows and other

4 The communications protocols are the sets of rules which govern the transmission of data between hardware and/or software products.

5 The application programming interfaces are the protocols and tools which may be used for the development of applications designed to run in a given operating environment.

6 The substance of these arguments is set out in several articles, including: Richard Stallman, *Patent Licenses Discriminate*, April 23, 2002 available at http://news.zdnet.com/2100-10532_22-298367.html (suggesting that both per copy fee and scope limitation in RAND licenses "*discriminate against the free software community*"); Mikko Välimäki and Ville Oksanen, *Patents on Compatibility Standards and Open Source – Do Patent Law Exceptions and Royalty-Free Requirements Make Sense?* Journal of Law and Technology 3/2005 (Volume 2, Issue 3), pp. 436-445 (noting that "Since many open source licenses do not allow the collection of even "reasonable" patent royalties, it may be impossible to implement a RAND standard in open source software. Thus, such policies may in fact discriminate against open source developers").

7 See Patrick Durusau, "Self Inflicted Discrimination and the GPL", April 2008 available at http://www.durusau.net/publications/gpl.pdf (noting that "*GPL has decided a priori that it must have the sublicensing provision [in Section 7] […]. GPL followers have chosen a particular license and software development model and should allow others to do the same.*")

8 See Nah Soo Hoe, "Free/Open Source Software and Open Standards", available at http://akgul.bilkent.edu.tr/iosn/foss-openstds-withcover.pdf ("*FOSS licenses, then, do differ with regard to the nature and degree of rights and obligations described. Consequently, licenses like the BSD allow the usage of technology available under RAND terms but GPL does not allow any GPL-based public distribution to include any technology available under a RAND license that is not royalty-free.*")

9 Thus, Google's Nexus One phone has a GPL (Linux) operating system but also implements GSM and mp3 in software that is included under other licenses. See at http://www.google.com/phone/static/en_US-nexusone_tech_specs.html.

Microsoft products).[10] Thus, while compliance with the GPL requirements creates some structural impediments for open source and proprietary software developers to enter into traditional trade secrets and patent license agreements that would make it possible for open source software developers to implement protected specifications of interoperability technology, the solution devised in the Microsoft-Novell agreement consists in passing over the open source software vendors and distributors and reaching out directly to their customers – the users of their products – to offer them the assurance that they will not be sued for IP infringements on account of their use of open source software products which implement patented technologies without the required IP licenses. This solution bridges the gap between open source and proprietary software – both groups of software developers can continue to develop and distribute their products, ensure that these products interoperate with each other, and remain faithful to the fundamental principles of their respective business models. Since then, a number of similar agreements have been concluded with other open source software vendors, including Linux platform provider Xandros and Turbolinux, a leading Linux client and server distributor in Japan and China.

8. More bridges have recently been established with the open source community. For certain classes of specifications (including the specifications of the Office binary file formats), Microsoft has made its patents available without charge to implementers. For instance, the Open Specification Promise is a unilateral, perpetual, worldwide, royalty-free promise publicly made by Microsoft not to assert patent claims against any developer and any user of products implementing (wholly or partially) over 200 technologies developed by Microsoft, including file formats, virtualization, Web services, and security technologies.[11] Thus, any developer – including any open source developer – and any customer is free to use these technologies or products implementing those technologies without running the risk of being sued for infringement of patents which are owned or controlled by Microsoft and are necessary to implement the relevant specifications.

9. Against this background, the following considerations point to a positive response to the question whether patents in the software industry enable the supply of innovative products at lower prices to customers:

→ Software vendors rely on various development and distribution models that span across a wide spectrum;

→ One of these models relies on IP value (including patent protection); another model at the far end of the spectrum (in its pure form) relies less, if at all, on IP. However, in addition to the cases at the edge, a large number of the software vendors in middle section of the spectrum rely on IP, including those who claim to be committed to "open source software," such as IBM[12] and Google;[13]

→ Competition and innovation in the software industry is best served by the co-existence of the open source and proprietary software development and distribution models, and the ability of customers to choose freely between the open source and the proprietary solutions;

→ From a regulatory perspective, interoperability is regarded as a critical factor to ensure that customers can exercise such free choice;

→ Creative solutions have been developed to promote interoperability between open source and proprietary software and enable consumers to choose and combine both products.

10. Interoperability is a two-way avenue. It requires cooperation between software vendors. This is true both from a technical and also a legal perspective. Hopefully, both proprietary and open source software vendors will continue to adhere to the solutions that have been developed to promote interoperability and resist attempts to undermine the positive results achieved in recent years.[14] ■

10 See at http://www.microsoft.com/presspass/press/2006/nov06/11-02MSNovellPR.mspx.

11 See at http://www.microsoft.com/interop/osp/default.mspx. For a number of additional specifications, Microsoft has made its patents available without charge to implementers pursuant to the Microsoft Community Promise (see at http://www.microsoft.com/interop/cp/default.mspx) or an earlier form of covenant not to sue (see at http://office.microsoft.com/en-us/products/HA102134631033.aspx). For an overview of the commitments made by Microsoft to promote interoperability with its most successful products, please see at http://www.microsoft.com/interop/principles/default.mspx.

12 This is obvious from IBM's response to the complaints filed with courts and with the European Commission by several companies claiming that IBM illegally refuses to license the technologies required to ensure interoperability with IBM-compatible mainframes. Thus, in response to the complaint filed by T3 Technologies, a maker of IBM-compatible mainframes for small- and medium-sized companies, with the European Commission, IBM stated that "IBM has not seen T3's alleged EU complaint. Nonetheless, IBM is confident that it is no violation of competition laws for IBM to rightfully seek to prevent another company from violating IBM's intellectual property rights. IBM has spent great time and expense developing its technology and will defend its intellectual property rights vigorously." See at http://news.cnet.com/8301-1001_3-10145734-92.html

13 The posting of a Google blog about the "openness" of Google technologies has attracted strong reactions from members of the IT community who have pointed out that Google's search results ranking algorithm (protected by patent and trade secrets) is excluded from its open source discourse and have questioned the validity of the explanations put forward by Google to justify this exclusion. See at http://www.businessinsider.com/google-should-open-source-what-actually-matters-their-search-ranking-algorithm-2009-12.

14 For some examples of such cooperation, please see at http://interopvendoralliance.com/default.aspx, http://microsoftontheissues.com/cs/blogs/mscorp/archive/2009/07/22/collaboration-competition-and-ip-in-the-real-world.aspx and http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1523940. As mentioned above, GPL v3 (see footnote 2 above), Section 11, 7th paragraph, may create obstacles to the conclusion of agreements such as the 2006 Microsoft-Novell agreement. It would be unfortunate that attempts to promote interoperability be derailed precisely by those who claim to be its most ardent defenders.

Intellectual property and competition: Complementary policies? The case of the software and pharmaceutical industries

David R. Schmidt
dschmidt@ftc.gov
Assistant Director, Bureau of Economics
*US Federal Trade Commission**

Abstract

This talk focused on results from the FTC's 2009 Interim Report on Authorized Generics. Some basic statistics about the impact of authorized generics on various market participants were discussed in the context of the impact of authorized generics on the incentives underlying generic entry decisions and also as authorized generics relate to patent settlement agreements.

Cette présentation présente les résultats du Rapport 2009 de la Federal Trade Commission relative aux médicaments génériques. Il analyse des statistiques concernant leur impact sur les différents acteurs dans le cadre des motivations ayant conduit à leur mise sur le marché. Il étudie par ailleurs des génériques ayant fait l'objet d'accords transactionnels en matière de brevet.

Authorized generics: How they relate to generic entry and patent litigation

1. I would like to thank the Portuguese Competition Authority for its warm hospitality and also for the opportunity to present some of this work that we at the Federal Trade Commission have been doing on pharmaceutical markets. I want to highlight the disclaimer at the bottom of my first slide that the views expressed are mine and not necessarily those of the Commission; although I will intend to show some results from research that we have been doing at the Commission.

2. My talk will focus on two strands of research that we've been pursuing at the Federal Trade Commission that are more connected than I anticipated when I first started looking at them. They both relate to pharmaceutical markets, obviously, so I want to get some terminology clear at the beginning. We've been undertaking a very long and in depth study of a practice in pharmaceutical markets of branded companies launching what we refer to as an authorized generic drug. An authorized generic is essentially the brand company taking their approval to sell a branded drug and, typically once they face generic competition, they start selling a generic version of the drug in addition to the branded version. It is literally the same product coming off the production line, but as far consumers are concerned, they can't tell the difference between a regular generic drug and the authorized generic drug. If you walk into the pharmacy, and the pharmacist asks you if you would like the prescription filled with a generic drug, they will fill it with the drug and you can't tell whether you are getting an authorized generic drug or an independent generic. So this is really what economists would refer to as a homogeneous product; indistinguishable from the consumer's perspective. Of course, the second issue that most of us are very familiar with, is pay for delay settlements. A typical way for generics to enter the market, at least in the United States, is to challenge the validity of a patent and if they are successful, the generic company that challenges the patent gets one hundred and eighty days to be the only generic drug on the market. That has lead to situations where branded companies will enter into settlements of these patent disputes with the generic company, and part of that settlement will typically be a date at which the generic drug is allowed to enter the market. Sometimes those settlements also involve some form of payment from the brand company to the generic company. Obviously, any time we see an incumbent supplier negotiating an entry date with a potential competitor, and money flowing from the incumbent to the potential entrant, we as competition authorities have to be at least somewhat concerned about that. There might be justifications for it, but it's certainly something to which we should pay attention. As I'm well aware, the European Commission has a very informative pharmaceutical sector inquiry report that brings up many of these issues[1] and we at the Federal Trade Commission just have released another report on this issue,[2] so it's a very lively topic.

* The views expressed are those of the speaker and do not necessarily represent the views of the Federal Trade Commission.

1 See European Commission, Competition DG, Pharmaceutical Sector Inquiry Final Report, http://ec.europa.eu/competition/sectors/pharmaceuticals/inquiry/index.html.

2 Federal Trade Commission, Pay-for-Delay: How Drug Company Pay-Offs Cost Consumers Billions, http://www.ftc.gov/os/2010/01/100112payfordelayrpt.pdf.

3. I think we are all pretty familiar with, and actually Frank pointed out is his talk, the welfare impacts of generic entry. Frank's talk was a very nice lead-in to this talk, because if we think the quantity of the drug is not changing much and prices come down when the generic entry occurs, then the welfare implications are pretty clear. From the economist's perspective if roughly the same profile of product is being consumed, then all that is happening when prices change is a transfer; merely a question of who is getting the money. Generics typically enter with lower prices than the brand. It's interesting when we have these situations where the generic entry happens through a patent challenge, what we find is that the generic price is typically lower than the brand price, but it's not as low as sometimes people say. It's typically like a 20 percent discount to the brand for the first six months when only one generic is on the market. [3] We do find that substantial sales quantity shifts to the generic, within the second month often more than half of market shifts to the generic. By the sixth month it's over 70 percent. So, the brand loses a huge amount of their market share very quickly, even when they are only competing with one generic. Obviously, the generic company likes getting this market share, the brand dislikes losing it, and consumers appreciate the lower prices. If an agreement delays entry of the generic, all the effects reverse. It's bad for consumers, and it would be bad for the generic company but it would be good for the brand. If there's a payment from the brand to the generic company to compensate the generic company for losing those entry benefits, then both the brand and the generic company can benefit and consumers are still left with the harm. And so obviously, that's our main concern in these situations.

4. Now, how does this tie with the authorized generic question? What typically happens with the authorized generic is if the generic company is allowed to come in and compete with the brand, the brand will typically launch their authorized generic, almost on the same exact day as the independent generic enters the market. And so you get some competition there in the generic market. The generics are competing obviously with the brand, but now we also have the authorized generic competing with the independent generic. This slide shows estimates of how much the price of the generic drug, and only the generic drug, changes when an authorized generic enters into the market. We see decreases in retail prices somewhere near 5 percent or less, and we see somewhat bigger wholesale price decreases. The main takeaway from this slide is that it appears that purchasers of drugs do benefit when an authorized generic enters the market; prices come down; more competition is good. That's something we would expect.

5. If it's good for consumers, what are the potential concerns? I can think of two main potential concerns. One of the concerns is that there might be competitive harm from the authorized generic entering the market. Another possible concern might be that there could be competitive harm from the authorized generic not entering the market. So, I think that covers most of the bases. The first concern we might have when the authorized generic enters the market is that they might be doing it in order to deter patent challenges. I'm an independent generic and I'm looking to challenge the patent of an incumbent brand. If I think I'm going to have the generic market to myself for the first one hundred and eighty days, that seems like it could be a very profitable situation. If I now anticipate that rather than having that generic market to myself, I'm going to be competing with the authorized generic, that's somewhat less attractive to me. How much less attractive is something we'll get to in our research. But at least on the margin, if I'm an independent generic thinking of challenging a patent and I'm just on the border line between challenging and not, you can well imagine that my expectation about whether I am going to be competing with the authorized generic or not would potentially impact that decision. So there is one concern that brand companies are doing this just to deter patent challenges. Obviously, the second concern is that, and this is something we'll provide evidence on, is that branded companies sometimes make a promise not to compete via an authorized generic as part of one of these patent dispute settlements. In lieu of a payment to the generic company, they can just say "If you delay your generic entry three years we will promise that at that time we will not compete with you with an authorized generic." So it can be a way to get some compensation to the generic company.

6. The results I'm presenting here come from an interim report that we've just released last June. We're working on a final report that will hopefully get into some of these questions in more detail, looking at the question of entry deterrence. Relating to patent challenge deterrence, one of the pieces of evidence that I found relatively convincing to suggest that perhaps there were other things going on, is to look at when we see authorized generics entering the market. We split the markets broadly into two categories, big markets and small markets, just by literally taking the biggest half of the markets and the smallest half of the markets. And looking to see how often we saw authorized generics entering these markets and what we see here is that authorized generics enter very frequently in the big markets, 80 percent of the big markets we saw an authorized generic entering, and whereas with the smaller markets, we see them entering only 33 percent of the markets. My interpretation of this, at least as far as it relates to the deterrence question, is that entering with a generic drug into these very large markets is extremely attractive to generic companies, even when they're facing competition from an authorized generic. In these big markets, that is still a very profitable proposition for the independent generic that gets to enter this market and compete with the authorized generic during the 180 days. If you were going to try to deter entry by deterring patent challenges, you wouldn't be likely to be able to deter challenges on the blockbuster drugs. Those are just so attractive that this threat of facing competition from the authorized generic would not be enough to deter a potential entrant from challenging the patent there. The place where you would be likely to be able to see some entry deterrence might be in the smaller markets, where markets are already relatively small if you are going to sink the cost of getting to patent litigation, you might need to have a monopoly on the generic market there for the 180 days in order to justify the expense of the patent challenge. Here we see relatively fewer authorized generics entering. This does not necessarily tell

3 This estimate and the estimates quoted throughout the rest of this speech come from: Federal Trade Commission, Authorized Generics: An Interim Report, issued June 24, 2009. http://www.ftc.gov/os/2009/06/P062105authorizedgenericsreport.pdf.

us all of what's going on, but if nothing else, it does suggest that branded companies have some incentive to enter with authorized generics that is not related to entry deterrence question, because they are not likely to be deterring anything in these blockbuster markets.

7. Just to put some numbers to some of these incentives, we went through and we split out the markets, where we saw authorized generics entering during these 180-day exclusivity period, so this would be a situation where there would be a brand drug on the market, and for these 180 days there would be an independent generic on the market and also the authorized generic. On the second row, only the brand and the independent generic are on the market. These numbers are the wholesale expenditures, used as a measure of the revenue of each of these companies, relative to the revenue of the brand prior to the generic entry. We are looking to see how the market changes relative to the pre-generic entry situation. Immediately what we see when the first-filer generic (the independent generic that challenged the validity of the patent) gets to come on the market with the 180-day exclusivity, during the 180 days they get revenue that is equivalent to 61 percent of the revenue that the brand was making prior to generic entry. When they face competition from an authorized generic, that number drops to 33 percent. So, more or less their revenue from entering this market gets cut in half when they face competition from an authorized generic. So it's a big impact on them. The authorized generic, obviously picks up some of that market share. And surprisingly, it does appear to have some impact on the brand's revenue, whether an authorized generic enters or not. This is literally the brand, not the brand company but the revenues associated with the brand drug. This is a result, incidentally we're looking further at in our final study to see what's going on here, this sort of surprising me, the magnitude of this, to be honest. But if we take these values for what they are, if we look at the brand company, the brand company launches the authorized generic, and they get the branded sales, the net effect on the brand company of launching the authorized generic, these estimates would tell us that their total revenues go up by ten percent. So, it's beneficial to them to launch this authorized generic. And again, this does not really close the loop on the entry deterrence question, but it does suggest that if we're thinking in terms of a predatory pricing sort of test, we want to know if they are losing money now in order to deter some action in the future. That doesn't appear to be the case. It appears that they are probably making money by selling the authorized generic.

8. The arguments surrounding patent litigation settlements are quite well-known. There certainly can be justifications for these sorts of settlements, but they obviously can be competition concerns. The last point about the incentives relating to authorized generics allows us to really add something new to this discussion, which is that, this previous slide showed that the brand company could promise not to sell an authorized generic and give up the equivalent of ten percent of their pre-generic entry sales. The benefit to the first filer generic from that would be that they would get extra revenues equivalent to 28 percent of the brand's pre-entry sales. If we want to think about this is a bargaining chip that the brand can offer to the generic challenger, it's a lot more efficient than offering them money. If I want to transfer the equivalent of 28 percent of my pre-entry sales to a generic company, just with a direct payment, I have to give up 28 percent. If I promise not to launch an authorized generic, I'll only have to give up something that is worth ten percent of my pre-entry sales and they get 28 percent. So the incentives present a relatively attractive way to offer compensation to the generic company. Not surprisingly, we see firms using this in settlements.

9. At the Federal Trade Commission we are in this nice position that pharmaceutical companies, when they reach these sort of patent dispute settlements, have to file the terms of that settlement with us. We did an analysis of those settlements. This graph indicates how many times we saw settlements that involved a promise by the brand company not to launch an authorized generic in exchange for some agreement about when the first filer would have entered with their own generic version of the drug. So, we see it pick up greatly in 2007 and then drop off somewhat in 2008. I'd hesitate to label this as a trend because these are relatively small numbers, we're talking about nine and five settlements in the later years. The point is that promises not to launch an authorized generic have gone from not being unheard of several years ago to being not uncommon now.

10. One pattern that clearly emerges when we look specifically at the drugs on which these types of settlements are reached, is that the settlements that delay authorized generic entry more than six months occur only on relatively small drugs. On the horizontal axis of this graph is the number of months that the brand company is promising to withhold its authorized generic from the market, starting from the day the independent generic enters the market. For instance, this ten indicates that ten months after the independent generic launches, the brand would be allowed to launch their authorized generic under the terms of the settlement. On the vertical axis are the sales of the drugs. This drug up here represents a little over 5 billion US dollar in sales, so it's a very big drug. I think there is an interesting pattern here. An awful lot of observations are right here on this vertical line, which happens to be six months, corresponding to the 180 day exclusivity period that we talked about. We see a lot of settlements where the brand will agree not compete with the authorized generic during this 180-day exclusivity period. Those agreements at 6 months are both the median and mode of the length of promises to not compete. The other pattern that really sticks out here is that the agreements for the relatively big drugs tend to be short in duration. We don't see 20 months or 30 months out here for any of these larger drugs, they are all 6 months or less. I think there's an obvious explanation for that. After 6 months typically on these really big drugs like this drug here, you would see ten other generics entering the market in the 7th month. And so a promise by the brand company not to launch an authorized generic in month 7 on this drug would really be of no value to the first filer, because there are going to be something like ten other generics in the market and whether there is an eleventh firm in the market is not something that would be terribly valuable to the first filer generic company. For these drugs out here, these relatively small drugs, it's not unusual to see only maybe one additional generic entering those markets;

sometimes no other generics will enter the market. It could very well be of interest to an independent generic to keep the authorized generic off the market for a long time here because, let's not forget, these authorized generics are sold by the branded companies that have already done the research how to make this drug, they already have their production set up. So they have virtually no entry cost to issue on the authorized generic version of the drug. They seem to be the most likely entrant in this market. Other independent generic companies would have to spend time and money getting authorization from FDA to market the drug and would have to source the materials and all this. The branded company has all of that taken care of, so they are the most likely entrant. And so on these smaller drugs here, it can really be a valuable promise to the independent generic that this most likely entrant isn't going to come in a market where we wouldn't necessarily expect to see that many entrants. So, I think, the pattern makes some sense.

11. Ok, so just to conclude, let me summarize where I come out on authorized generics so far. First, the most obvious thing, I think, for many of us is competition authorities is that entry by authorized generics appears to lower prices and that's a good thing for consumers. We want lower prices. There is some harm to a competitor, the independent generic; they do lose revenue from this. We typically as competition authorities are not concerned about harm to competitors, but harm to the competitive process. In this instance it appears to be profitable for the brand to enter in this way, so entry deterrence does not appear to be the sole motivation for launching authorized generics. And then finally, these promises not to launch an authorized generic do seem to be a component of the pay for delay sorts of settlements. As we continue our investigation into those settlements we have to be mindful to these impacts. Thank you for your time. ■

III Lisbon conference on competition law and economics
PANNEL V
14-15 February 2010

Competition policy and single firm conduct: What consequences in terms of enforcement?

Bo Vesterdorf
bovesterdorf@gmail.com
Consultant

Abstract

In this paper the complexity of the interpretation and application of article 102 TFEU is discussed to underline the difficulties undertakings may have in trying to predict whether their conduct might violate article 102. In the light of such an degree of uncertainty regarding possible violation, undertakings may in order to be on the safe side prefer to abandon competitive conduct which after all may not be in violation of the competition rules and this may lead to stifle competition unnecessarily. Furthermore, it is submitted that for the same reason fines should not be imposed unless intent or gross negligence is clearly demonstrated, in other words no fines in case of simple negligence which the undertakings may easily become guilty of due to the unclear and imprecise scope of article 102.

1. Thank you for inviting me once again to speak at a conference organised by the Portuguese Competition Authority. It is a pleasure to be back in Lisbon and to have to talk on a subject which I find particularly interesting, namely single firm conduct, which means on art. 102 TFEU as it is now called.

There are several reasons why I find art. 102 to be of particular interest.

2. First, it is rule of law which is difficult to interpret and apply, but the application of which may have and, indeed, most often has a major impact on the market, market players and in the end consumers; second it is being used more and more and the cases in which it has been used recently illustrate its difficulty and have given rise to important and interesting questions of law; third, it is an article which easily lends itself to be abused; and fourth and finally, sanctions for violation of art. 102 have become extremely high, at least numerically, and in that connection the question must be asked if and when sanctions for violation of art. 102 are appropriate and just.

3. Let me now turn to the question of why I think that it is a difficult rule of law. The two decisive conditions which must be met for the article to be applicable, namely dominance and abuse, are not at all clear legal notions; they may on the surface seem to be reasonably simple to interpret, as in fact the Court of Justice has done in its case law, in for example the seminal Hoffman-La Roche and United Brands judgments. The problem is, however, that neither of the well-known definitions of the two notions given by the ECJ really helps us in such a way that it has become easier in practice to know with a sufficient degree of certainty whether the article is applicable in a particular set of circumstances, circumstances which may, and indeed very often, differ considerably from the circumstances of cases decided in earlier practice either by the Commission or on appeal by the EU-courts.

4. In Hoffman-La Roche, the ECJ stated that dominance exists when an undertaking to an appreciable extent is able to act independently of its competitors. Unfortunately, that does not tell us how independent the undertaking must be before it is to an appreciable extent. In other words, this needs to be examined and appraised in each individual case on the basis of the particular, normally complex factual situation. Furthermore, even before that, the undertaking in question needs to know which product market is the relevant one, at least according to the competition authorities. Even with the assistance of the Commission guidelines on dominance, that again is not in all circumstances easy to determine and predict with a sufficient degree of certainty. In order to be on the safe side, the undertaking may decide to base itself on the most narrow possible product market and thus, if mistaken, in reality wrongly consider itself to be dominant on the basis of a too large market share. If that happens, it may well impose unnecessary constraints on itself and thus contribute to stifle competition which is undesirable, which is what we want. On the contrary, we want even big and perhaps dominant companies to compete, even vigorously, as long as they do not recur to conduct that may be considered to be abusive.

5. However, the same difficulties apply to the interpretation of the notion of abuse. When the ECJ in for example United Brands underlined that an important element in that connection is to find out whether the undertaking in question has had a competitive conduct which cannot be called normal competition on the merits, this does not suffice because one has to determine what is "normal" competition and

which elements fall under the notion of the "merits". What is normal and which are the merits? When does it become not normal and where more precisely are the limits of what may be called merits?

6. Take the example of rebates. Using rebate systems are and have always been a very widespread and one of the most normal and natural elements of competition on price. There are very many ways in which a rebate system can be construed, but they all have the same purpose, namely that of attracting customers and hopefully have them come back, thus to create a certain loyalty. Also the most ordinary rebate system, where the rebate is only determined by the quantity to be bought on the occasion, has the aim of attracting customers and hopefully have them come back for more, thus to create a certain loyalty. If the undertaking is dominant, it follows from the case law that it may not use loyalty inducing rebate systems. However, whether or when a rebate system crosses the line between what is a legal rebate and what is not, is not always easy to predict, even with assistance of the Commission paper on article 102 and exclusionary practices and its suggestion to examine the question on the basis of the as-efficient competitor test. The result will in the end depend on the individual and subjective economic analysis made by one competition authority but perhaps not by another one, depending on how actively and intensively this or that authority wants to monitor and regulate the market.

7. In brief, article 102 is not a simple article to interpret and apply and in many situations it will be difficult for the undertaking to know whether its conduct may violate art. 102. This does not create legal certainty. That this is so, is perhaps best illustrated by the very Decisions of the Commission in application of art. 102. In these decisions, the analysis of and determination of the relevant product market and the question of dominance will normally run into dozens and dozens and some times hundreds of pages and the discussion on the alleged abuse will normally also cover a very large number of pages. It will very often entail complicated and some times very complicated economic analysis of products and markets, analysis that one would be hard pressed to presume or reasonably expect that undertakings would or should or even may be able to carry out in the course of their business activity. The recourse of the Commission to closely examine the economics of each case before taking a decision in order to avoid taking unnecessary decisions in cases where no market structure or consumer harm appears to be the demonstrated or possible effect of the conduct in question is, also in my submission, good and wise but it certainly does not make it easier for undertakings to predict what legal consequences their market conduct may have.

8. Secondly, the art. 102 cases which the Commission has decided during the last years have been cases of great interest, the outcome of which has had very important economic and/or legal consequences both for the undertaking in question and for the market. The Microsoft decision and judgment by the CFI, now the EGC, is of course a case in point, dealing with the difficult delineation between competition policy and IP-rights and with the highly important question of the distinction between legal product development and illegal tying. It is in no way simple to know in advance when you as an IP-right holder may be forced to give a license to a competitor because of the existence of so called extraordinary circumstances – whether or not such circumstances exist depending on the ex-pos facto appreciation of a competition authority – nor is it simple and easy to predict when the competition authorities may decide that what the undertaking thought was natural and logical product development is from the point of view of the particular competition authority illegal tying. In the recent Intel Decision by the Commission important questions regarding rebate systems have been decided, some of the questions apparently being relatively straight forward questions of evidence, others of a much more legally interesting character. It is going to be very interesting to see how the appeal in that case will de decided by the EGC.

9. My third point is that art. 102 is an article which easily lends itself to be used or even abused in a way in which it was not meant to be used. I have the impression that it happens, perhaps quite frequently, that competitors to a dominant undertaking suddenly find it useful to make a perhaps slightly frivolous or even fake complaint to a competition authority in the hope that the authority gets hooked and starts examining the case. That may be enough to make the perhaps dominant and perhaps abusing undertaking try – in order to avoid all the hassle - to settle the case with the complainant who may be very satisfied to get a nice, perhaps really unmerited, settlement payment and then as a consequence drop the complaint. And if the case does not get dropped, at least the complainant has created problems for the dominant undertaking which may also be fined heavily. Furthermore, in the light of the imprecise character of art. 102 and the unclear limits of its scope of application, it gives the competition authorities a very powerful tool, a tool that can in fact – if used too liberally – stifle competition because undertakings for fear of violating the article if found to be dominant submit themselves to perhaps unnecessary competitive constraints.

10. Fourth, what about enforcement measures? Clearly the approach of the European Commission is to fine and – it seems – to fine as heavily as possible. It certainly has handed out huge fines in recent cases. Even though percentage wise the fine, in for example Intel, may not have been anywhere near the ten percent limit, it was a huge, huge fine. It is obviously now for the EU-courts to decide if that fine was appropriate under the circumstances of the case.

11. Are fines the best weapon or are they the only weapon? They may be a good weapon, but they are in my opinion not the only one and perhaps not even the most appropriate in many art. 102 cases.

This brings me to what I think really needs to be stressed and taken into consideration by authorities when deciding whether or not or how to sanction abuse of dominant position.

12. When we are talking about fines, which as has been seen can be of enormous amounts and on top of that we are having to do with sanctions which are at least of a quasi penal character, it is in all countries, where the principle of

the rule of law is fundamental, a precondition for sanctioning violation of a rule of law, that that rule of law is so clear and precise that it does not leave people in reasonable doubt as to what is prohibited.

13. Except for a few types of abusive conduct, for example clear discrimination or evident predatory pricing, this is not the case of article 102 as I have tried to explain in my first point on the interpretation of that article.

14. In respect of a rule of law that is not sufficiently clear and thus lends itself to difficult questions of interpretation, it is inappropriate to impose sanctions for its violation except in very clear cut cases where intent or at the least gross negligence has been demonstrated. Under the present regime, according to Regulation 1, sanctions may be imposed even in the case of simple negligence. This I do find inacceptable, at least for first time offences.

It is consequently my submission that violation of article 102 normally only should be sanctioned by fines in clear cut cases and where as a consequence either intent or at least gross negligence can be clearly demonstrated. If, however, an undertaking which has been found guilty of violating art. 102 once but not fined, repeats the abuse, then it should obviously be fined and indeed fined severely for the new offence.

15. Furthermore, in my submission it follows form the complicated and imprecise character of article 102 that competition authorities should apply that rule of law with great caution. This is the more so since in many if not most of the cases the market can be expected to react and correct the situation in case of abusive conduct. New entrants to the market will appear unless there are too high barriers for entry, or existing competitors will step up their endeavours and thereby exert sufficient competitive pressure. The market is, I think, often better able to regulate itself than regulators.

16. The really worrying situation of abuse is where the dominant undertaking enjoys legal or de facto monopoly or near monopoly, in other words where there are no competitors or those that are there are so small that they cannot really exert any or sufficient competitive pressure on the dominant undertaking. However, even in such cases the market may well correct itself if the dominant undertaking only has a de facto monopoly or near monopoly. If such an undertaking abuses the dominance for example by raising prices excessively, we may normally expect new entrants to the market or customers may even react by stopping to buy unless of course the product is indispensable. This is where competition authorities must step in and react if necessary.

17. What I have said today must not be interpreted to mean that I disagree fundamentally or even in a more important way with the Commission with regard to its enforcement of the competition policy, but that I advice a higher degree of caution as to its application of art 102 in the future. It should consider more carefully whether there is a sufficiently real need to intervene and if it does intervene, only impose fines on the undertaking if it finds sufficient evidence of gross negligence or intent, intent naturally to be fined more severely than gross negligence. ■

Policing unilateral conduct

Thomas Barnett
tbarnett@cov.com
Co-Chair, Antitrust and Consumer Law Practice
Washington, DC

Abstract

Unilateral action presents some of the most difficult challenges in the enforcement of the competition laws. Agencies, courts, and academic scholars have expended significant effort in recent years to provide greater clarity and guidance on competition law compliance in this area. These remarks review recent developments in the European Union and in the United States, concluding that the EU has made progress while uncertainty may have increased in the U.S. Finally, two hypothetical negotiations between a customer and a dominant supplier are presented to illustrate some of the difficulties inherent in characterizing conduct as lawful or unlawful and in crafting effective remedies.

Les comportements unilatéraux constituent un des défis les plus difficiles à relever en droit de la concurrence. Ces dernières années, autorités, tribunaux et universitaires ont déployés des efforts considérables pour apporter plus de clarté à la matière. Les remarques qui suivent soulignent les progrès réalisés en Europe et l'incertitude croissante aux Etats-Unis en prenant appui sur de récentes affaires. Enfin, deux cas de négociations hypothétiques entre un client et un fournisseur en position dominante sont présentés afin d'illustrer certaines des difficultés inhérentes à la qualification de la conduite abusive et à l'élaboration de remèdes efficaces.

1. I appreciate the opportunity to address this audience – and an excuse to visit Lisbon is always welcome as well. It is a beautiful city, and this is a terrific conference. I will talk first about recent developments and then address from a bigger picture perspective some of the issues that I see in the unilateral conduct area. I am going focus in particular on how to characterize conduct as lawful or unlawful, good or bad, pro-competitive or anti-competitive and then offer some general observations.

2. This is a hard area. It is one of the least well defined areas of competition enforcement, but it is nonetheles an important area for enforcement. I previously have said that it's a big challenge, but it's a challenge that we need to and should undertake. It's hard because it is hard to draw the line in an appropriate place. And by appropriate I mean the line that is going to maximize consumer welfare. We want to leave companies room to compete vigorously, to innovate, and to be aggressive, but at the same time not let them, if you will, abuse their dominant position, restrict competition, and harm consumer welfare. Because it is a hard line to draw, I think it is incumbent on all of us – the enforcement community in particular, but for all of us in general – to try to develop better guidance in this area.

3. Over the last five years or so, there has been a tremendous amount of time, effort, and resources put into trying to develop better guidance. One point has become clear through these efforts: We do not yet have a consense on a single, general test for determining when unilateral conduct violates the competition laws. It is not enough to say that, if you have a large share of the market and you do something that harms a competitor, you could be in trouble under Article 102 or Section 2. A prior speaker noted that the European Commission has not settled on a general test for defining where to draw the line for finding violations. There was reference earlier to the report – the section 2 report we call it – that the Department of Justice issued, which reflected years of effort by both the FTC and U.S. Department of Justice. The conclusion that the DOJ also reached was that we do not yet have a general test we can advocate. The report suggested that, at least until we learn enough to develop a satisfactory general test, it would be better to focus on less ambitious conduct-specific tests.

4. Let me turn now to the current status of these isssues in Europe. There has been substantial progress that has been made, and I credit the European Commission in particular for a number of steps it has taken in this regard. The Commission bears the bulk of the responsibility for developing better guidance on dominance issues. The Court First Instance – it is now the General Court – decision in the Microsoft case does not provide clear guiding principles for determining violations. To be clear, I understand why the court was not in a position to offer more specific guidance. After all, as we have been discussing, it is a very difficult goal to achieve.

5. The Commission has issued a paper on unilateral conduct issues that provides useful guidance. The evolution in the Commission's approach on this front, however, illustrates some of the challenges to providing clarity while adhering to a legal standard that will benefit consumer welfare. The Commission used to be much more clear about where the line was in many instances because it had issued a list of absolute prohibitions, a black list. Although that list provided clarity, I would suggest it was not drawing the line in the appropriate place. By putting setting forth such broad prohibitions, the Commission almost certainly was prohibiting conduct that could be beneficial for consumers. Recognizing this concern, the Commission has undertaken a major shift to an effects-based analysis in which it finds a violation only if there is evidence to demonstrate an anti-competitive effect (*i.e.*, harm the consumers).

6. The move to an effects-based analysis is, in the long run, a beneficial change, but it presents greater challenges in terms of providing guidance. The European Commission is struggling with this issue. A pure effects-based analysis can lead

to the unhappy perception that it all depends on what the economists decide at the end of the day, which is not enough guidance for the business community. In my view, many of the remarks you have heard today reflect an effort to point to, if not a safe harbor, something in that direction that can give the business community a degree of clarity and comfort. These efforts are good and useful and should be encouraged. Nevertheless, there is still a lot to be done.

7. On the US side, I would describe our situation as one of decreasing guidance and greater undertainty. The area is more confused today than it has been for many years for several reasons. The first reason is that the Department of Justice issued the section 2 report I referenced above to try to provide greater guidance. The report tried to describe what courts have said, tried to describe the various views in the overall debate, and, where possible, to describe where the DOJ thought the law ought to go. I am not here to defend what the report said or didn't say about where lines should be drawn – that is not my point. My point is to note the effort to enhance the clarity and predictiability of the lines, wherever they might be drawn, under an effects-based approach.

8. My successor withdrew the report in one of her first acts. If she disagreed with some of the conclusions, she certainly acted appropriately in making public that she did not agree with portions of the report. What she did not do, however, was set forth an alternative view of how to assess liability under section 2. Further, in her remarks, she cited to cases that had been decided decades ago without also addressing more recent Supreme Court cases that addressed related issues. The net effect was to create uncertainty at least as to how the DOJ would seek to enforce section 2. It is important to put these observations in the proper context. It takes time to work through these issues. I am not saying that the new administration necessarily should have issued a new report or guidance statement by now; I am merely observing that it is now less clear where the Department of Justice would draw the line with respect to unilateral conduct cases.

9. Similarly, the US Federal Trade Commission did not join the section 2 report at the time it was issued, but it has not issued its own guidance document in this area. At the same time, the FTC has been working to expand the application of Section 5 to unilateral conduct cases. Let me explain briefly what that means. The U.S. Federal Trade Commission does not enforce directly the same competition statutes enforced by the Department of Justice. The Department of Justice, for example, enforces the Sherman Act, which is equivalent your Articles 101 and 102. The FTC enforces Section 5 of the FTC Act, which prohibits unfair trade practices. It is a broad delegation of authority. As an example, the whole area of consumer protection that has to do with fraud and deception is addressed under the same statute. Those are not competition cases, but they fall under the same statute. Traditionally, when the FTC has dealt with a competition-related complaint, it has looked to the principles of the Sherman Act in deciding whether or not the challenged conduct was an unfair trade practice.

10. The FTC has always maintained that it has the authority to pursue competition cases under Section 5 even though the challenged conduct would not violate the Sherman Act, but it has rarely exercised that authority. The FTC currently is exploring ways to expand the exercise of its authority in competition cases.

11. As an example, you will see if you look at the complaint issued against *Intel* and the statements by Chairman Leibowitz as well as the concurring statement by Commissioner Rosch, the FTC expressly included a pure Section 5 claim in additino to a competition-based claim. Again, my point is not to debate whether it is a good direction or a bad direction, but to illustrate that, if you are a company trying to decide what actions might violate the competition laws, you now have greater uncertainty. You now have to deal with a potential stand-alone Section 5 claim by the FTC that is not tied to the traditional notions of competition analysis and that has not been fleshed out in any detail.

12. You will see speeches from FTC commissioners acknowledging this uncertainty and acknowledging the responsibility to set out limiting principles for this approach. As of now, they have not achieved this goal. The bottom line result is that we have great uncertainty in the United States, particularly with respect to how the competition enforcement agencies will address unilateral conduct issues. We likely have greater certainty in the U.S. with respect to how the courts will address these issues. Our Supreme Court has addressed a number of unilateral conduct issues and has set forth relatively clear guidance in a certain areas, such as predatory pricing and price squeeze claims.

13. Now, I will turn to some of the challenges inherent in policing single-firm conduct. To illustrate some of the issues, I am going to talk through two hypothetical negotiations. My disclaimer is that similarities to actual companies are intentional, but these are stylized facts that do not accurately depict any particular case.

14. My first hypothetical is the "aggressive customer" case. We start with Supplier A, which sells widgets. It has an 80% share of the relevant market, and we will assume for present purposes that there are additional factors that make it a dominant company. Supplier B accounts for the remaining 20% of the relevant market. Next, we have Customer X, a big purchaser that buys a million widgets a year at a current price of $100 dollars per widget for a total of $100 million dollars per year. Customer X hires a new CEO, and he announces a new program: "*I am going to increase profitability to please our shareholders. A key element of my plan is to cut costs*." The CEO directs his or her managers to go to all of their significant suppliers and say the following: "*If I do not obtain at least a 10% price reduction on my procurement, I will lose my job*". Customer X now goes to Supplier A and says: "*We've been paying $100 per widget, if you don't cut the price to $90, we are going to take all of our business away.*" Meanwhile, the managers of Customer X had quietly approached Supplier B to try to gain some bargaining leverage and said "*We would like to shift our business to you, provided that you will sell to us for $90 per widget*." Supplier B had responded: "*The best price we are going to be able to offer is $110, and we can only sell you up to 250,000 units because we can't make anymore.*" So Supplier B is not an alternative for this Customer X.

An important fact to note is that Supplier A does not know about the cost and capacity limitations faced by Supplier B. As a result, Supplier A caves to what is a bluff by Customer X, reduces its widget price to $90, and continues to supply Customer X with all of its widget needs.

15. My second hypothetical is called the "aggressive supplier" hypothetical. This scenario involves the same basic set up with Supplier A and Supplier B as suppliers with the same market shares and Customer X as a big purchaser. Customer X does not, however, have a new CEO. Instead, as part of a new sales program, Supplier B approaches Customer X and offers to sell a million widgets at a price of $95 each. Customer X notifies Supplier A that it may be reducing its widget purchases. Supplier A responds by saying: "*Please wait and talk to me before you make a decision. We have had a great relationship for many years, and I want to continue to earn your business.*" Supplier A continues: "*If we reduce our price to $90, will you continue to purchase from us?*" Customer X decides that it likes the offer from Supplier A and accepts.

16. I observe that the result under both the "aggressive customer" and "aggressive supplier" scenarios is the same. Customer X is buying a million widgets from Supplier A at $90. Supplier B sells nothing to Customer X, either before or after. The question is whether you see some difference in the competition analysis of the scenarios. There would seem to be three possibilities: (i) both scenarios involve only lawful activity; (ii) both scenarios show anticompetitive behavior; and (iii) one illustrates lawful behavior and the other unlawful behavior. My intention in presenting the hypotheticals is to suggest that some may perceive a difference between the two scenarios. The difference would not be with respect to results in the market place because the results are the same: Supplier A supplying Customer X with all of its widget needs.

17. If the "aggressive customer" scenario seems to present less of a competition concern, is that because Supplier B was not an option for the customer? Supplier B quoted a higher price, $110 a unit, and had capacity constraints. If these facts are the basis for a distinction, however, it bears repeating that Supplier A did not know what Supplier B had said. This goes to the issue of how Supplier A can know when it is crossing the line. There also is a potential distinction based on motive. Supplier A in the first scenario was portrayed as something of a victim of a large and aggressive cutomer. In the second instance, Supplier A appears much more as an aggressor, reaching out to snatch a business opportunity away from Supplier B.

18. My point in going through these scenarious is to illustrate at least a flavor of what goes on in the real world. There may be instances where a company analyzes whether there is a contestable portion of its market, whether it has an opportunity to deprive a competitor of minimum viable scale, and whether it can adopt a rebate scheme to achieve such a result. Even if such instances occur, however, the scenarios similar to those that I described above also are likely to occur.

19. Trying to characterize what goes on in the real world is rarely neat and clean. It is not even clear in the two scenarios that Supplier A has an exclusive supply agreement with Customer X. We could further tweak the examples by saying that Customer X discussed its strategy internally. Its managers might discuss whether to buy from Supplier A or B and speculate that: "*if we reduce our purchases from Supplier A, we are afraid that Supplier A will increase its price. They will penalize us for giving business to Supplier B.*" This is all internal to Customer X. Supplier A may never know that this discussion took place. How should one take into account such evidence? Should it be given any weight in making a case against Company A in terms of trying to prove exclusionary or predatory behavior? It is sufficient for my purposes today to sya that the answer is not clearly yes.

20. I turn now to two final observations about policing unilateral conduct. First, I have said before and I willl say it again that remedy is critical. If you are going to define particular activity as a violation of the competition laws, you have to decide what remedy will address the violation. In my second scenario, for example, if you think that Supplier A violated the competition laws, are you going to constraint Supplier A from selling below a certain price level, such as $100 dollars a unit? If so, because Supplier A is the dominant supplier, many of the customers in the market may pay a higher price as a result, including Customer X. Can Supplier A condition a discount on a minimum purchase volume? If not, Supplier A likely could not price discriminate across customers and likely would charge a higher price than it otherwise would have charged to some customers.

21. Second, I want to echo one of the important points made by an earlier speaker. If we all agree that this is a hard area and that we are struggling to provide clear guidance to the business community, we also should be concerned about the possibility of excessively severe sanctions in this area. Without clear guidance and with competitively ambiguous conduct, severe sanctions can cause affirmative harm to consumer welfare as well as present fairness issues.

22. Thank you for the opportunity to speak. I look forward to our panel discussion. ■

Concurrences est une revue trimestrielle couvrant l'ensemble des questions de droits communautaire et interne de la concurrence. Les analyses de fond sont effectuées sous forme d'articles doctrinaux, de notes de synthèse ou de tableaux jurisprudentiels. L'actualité jurisprudentielle et législative est couverte par dix chroniques thématiques.

Concurrences

Editorial

Elie Cohen, Laurent Cohen-Tanugi, Claus-Dieter Ehlermann, Ian Forrester, Eleanor Fox, Laurence Idot, Frédéric Jenny, Jean-Pierre Jouyet, Hubert Legal, Claude Lucas de Leyssac, Mario Monti, Christine Varney, Bo Vesterdorf, Louis Vogel, Denis Waelbroeck...

Interview

Sir Christopher Bellamy, Dr. Ulf Böge, Nadia Calvino, Thierry Dahan, Frédéric Jenny, William Kovacic, Neelie Kroes, Christine Lagarde, Mario Monti, Viviane Reding, Robert Saint-Esteben, Sheridan Scott, Christine Varney...

Tendances

Jacques Barrot, Jean-François Bellis, Murielle Chagny, Claire Chambolle, Luc Chatel, John Connor, Dominique de Gramont, Damien Géradin, Christophe Lemaire, Ioannis Lianos, Pierre Moscovici, Jorge Padilla, Emil Paulis, Joëlle Simon, Richard Whish...

Doctrines

Guy Canivet, Emmanuel Combe, Thierry Dahan, Luc Gyselen, Daniel Fasquelle, Barry Hawk, Laurence Idot, Frédéric Jenny, Bruno Lasserre, Anne Perrot, Nicolas Petit, Catherine Prieto, Patrick Rey, Didier Theophile, Joseph Vogel...

Pratiques

Tableaux jurisprudentiels : Bilan de la pratique des engagements, Droit pénal et concurrence, *Legal privilege, Cartel Profiles in the EU...*

Horizons

Allemagne, Belgique, Canada, Chine, Hong-Kong, India, Japon, Luxembourg, Suisse, Sweden, USA...

Droit et économie

Emmanuel Combe, Philippe Choné, Laurent Flochel, Penelope Papandropoulos, Etienne Pfister, Francisco Rosati, David Spector...

Chroniques

Ententes

Michel Debroux
Laurence Nicolas-Vullierme
Cyril Sarrazin

Pratiques unilatérales

Frédéric Marty
Anne-Lise Sibony
Anne Wachsmann

Pratiques restrictives et concurrence déloyale

Muriel Chagny
Mireille Dany
Marie-Claude Mitchell
Jacqueline Riffault-Silk

Distribution

Nicolas Ereseo
Dominique Ferré
Didier Ferrié

Concentrations

Olivier Billiard, Jacques Gunther, David Hull, Stanislas Martin, Jérôme Philippe, Igor Simic, David Tayar, Didier Théophile

Aides d'État

Jean-Yves Chérot
Jacques Derenne
Christophe Giolito

Procédures

Pascal Cardonnel
Christophe Lemaire
Agnès Maîtrepierre
Chantal Momège

Régulations

Joëlle Adda
Emmanuel Guillaume
Jean-Paul Tran Thiet

Secteur public

Bertrand du Marais
Stéphane Rodrigues
Jean-Philippe Kovar

Politique internationale

Frédérique Daudret-John
François Souty
Stéphanie Yon

Revue des revues

Christelle Adjémian
Umberto Berkani
Alain Ronzano

Bibliographie

Centre de Recherches sur l'Union Européenne (Université Paris I – Panthéon-Sorbonne)

Tarifs 2010

Revue Concurrences | *Review Concurrences*

	HT *Without tax*	TTC *Tax included (France only)*
❑ Abonnement annuel - 4 n° (version papier) *1 year subscription (4 issues) (print version)*	445 €	454,35 €
❑ Abonnement annuel - 4 n° (version électronique + accès libre aux e-archives) *1 year subscription (4 issues) (electronic version + free access to e-archives)*	395 €	472,42 €
❑ Abonnement annuel - 4 n° (versions papier & électronique accès libre aux e-archives) *1 year subscription (4 issues) (print & electronic versions + free access to e-archives)*	645 €	771,42 €
❑ 1 numéro (version papier) *1 issue (print version)*	140 €	142,94 €

Bulletin électronique e-Competitions | *e-bulletin e-Competitions*

❑ Abonnement annuel + accès libre aux e-archives *1 year subscription + free access to e-archives*	575 €	687,7 €

Revue *Concurrences* + bulletin *e-Competitions* | *Review Concurrences + e-bulletin e-Competitions*

❑ Abonnement annuel revue (version électronique) + e-bulletin *1 year subscription to the review (online version) and to the e-bulletin*	745 €	891,02 €
❑ Abonnement annuel revue (versions papier & électronique) + e-bulletin *1 year subscription to the review (print & electronic versions) + e-bulletin*	845 €	1010,62 €

Renseignements | *Subscriber details*

Nom-Prénom | *Name-First name :* . . . . . e-mail *:* . . . . . .

Institution | *Institution :* . . . . . .

Rue | *Street :* . . . . . Ville | *City :* . . . . . .

Code postal | *Zip Code :* . . . . . Pays | *Country :* . . . . . .

N° TVA intracommunautaire/ *VAT number (EU) :* . . . . . .

Formulaire à retouner à | *Send your order to*

Institut de droit de la concurrence
25 rue Balard - 75 015 Paris - France | contact: webmaster@concurrences.com
Fax : + 33 (0)1 42 77 93 71

Conditions générales (extrait) | ***Subscription information***

Les commandes sont fermes. L'envoi de la revue ou des articles de *Concurrences* et l'accès électronique aux bulletins ou articles de *e-Competitions* ont lieu dès réception du paiement complet. Tarifs pour licences monopostes; nous consulter pour les tarifs multipostes. Consultez les conditions d'utilisation du site sur www.concurrences.com ("Notice légale").

Orders are firm and payments are not refundable. Reception of Concurrences *and on-line access to* e-Competitions *and/or* Concurrences *require full prepayment. Tarifs for 1 user only. Consult us for multi-users licence. For "Terms of use", see www.concurrences.com.*

Frais d'expédition *Concurrences* hors France : 30 € | *30 € extra charge for sending hard copies outside France*