2011
CPI Journal
Vol 7 • Number 2 • Autumn

CPI COMPETITION POLICY INTERNATIONAL

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

"POTENTIAL" DOWNSTREAM MARKETS IN EUROPEAN ANTITRUST LAW: A CONCEPT IN NEED OF LIMITING PRINCIPLES

John Temple Lang

Cleary Gottlieb

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

"POTENTIAL" DOWNSTREAM MARKETS IN EUROPEAN ANTITRUST LAW: A CONCEPT IN NEED OF LIMITING PRINCIPLES

Dr. John Temple Lang*

ABSTRACT

Under European Union competition law, a dominant company has a duty to provide important inputs to its competitors. The leading cases involved vertically integrated dominant companies, which operated both harbors and car ferry companies. They were ordered to give access to their downsteam competitors, the other car ferry companies that needed access to the harbors. In these cases it was clear that there were two markets: a market for the supply of harbor services to ferry companies, and a separate market for the supply of ferry services to travelers. If all the other conditions for a duty to contract are fulfilled, the dominant company cannot avoid the duty merely by arguing that it has never granted access before. This led to the statement that it is enough if there is a "potential market" for the supply of the input in question by the dominant company, if the other conditions are fulfilled.

This phrase has led to arguments by competitors requesting one of several products sold only in combination by the dominant company, or one specific input out of the dominant company's integrated operations, or the dominant company's principal competitive advantage.

In some cases competitors have claimed the right to use the dominant company's intellectual property rights, to produce or use the dominant company's products. In all these cases one important question is whether there is in any sense a "market" for an input that is used by the dominant company in the course of its activities. Since not everything that could be licensed or sold must be licensed or sold, there must be principles limiting the rights of competitors to demand access to the parts of a dominant company's operations that they need.

A number of substantive questions, and some procedural questions, arise in such cases. The European Commission's Guidance paper on exclusionary abuses makes it clear that there must be an "upstream" and a "downstream" market, but does not discuss or even fully list the other conditions of a duty to contract. This article argues that the "potential market" phrase means only that it is not a defense to show that the dominant company has never before made a contract of the kind suggested. If there is only one market on which the dominant company sells, a potential competitor has no right to insist on being given access to whatever inputs it needs to compete effectively on that market.

Access may be ordered only if an identifiable abuse of the dominant position has been committed. To prove an abuse, harm to consumers, and not only to competitors, must be shown. The duty to contract must be the appropriate remedy to put an end to the abuse. If no duty to contract can be shown, there cannot be a duty to contract on the basis of a tying argument, among other reasons because in tying cases the competitor wants to sell its products to third parties, and complains that tying prevents it from doing so. In the cases discussed here, the competitor itself wants to be supplied, so that it can produce the products that it wants to make.

* Cleary Gottlieb Steen and Hamilton, Brussels and London; Professor, Trinity College, Dublin; Senior Visiting Research Fellow, Oxford. Financial support from IBM is gratefully acknowledged by the author.

A number of cases have recently arisen in which competitors have sought access to inputs controlled by supposedly dominant companies to which the dominant companies have never previously given access. Competitors rely on the argument that the inputs could be the subject of a "potential market" under EU competition law. In some of these cases the dominant company sells a combination of two products that must work together, and the competitor wants to buy or to get the right to produce one of them, for sale together with its version of the other product. In other cases the competitor needs to obtain one specific input from what appears to be unified seamless production or distribution operations of the company that is said to be dominant. In what may be regarded as a third group of cases, the competitor wants the right to use the dominant company's principal competitive advantage, to use it in combination with other inputs already available to the competitor. This article considers the implications of the idea of "potential markets" in the context of EU competition law principles on the duty of dominant companies to contract.

Related questions arise under European competition law when a competitor or potential competitor of a company that is said to be dominant claims to be entitled to a compulsory license of an input consisting essentially of intellectual property rights, in order to use products or services produced by the dominant company. This article also considers some of those questions, in particular those which arise before the validity of the intellectual property rights in question is finally determined.

These questions may arise in proceedings for patent infringement brought by the company that is said to be dominant, or in a competition procedure before the European Commission or a national competition authority of an EU or EEA Member State. Some of the questions discussed here arise primarily because some courts that have jurisdiction to decide patent infringement cases have no jurisdiction to decide the validity of the intellectual property rights that are the subject of the proceedings. Corresponding questions arise in procedures before competition authorities, none of which have competence to decide the validity of intellectual property rights.[1]

Under certain circumstances, not yet very clearly or fully defined in the judgments of the European Court of Justice in several well-known cases, European competition law imposes on a company that has been found to be dominant a duty to grant a compulsory license of intellectual property rights.[2] If those circumstances do not exist, European competition law imposes no obligation to license (except in standards cases under Article 101 TFEU, which raise different issues, not considered here[3]), and the conflict between the supposed intellectual property right and European competition law does not arise.

National competition law under Regulation 1/2003 may be stricter than Article 102 TFEU, that is, it may impose more onerous obligations on a dominant company than those imposed by EU law.[4] But even in a Member State with stricter rules on unilateral conduct of dominant companies, the issues discussed here are likely to arise.

I. "POTENTIAL MARKETS"

The duty to contract is normally considered to arise primarily in situations in which there is an upstream market producing inputs, services, or raw materials, which are then sold to companies for use in a separate downstream market. The original examples were harbor operations that provided harbor facilities to car ferry companies and other transport operations.[5] In such cases the two markets are clearly distinct: they involve different products and services, and the buyers in the two markets are different.

These cases, in which the phrase "essential facility" was first used officially in European competition law, were all cases in which the abuse alleged consisted essentially of discrimination by the harbor operator in favour of car ferry or other shipping companies associated with it. It was not until later that cases arose in which a competitor wanted access to something that the vertically integrated company had never before supplied outside its own group. These cases involved the Commission applying Article 102(b) TFEU (on foreclosure) instead of Article 102(c) TFEU (on discrimination), but the significance of this does not seem to have been fully understood. The Commission paid little attention to abuse of dominant positions until the Discussion Paper was adopted in 2005.[6]

Cases of first refusal to contract had been approached without an adequate intellectual framework.

The idea of a "potential market" arose in situations in which it was said that a dominant company had operations which, although at first sight appeared unified, should be analyzed as consisting of an upstream stage producing an input or facility and a downstream stage using the input or facility.[7] Competitors wishing to enter the supposed downstream market, or to obtain advantages for use in that market, argued that the fact that the dominant company in question had never given access to the input or facility to any user not associated with it should not be a defense. The difficulty, of course, is that many companies that are not usually thought of as vertically integrated have operations that consist essentially of producing a raw material, an intermediate product, or a component for incorporation in a final product; combining hardware and software; or selling a complex final product, such as a car, consisting of a great number of components designed, manufactured and assembled in a particular way.

Some cases were relatively clear. The fact that one particular dominant harbour operator had never given access to any car ferry company that was not associated with it would not be a justification for refusing access if the other conditions required by Article 102 were fulfilled, because many other harbour operators do so (and also because the two markets are so clearly distinct). But if no company resembling the supposedly dominant company had ever given access to outside interests anywhere in the world, and if the operations producing the supposed input had never been considered separate or downstream from the rest of the company's activities, it was not easy to see what principles, if any, should be applied.

A) IMS HEALTH

The facts of the IMS Health case are well known, and have garnered much commentary.[8] IMS Health had compiled a specialized map of Germany designed to relate the places where pharmaceutical products are prescribed to the places in which they are bought. Pharmaceutical companies used this map to estimate the effectiveness of their sales representatives, who talk to doctors and hospitals, and not to the patients who buy the medicines. The sales data analyzed using this map were available to any company that wanted them, but NDC Health, a competitor, complained that the map was copyrighted and that the pharmaceutical companies preferred the IMS map to any other. IMS Health had never given a copyright license to anyone.

The question was whether it could have any obligation under what is now Article 102 TFEU to do so.

Advocate General Tizzano first recalled that in RTE-ITP[9] and Bronner[10] the supposedly dominant companies had never previously sold or licensed the input requested separately. He said,

"Thus in applying the case law cited on the refusal to grant a license I consider it to be sufficient that it is possible to identify a market in upstream inputs, even where the market is a 'potential' one only, in the sense that operating within it is a monopoly undertaking which decides not to market independently the inputs in question (notwithstanding that there is an actual demand for them) but to assert exclusive rights over a downstream market by restricting or eliminating all competition on that market.

"To take a classic example of the essential facility doctrine, it is instructive to consider the case where access to a port is indispensable in order to be able to provide maritime services in a given geographical market. For the purposes of such a case it may be assumed that the owner of the port uses that infrastructure on an exclusive basis in order to secure a monopoly over the market for maritime transport services refusing without any objective justification to provide the necessary port services to arms'-length undertakings... In such a case the case law on the refusal to grant a license must apply irrespective of the fact that the port services are not offered on the market... by its conduct it would be eliminating any competition on the secondary market."[11]

Tizzano continues:

"Since... in order to be able to identify a market for upstream inputs it is not necessary for them to be marketed independently by the undertaking controlling them.... [S]uch a market may always be identified where (a) the inputs in question are essential (since they cannot be substituted or duplicated) to operating on a given market (b) there is an actual demand for them on the party of undertakings seeking to operate on the market for which those inputs are essential."[12]

He goes on to say that there is no duty to license when the competitor plans only to produce goods or services duplicating those of the dominant company.The Court of Justice said,

"It appears therefore, as the Advocate General set out in points 56 to 59 of his Opinion, that, for the purposes of the application of the earlier case law, it is sufficient that a potential market or even a hypothetical market can be identified. Such is the case where the products or services are indispensable in order to carry on a particular business and where there is an actual demand for them on the part of undertakings which seek to carry on the business for which they are indispensable. Accordingly it is determinative that two different stages of production may be identified and that they are interconnected, inasmuch as the upstream product is indispensable for the supply of the downstream product."[13]

A paragraph in a judgment in a case under Article 267 TFEU should not be treated as if it were legislation. Judgments in Article 267 cases serve only to answer the specific question that has been asked, in the context of the specific facts from which the question has come before the Court. In Article 267 cases the Court does not usually set out to state the law comprehensively, and certainly not on issues that have not been argued and that do not need to be decided. It seems clear that it would be too simple, and indeed unjustifiable, to suggest that there are only three conditions for a duty to supply. These three conditions, (1) two interconnected stages of production; (2) indispensability; and (3) actual demand, would ignore other requirements that are equally well established in the case law of the Court.

It would be surprising if every input that resulted from a first "stage" of production could be demanded by any competitor or complainant who needed the input.

B) SUBSTANTIVE QUESTIONS

The Commission's Guidance paper14 on the Commission's enforcement priorities in applying what is now Article 102 TFEU says that,

"Typically competition problems arise when the dominant undertaking competes on the 'downstream' market with the buyer whom it refuses to supply. The term 'downstream market' is used to refer to the market for which the refused input is needed . . . This section deals only with this type of refusal."[15]

Having established that it is dealing only with two market situations, the Guidance goes on to say that the Commission will regard a refusal to supply case as a priority if:

(1) the refusal relates to a product or service that is objectively necessary to compete effectively on a "downstream" market;
(2) the refusal is likely to lead to the elimination of effective competition on the downstream market; and
(3) the refusal is likely to lead to harm to consumers.[16]

Although the Guidance paper clearly does not exhaustively list the conditions that are required for a refusal to contract to be contrary to Article 102, it is convenient to begin by discussing the conditions discussed in the Guidance paper.[17]

C) THE EXISTENCE OF A DOWNSTREAM MARKET: TWO PRODUCTS AND TWO STAGES OF PRODUCTION

The typical case involves a vertically integrated company that supplies an input for its own downstream operations, and is then also asked to supply the same input to a potential competitor of the dominant company's downstream operations. There must therefore be both a market for the supply of the input and a distinct market for which that input is necessary.

The Guidance paper identifies the possibility of an abuse even if the product or service refused has never been traded, if there is a "potential market."[18] This phrase requires explanation. The Court in IMS Health[19] did not need to explain it, because it was writing only in the specific context of that case, but the Commission should have done so, as the Guidance paper is intended to be generally applicable. Almost anything can, in theory, be leased, licensed or sold, and therefore anything might be a "potential market." Any owner of moveable or immovable property could sell, license or lease it, if it made sense for it to do so, but this cannot mean that there is a "potential market" for competition law purposes in every item of property in all circumstances. The mere existence of a demand cannot automatically create a duty to supply. If it did, the greater the competitive advantage given by the input in question, the greater would be the duty to supply it and share it with competitors, which would be irrational. Thus limiting principles are needed to identify true potential markets.[20]

The Court's words "two different stages of production" are helpful. There must be two separate and identifiable stages, rather than a continuous process. There must be an identifiable product or service at the end of the first "stage" that could be and usually is sold or licensed separately.

But the law must also answer the question of how to treat situations in which two components are produced simultaneously and then put together and sold in combination.

It would be irrational if there were a duty when the "stages" were consecutive, but not when they are simultaneous. The way that the manufacturing process is organized can hardly be the crucial question. Also, it cannot be enough that the end of the first "stage" is an intellectual property right. If it were enough, dominant companies would always be obliged to license all their intellectual property rights to every competitor that needed them, which could not be correct. A more precise or more limited concept is needed of the kind of input that can be a "potential market."

The key issue, it is suggested, is whether it would make sense—that is, whether it would objectively be economically rational—for the owner of the input requested, in the context of the business in which the owner is engaged and the use that it is making of the input, to sell it or license it to third parties. It may be rational to share the cost of an upstream facility, even with downstream competitors, particularly if the capacity of the facility is greater than is needed for the dominant company's downstream operations, or if the product to be sold or licensed is a by product ancillary to the main activities of the company.[21] It is not normally economically rational for a company to supply an asset that is used in its business to a "horizontal" competitor, that is, a direct competitor in the same market. A downstream market is needed for Article 102 to apply in refusal to supply cases because the dominant company's operations must consist of two separate stages: the supply of the input that is required, and its use to provide other, different, products or services to other buyers. In such situations the refusal may enable it to monopolize the downstream market.

If a company operates in only one market and has only one unbroken manufacturing process, however complicated, and only one product or set of products, there is no meaningful sense in which there is a "potential market" for sharing its assets or inputs with its direct competitors.

Common sense and case law confirm that there might be a potential market for sharing a byproduct of the dominant company's principal activities, or sharing the use of a facility with spare capacity, but not its most important inputs. In RTE-ITP, the information needed by the magazine was an incidental result of the television broadcasting, not the television companies' main activities. The information needed could be easily provided (and indeed, was being provided to daily newspapers) because the Magill magazine was in a market entirely different from that for television broadcasting.[22]

In Microsoft,[23] the information that it was ordered to provide concerned only interoperability, and not the core functions of the Microsoft products.

Because there is no duty to supply or license if there is no separate downstream market, a complainant needs to prove that the supposedly dominant company's operations consist of two parts. That situation might arise in a case not considered by the Guidance paper, in which the dominant company is horizontally integrated, producing two products or services that are linked to one another, and sells them both to the same buyers. Suppose that these two (or more) products or services are both needed by users for simultaneous use: neither works without the other. And suppose that the complainant plans to provide its version of one of these products, but wants to buy the other from the dominant company, or get a license to produce the latter product, using the dominant company's technology.

Again, the key question is whether there is in any sense a separate "market" for the latter product when it is produced by the dominant company. The answer seems clear. It is not normally economically rational for a company that sells a combination of two products to its customers to sell one of them to a competitor (or to license the competitor to produce it) merely to allow the competitor to combine it with the competitor's own version of the other product. That would make sense only in the context of a joint venture, or if the supposedly dominant company had a shortage of production capacity, or in anticipation of a merger.

So a horizontally integrated company is not in a situation essentially different from that of a vertically integrated company for competition law purposes, in this respect.

What may be another way of arriving at the same conclusion is to say that a duty to contract may not be imposed, even if the dominant company is vertically or horizontally integrated, if it would oblige the dominant company to share its principal competitive advantage and to lose its incentive to invest in the asset or input being shared.[24]

It could not be right to say that a competitor has a right to select the dominant company's principal competitive advantage or its principal asset and insist on getting the right to use it. That would mean that competitors would have the right progressively to take away the dominance of the company in question, which Article 102 clearly does not allow. This seems to be a more useful test than trying to analyze the stages of production in the dominant company's operations.[25] This approach is confirmed by considering the enormous difficulties of devising an appropriate payment if a dominant company's principal advantage was being shared on a compulsory basis, initially with one competitor, later perhaps also with others (because of the duty not to discriminate).

How much difference would it make if the only input needed was a license of an intellectual property right? Since the economic significance of a license would be to enable the complainant to use an asset or technology owned by the dominant company, the fact that formally only a license would be required would be unimportant. The license would simply be the means of giving access to the asset or technology in question.

Apparently similar issues can arise in the pharmaceutical industry with compound medicines, which are medicines that consist of two effective ingredients taken together. A complainant producing one ingredient may claim that the other ingredient is an essential facility, and is therefore needed to enable it to produce the compound medicine.

However, a distinction must be drawn between the case where a complainant wants supplies of a single product that is already produced and sold by the dominant company, and cases in which it wants a part of the dominant company's product or production process which is not sold separately, and for which there is therefore at first sight no identifiable market in existence. The mere fact that the dominant company sells a combination of two products and that a rival is able to produce only one of them is not an abuse, and no order to contract can be made.

A fortiori, it is not an abuse for a dominant company merely to refuse to share an asset or other part of its overall operations just because the competitor is unable to obtain the part it needs for its own activities.

The conclusion suggested is that if the "potential market" concept merely means that it is not a defense for a dominant company to show that it has never granted a license before, it is certainly correct. This is what the Advocate General said in IMS Health. Yet if the phrase is thought to mean more than that, it is hard to see what it could mean, and some limiting principles would clearly be needed. Any other meaning would be inconsistent with legal certainty.

II. ELIMINATION OF EFFECTIVE COMPETITION

In theory, if there is no "downstream" or other separate market for which the product or service is an input, the second condition stated by the Commission—the elimination of effective competition in that market—does not arise. It is nevertheless useful to analyze the connection between the refusal to license and competition. The refusal to supply or license may eliminate all competition from the complainant, if the input truly is essential to its operations. But the dominant company may be exposed to competition in the market in which it sells, even if that market is for the combination of two products, from other companies that produce them both. The question in refusal to license cases is not whether competition from the complainant is eliminated by the refusal, but whether all competition from all sources is eliminated.[26] If other companies individually or together produce, or have access to, the input that is said to be essential for the complainant, or to satisfactory alternative inputs, it is clear from the Bronner judgment that there is no duty to contract.[27] A dominant company is never obliged to remedy weaknesses in an individual competitor's business plan unless the dominant company has caused those weaknesses in some way.

If there is clearly only one market on which the dominant company sells, and there is no competition in that market, a potential competitor has no right to insist on being given access to whatever inputs it needs to compete effectively in that market. This is obvious, once it is stated. But its omission in the Commission's Guidance paper makes its conclusions seriously incomplete.

It is well-established in the EU case law that it is not an abuse to refuse to license an intellectual property right: there must be some "additional abusive conduct," a separate abuse.[28] This is so even if the effect of exercising the intellectual property rights is the creation of a monopoly. The fact that there will be no competition if a license of intellectual property rights is refused is not "additional abusive conduct," nor is it an "exceptional circumstance" justifying an order to license, as mistakenly determined by the Commission in its IMS Health interim measures decision.

A) HARM TO CONSUMERS DUE TO THE REFUSAL

Article 102(b), which is the principal and probably the only legal basis for the prohibition of foreclosure and exclusionary abuses (as distinct from discrimination cases) expressly applies only if there is harm to consumers. It is not sufficient for the complainant to claim that if it got a license or a contract, there would be one more competitor. If that were enough, there would always be a duty to license, which runs counter to established law. To say that one more competitor would be enough to justify a compulsory license would be to look only at static competition.

> *In all refusal to contract cases, it is is essential to look at dynamic competition.[29]*

Any duty to contract inevitably has implications for the incentives for further investment of both the dominant company and the companies with which it may be obliged to contract. It discourages the dominant company from investing, since the company will fear that success will require sharing the fruits of its investment. A duty to contract also discourages the companies contracting with the dominant company from investing, because such companies no longer need to invest in developing alternatives; instead, they can "free-ride." An important finding by the Commission in the Microsoft case was that compulsory disclosure of interoperability information would not reduce the incentives of Microsoft to invest, since Microsoft was obliged to disclose only the information needed for interoperability, and could continue to develop its systems. Nor would disclosure reduce the incentives of other companies to invest, because they would continue to be under competitive pressure from Microsoft and rival firms.[30]

Several of the leading judgments have considered whether the complainant can show that it plans to produce a new kind of product or service for which there is a clear and unsatisfied demand, which the dominant company is unable or unwilling to produce. This was the situation in the RTE-ITP case, involving an integrated weekly television programs guide.[31] It was not the situation in Bronner[32] or in IMS Health.[33]

If the complainant can make such a showing, the harm to consumers caused by preventing the development of the new kind of product is sufficient to constitute an abuse, provided the other conditions are met.

However, if the complainant plans to produce only a combination or a product that is essentially a copy of the dominant company's product, there is insufficient harm to consumers. Similarly, if there is no scope for non-price competition in the downstream market, there is no justification for a duty to contract. Harm to consumers must always be proved under Article 102(b) TFEU if the abuse consists of foreclosure or exclusion of a competitor. Yet it is important to recognize that harm to consumers, if it is serious, may be enough to create an abuse. So in RTE-ITP, the mere refusal to provide television program information was an abuse, because it made it impossible to provide consumers with a product for which there was a clear and unsatisfied demand.[34]

B) ARTICLE 102(b) TFEU: FORECLOSURE AND EXCLUSIONARY ABUSES

Article 102(b) TFEU prohibits conduct limiting the production, markets or technical development of competitors35 of the dominant company, if consumers are harmed. This is the Treaty definition of foreclosure and exclusionary abuse. Similarly, the Court in Microsoft said that this clause is not limited to cases involving a new kind of product, but also applies when, in effect, the dominant company's conduct imposes a permanent handicap on its competitors.[36]

The Court in the GlaxoSmithKline case[37] under Article 102 TFEU also relied on Article 102(b). If this handicap limits competition in a market for a new or improved product that competitors were already producing (or would produce, if the evidence that they would do so is strong enough), and which they would be under continuing competitive pressure to improve, there may be an abuse. It seems clear that a dominant company never has a duty to share, or part with, its principal competitive advantage, since that would deprive both it and its competitors of their respective incentives to invest and innovate.

> ***But it is nonetheless difficult, if not impossible, to visualize an abuse for which the appropriate remedy would be an order to share the dominant company's principal competitive advantage.***

An instance in which such a permanent handicap would be imposed is if a dominant company regularly makes changes in its products that causes them to work unsatisfactorily with competitors' products, and then refuses to provide new interoperability information promptly. This was found to be the situation in the Decca Navigator case,[38] and was thought to be the situation in the original IBM case brought by the European Commission.[39] Similar handicaps were imposed, according to the Commission, by AstraZeneca on its generic competitors by the withdrawal of the listings for some of its patents.[40]

C) NO DUTY TO CONTRACT WITHOUT AN IDENTIFIABLE ABUSE

Although it has been insufficiently emphasized by both the Commission and the Court, it is important to note that Article 102(b) can impose a duty to contract only when an abuse has been committed.[41] It prohibits only conduct creating a handicap or difficulty to which the competitors would not otherwise be subject. It does not create a duty to help competitors to overcome difficulties not caused or increased by the conduct of the dominant company. It is not illegal foreclosure merely to refuse to help a competitor.

In Bayer[42] the Court said:

"Under Article [102], refusal to supply, even where it is total, is prohibited only if it constitutes an abuse. The case law of the Court indirectly recognises the importance of safeguarding free enterprise when applying the competition rules of the Treaty where it expressly acknowledges that even an undertaking in a dominant position may, in certain cases, refuse to sell or change its supply or delivery policy without falling under the prohibition laid down in Article [102]."

This failure to distinguish between free enterprise and abuse is one of the most important omissions from the statements made by the Court in IMS Health[43] and by the Commission in the Guidance paper.

It is elementary, and it should be obvious, that Article 102 TFEU applies only when an abuse has been committed. No compulsory license or other remedy can be ordered under Article 102 TFEU unless an identifiable abuse has been proved. There is no duty to license merely to create one more competitor. It is not an abuse to refuse to license merely because there may otherwise be no competition in the short term, because that may often be the result in cases involving intellectual property rights.

If the dominant company has done nothing to make the market less competitive, it cannot be ordered to make it more competitive, and obtaining intellectual property rights for one's own inventions does not make the market less competitive. If the dominant company has done nothing to create a handicap or difficulty for competitors to which they would not otherwise have been subject, there cannot be a duty to contract. In other words, anticompetitive foreclosure must be proved before any remedy can be ordered, and the mere refusal to help a competitor is not anticompetitive.

Intellectual property rights cases, more so than in any other kind of case, recognize that the mere refusal to license a property right is not an infringement of Article 102 TFEU. There must be some "additional abusive conduct"[44] that constitutes a distinct and separate abuse, distinct from, and in addition to, the refusal to license.

D) A DUTY TO CONTRACT ONLY AS A REMEDY FOR AN IDENTIFIABLE ABUSE

This important and undeniable principle suggests another and better approach of looking at the case law that goes far to put everything into perspective. The fist question to ask is whether an abuse exists.

Once an abuse has been identified and proved, it is easier to answer the next question, which is whether a duty to contract—whether to sell, license or lease—would be the appropriate and proportional remedy for the abuse in question. This explains RTE-ITP,[45] Commercial Solvents,[46] and the discrimination cases. It entirely avoids the insuperable difficulties of basing all crucial distinctions on the nature of the "stages" in the production process.

There are a number of arguments based in law, economics and policy for this approach, which cumulatively are extremely strong:[47]

- This approach is based on the express words of Article 102(b): "limitation" (of the possibilities of rivals) and "prejudice" to consumers. Conduct which limits possibilities of rivals only in ways in which they would be limited anyway cannot be illegal. Rivals are already limited by having to respect intellectual property rights. The approach involves no new rules or concepts.

- It provides a rational, coherent and comprehensive basis for the relevant legal and economic principles, which should be broadly acceptable to competition lawyers and economists, and to intellectual property lawyers.

- It confines the concept of "abuse" under Article 102 to the three correct, useful and traditional categories under European competition law: exploitative abuses (Article 102(a), foreclosure of competitors (Article 102(b)), and unjustified discrimination between companies not otherwise associated with the dominant company (Article 102(c)).

- It seems reasonable to say that an abuse always involves some conduct of the dominant company. Mere inaction is not an abuse. Therefore a remedy must offset or eliminate the consequences of some positive action.

- It answers the following two questions: what "additional abusive conduct" is enough? If refusal to give access is only illegal when linked to such conduct, why not simply prohibit the separate abuse? The answer is that a compulsory license, when appropriate, is a more effective remedy.

- It avoids the insuperable difficulties of "balancing" the incentives to invest of the dominant company and its downstream competitors in the future. The Court in Microsoft carefully avoided undertaking this task, and it seems wise to avoid it.

- It harmonizes the interpretation of Article 102 TFEU with the well-established duty of parties to patent pools, joint ventures and standard setting agreements to license essential patents to non-parties.48

- It encourages use of a market-based remedy requiring little competition law supervision.

- It states a rule with built-in limiting principles, which are needed because of the vagueness and potentially broad scope of the concept of "potential markets." There would be no new concept of abuse and, as in the case of any other remedy for an abuse under Article 102 TFEU, the remedy must be an appropriate, proportionate and effective remedy to put an end to the abuse. The question of the appropriateness of a remedy arises on any view of the law. A remedy must be enough to put an end to the abuse effectively, but go no further.

- It provides a basis for distinguishing three types of cases from each other. The first is where the dominant company developed the property itself, when normally no duty arises to give a first license, and there is no duty except under Article 102(c) TFEU. The second category of cases is when a dominant company acquired the property and then deprived its competitors of access to it. In this group of cases, a duty to license is appropriate if the dominant company is substantially restricting competition.

The third group is dynamic competition cases, where the dominant company harms consumers by foreclosing potential competition to protect itself against technical development or against a new kind of product for which there is a clear and unsatisfied consumer demand.

- It gives the phrase "additional abusive conduct" a clear meaning, that of "abuse."

- It confirms that, in principle, it is never illegal in itself to refuse to license an intellectual property right, as the Court has repeatedly affirmed.[49]

- It has the important advantage of avoiding consequences contrary to policy. It would not lead to protecting competitors rather than competition, using competition law for regulatory purposes, or discouraging investment or innovation. These are serious risks to which European Union law has been exposed in recent years.

- It provides a rational basis for saying that a dominant company has no duty to facilitate companies which wish to copy, add on or imitate devices, unless it has taken steps to exclude them or create difficulties or handicaps for them.

- It allows a variety of justifications for refusal to license (including the defense that the dominant company will soon produce the new kind of product itself).

- It seems to be an approach on which European and US law could agree. This is important because it is often said that intellectual property rights are an area on which the two jurisdictions diverge.[50]

- It allows a distinction to be drawn between a compulsory license in a single market situation—which can be appropriate only if the abuse is in that market and which requires a very strong justification (since it would lessen dominance, as distinct from ending abuse)—and a compulsory license in a second distinct market, which is more likely to be proportional.

- It does not involve trying to use competition law to correct any defects which may be thought to exist in intellectual property law.[51]

- It provides a relatively uncontroversial rationale for the results in RTE-ITP52 and Microsoft.[53]

- As Mr. Justice Laddie said in Philips Electronics v. Ingman and Video Duplicating,

"The existence of the intellectual property rights may facilitate anti-competitive behaviour, but such behaviour consists of abusive interference with the market for a product . . . In prohibiting the conduct the court may have the power to intervene in the manner in which the intellectual property rights are exploited by the proprietor. This is to ensure that the proprietor does not continue the abusive conduct in relation to the products by the back door route of using the intellectual property rights."[54]

In short, a refusal to contract or to license is never an abuse in itself, but a duty to contract may be the correct remedy for some other abuse, once the abuse has been identified and proved. All the cases in EU law in which access has been ordered have involved identifiable abuses.

The abuse, once identified, and the duty to contract must be related in some way.

The only way in which they could be related is when the duty is a remedy to end the abuse. Imposing a duty to contract, even if no abuse had been committed, merely to create more competition, would be a regulatory rule unjustified by competition law principles.

This approach has another advantage. It largely avoids weighing up the effect of imposing a duty to contract on the incentives to innovate of the dominant company and of the competitors. Under this approach, such an inquiry arises only when the competition authority is considering whether an order to contract is proportional. It is easier and more appropriate in a judicial context to answer that question than to try to weigh up what sounds like a policy question of the relative importance of the two sets of incentives in the future.[55]

E) ACQUIRING THE ONLY EFFECTIVE ALTERNATIVE TECHNOLOGY

Two cases in which a duty to contract may be appropriate, but which fall outside the types of cases discussed above, should be mentioned.

The principle that a duty to contract must be the appropriate remedy for an identified abuse is illustrated by a situation in which a dominant company acquires the only effective technology which is an alternative to its own technology, or the only useful alternative input, and the acquisition is an abuse.[56] Competition, or potential competition, is suppressed. The dominant company may wish to suppress the alternative technology, or to use it to strengthen its own dominance. Even if the alternative were not used, its existence might constrain the dominant company, provided that it was owned by a non-associated company. So if a dominant company acquires the only significant alternative technology, the appropriate remedy is to order the company to sell or license it to a direct competitor. The dominant company might also need to be ordered not to use it for its own purposes.

F) LIMITING SUPPLIES IN ORDER TO RESTRAIN PARALLEL IMPORTS

In the *GlaxoSmithKline* judgment[57] under Article 102 TFEU, the Court held that a dominant supplier of medicinal products was not entitled to refuse to meet ordinary orders from wholesalers in order to prevent parallel imports into higher price countries.

The dominant supplier could only refuse to meet orders that are "out of the ordinary in terms of quantity." It could therefore be ordered, if necessary, to supply ordinary quantities, on the grounds that a refusal to sell would limit markets to the prejudice of consumers and would amount to discrimination that might ultimately eliminate a trading party from the market.

III. ARE THERE DIFFERENT RULES FOR INTELLECTUAL PROPERTY AND OTHER KINDS OF PROPERTY?

The question arises whether there are different legal rules on the duty to contract for intellectual property. The basis for such a distinction rests on the Court's repeated statement that a refusal to license an intellectual property right is not an abuse, and that there must be additional abusive conduct if there is to be a duty to license. The Court has not been required to articulate what differences, if any, there may be for other kinds of property. It is understandable that the Court's comments concerned only intellectual property rights, since they formed the subject of the cases involving first refusals to contract.[58] Intellectual property rights create legal monopolies (though not necessarily economic monopolies), and it is obvious that if there were always a duty to license an intellectual property right that created a legal or an economic monopoly, the rights given by intellectual property legislation would be completely transformed into mere rights to royalties. The Court thus needed to say that refusal to license such a right was not, in itself, an abuse. But that left unanswered the question whether corresponding rules apply to other kinds of property.

This question is easier to answer in the light of the fundamental rule explained above, that there is never a duty to contract or license unless it has been proved that an identifiable abuse, contrary to Article 102 TFEU, has been committed. If the abuse is discrimination, contrary to Article 102(c) TFEU,

the duty to end the discrimination may involve a duty to grant access on the same terms as those on which access has already been given.

In this respect, there is no reason to differentiate between intellectual property and other kinds of property.[59]

Cases of first refusal to give access to property or inputs other than intellectual property rights are unusual, simply because other kinds of property or inputs do not usually involve anything resembling a monopoly. But Commercial Solvents[60] and Bronner[61] did involve what were said to be monopolies, and RTE-ITP involved a monopoly of the television program information (and only incidentally a copyright).[62]

In Bronner the Court held that there was no duty to contract because the complainant had not proved that no alternative economic distribution system could be set up,[63] but the judgment seems to imply that if no other system were possible (and if the other conditions for a duty to contract were fulfilled), there would have been foreclosure, and a duty to contract.

In Commercial Solvents[64] and RTE-ITP[65] the Court held that abuses had been committed, and although the words were not used, it is easy to see that in each case they were foreclosure or exclusionary abuses. In Microsoft[66] the relevant duty was to provide the information needed for interoperability and, as in RTE-ITP, the intellectual property right was merely incidental.[67] There is nothing in any of these judgments to suggest that a refusal to give access to any kind of property, input or service can be an abuse in itself, without proof of any other abuse.

Almost all the reasons outlined above for saying that a refusal to license an intellectual property right is not in itself an abuse apply also to refusals to give access to all other kinds of property. The conclusion therefore is that intellectual property is not a special case, and that the rules discussed here apply equally to all kinds of property and inputs. Certainly, as far as "potential markets" are concerned, it is equally appropriate to use the concept in connection with both kinds of property. Intellectual property rights are no more or less easily sold or licensed than other kinds of property.

When the supposed abuse is foreclosure rather than discrimination, Article 102(b) TFEU does not suggest that different kinds of property should be differently treated.[68] Mere ownership of property, normally giving the dominant company an exclusive right to use it, is not an abuse, because it does not "limit" the markets, production or technical development of competitors.

IV. TERMINATION OF EXISTING SUPPLY ARRANGEMENTS

In cases involving termination of existing supply arrangements, since the dominant company has already supplied or licensed in the past, there are two stages in the production chain, and there is no need for an analysis of "potential" markets. Previous contracts show that there is both a market for the supply of the input and a market where that input is used in a distinct product.

The Commission's Guidance paper says that the Commission will apply the same criteria in cases of termination of existing supply arrangements as in cases where the dominant company refuses to supply a good or service which it not previously supplied to others.[69] It adds, however, that the termination of an existing supply arrangement will more likely be found an abuse of a dominant position than a de novo refusal to supply.

The Guidance paper gives two unconvincing reasons for treating a termination of existing supply more strictly than a de novo refusal to supply.

First, the company previously supplied could have made relationship-specific investments. This argument cannot be accepted, since it is not a competition law consideration, but a commercial or contractual one. Termination of supply could be a breach of contract and the contracting party could request damages for its investments, but this does not mean that it is necessarily easier to find an abuse from a competition law perspective.

The second argument in the Guidance paper is that in the past, the owner of essential input has found it in its interest to supply. According to the Commission, this indicates that supplying the input does not imply any risk that the owner receives inadequate compensation for the original investment.

The argument of inadequate remuneration for the dominant company could be a justification for termination, if it was objectively shown, but it does not explain why termination would be illegal in the absence of inadequate remuneration. The mere fact of having supplied once cannot create a duty under competition law to continue supplying indefinitely.

Under competition law, the effects of the termination on competition and on consumers should be the criteria for assessing whether or not the termination constitutes an abuse of dominance.

These cases can arise under Article 102(b) TFEU (foreclosure) or 102(c) TFEU (discrimination). If the complainant is the only one cut off from supplies, there may be discrimination. If everyone is cut off, there may be foreclosure on the downstream market. There may be ill effects for justifications for competition and for consumers, but whether the termination is justified will depend on the dominant company's reason for the termination.

Certainly, it should not be presumed that a refusal to supply is an abuse merely because a contract has been made already, as the Commission seems to believe. Such a result implies that a dominant company could be locked into a contractual arrangement. This would discourage such a company from supplying in the first place, which would be damaging to the economy.[70]

There are, however, some differences between de novo refusal to deal cases and termination of supply cases.

A) ELIMINATION OF EFFECTIVE COMPETITION FROM STOPPING SUPPLY

Clearly, existing competition is different from potential competition. If the dominant company stops supplying a player in the downstream market, there will be one fewer competitor on that market, if the input is essential and there is no other source of supply.

It could be harder for the dominant company to prove a valid business justification for the termination of the supply than it would be in cases of a de novo refusal to deal, since the dominant company found it economically rational to supply the complainant in the past.

But keeping an inefficient competitor in the market, or mere duplication or imitation of an existing product, is not sufficient to create a duty to resume supplies. The fact that it is sometimes pro-competitive to deal with a competitor does not mean that to stop doing so is necessarily anti-competitive.

B) HARM TO CONSUMERS DUE TO REFUSAL

A duty to resume supply should only be ordered if there would otherwise be harm to consumers.

In termination of existing supply cases it is appropriate first to look at the effect on existing (static) competition. There will always be some lessening of competition, if the dominant company cuts off supplies to at least one competitor in the downstream market. If many competitors remain in the downstream market, the consequences of the termination of one might be negligible.

If it is useful to look at dynamic competition, the effect of the termination on competition in the future is not likely to be significantly different from that in cases of first refusal to contract.

C) JUSTIFICATIONS FOR STOPPING SUPPLY

The mere fact that the company is dominant and terminates a contract, whether or not in accordance with contractual rules, is not sufficient to constitute an abuse. Either a separate positive act or the factual circumstances surrounding the termination can constitute the abuse.

The dominant company will always have a specific reason to terminate the contract. There might be acceptable reasons to terminate an existing supply arrangement, but it is also possible that the dominant company's motive is to reinforce its dominance or to extend it into the downstream market, as occurred in Commercial Solvents.[71]

If a dominant company wants to integrate forward and penetrate the downstream market, it is likely to commit an abuse if it wishes to monopolize the downstream market and, by terminating the contracts, is trying to eliminate its competitor(s) in this market. If the dominant company is already present in the downstream market, it might want to cut off supplies from its main competitor in the downstream market. In other words, if the only reason for termination is to eliminate competition in the downstream market, the termination is illegal.[72]

D) JUSTIFICATIONS DUE TO CHANGED POLICIES, TECHNOLOGIES OR CIRCUMSTANCES

There can be a wide variety of changed circumstances which lead the dominant company to terminate the existing arrangements. If refusal to make a first contract were justified, termination would normally be lawful under competition law. Any other approach would imply a presumption that termination is contrary to competition law, an unjustified conclusion.

One key problem arises where the dominant company wishes to terminate because it wants to go into the downstream market, and there is little scope for competition between it and the other contracting party. This would mean that if it continues to supply the input, the dominant company would need to avoid imposing a margin squeeze on the other party.[73] Since in those circumstances consumers would not benefit from significant competition between the companies, it seems unlikely that competition law should impose a duty to continue to supply.

Another set of issues arises if the dominant company wants to integrate forward into the downstream market, but lacks sufficient capacity to produce the input for both companies. It is generally assumed that a dominant company never has an obligation to expand its production to supply a downstream competitor. The analysis might depend on whether the total demand in the downstream market for the end product was stable (in which case the dominant company would take away some of the competitor's sales even if it continued to supply), or was likely to expand. In the latter case the dominant company might presumably use its total production of the input for its own sales, leaving its competitor with nothing, but consumers would presumably benefit from the expansion of the market.

A second type of case arises when the dominant company has developed a new and better technology, or a new or cheaper input for use in the downstream market. If it is not obliged to give the other party a contract to supply the more efficient input, under the principles applying to first contracts,

it seems unlikely that competition law should impose a duty to continue supplying the less efficient input,

since a company relying on it will leave the market anyway in due course. This would also be the position if the dominant company adopted a new more efficient technology under which there were no longer two production stages. Another situation arises if the dominant company develops a new use for the input, and the other party's use of it would endanger the new use. In a U.K. Office of Fair Trading case, Du Pont v. Op Graphics (Holography)[74] a refusal to continue supplying a firm for graphics arts purposes was held to be justified, because Du Pont was withdrawing from the graphic arts market in order to use the technology only for security purposes, which might have been endangered if the same technology was also being used for graphic arts by companies unconcerned with security issues.

Termination would be justified if, without any other change of circumstances, it became clear that the other party's activities threatened the efficiency of the dominant company's operations in either market or interfered with their expansion or development, if continued production of the input was no longer economic, or if the other party is no longer creditworthy or no longer has the expertise needed to share the facility.

Similarly, if there is a fall in the supply of a raw material needed for the production of the input in question, the dominant company may give preference to customers with long-term contracts, and presumably also to its own downstream operations with which it has permanent relationships.[75]

E) REMEDIES

The appropriate remedy in cases of unjustified termination of supply could be an order to resume supply. In termination of supply cases, it will be easier to determine the price of the input, and reasonable and non-discriminatory terms of the supply, since there used to be a business relationship indicating what the normal terms of the contract might be. The dominant company should, however, always have the possibility to prove that circumstances have changed in the meantime.[76]

1. When Is There a Right to Imitate a Dominant Company's Product?

As explained above, there is normally no right to copy the product of a dominant company.[77]

This analysis is confirmed by the Commission's action on one feature of the Microsoft case. Sun initially asked both for interoperability information, and for the right to use programs written by Microsoft together with operating systems on Solaris. The Commission refused the second claim because it would have created software copying Microsoft's platform on the basis of Solaris. In other words, the claim was for the right to produce a copy of the Microsoft product, and not merely for interoperability. It was therefore unjustified. The key distinction is between making the competitor's product work with the dominant company's product, when that is necessary, and being able to copy the dominant company's product itself.

In IMS Health, a competitor claimed a right to an intellectual property license to enable it to copy the product of the supposedly dominant company. The Court of Justice said,

"the refusal by an undertaking in a dominant position to allow access to a product protected by an intellectual property right, where that product is indispensable for operating on a secondary market, may be regarded as abusive only where the undertaking which requested the license does not intend to limit itself essentially to duplicating the goods or services already offered on the secondary market by the owner of the intellectual property right, but intends to produce new goods or services not offered by the owner of the right and for which there is a potential customer demand."[78]

The Advocate General in the same case made the same points.[79] However, the Court of First Instance in Microsoft went a little further. It said,

"The circumstance relating to the appearance of a new product, as envisaged in Magill and IMS Health, cannot be the only parameter which determines whether a refusal to license an intellectual property right is capable of causing prejudice to consumers within the meaning of Article [102(b)]. As that provision states, such prejudice may arise where there is a limitation not only of production or markets, but also of technical development."[80]

It would usually be difficult for a competitor to argue that its technical development was improperly limited merely by being prevented from copying the dominant company's products, and that consumers were harmed, in the absence of any "additional abusive conduct" other than the refusal to license.

This conclusion would be further reinforced when, as in some situations, two compulsory licenses would be needed. If the downstream products or services that the competitor wishes to produce infringe intellectual property rights of the dominant company, there could be no purpose for a compulsory license of part of the capital equipment to enable the competitor to produce the infringing products, unless there were also a right to a compulsory license of the rights over the products. Competition law cannot prohibit enforcement of intellectual property rights unless there is a duty to license them. Competition law allows foreclosure of a product that is being unlawfully sold.

2. Can There Be a Duty to Supply, If There is No Duty to License?

Assuming in a given situation that there is no duty to license competitors to use intellectual property rights, could there instead be a duty to supply the competitors with the products to which those rights apply? The competitors' argument would be that there can be a duty to supply products, even if they happen to be patented, and even if there is no "additional abusive conduct" necessary for a duty to license the intellectual property itself.

This argument raises the question whether the rules on compulsory access to intellectual property are the same as those regulating access to other kinds of property, mentioned above.

At first sight, it would be odd and irrational if there were a duty to supply a competitor with the dominant company's finished products if there was no legal duty to license the same competitor to manufacture similar products itself.

If there is no right to make a copy, there can be no right to buy the original.

In both situations, the intellectual property right would be transformed into a right to receive payment. This leads to the problem of how much a direct horizontal competitor should pay would arise, and whether the dominant company would be obliged to provide the competitor with a minimum gross profit margin, and if so, on what that margin could be based. The basic principle that there must be some identifiable conduct other than the refusal itself for there to be an abuse, and for a dominant patent owner to have a duty to share the benefit of its property, applies equally to both situations. With the possible exception of "additional abusive conduct," all the requirements for a duty to contract are likely to apply equally in both.

The fact that the competitors would benefit similarly from a duty to supply and from a duty to license suggests that the legal requirements for a duty should be the same in both cases.

There could hardly be a benefit to consumers from imposing a duty to supply patented products, since consumers can already buy them from the dominant company without paying for an intermediary. Neither consumers nor competition would benefit more from a duty to supply than from a duty to license.

In fact, they would generally benefit less, for several reasons. If there was scope for substantial competition in the downstream market (presumably in related services if the competitor planned to copy the dominant company's product), there would be less competition if the competitors were buying the products than if they were getting a license to manufacture the products themselves.

The scope for "follow-on" innovation, or for product differentiation, by a competitor buying the dominant company's product would be less than if it obtained a license of the relevant technology. A competitor buying the products could not take advantage of having lower production costs, which might be relevant if it were entitled to a license.

In short, if there is no duty to license to enable the competitor to produce certain products, it is difficult to imagine circumstances in which there would be a legal duty to supply the products themselves.

The economic arguments do not depend on the legal nature of the contract desired by the competitor. This conclusion does not depend on the rules for intellectual property and other property being the same.

V. IF THERE IS NO DUTY TO LICENSE OR TO SUPPLY, CAN THERE BE ILLEGAL TYING OR BUNDLING AS A RESULT OF REFUSAL TO SUPPLY SEPARATELY?[81] FORECLOSURE IN TYING CASES

The competitors might argue that the requirements for the abuse of exclusionary tying or bundling are less strict than the requirements for the abuse of refusal to license or to supply, and that the dominant company's conduct constitutes illegal tying or bundling.

In theory, tying can be illegal only if the two products are distinct. At least two tests of distinctiveness can be suggested. The first test asks if there is an independent demand for the tied product to be sold separately. The second test looks to see if there is a demand for the tying product to be sold separately. According to the Court in Microsoft, the first question correctly suggests that only when the advantages of tying or bundling are outweighed by the benefits of choice will consumers make separate purchases, if they are able to do so.[82] The Court said, ". . . the distinctness of products for the purposes of an analysis under Article 82 EC has to be assessed by reference to customer demand . . . in the absence of independent demand for the allegedly tied product, there can be no question of separate products and no abusive tying."[83] Under the second test, even if there were no demand for the tying product to be sold alone, tying it with the dominant company's tied product might force the consumer to buy, thereby denying choice to the consumer. Thus the second test is not the right approach, when the issue is foreclosure. Tying can be unlawful for two distinct reasons: that it is exclusionary foreclosure, keeping competing suppliers of the tied product from selling it, and that it is exploitative, forcing buyers to pay for the tied product that they do not want to buy from the dominant company[84] or perhaps at all.

There cannot be illegal tying or bundling unless the products are separate, but the key questions in most tying cases are whether there is illegal foreclosure and, if there is, whether there is sufficient justification. Foreclosure in this context must have the same meaning as in exclusionary abuse cases under Article 102(b)[85]: has the conduct of the dominant company created or increased a handicap or difficulty for competitors to which they would not otherwise have been subject? In the context of tying as elsewhere, mere failure or refusal to help a competitor is not illegal foreclosure, and does not need a justification.

It is of course correct that the requirements of the abuse of tying or bundling are different from those for a compulsory license. According to the Commission's Guidance paper[86], tying is illegal if the company is dominant, the products are distinct, the tying is likely to lead to "anticompetitive foreclosure" (that is, it is exclusionary), and there is no objective justification for tying. Complementary products (products that must be used together, such as nail guns and nails) can be separate products if there is a separate demand from consumers for competitors' versions of the complementary products. However, if the products in question are purely functional and the competitors' products are identical to those of the dominant company, it is not clear why there would be a separate demand for them from consumers, except for price reasons. If there is no reason for a separate demand for the tied product, tying is not exclusionary.

Unfortunately, the Commission has not explained or defined "anticompetitive foreclosure," except in the specific and unusual circumstances of the Microsoft tying case, and has not relied on what is in effect a definition of anticompetitive foreclosure in Article 102(b).

"Foreclosure" is not necessarily "anticompetitive."

Unfortunately, the Commission has not explained or defined "anticompetitive foreclosure," except in the specific and unusual circumstances of the Microsoft tying case, and has not relied on what is in effect a definition of anticompetitive foreclosure in Article 102(b).

Competitors may be legitimately foreclosed, that is, progressively pushed out of the market, if the dominant company consistently sells better products or charges lower prices than they do. Foreclosure can be anticompetitive only if the conduct causing it is not merely offering better bargains or some other result of procompetitive conduct, but if it involves creating a handicap or difficulty for competitors without any corresponding or off-setting benefit to consumers or competition.

In Microsoft, the Court concluded that there was illegal foreclosure as a result of tying on a series of factual grounds[87]:

- The company sold Windows only bundled with Windows Media Player;

- There was no extra charge for the Media Player;

- It was not possible to remove the Media Player;

- OEMs were understandably reluctant to add a second media player, increasing the price and using additional capacity;

- The Media Player automatically got the benefit of the worldwide market penetration of the Windows operating system, without having to compete on its merits as a media player;

- Downloading via the internet was less effective as a method of distribution than pre-installation by OEMs;

- Competitors' products were at a disadvantage even if they were better than Microsoft's product;

- The bundling increased the barriers to entry of competitors;

- Bundling allowed Microsoft to expand its position in adjacent media-related software markets; and

- Content providers and software developers primarily used the Media Player because that allowed them to reach the largest number of PC users in the world.[88]

In short, competitors were foreclosed for a number of reasons, all due directly or indirectly to Microsoft's conduct, that were not the direct results of Microsoft's intellectual property rights and not the result of Microsoft offering better products or lower prices. These factors taken together created a handicap for competitors, to which they would not otherwise have been subject, with no off-setting advantages for competition, consumers or competitors.

A) "DISTINCT PRODUCTS" IN TYING CASES

On distinct products, the Guidance says "whether the products will be considered . . . to be distinct depends on customer demand." Products are distinct if, in the absence of tying or bundling, "a substantial number of customers" would buy the tying product without buying the tied product from the same supplier. There may be indirect evidence of distinctness if there are companies specialized in manufacturing or selling the tied product without the tying product or without "each of the products bundled by the dominant undertaking," or if companies with little market power tend not to tie or bundle the products.

This description of "distinct" products is less useful if the dominant company has never sold the supposedly distinct products separately.

In particular, if new companies are set up to manufacture the tied product without the tying product, their emergence could hardly be enough in itself to make the dominant company's conduct illegal, even if they allege that they are unable to sell their new ("tied") products because the dominant company is selling its version of those products with the tying product. If customers have never had an opportunity to buy the tying product without the tied product, it is difficult and perhaps impossible to say whether a "substantial number" of them would choose to buy the tying product without buying the other product from the same source.

According to the Guidance, the competition authority may:

(a) Decide whether a substantial number of customers would buy the products separately, in circumstances that have never arisen:

(b) Deduce from its answer to this hypothetical question whether the products are or ought to be considered "distinct";

(c) Then determine whether there is "anticompetitive" foreclosure as a result of the sale of the two products together; and

(d) Assess the possible efficiency or other justifications for the conduct.

If the dominant company had never sold the tied products without the tying products, this exercise would be undesirably speculative. There would be a risk that the competition authority, without evidence about

what customers would do, would form its own opinion about whether the products ought to be considered distinct, and therefore whether customers ought to be enabled to choose whether to buy them separately. It is difficult to state the issues without using language about "tying" and "tied" products that begs the question by implying that they are in some sense separate products. Furthermore, this approach is likely to lead to regulatory action rather than actions based on competition law. Regulatory action, if duly authorized by legislation, allows regulatory authorities to impose new obligations to make a lawful but uncompetitive market more competitive. Competition law, on the other hand, allows official action only to end identifiable infringements.

The Microsoft Court identifies an additional complication: "the IT and communications industry is an industry in constant and rapid evolution, so that what initially appear to be separate products may subsequently be regarded as forming a single product, both from the technological aspect and from the aspect of the competition rules."[89]

The Guidance paper asserts that if there are not enough customers to buy the tied product separately, tying can lead to higher prices.[90] Although this comment appears in the context of anticompetitive foreclosure, it seems more relevant to the question of distinctness. Regardless of its application, the comment seems incorrect, because in that situation tying would provide economies of scale.

The practical conclusion seems to be that if the dominant company has never sold the tying product without the tied product, the products should not be considered distinct, unless there is clear evidence that other companies previously sold the products separately, and that there is a significant consumer demand for separate sales. Even if there were such evidence, it would be necessary to consider whether "constant and rapid evolution" had made them into a single product. If it seemed that such evolution had occurred, the question of distinctness would merge into the question of the reasons for the evolution. The word "justification" is not appropriate unless there is some apparently unlawful conduct that needs justifying.

B) "THESE ARE NOT NORMAL TYING CASES

It is important to be clear about the difference between the situations discussed here and a normal tying or bundling case.

In these situations the competitor wishes to buy the goods in question from the dominant company, and is arguing that it has a legal right to be supplied with them. In a normal tying case, the competitor wants to sell its own product to third parties, and complains that it cannot do so because the third party is obliged to buy that product from the dominant company. In short, these situations are in fact duty to supply cases, not tying cases.

Alternatively, the competitor might claim that it does not need to buy the product in question itself, provided that the dominant company offers its other product separately to third parties. That would help the competitor to sell whichever of the two products it was able to produce. But it would do nothing to enable the competitor to offer a package consisting of both products if it is not able to provide them both. The competitor needs to buy from the dominant company in order to offer a combined package. Therefore, once again, this is a duty to supply situation, not a tying case.

If the principal or only difficulty for competitors is due to the fact that the dominant company has intellectual property rights, bundling or tying (if those words were thought appropriate) of goods produced using those rights would not be anticompetitive foreclosure.

It is legitimate competition for a dominant company to obtain and exercise intellectual property rights for its own invention, even if it may be an abuse for it to acquire exclusive rights to the only effective competing technology.

If the intellectual property rights were legitimately acquired, and provided that there is no other abusive conduct, it is lawful for the dominant company to exercise them.

This is essentially merely another way of stating three general points made previously in this paper. First, conduct forecloses illegally only if it is not "competition on the merits"[91]—i.e., the conduct does not offer better bargains—and if it creates a handicap which competitors would not otherwise have been under. Second, a competitor cannot have a right to sell products that infringe intellectual property rights, unless it has a right to a compulsory license of those rights, under either Article 101 (in standard cases, not discussed

here) or Article 102. Third, there is no duty to supply the dominant company's finished products to competitors for simple resale.

C) EFFICIENCY BENEFITS IN TYING AND BUNDLING CASES

Efficiency benefits in tying and bundling cases are essentially economies of scale or scope, either in production, consumption, or use. However, such benefits can often be obtained without tying or bundling, whether contractual or technological.

The Advocate General in Tetra Pak I said,

"the undertaking in a dominant position may... strive through its efforts to improve its market position and pursue its legitimate interests. But in doing so it may employ only such methods as are necessary to pursue those legitimate aims. In particular it may not act in a way which, foreseeably, will limit competition more than necessary."[92]

In short, a justification, if one is needed, must be objective, proportionate, and appropriate. The requirement of appropriateness might mean that the justification in a tying case might not be quite the same as a justification in a case involving refusal to supply or to license, but there is no reason to think that justifications would be easier or harder to prove in a tying case. Both the Commission and the Court should try to be consistent across the whole range of abuses under Article 102.

D) "BALANCING" IN TYING CASES

On this analysis it is not necessary, as it sometimes may be in tying or bundling cases, to "balance" exclusionary or anticompetitive effects of the conduct in question against procompetitive effects, although it is very difficult to develop a convincing way of offsetting or balancing them. The supposedly anticompetitive effects are merely the result of the exercise of intellectual property rights, which cannot, without "additional abusive conduct," be contrary to Article 102. The exercise of intellectual property rights in itself is, as a result of legal principle, presumed to be procompetitive, because they are created by legislation to promote innovation in the long term. In the situation under discussion, therefore, there are no anticompetitive effects, and no anticompetitive foreclosure.

This analysis is therefore consistent with the more complicated factual analysis of the Microsoft case by the Court, which also carefully avoided "balancing" anticompetitive and procompetitive effects, although the Commission had claimed to balance them.

The only situation in which "balancing" might perhaps be necessary would be if the competitors wanted to be licensed for or supplied with products that were necessary for the supply of services.

Again, the question would be whether the competitors would be offering essentially the same kinds of services as the dominant company. If there were little scope for added value in the services market, the competitors would presumably be offering the same, or almost the same, kinds of services as the dominant company. It might then be necessary to see whether the competitors had advantages that the dominant company lacked, of which consumers would be deprived if the competitors were unable to provide the products needed. It would presumably be necessary to see whether these advantages were sufficient to outweigh the dominant company's economies of scale and scope, and the advantages of bundling for consumers.

The principal efficiency that might need to be taken into account in carrying out a balancing test would be the economies of scale and scope of the dominant company. A dominant company almost always has economies that are not available to competitors. This is particularly likely to be true of spare parts, but it is also likely to apply to production of consumables.[93]

One objection to the idea of "balancing" on these lines (apart from the difficulty of doing it in any objective way) was stated by the Court in Deutsche Telekom.[94] The Court said,

"If the lawfulness of the pricing practices of the dominant undertaking depended on the particular situation of competing undertakings, particularly, their cost structure, - information which is generally not known to the dominant undertaking - the latter would not be in a position to assess the lawfulness of its own activities."

This principle cannot be confined to the pricing practices of the dominant company. It must apply to all possibly abusive conduct. The principle of legal certainty

means that the law must not prohibit conduct which is unlawful only because of something that the dominant company cannot be expected to know.
Strictly speaking, when two kinds of goods are sold together because they are linked by their nature or by normal commercial usage, there is no "tying," and therefore no need for efficiency justifications. However, tying cases are not easily or satisfactorily resolved by arguments about how separate the products are, and

it seems more useful to look directly at the effects of the products being sold together,

as is done here, and for any justifications that may need to be considered.

The conclusion reached is that if the dominant company has intellectual property rights and has no duty to license them or to supply the products, it is not illegally tying if it exercises its intellectual property rights and refuses to sell the products separately. There is no "additional abusive element" because there is nothing that causes foreclosure except the exercise of the intellectual property rights itself.

E) IS HARM TO CONSUMERS NECESSARY IN TYING CASES?

Article 102 expressly requires harm to consumers as a necessary element in exclusionary abuses involving "limiting production, markets or technical development." However, harm to consumers is not expressly required for "unfair" purchase prices, discrimination, or tying under Article 102(d).

There are a number of strong reasons for believing that harm to consumers should be regarded as a necessary element in all abuses under Article 102.[95] Harm to consumers is always relevant under Article 101. The Advocate General in Bronner said, "the primary purpose of Article [102] is to prevent distortion of competition and in particular to safeguard the interests of consumers rather than to protect the position of particular competitors."[96] The Commission and the Court in Microsoft both considered it necessary to assess carefully the effect of Microsoft's conduct on consumers. It would produce odd and irrational results if harm to consumers was needed in some kinds of abuses but not in others. In discrimination it is particularly important to distinguish cases where there is harm to consumers from cases where it is procompetitive.

Whether tying in any particular case is regarded as exclusionary and harmful to competition or as coercion of customers, harm to consumers seems essential to consider. The fact that tying is lawful if there is insufficient demand from consumers for the tied product to be sold separately also shows that harm to consumers is a crucial question in tying and bundling cases. When the objection to tying is that it causes foreclosure, harm to consumers must be necessary as it is in all other foreclosure cases.

F) THE PRICING ISSUE

If the tying argument were accepted, the dominant company would be obliged to sell the products separately. It would presumably wish to sell the secondary products to its competitors at the same price at which it sold them to its customers. This would raise the difficulty mentioned above, that customers would have no reason to buy the secondary products from competitors when they could get them from the dominant company directly. To provide a benefit to consumers, and indeed to provide an advantage to the competitors, the dominant company would have to be ordered to sell to the competitors at a reduced price, to provide them with a profit margin. But there does not seem to be any basis in competition law for ordering a dominant company sell its final product to a direct horizontal competitor at a reduced price. (Margin squeeze cases concern sales of an input to a downstream competitor.) In other words, the tying argument would be open to all the same objections as the argument that there is a duty to license or a duty to supply. The practical problems would be identical, even though the legal arguments would be different.

G) THE COMMISSION'S DISCUSSION PAPER IN 2005 AND ITS GUIDANCE PAPER IN 2009

The Commission in its Discussion paper in 2005 wrote:

"If a dominant position on an aftermarket has been established... the Commission presumes that it is abusive for the dominant company to reserve the aftermarket for itself by excluding competitors from that market. Such exclusion is mostly done through either tying or a refusal to deal. The tying can come about in the various ways described in the section on tying. The refusal to deal may, for instance, involve a refusal to supply information or products needed to provide products or services in the aftermarket; a refusal

to license intellectual property rights; or a refusal to supply spare parts needed in order to provide aftermarket services."

This statement has probably been superseded by the Commission's later Guidance paper, and is certainly surprising. Such a presumption would, as O'Donoghue and Padilla have pointed out,[97] lead to a standard on tying in aftermarkets that is stricter than the tests applied in Hilti, Tetra Pak, and very fully and carefully by both the Commission and the Court in Microsoft.

As tying is usually procompetitive, abuse cannot be presumed or established without careful analysis. The comment on tying in aftermarkets is not even consistent with the Discussion paper's own comments on tying in other kinds of markets. The statement quoted is inconsistent with the Commission's later Guidance paper, which says nothing about aftermarkets, and which suggests a much more careful economic analysis of tying, where a series of factors "are generally of particular importance for identifying cases of likely or actual anti-competitive foreclosure."[98] This indicates that

the sweeping and unexplained presumption suggested by the Discussion paper has been abandoned.

The Guidance paper expressly contemplates the possibility of efficiencies, which was not even mentioned in the earlier paper. It seems that the Commission has now accepted that the Discussion paper was wrong, which is the correct position. However, the factors mentioned by the Commission in the Guidance paper are not especially helpful because they concern the extent of the economic effects of the conduct, rather than whether it is anticompetitive for foreclosure, which is the key issue. The factors are: whether the dominant company's tying or bundling strategy is lasting; whether it is dominant for more than one of the products, and; "if there is not a sufficient number of customers who will buy the tied product alone to sustain competitors of the dominant undertaking in the tied product, the tying can lead to those customers facing higher prices."[99] But if there are not enough customers who want to buy the tied product separately, that suggests either that the products are not really separate or that there is no consumer harm resulting from the tying. In addition, companies cease to produce products for which there are not enough buyers.

VI. THE PROCEDURAL POSITION OF THE COMPANY SAID TO BE DOMINANT: INTERIM MEASURES

A dominant company is free to acquire and exercise intellectual property rights for inventions that it has developed. Except in very rare circumstances outlined in ITT Promedia,100 the company could not be accused of vexatious litigation if it brings proceedings for infringement of its rights. The competition authority cannot prevent the dominant company from exercising its rights unless the authority finds that their exercise, or the refusal to license them, is an abuse. In theory, the competition authority might adopt an interim measures decision to prevent their exercise, but there are a number of reasons why this would be inappropriate (except perhaps in discrimination cases).

First, the President of the Court of First Instance in IMS Health ruled that interim measures to prevent the exercise of intellectual property rights are rarely justified.[101] Second, it would be inappropriate to order a compulsory license on an interim basis, because of the inconvenience and confusion that would result if it were finally determined that no license was justified.

Third, with the possible exception of discrimination cases, the conditions making it appropriate to impose a duty to contract are so difficult to apply, even in cases in which it seems likely that there is a duty to contract, that it is unwise and inappropriate to deal with them in an interim measures decision. The IMS Health interim measures decision shows how badly a competition authority can go wrong in an interim measures decision (and the Commission has adopted hardly any interim measures decisions, in spite of its power to adopt interim measures under Regulation 1/2003).[102] Fourth, it would clearly be inappropriate to adopt an interim measures decision finding that a dominant company was engaged in vexatious litigation, since the conditions for such a finding are not clear, and are rarely fulfilled. The national court dealing with the litigation would be much better placed than the competition authority to decide whether the infringement claim was justified or not. Fifth, as explained above, the mere exercise of an intellectual property right is never an abuse in itself. There must be some other identifiable conduct that constitutes an abuse, and for which a compulsory license or an order to contract is the appropriate remedy.

An interim measures decision, therefore, would have to consider whether another abuse had been committed, and if so, whether an order to contract was the right remedy for that abuse.

An interim measures decision, therefore, would have to consider whether another abuse had been committed, and if so, whether an order to contract was the right remedy for that abuse.

That would involve a substantial analysis of the facts, which would be inappropriate in an interim measures procedure. Sixth, any duty to contract must specify the terms of the contract. That would be difficult enough in a definitive decision, but inappropriate in an interim measures decision. The Commission's interim decision in IMS Health, which merely said the terms should be "reasonable and non-discriminatory,"[103] was clearly an abandonment of the Commission's responsibilities.

A) PROCEDURAL ISSUES

Since neither the Commission nor a national competition authority has any competence to decide the validity of intellectual property rights, a ruling entity has several possibilities when deciding how to deal with the request for a compulsory license, if the validity of the intellectual property right has not been finally determined. It could simply adjourn the case, without doing anything, and wait for the final result of the litigation to determine the validity of the right. That is a straightforward approach, and would normally be correct.

In at least some cases, the complainant may in effect be seeking two contracts from the dominant company: a license of the intellectual property right, if it is valid, and access to information or something else in the possession of the dominant company, which would not become available automatically even if the intellectual property right was declared invalid. In RTE-ITP,[104] the magazine Magill needed each of the television stations to provide details of the programs to be broadcast each week. The television companies argued that these programs were protected by copyright under U.K. and Irish law. The Court decided that even if that were correct, the companies still had a duty to give Magill the information, on reasonable terms as the Commission had required.

The point made here is that even if it had been clear that there was no copyright in the weekly program lists, Magill would still have needed, and been entitled to, the information.

Therefore the competition authority confronts two questions. The first question is whether, even if the intellectual property right is not valid, there is a duty to give access to the information or whatever else it is that the complainant says it needs. The second is whether, if the intellectual property right is valid, the dominant company has a duty to license it. Generally, if there is no duty to provide access, there will be no duty to license either. It is impossible to think of a situation in which there might be a duty to grant a license even if there was no duty to provide access, although there are of course situations in which the information is already public, and only a license of the right to use it is needed.

It follows that the competition authority might consider that it should answer the first question. If there is no duty to provide access, the second question does not arise. If there is a duty to grant access, there may be a duty to grant a license of any intellectual property rights that may be needed to make the access effectively available. In at least some cases the two questions are not really separate. The RTE-ITP case was unusual because the information was what was sought, and the copyright license was merely incidental. The issues in that case could have been separated.[105] In other cases such as IMS Health,[106] the only thing that is really needed is the license of the intellectual property right. If there is no valid right, there is no need for a license (and nothing to license). So the first question is, in effect, whether there is a duty to license the right, assuming that it is valid. That question is not a procedural question, but one of substance.

B) REMEDIES IN A DEFINITIVE DECISION BY A COMPETITION AUTHORITY

In theory, a competition authority might adopt a definitive decision determining whether there had been an abuse, contrary to Article 102 TFEU, for which a duty to contract was the correct remedy, independently of whether or not the intellectual property right was ultimately determined to be valid by a competent court. As already explained, if the authority decides that there is no abuse and therefore no duty to contract, which would dispose of the case. However, if the authority intends to find that there is a duty to contract, it would

be difficult to state the terms of the contract if the authority did not know whether the intellectual property right was valid. In practice, the authority would find it wise to adjourn the case.

If the right is finally held to be invalid, in theory, competitors are free to use the invention royalty-free. But if the competitors need something in addition to the right to use the invention, the authority would need to determine exactly what they were entitled to get access to. In the RTE-ITP case, there was no difficulty in defining that what the magazine needed from each television company was merely the next week's programs.[107] But in a more complicated case, this finding would be much more difficult.

It would be particularly difficult in a "potential market" case, in which no contract of the kind in question had ever been entered into by anyone.

Even if there were a "potential" downstream market, the authority would need to determine exactly what it consisted of, what should be made available, and on what terms. Also, it would be difficult, if not impossible, to decide what one direct competitor should pay to another for an important input or competitive advantage, as a matter of competition law. It might be easier under a regulatory regime, in which the regulatory authority could impose new obligations, and is free to act on new policy aims. But competition law is not a regulatory regime in this sense.[108]

Further complications are likely to arise in a market in which the dominant company sells two products to be used in combination. The competition authority would have to determine what information had to be given with the products to be delivered to the complainant. The dominant company might have contracted with buyers of the combination that they would not use either product together with competitors' versions of the other. Since the two products, in the kind of situations visualized, have to work with one another, such a restriction on use would probably be valid, and certainly the competition authority could not declare it invalid merely to facilitate the order to contract with, or supply to, the complainant. If the patents were valid, a license to customers to use each of the combined products only with the other would be a field of use restriction, and almost certainly valid (a limited license of an intellectual property right is not subject to the same constraints under competition law as a contractual restriction that may fall under Article 101 TFEU).

The competitors would presumably argue that a dominant company cannot use its refusal to allow customers to use one of its products with competitors' versions of the other as an indirect way of enforcing its intellectual property rights. The questions are nevertheless almost certain to be distinct. The competitor wishes to get the right to use the dominant company's intellectual property rights so that it can supply a competing version of one of the dominant company's products. If there was an alternative source of the other product, the competitor could combine its product with that of the third party. But if the competitor needs to combine its product with the other product produced by the dominant company, the latter is surely entitled to insist by contract that its customers use only a combination that it can guarantee will work properly; it has always been recognized as a justification for refusal to contract to show that use of the competitor's product would lessen the efficiency of the dominant company's products or services. The authority could override this insistence only if it could be certain that the two companies' products would work satisfactorily in combination.

The justification for the contractual limitation on customers would be entirely independent of the question of the duty to contract with competitors.

This seems likely to be the result in most if not all of the range of situations considered here. In Consten-Grundig,[109] the Court held that companies cannot use intellectual property rights to reinforce illegal contractual restrictions on parallel imports with their competitors, when the trademarks in question had been created artificially for that purpose. Companies cannot defend a restrictive agreement merely on the grounds that it restricts the other competitor no more than it is restricted anyway by intellectual property rights. But situations like Consten-Grundig are quite different from the cases discussed here. In the circumstances considered in this article, the intellectual property rights are not obtained artificially or collusively, and the supposedly restrictive agreements are with third parties, the customers of the dominant company. If the agreements with the customers were field of use restrictions, it would be even more clear that they

could be found invalid, if at all, only on entirely different grounds, which are difficult to imagine.

C) COMMITMENT DECISIONS

As the law is relatively complicated, a competition authority may be tempted to send a dominant company a short and superficial "preliminary assessment" of its "concerns," for the purpose of getting the company to negotiate a commitment that would make it unnecessary for the competition authority to analyze the case thoroughly. But the phrases quoted from Article 9 of Regulation 1/2003 are not intended to allow the company concerned to be deprived of the right to know clearly the arguments against it. Indeed, they are intended to ensure that the company gets something substantially equivalent to a statement of objections. A distinction should therefore be drawn. In any refusal to contract case, the company should insist on getting a carefully written and fully reasoned statement of objections, to see whether all the conditions discussed above are fulfilled and an identifiable abuse has been committed. However, if there has been an abuse, and if a duty to contract seems to be the appropriate remedy, it might be appropriate to negotiate the terms of the contract and to embody them in a commitment decision, if necessary. A competition authority should certainly be expected to write a detailed and clear statement of objections, or the equivalent, but may be excused if it prefers to work out the detailed terms of a duty to contract in the form of a commitment, once it has been proved that a duty to contract exists.

VII. A COMPREHENSIVE SUMMARY OF THE LEGAL RULES

In the light of this analysis, the legal rules on the duty to contract under Article 102 TFEU are more restrictive and more complicated than appears from the Commission's Guidance paper. They can be summarized as follows.

(1) A duty to contract under Article 102 TFEU can arise only when an identifiable abuse has been found. At least in the case of an intellectual property right, and probably in all cases, there must be an abuse in addition to the refusal to license.

(2) Under Article 102(b) TFEU, it is foreclosure and an abuse to "limit" the markets, production or technical development of competitors of the dominant company, if harm is caused to consumers. The mere exercise of intellectual property rights is never an abuse. Under Article 102(c) TFEU, it may also be an abuse for a dominant company to discriminate unjustifiably, if harm is caused to consumers. It may be contrary to both clauses of Article 102 TFEU to supply less than "ordinary" quantities to wholesalers, in order to prevent parallel imports. A dominant company is not obliged to confer an advantage on competitors, but it must not impose a handicap.

(3) As in the case of all other abuses, harm to consumers resulting from the abuse identified, must be shown. That harm may be preventing the development of a new kind of product for which there is a clear and unsatisfied demand, or imposing a continuing handicap on competitors in a dynamic market. Preventing a competitor from producing what would essentially be a copy or duplicate of the dominant company's product or service is not sufficient to justify a duty to contract. It is also an abuse to acquire the only competitive alternative to the dominant company's technology in order to suppress it or to use it to reinforce dominance, because it can be assumed that the alternative would otherwise be used to create competition.

(4) There must be two identifiable and separate markets, for an input and for an end- or "downstream" product. However, the fact that the dominant company in question has itself never made the input available to anyone is not a defense.

(5) If an abuse has been committed, a duty to contract may be the appropriate and proportionate remedy. There must be a link between the abuse and the duty to contract, which makes the duty the appropriate remedy for the abuse. However, no duty to contract may be imposed that would oblige the dominant company to share its principal competitive advantage, or deprive it of the incentive to invest in its principal activities, because that would end its dominance, an unjustifiable outcome under EU competition law.

(6) A duty to contract can arise only when competition would otherwise be eliminated. The fact that competition might otherwise be more difficult is not enough, as the Court made clear in Bronner.[110]

(7) Therefore there can be a duty to contract only when the product or service to be provided is objectively essential for competition in an identified market, and cannot be produced or otherwise obtained by any competitor or combination of competitors.

(8) There must be scope for non-price competition in the downstream market. This is more likely if the input required is a relatively small proportion of the total cost of producing the products or services for the downstream market, so that there can be effective competition using, or in connection with, the other factors of production.

(9) There may be justifications for refusal to contract that might make the conduct lawful, or that might make an order to contract an inappropriate remedy, even if an abuse had been committed.

(10) There is no duty to contract merely to create one more competitor. Nor is there a duty merely because competitors are unable to obtain or produce an input that they need elsewhere

(11) The terms of a duty to contract are relatively clear when the abuse is discrimination: the duty is to give access on non-discriminatory terms. If the abuse was unjustified termination of supply, the remedy would be an order to resume supply, presumably on the same or similar terms to be adjusted as necessary. In the case of a first contract or license, terms are much more difficult to determine. A competition authority cannot avoid its responsibilities by ordering a contract on "reasonable and non-discriminatory" terms, without giving proper guidance. An order in such vague terms would be void for legal uncertainty, particularly in the case of first refusal, when the "non-discriminatory" obligation would be meaningless.

(12) If there is no duty to license a competitor to produce a product, there is no duty to supply the product, and it is not illegal trying to sell the product only together with another product. Article 102 should be interpreted and applied consistently across the whole range of possible abuses..

---

1 Herbert Hovenkamp, Mark D. Janis & Mark A. Lemley, *Unilateral Refusals to License*, 2 J.L. & Econ. 1 (2006); Brian Sher & Daniel Wall, *Is Intellectual Property Just Property? Refusals to License IPRs in the U.S. and Europe*, International Bar Association Conference, San Francisco 2003. Daniel Beard in his article Microsoft: What Sort of Landmark, 4(1) Competition Pol'y Int'l 33, 48 (Spring 2008) says that the IMS test has led to "inevitable uncertainty as to when dominant undertakings will be required to supply IPRs and other sensitive information."

2 *See, e.g.*, Joined Cases C-241/91 P and C-242/91 P, RTE and ITP ("Magill"), 1995 E.C.R. I-743. These cases are distinct from those involving consumables and spare parts—Case 22/78, Hugin v. Commission, 1979 E.C.R. 1869; Case C-53/928, Hilti v. Commission, 1994 E.C.R. I-667; Case C-333/948, Tetra Pak v. Commission, 1996 E.C.R. I-5951—which raise issues about tying and handling, discussed below.

3 John Temple Lang, *Eight Important Questions on Standards under European Competition Law*, 7 Comp. Pol'y Int'l 32 (Spring 2011); John Temple Lang, *Abuse Under Article 82 EC: Fundamental Issues and Standard Cases*, in Neueste Entwicklungen Im Europäischen Und Internationalen Kartellrecht 95 (Carl Baudenbacher ed., 2007).

4 Council Regulation (EC) No. 1/2003, 2003 O.J. (L 1) 1, at art. 3(2).

5 *See, e.g.*, Sea Containers Ltd v. Stena Sealink, 1994 O.J. L15/8, 1995 4 CMLR 84; Port of Rodby v. Denmark, 1994 O.J. L55/52, 1994 5 CMLR 457; and Morlaix (Port of Roscoff), 1995 5 CMLR 177. *See generally* Robert O'Donoghue & A. Jorge Padilla, The Law and Economics of Article 82 EC ch. 8 (2006); John Temple Lang, *Anticompetitive Non-Pricing Abuses under European and National Antitrust Law*, in Annual Proceedings of the Fordham Corporate Law Institute: International Antitrust Law & Policy 2003, 235-40 (B.E. Hawk ed., 2004).

6 *The Application Of Article 82 of the Treaty to Exclusionary Abuses* (DG Competition, Discussion Paper December 2005).

7 Luca Prete, *From Magill to IMS: Dominant Firms' Duty to License Competitors*, 15(5) Eur. Bus. L. Rev. 1071, 1080 and 1083 (2004).

8 *See* Maurits Dolmans & Nicholas Levy, *EC Commission v. Microsoft: Win, Lose or Tie?*, Commercial Lawyer, January 2002, 1-5; O'Donoghue & Padilla, *supra* note 5. The case is reported as Case C-418/01, IMS Health v. NDC Health, 2004 E.C.R. I-5039; Case T-184/01R, IMS Health v. Commission, 2001 E.C.R. II-3193; and Case C-481/01 PR, 2002 E.C.R. T-3405.

The IMS Health case has given rise to a great deal of comment and criticism. *See* Andreas Schwarze, *Der Schutz des geistigen Eigentums im europäischen Wettbewerbsrecht*, EUZW 75-81 (2002); Pitovsky et al., *The Essential Facilities Doctrine Under US Antitrust Law*, 70 Antitrust L.J. 443 (2002); Paul D. Marquardt & Mark Leddy, *The Essential Facilities Doctrine and Intellectual Property Rights: A Response to Pitovsky, Patterson & Hooks*, 70 Antitrust L.J. 847, 847-873 (2002); Estelle Derclaye, *Abus de position dominante et droits de propriété intellectuelle dans la jurisprudence de la Communauté européenne: IMS survivra-t-elle au monstre du Dr. Frankenstein?*, 15 Les Cahiers De Propriété Intellectuelle 21, 21-55 (2002); Christian Koenig, Andreas Bartosch & Jens-Daniel Braun, EC Competition and Telecommunications Law 133-134, 151-157 (1st ed. 2002); Sergio Baches Opi, *The Application of the Essential Facilities Doctrine to Intellectual Property Licensing in the European Union and the United States: Are Intellectual Property Rights Still Sacrosanct?* 11 Fordham Intell. Prop. Media & Ent. L.J. 409, 409-506 (2001); David W. Hull, James R. Atwood & James B. Perrine, *Compulsory Licensing*, Eur. Antitrust Rev. 36-39 (2002). Matthias Casper, *Die wettbewerbsrechtliche Begründung von Zwangslizenzen*, ZHR 166, 685-707 (2002); Donna M. Gitter, *The Conflict in the European Community Between Competition Law and Intellectual Property Rights: A Call for Legislative Clarification of the Essential Facility Doctrine*, 40 Am. Bus. L.J. 217 (2003);

Maurits Dolmans & Daniel Ilan, *A Health Warning for IP Owners: The Advocate General's Opinion in IMS and Its Implications for Compulsory Licensing*, 11 Comp. L.I. 12 (2003); Alessandra Narciso, *IMS Health or the Question Whether Intellectual Property Still Deserves a Specific Approach in a Free Market Economy*, 4 I.P.Q. 445 (2003); Estelle Derclaye, *Abuses of Dominant Position and Intellectual Property Rights: A Suggestion to Reconcile the Community Courts Case Law*, 26 World Competition 685 (2003); Bruno Lebrun, *IMS v. NDC: Advocate General Tizzano's Opinion*, 26(2) E.I.P.R. 84 (2004); David Aitman & Alison Jones, *Competition Law and Copyright: Has the Copyright Owner Lost the Ability to Control His Copyright?* 26(3) E.I.P.R. 137 (2004); Beatriz Conde Gallego & Dimitris Riziotis, Comment, 5 Int'l Rev. Intell. Prop. & Comp. L. 564 (2004).

9 RTE-ITP, 1995 E.C.R. I-743.

10 Case C-7/97, Oscar Bronner, 1998 E.C.R. I-7791.

11 IMS Health, 2004 E.C.R. I-5039 at ¶ 57.

12 *Id*. at ¶ 59.

13 *Id*. at ¶ 44-45.

14 Communication from the Commission—Guidance on the Commission's enforcement priorities in applying Article 82 of the EC Treaty to abusive exclusionary conduct by dominant undertakings, O.J. No. C-45/7 (Feb. 24, 2009).

15 *Id*. at ¶ 76.

16 *Id*. at ¶ 81.

17 John Vickers said that the Microsoft case needed "robust limiting principles," and that the Commission should provide them. John Vickers, *A Tale of Two EC Cases: IBM and Microsoft*, 4(1) Comp. Pol'y Int'l 3, 23 (Spring 2008). Unfortunately it did not do so.

18 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 79.

19 IMS Health, 2004 E.C.R. I-5039.

20 Dolmans & Ilan, *supra* note 8, at 14.

21 Purple Parking v. Heathrow Airport, 2011 E.W.H.C. 987 at ¶¶ 168-178.

22 RTE-ITP, 1995 E.C.R. I-743 at ¶ 47; Burton Ong, *Anti-competitive Refusals to Grant Copyright Licenses: Reflections on the IMS Saga*, 26(11) E.I.P.R. 505, 507 (2004).

23 Case T-201/04, Microsoft v. Commission, 2007 E.C.R. II-3601 at ¶¶ 241 and 657.

24 Yet another way of reaching the same conclusion may be to say that there is no duty to contract if the purpose is to make the competitor merely produce a copy of the dominant company's product or service. This question is discussed below.

25 Dolmans & Ilan, *supra* note 8, at 14.

26 Maurits Dolmans, Paul-John Loewenthal & Robert O'Donoghue, *Article 82 EC and Intellectual Property: The State of the Law Pending the Judgment in* Microsoft v. Commission, 3(1) Comp. Pol'y Int'l 106, 129 (2007). In discrimination cases under Art. 102(c) there can be an abuse if competition is distorted even if it is not eliminated, see Purple Parking, 2011 E.W.H.C. 987.

27 Bronner, 1998 E.C.R. I-7791 at ¶¶ 41-47; Prete, *supra* note 7, at 1075 and 1081.

28 *See e.g.* RTE and ITP, [1995] E.C.R. I-743 at ¶¶ 49-50; Case T-198/98, Micro Leader Business v. Commission, 1999 E.C.R. II-03989 at ¶¶ 56-57; and IMS Health v. NDC Health, 2001 E.C.R. II-3193 at ¶¶ 34-35.

29 Cristophe Humpe & Cyril Ritter, *Refusal to Deal*, GCLC Research Papers on Article 82 EC, 4-5 (July 2005).

30 Microsoft, 2007 E.C.R. II-3601, at ¶¶ 657, 697 and 702.

31 RTE-ITP, 1995 E.C.R. I-743 at ¶ 54.

32 Bronner, 1998 E.C.R. I-7791

33 IMS Health, 2004 E.C.R. I-5039.

34 RTE-ITP, 1995 E.C.R. I-743 at ¶¶ 52-54 and Ong, *supra* note 22, at 507.

35 The cases in which it has been held that Art. 102(b) prohibits limiting the possibilities open to competitors are listed in Temple Lang, *supra* note 3, at 99.

36 Microsoft, 2007 E.C.R. II-3601, at ¶¶ 647-649, 659 and 665.

37 Joined Cases C-468/06 to C-478/06, Lélos v. GlaxoSmithKline, 2008 E.C.R. I-7139.

38 Commission decision of 21 December 1988, Decca Navigator System, 1989 O.J. L. 43/27.

39 *See* Commission's Fourteenth Report on Competition Policy, at ¶¶ 94-95 (1984); Steven Anderman, *Does the Microsoft Case Offer a New Paradigm for the 'Exceptional Circumstances Test and Compulsory Copyright Licenses Under EC Competition Law?*, 1(2) Comp. L. Rev. 7, 17 (2004).

40 Case T-321/05, AstraZeneca v. Commission, 2010 E.C.R. II-____; Bellamy & Child: European Community Law of Competition 1-1679, 1028-1029 (Peter Roth QC & Vivien Rose eds., 2007).

41 Dolmans, Loewenthal & O'Donoghue, *supra* note 26, at 123 and 125.

42 Case T-41/96, Bayer AG v. Commission, 2000 E.C.T. II-3383 at ¶ 180.

43 IMS Health, 2004 E.C.R. I-5039.

44 RTE-ITP, 1995 E.C.R. I-743.

45 *Id*.

46 Joined Cases C-6/73 and C-7/73, Istituto Chemioterapico Italiano SpA and Commercial Solvents Corp v. Commission, 1974 E.C.R. 223.

47 John Temple Lang, *European Competition Law and Compulsory Licensing – A Comprehensive Principle*, 4 Europarättslig Tidskrift 558, 558-588 (2004); Temple Lang, *supra* note 5; John Temple Lang, *European Competition Law and Intellectual Property Rights - A New Analysis*, 11 ERA Forum 411, 411-437 (2010).

48 Commission Guidelines on the applicability of Article 101 of the TFEU to horizontal co-operation agreements, 2011 O.J. C. 11 ch. 7; *see* Commission of the European Communities, Eleventh report on competition policy, ¶ 94 on IGR Stereo Television ("Salora"), (1982).

49 *See, e.g.*, RTE-ITP, 1995 E.C.R. I-743 at ¶ 49.

50 Coco Rita, *Antitrust Liability for Refusal to License Intellectual Property: A Comparative Analysis and the International Setting*, 12(1) Marq. Intell. Prop. L. Rev. 1 (2008).

51 Wolfgang Kerber & Claudia Schmidt, Microsoft, *Refusal to License Intellectual Property Rights, and the Incentives Balance Test of the EU Commission* 19-20 (Working Paper, November 8, 2008), *available at* http://ssrn.com/abstract=1297939 (Kerber & Schmidt, however, favor a test balancing incentives).

52 RTE-ITP, 1995 E.C.R. I-743.

53 Microsoft, 2007 E.C.R. II-3601.

54 Philips Electronics v. Ingman and Video Duplicating, (1998) Ch. 1997 P.No. 4100 at ¶ 37.

55 Anneleen Straetemans, The EU Microsoft Case – Not as Soft a Case, 44(4) JURA FALC. 563, 580-581 (2007-2008).

56 Case T-83/91, Tetra Pak International SA v. Commission, 1994 E.C.R. II-755 at ¶¶ 212-213; *cf.* SCM Corp. v. Xerox Corp., 645 F.2d 1195 (2d Cir. 1981).

57 GlaxoSmithKline, 2008 E.C.R. I-7139 at ¶¶ 49, 70-71 and 77.

58 *See, e.g.*, Case 53/87, CICRA v. Renault ("Volvo v. Veng"), 1988 E.C.R. 6039.

59 Humpe & Ritter, *supra* note 29, at 10-11.

60 Commercial Solvents Corp, 1974 E.C.R. 223.

61 Bronner, 1998 E.C.R. I-7791.

62 RTE-ITP, 1995 E.C.R. I-743 and Ong, *supra* note 22, at 507.

63 Bronner, 1998 E.C.R. I-7791 at ¶¶ 41-48 and Sher & Wall, *supra* note 1, at 13.

64 Commercial Solvents Corp, 1974 E.C.R. 223.

65 RTE-ITP, 1995 E.C.R. I-743.

66 Microsoft, 2007 E.C.R. II-3601.

67 RTE-ITP, 1995 E.C.R. I-743; Dolmans & Ilan, *supra* note 8, at 15; and Ong, *supra* note 22, at 507.

68 Contra Estelle Derclaye, *The IMS Health Decision: A Triple Victory*, 27(3) World Compet. 397, 403 (2004).

69 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 84.

70 Cyril Ritter, *Refusal to Deal and "Essential Facilities": Does Intellectual Property Require Special Deference Compared to Tangible Property?*, 28(3) World Compet. 281, 284-285 (2003).

71 Commercial Solvents Corp, 1974 E.C.R. 223.

72 Anderman, *supra* note 39, at 7 and 16; Bellamy & Child, *supra* note 40, at 1006.

73 *See* Case C-52/09, Konkurrensverket. v. TeliaSonera AB, 2011 ECR I-____ (Feb 18) (a margin squeeze may be an abuse even if there is no duty to supply).

74 Case CP/1761/02, E.I. du Pont de Nemours & Co. and Op. Graphics (Holography) Ltd., Decision of the Office of Fair Trading on Sept. 9, 2003, at ¶¶ 34-35.

75 *See* Case 77/77, BP v. Commission, 1978 E.C.R. 1513.

76 Ritter, *supra* note 70, at 285.

77 IMS Health, 2004 E.C.R. I-5039 at ¶ 49.

78 *Id*.

79 IMS Health, 2004 E.C.R. I-5039 at ¶¶ 62 and 66.

80 Microsoft, 2007 E.C.R. II-3601 at ¶ 647.

81 *See* O'Donoghue & Padilla, *supra* note 5, at 508-509 and ch. 9 generally. The leading cases are Case C-53/928, Hilti v. Commission, 1994 E.C.R. I-667 and Tetra Pak II, 1996 E.C.R. I-5951. Neither deals with technological tying (product integration).

82 Kelyn Bacon, Tying after Microsoft: the Step Forward and Two Steps Back, 4(1) Comp. Pol'y Int'l 65 (Spring 2008).

83 Microsoft, 2007 E.C.R. II-3601 at ¶¶ 917-918. However, competition may be restricted even if there is no direct harm to consumers as a result of the tying, see ¶ 960 ff.

84 Id. at ¶¶ 865, 962-963. At ¶ 867 the Court said, ". . . in principle conduct will be regarded as abusive only if it is capable of restricting competition." In any case, one would expect the basic requirements for all kinds of abuses to be the same.

85 Kai-Uwe Kühn, Robert Stillman & Christina Caffara, *Economic Theories of Bundling and Their Policy Implications in Abuse Cases: An Assessment in the Light of the Microsoft Case*, 1 Eur. Comp. J. 85, 86 (2005) say "[t]he most important criticism of the European Commission in cases like Tetra Laval/Sidel and GE/Honeywell has been that the specific and competitive mechanism was never clearly identified and therefore that there was no clear set of evidence that could have led to the conclusion that bundling was anti-competitive."

86 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 47-62.

87 Eleanor M. Fox, Microsoft *(EC) and Duty to Deal: Exceptionality and the Transatlantic Divide*, 4(1) Comp. Pol'y Int'l 25 (Spring 2008) (criticizing the Microsoft judgment for relying too heavily on facts without clearly stating principles).

88 The Court summarized its agreement with the Commission that the bundling had anticompetitive effects at ¶ 1088. See also ¶ 1038-1054.

89 Microsoft, 2007 E.C.R. II-3601 at ¶ 913.

90 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 55.

91 AstraZeneca v. Commission, 2010 E.C.R. II-____ at ¶¶ 335, 812, 817, 824, 845.

92 Case T-51/89, Tetra Pak International SA v. Commission, 1990 E.C.R. II-312 at ¶ 68.

93 Herbert Hovenkamp, Tying Noncompetitive Goods (U Iowa Legal Studies, Research Paper No. 11-19, 2011), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1785586.

94 Case T-271/03, Deutsche Telekom AG v. Commission, 2008 E.C.R. II-477 at ¶ 192.

95 O'Donoghue & Padilla, *supra* note 5, at 224-225; Temple Lang, *supra* note 5, at 235-340, 245, 250-253, 311-315.

96 Oscar Bronner, 1998 E.C.R. I-7791 at ¶ 58.

97 O'Donoghue & Padilla, *supra* note 5, at 508-509.

98 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 52. *See* Centre for European Policy Studies Task Force, Treatment of Exclusionary Abuses under Article 82 of the EC Treaty 58-66 (2009).

99 Guidance on the Commission's Enforcement Priorities, *supra* note 14, at ¶ 55.

100 Case T-111/96, ITT Promedia NV v. Commission, 1998 E.C.R. II-2937 at ¶¶ 72-73.

101 IMS Health, 2001 E.C.R. I-3193 at ¶ 144.

102 Atwood, Hull & Perrine, *supra* note 8, at 38.

103 Case COMP D3/38.044, NDC Health v. IMS Health, 2002 O.J. L59, 18 and Christopher Stothers, The End of Exclusivity?: Abuse of Intellectual Property Rights in the E.U., 2 Eur. Intellect. Prop. Rev. 86, 90 (2002).

104 RTE-ITP, 1995 E.C.R. I-743.

105 *Id*.

106 IMS Health, 2004 E.C.R. I-5039.

107 RTE-ITP, 1995 E.C.R. I-743.

108 John Temple Lang, European Competition Policy and Regulation: Differences, Overlaps and Constraints, in Antitrust and Regulation in the EU and US: Legal and Economic Perspectives 20 (François Lévêque & Howard Shelanski eds., 2009).

109 Joined cases 56/64 and 58/64, Consten SA and Grundig-Verkaufs-GmbH v. Commission, 1966 E.C.R. 299.

110 Oscar Bronner, 1998 E.C.R. I-7791 at ¶¶ 38-45; Prete, *supra* note 7, at 1075 and 1081.

FROM THE EDITOR
David S. Evans

The financial crisis began in 2007, deepened with the collapse of Lehman Brothers in September 2008, and appears likely to continue given the sovereign debt woes spreading across a shaky European Union. The forces battling the crisis have mainly included banking regulators, financial markets experts and macroeconomists. But the antitrust profession has gotten some work, too.

Some of that work is fortuitous. Sir John Vickers and Mario Monti were enlisted because of their sidelines in banking and monetary affairs. Ex-U.K. OFT head Vickers chaired the U.K.'s Independent Commission on Banking in 2011. Former EU Competition Policy Commissioner Monti (and CPI Editorial Board Member) was appointed Prime Minister of Italy in November 2011 to help dig the country out of its dismal economic condition. Some of it is part of antitrust's day job—the Directorate General for Competition Policy has kept busy in 2008 and 2009 examining whether bank bailouts were consistent with State Aid rules, and of course many lawyers have worked on the fallout from those inquiries. Still others heard the phrase "Too Big to Fail" as a rallying cry for the antitrust profession to say, not so fast. Beyond this, the financial crisis and proposed solutions to it have raised antitrust questions from the state of competition in an increasingly consolidated banking system to possible creation of market power in central clearing houses for derivatives.

This Autumn 2011 issue of Competition Policy International focuses on the intersections between antitrust, financial regulation, and the crisis overall. It is a good time to do this. The US and European authorities have been dealing with the crisis for more than three years. Enough time has passed for us to take a look at what has been done. And yet the same time, governments are still grappling with financial reform. Going forward there is much to analyze how competition policy fits into these efforts.
We begin with a symposium on some general issues. Gert-Jan Koopman, Deputy Director for State Aids at the European Commission, leads off with a survey of how the Commission has handled state aid involving the financial sector during the crisis. Professor Abel Mateus, a Director of the European Bank for Reconstruction and Development and former head of the Portuguese Competition Authority, examines the Independent Commission on Banking and other proposed regulatory reforms. Three Allen & Overy lawyers—Todd Fishman, Olivier Fréget & David Gabathuler—look at how the new financial regulations in the United States and European Union could constrain the enforcement of competition policy.

Our second symposium concerns the regulation of the payment industry. Concern over this industry by antitrust and banking regulators predated the crisis. But in the United States, at least, the financial crisis provided momentum to efforts to regulate aspects of these cards. Columbia University Law School Professor Ronald Mann argues that efforts to regulate credit and debit cards have reduced competition. The next two articles focus on efforts to regulate interchange fees. Professor Richard Epstein of New York University Law School examines the provision of the Dodd-Frank Act that required the Federal Reserve Board to regulate debit card interchange fees and posits that it should be viewed as an unconstitutional taking of property. My article concludes this symposium with a look at how reducing the fees that the card business can receive from the merchant-side of this two-sided business could affect the pass of innovation.

Right on the heels of the U.S. Department of Justice's approval of the NYSE Euronext and Deutsche Borse merger is our article by Craig Pirrong, who reveals the efficiencies in vertically-integrated financial exchanges.

We have, as our break from financial regulation, an article by John Temple Lang, a partner at Cleary Gottlieb and a professor at Trinity College in Dublin, with a new twist on a well-trod topic. The well-trod is compulsory access to property under the antitrust laws. The new twist concerns access to property that resides in a potential rather than actual market.

Our Classic concludes the Fall 2011 issue. We have chosen William Baxter's article on interchange fees, which was published in 1983. While there is much to

criticize in his article in hindsight, it provided some of the early groundwork for the vibrant literature on multi-sided platforms that started around 2000, and for the related literature on the economics of interchange fees. Thomas Brown, a partner at O'Melveny and Myers and former counsel to Visa, introduces the article and explains its importance. In selecting this classic we also honor the late Professor Baxter, who headed the Antitrust Division of the U.S. Department of Justice from 1981 to 1983 and made a number of seminal contributions to antirust, including spearheading the basic framework for modern merger analysis.

Readers will see that we have made some changes to the design of the Journal. In Spring 2011 we introduced the new e-book format, which allows us to do a number of things including incorporating audio and video. With this issue we've moved to a new design that we believe

David S. Evans
Global Economics Group, Boston, MA, U.S.A.
and University of Chicago, Chicago, IL, U.S.A.

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

STABILITY AND COMPETITION IN EU BANKING DURING THE FINANCIAL CRISIS: THE ROLE OF STATE AID CONTROL

Gert-Jan Koopman

European Commission

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

STABILITY AND COMPETITION IN EU BANKING DURING THE FINANCIAL CRISIS: THE ROLE OF STATE AID CONTROL[1]

Gert-Jan Koopman*

ABSTRACT

The available evidence suggests that the European Commission's State Aid ("SA") control of public assistance to the financial sector in the European Union during the period 2008-2010 has had a positive impact on both financial stability and competition in the EU's internal banking market. The particular features of the crisis regime dedicated to assessing State Aid not only allowed the disbursement of unprecedented amounts of aid, often in record time, but also rendered the aid more effective by ensuring that aid recipients, where necessary, were restructured or liquidated. The conditions imposed on banks receiving large amounts of aid have generally led to highly significant restructuring and addressed fundamental weaknesses in business models, helping to avoid the creation of "zombie banks." At the same time, where aid amounts were relatively small and banks were sound, these rules allowed financial institutions to be aided without requiring changes in their business model.

SA control has ensured that the large amounts of aid granted did not lead to major distortions in the Internal Market. Absent this control, these public interventions could have triggered a fragmentation of the Internal Market itself.

While all substantial aid is likely to have a distortive effect, available indicators suggest that SA control has effectively mitigated these consequences. There is little evidence of retrenchment behind national borders and aided banks have generally not seen their market shares increase. Moreover, the crisis framework is likely to have had a strong signalling function to financial institutions with respect to moral hazard going forward.

In the absence of EU-wide rules for bank resolution, the SA crisis regime also presently acts as the de facto EU-wide resolution framework. However, it is an imperfect tool resolution compared to a full-fledged regulatory framework that helps avoid recourse to aid in the first instance and can provide clear ex ante guidance for all market players (which in itself is confidence-enhancing).

The re-emergence of serious tensions in the EU banking sector from the summer of 2011 onwards is largely linked to concerns about the sustainability of public finances in a number of EU Member States feeding through to concerns about assets on banks' balance sheets. To remedy this, stability-oriented macroeconomic--especially fiscal--policies are required, and appropriate regulation of banking is needed. A key challenge for State Aid control in EU banking, therefore, is to ensure appropriate coordination with regulatory and macroeconomic policies as they are further developed.

* Deputy Director General for State Aid, Directorate General for Competition, European Commission.

I. INTRODUCTION

The economic and financial crisis triggered by the bankruptcy of Lehman Brothers unleashed tensions in banking systems across the globe that, in terms of scale and impact, are unparalleled in modern history. Although the crisis was triggered by a shock in the United States, it spread rapidly across borders. It strongly affected European financial institutions that held many "toxic" assets originating in the United States on their balance sheets and enjoying close relationships with their U.S. peers. Preexisting weaknesses of EU banks also played a role; some had too-high leverage ratios and overly relied on wholesale markets for their funding.

Finally, the fragmented regulatory framework in the European Union clearly also played a major role in allowing these unsustainable trends to build up.[2] In the fall of 2008, a coordinated approach in the European Union was put in place to safeguard macro-financial stability through the provision of unprecedented resources by the European Union's Member States to their banks. In parallel, the European Central Bank ("ECB") and other central banks provided ample liquidity while a macro-economic stimulus package was launched to maintain demand in the EU economy. The collapse of Europe's banking system was thereby avoided, even though the system remains under severe pressure on account of the EU sovereign debt crisis.

These initiatives revealed the challenges of coordinating Member State actions in the context of a systemic crisis where macro-financial stability concerns were pursued through Member States' actions in an internal European market. This market is one where banks are free to operate across borders, requiring cross-country competition concerns to be factored into the design of the strategy.

The crisis also suggests that some large financial institutions had taken unwarranted risks on the back of an implicit state guarantee that they would not be allowed to fail.

Moral hazard thus had to be addressed. Lastly, most Member States had no resolution framework for banks, nor did a dedicated EU bank resolution framework exist.

The European Union does, however, have a system of centralized State Aid ("SA") control established by the Treaty on European Union,[3] whereby the European Commission ("Commission" or "EC") vets all SA that Member States intend to grant. The Commission can approve this aid unconditionally, approve it under certain conditions (e.g. by requiring restructuring), or reject aid applications. The European Commission can also establish guidelines and frameworks that clarify the rules it will apply to individual cases. This supranational set-up is unique in the world and reflects the need to ensure common rules for State intervention in an internal market composed of Member States that enjoy significant national economic powers.

In order to deal with the challenges of safeguarding competition in the internal market, addressing moral hazard and providing a degree of coordination with regard to bank resolution, the European Commission has developed a crisis State Aid framework for financial institutions since October 2008. The framework became a de facto key microeconomic coordination framework complementing fiscal and macro-financial stability-oriented policies. Apart from the role played by the ECB and other European central banks, the latter policies were largely coordinated through the European Council and the Council of Finance Ministers ("ECOFIN") on the basis of broad views reflecting a consensus-seeking approach among Member States.

This paper briefly describes the approach to SA control taken by the European Commission in this context and provides a concise evaluation of its effects. In particular, it assesses whether, in practice, there has been a trade-off between competition and financial stability.

II. THE EUROPEAN COMMISSION'S APPROACH TO STATE AID CONTROL IN THE CRISIS

The European Commission decided at the beginning of the crisis that State Aid control would have to complement, and indeed, support, macro-financial stability-oriented policies in order to preserve the internal market. More fundamentally, since it was the only tool available at the EU level to address moral hazard and impose restructuring of unviable business models or the liquidation of banks, the European

Commission considered that SA control could also be helpful from a macroeconomic point of view.[4] Furthermore, lessons learned from the financial crisis in Japan were taken to heart[5]: undercapitalized banks with unsound business models ("zombie banks") require appropriate restructuring because without it they could drag down growth for a very long period.

State Aid control was therefore seen as part of the solution from the very beginning. Not all Member States welcomed this, and some feared that unduly rigorous application of competition rules would clash with stability-oriented policies. The economic literature on this matter does not provide unequivocal answers to the question whether there is a trade-off between financial stability and competition. A more traditional strand of the literature holds that competition results in smaller and less diversified banks that are less able to withstand shocks. This suggests that the promotion of competition in banking could endanger financial stability.[6] However, many of these drawbacks could be addressed by appropriate regulation and supervision. More recent analysis shows that large banks in concentrated banking systems may create adverse selection issues[7] and could also lead to "too big to fail" dilemmas, creating significant mispricing of risk and moral hazard. A very concentrated banking sector itself could increase contagion risk,[8] which, in turn, could make it more difficult to supervise and regulate the sector appropriately. As recognized by the U.K. Banking Commission,[9] the literature points to different mechanisms that affect the interplay between competition and stability oriented policies. No structural trade-off between financial stability and competition can be identified. However, the design of both policies needs to be sensitive to spillover effects and should, especially in a crisis environment, be taken forward in an integrated manner to allow interactions to be internalized as best as possible. The European Commission recognized this in 2008 when it decided to adapt its state aid policy in the banking sector to the needs of such an approach, given the systemic vulnerabilities in the banking sector.

The European Commission was sensitive to these concerns. It designed a dedicated set of rules that took account of the need to respond to a horizontal shock to the banking system requiring the disbursement of large amounts of aid in record time to prevent a major economic crisis, while also recognizing that there were significant differences across affected banks. The approach was therefore from its inception based on the principle of proportionality.

To develop adequate rules, the European Commission adapted the preexisting rescue and restructuring guidelines[10] to fit a situation where large amounts of support for banks were required for stability reasons. This framework was set up on the basis of European Treaty Article 107(3)(b), which specifically allows State Aid to be granted to deal with a severe economic crisis. The four Communications that are at the heart of this framework are briefly described in Chart 1.

Chart 1 - The EU Crisis SA Framework for Financial Institutions

Date	Communication	Main Principles
October 13, 2008	The Application of State Aid Rules to Measures Taken in Relation to Financial Institutions in the Context of the Current Global Financial Crisis[11] (Banking Communication)	Adapting certain principles of Rescue & Restructuring guidelines to financial cases, i.e. allowing capital injections, distinguishing between fundamentally sound and distressed institutions.
December 5, 2008	The Recapitalisation of Financial Institutions[12] (Recapitalisation Communication)	- Guidance on pricing of capital injections based on ECB recommendation (7 percent to 9.3 percent); - Threshold for in-depth restructuring requirement (2 percent aid/0 percent Risk Weighted Assets, as of Jan. 1, 2011).
February 25, 2009	The Treatment of Impaired Assets in the Community Banking sector[13] (IAC)	Valuation and assessment guidelines for transfer or guarantee by the State of toxic assets.
July 23, 2009	The Return to Viability and the Assessment of Restructuring Measures[14] (Restructuring Communication)	Principles of restructuring for rescued financial institutions: - Restoring the long-term viability without SA; - Burden-sharing; - Measures to address distortion of competition.

In practice, the crisis framework allows speedy rescues—often within 24 hours—that are temporarily approved on the basis of their compliance with the framework, that is, entry conditions. Temporary approval is followed by a final decision verifying compliance with the rules on restructuring and exit from State Aid. Exit is based on mandatory restructuring plans, initially in cases where recapitalization and/or asset relief aid was "significant," and from January 1, 2011 onwards, for all recapitalization and asset relief aid. The implementation of the conditions set out in the restructuring plans, which can have periods of up to 5 years, is monitored by the European Commission and its dedicated "monitoring trustees." The financial institutions concerned are thus subject to effective control from the implementation of agreed restructuring measures by the Commission for a prolonged period.

Under the crisis framework, Member States can notify the European Commission of either aid schemes or individual cases. The advantage of schemes is that once the conditions are agreed upon by the Commission, they can be used by Member States without subsequent need for agreement by the Commission. Recapitalization, asset relief, and guarantee schemes were thus established, and as of November 1, 2011, ten schemes are still in place.

The rules require that public support (whether through guarantees, recapitalizations or impaired asset measures) must be remunerated, is subject to common pricing rules to avoid distortions in the internal market, and must provide incentives for exiting aid. Moreover,

restructuring plans are assessed on the basis of viability, burden-sharing and the avoidance of distortions on competition.

The Commission, through its binding decisions, has often required significant adjustments in the banks' restructuring plans in order to minimize distortions of competition, including closing unprofitable businesses or selling assets. The framework itself exemplifies a pragmatic approach based on the proportionality principle, marrying policies protecting macro-financial stability with the established principles of competition policy for rescue and restructuring aid.

Internally, the European Commission set up a Financial Sector Task Force to pool expertise across Commission services, drawing in a small number of external financial sector specialists who developed the necessary consistency in case practice through novel management structures and processes. At the height of the crisis the Task Force comprised about 40 members.

III. THE APPLICATION OF THE CRISIS FRAMEWORK: A REVIEW IN OUTLINE

Member States injected unprecedented volumes of aid into the financial sector. Before the financial crisis, total State Aid in the European Union hovered around 0.5 percent of Gross Domestic Product ("GDP"). Then, from October 1, 2008 to October 1, 2011, the Commission approved € 4506.5 billion (36.7 percent of EU GDP) in aid. The bulk of the aid was authorized in 2008, when € 3457 billion (27.7 percent of EU GDP) was approved, mainly in the form of guarantees (i.e. contingent liabilities for the State). After 2008, the approved aid shifted focus to recapitalization of banks and impaired asset relief.[16]

Member States, however, did not use their full quota of approved aid. The overall amount of aid used in 2008-2010 stands at € 1608 billion (13.1 percent of EU GDP).[17] Guarantees and liquidity measures account for € 1199 billion, or roughly 9.8 percent of EU GDP. The remainder went toward recapitalization and impaired assets measures amounting to € 409 billion (3.3 percent of EU GDP). Slightly over 72 percent of the aid used has been granted through schemes; the rest was provided on ad hoc basis. While the aids granted for recapitalizations and impaired asset measures have led to actual expenditure by the state, the guarantees have to date not been called.

Expressed as a percentage of the size of the EU banking sector (approximately € 42 trillion), this equates to some 2 percent of banking sector assets given as guarantees and other liquidity measures, and about 1 percent in capital injections and asset relief measures.

In the period between October 1, 2008 and October 1, 2011, the Commission took a total of around 250 decisions in the financial services sector under the crisis rules. These decisions authorized, amended or prolonged more than 30 schemes and addressed the situation in 37 financial institutions in the form of individual decisions. The Commission has so far taken only one prohibition decision. Financial crisis measures were taken in all Member States, except Bulgaria, the Czech Republic, Estonia, Malta and Romania.

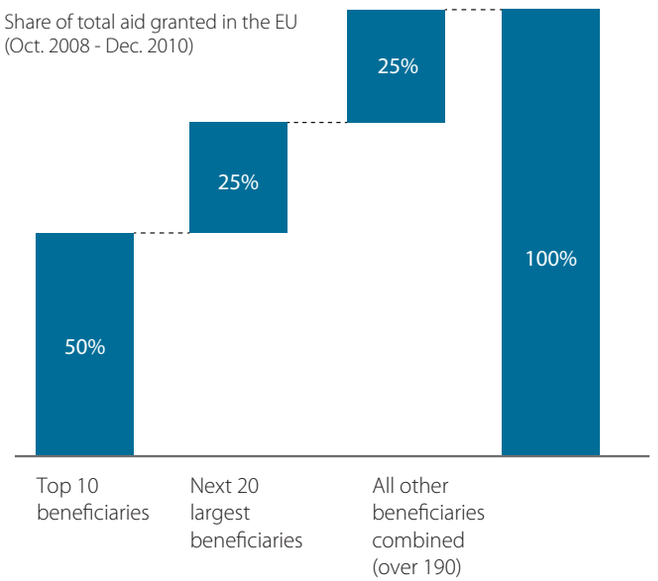
An interesting feature of the distribution of SA is the strong concentration in certain Member States and financial institutions. Banks in Germany, the United Kingdom and Ireland received about 60 percent of total aid. However, there was considerable variation in the relative importance of aid, i.e. as a percentage of the banking sector's size.

While Member States granted aid to, on average, 3 percent of the assets of their national banking sector, Greece and Ireland granted more than 8 percent.

The concentration of aid by bank was much more pronounced. Of the 215 Institutions receiving aid in the crisis until December 2010, 10 institutions were responsible for 50 percent of the aid; the next 20 took 25 percent of the aid. With the exceptions of Denmark and Spain, in all other Member States the top 3 beneficiaries received more than 50 percent, and in many cases more than 80 percent, of the aid.

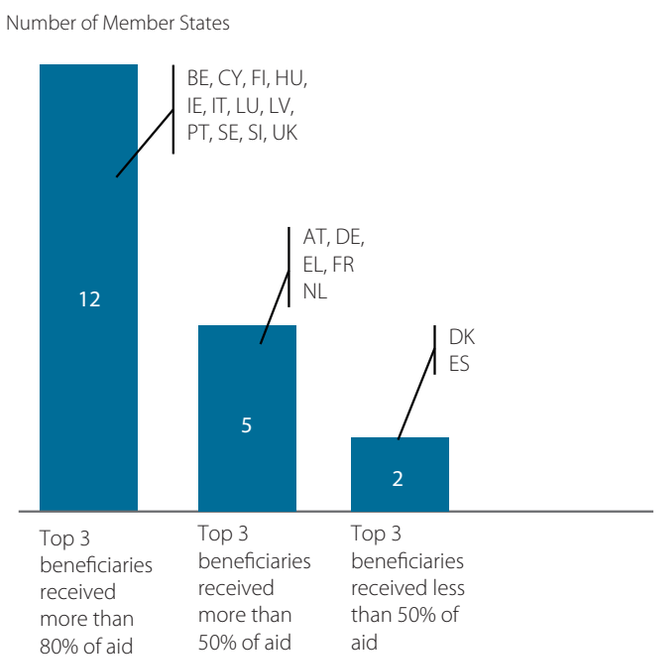
Figure 1: Concentration of Aid by Member State, October 2008 – December 2010

In the Single Market as a whole, 50% of aid was granted to 10 financial institutions.



Source: Comission Services

In most Member States, aid was concentrated on a few financial institutions



Source: Comission Services

Although the crisis was triggered by a horizontal and systemic shock, there were significant differences in the vulnerability of individual banks, often reflecting the strength of underlying business models. In fact, the Commission's decisional practice demonstrates that it believes only a small minority of banks was truly inherently vulnerable to the effects of the systemic shock on account of preexisting weaknesses.[18]

It was the uncertainty surrounding the precise situation of all the banks that subsequently led to system-wide contagion.

Addressing the root causes of the problems in weak institutions therefore had to be an essential component in any effective strategy to restore confidence in the banking system and to promote macro-financial stability.

A framework for access to aid to be applied all throughout the European Union needed to be coordinated to avoid stability-oriented policies by individual Member States that would be at the expense of other Member States. For example, the initial conditions of the proposed Irish guarantee system were such that a deposit outflow from UK and foreign banks located in Ireland (which were not covered) was triggered.

The Commission intervened to amend the scheme ensuring that all banks located in Ireland were covered.[19] This, in turn, also underlines the synergies between competition and stability policies as pursued by the European Commission.

Many of the largest recipients of aid had fundamentally unsound business models, were characterized by excessive risk taking, and often relied on excessive wholesale and short-term funding. The largest recipients of aid were all (relatively) large banks in their Member State of origin relying on an implicit state guarantee that, together with their funding model, led to a significant mispricing of risk.

The 15 largest beneficiaries of State Aid in the form of asset support during the reporting period have been restructured following a decision by the Commission, or submitted a restructuring plan that is still being assessed by the Commission. Those heavily-aided institutions originate from a few Member States: the United Kingdom (RBS and Lloyds Banking Group),

Ireland (Anglo Irish Bank, Allied Irish Banks), Belgium (Fortis, supported together with the Netherlands and Luxemburg; Dexia, supported together with France and Luxemburg; KBC), Germany (Bayern LB, Commerzbank, HSH Nordbank, IKB, LBBW and West LB), and the Netherlands (ING and ABN Amro).[20]

Given these facts, addressing moral hazard is of key importance in the case practice of the European Commission. Moreover, they further underscore the role of competition policy in the context of stability oriented financial assistance policies.

Chart 1: The EU Crisis SA Framework for Financial Institutions

Date	Member State	Restructured Institution	Date of decision	Type of decision	Aid received as % of RWA (capital injections and asset relief)
2008	Germany	IKB	21/10/2008	Restructuring	26%
	Denmark	Rosklide Bank	5/11/2008	Restructuring	-
2009	Germany	Commerzbank	7/05/2009	Restructuring	8.2%
	Belgium, Netherlands and Luxembourg	Fortis	12/05/2009	Restructuring	4.1%
	Germany	West LB*	12/05/2009	Restructuring	18.0%
	Luxembourg	Kaupthing Banl Luxembourg	9/07/2009	Liquidation	-
	Latvia	Parex Banka	15/09/2009	Restructuring	29%
	United Kingdom	Northern Rock	28/10/2009	Restructuring	>14.4%
	Netherlands	ING	18/11/2009	Restructuring	5.0%
	Belgium	KBC	18/11/2009	Restructuring	5.1%
	United Kingdom	Lloyds Banking Group	18/11/2009	Restructuring	4.1%
	United Kingdom	Royal Bank of Scotland	14/12/2009	Restructuring	19.6%
	Germany	LBBW	15/12/2009	Restructuring	8.3%
2010	United Kingdom	Bradford & Bingley	25/01/2010	Liquidation	-
	United Kingdom	Dumfermline Building Society	25/01/2010	Liquidation	-
	Netherlands	SNS REAAL**	28/01/2010	Restructuring	<2%
	Belgium, France and Luxembourg	Dexia	26/02/2010	Restructuring	5.5%
	Sweden	Carnegie Investment Bank	12/05/2010	Restructuring	-
	Belgium	Ethias	20/05/2010	Restructuring	13.8%
	Spain	Caja Castilla - La Mancha	26/06/2010	Restructuring	15.1%
	Austria	BAWAG	30/06/2010	Restructuring	2.4%
	Ireland	Bank of Ireland*	15/07/2010	Restructuring	4.8%
	Netherlands	Aegon	17/08/2010	Restructuring	3.8%
	Germany	Sparkasse Koln/Bonn	29/09/2010	Restructuring	3.3%
	Denmark	Fionia Bank	25/10/2010	Liquidation	-
	Spain	Caja Sur	8/11/2010	Restructuring	19.0%
2011	Austria	Kommunalkredit	31/03/2011	Restructuring	18.4%
	Netherlands	ABN Amro Group	05/04/2011	Restructuring	2.75%-3.5%
	Greece	Agricultural Bank of Greece	23/05/2011	Restructuring	8.3%
	Denmark	Eik Banken	06/06/2011	Liquidation	-
	Ireland	Anglo Irish Bank - INBS	29/06/2011	Liquidation	~50%
	Germany	Hypo Real Estate	18/07/2011	Restructuring	31.5%
	Germany	HSH Nordbank	20/09/2011	Restructuring	11.6%
	Ireland	Quinn Insurance Ltd	12/10/2011	Restructuring	-

* Both Institutions received State aid after the restructuring decision and are thus in the process of submiting an amended restructuring plan.
** Aid to SNS REAAL did not exceed 2% of RWA and therefore the Comission's decision is based on a viability review.
"-" Indicates that only liability support was provided

As illustrated in Figure 2, of the 250 institutions receiving State Aid until November 1, 2011, only the banks receiving the proportionally largest SA were subject to restructuring decisions. This reflects the proportionate approach the European Commission follows. Recipients of SA in excess of 5 percent of their risk weighted assets ("RWA") were typically required to undertake a wide set of restructuring measures to ensure viability, burden-sharing and minimization of competition distortions, including closing of unprofitable activities, sale of subsidiaries, acquisition bans, and prohibitions on paying out dividends or interest on capital instruments. In some cases, the set of restructuring measures led to significant downsizing of the institution, of at least 50 percent or more.[21] On the other hand, no restructuring decisions were imposed on the vast majority of institutions that benefited from small recapitalization aid amounts or guarantees.

Consistent risk-based pricing of these guarantees across EU banks ensures that sufficient coordination is achieved.

In taking restructuring decisions, the European Commission explicitly weighs the risk that divestments of foreign subsidiaries would fragment the internal market. In a number of cases, the Commission requested that banks divest assets in domestic markets instead,[22] with a view toward ensuring competitive market conditions therein. The business models of many banks were de-risked in this process, leading to greater viability. Of the 34 restructuring decisions taken by the Commission between October 1, 2010 and November 1, 2011, 6 ended up in a formal liquidation. In all, Member States also resolved a number of banks without resorting to State Aid, but the absence of resolution frameworks led to far fewer banks being liquidated in the European Union than in the United States, where the Federal Deposit Insurance Corporation resolved hundreds of (predominantly smaller) banks by relying on its federal resolution powers.[23]

In deciding aid applications, the Commission systematically applied the crisis framework to ensure a consistent treatment of all banks in all Member States. For example, the Commission Communications set out in Chart 1 require aid schemes to allow for non-discriminatory coverage of banks and financial institutions have to pay for the aid on the basis of EU-wide pricing rules. The key principles of the restructuring communication–long-term viability, burden-sharing and measures to limit distortions of competition— were applied to all institutions undergoing restructuring in the following ways:

- The Commission pursued restoring the **long-term viability** of banks through requirements relating to their business models. This often involved the divestment of weak subsidiaries and limitations on future investments (i.e. acquisition bans), when they would go at the expense of capital positions. Corporate governance changes were often essential to ensure a return to viability, including, where necessary, changes of management.

- **Burden-sharing** is achieved through management changes, dilution of ownership and control (which, in some significant cases like Northern Rock and ABN Amro, led to bank nationalizations), and dividend and coupon bans. Capital operations—buybacks of existing shares, exercising call options on hybrid capital instruments, or early redemption of subordinated debt at nominal value—are typically not allowed for the duration of the restructuring plan. The remuneration of management was also addressed, by requesting compliance with the Commission and G20 guiding principles.

- **Measures to limit distortions of competition** are introduced to mitigate the consequences on the competitive position of the aided bank. These measures comprise the sale of profitable subsidiaries or changes in the balance sheet that seek to promote more equitable conditions of competition. Behavioral measures such as price leadership bans and minimum return on capital standards for new loans have also been taken in a number of cases,[24] particularly where no relevant structural measures could easily be identified.

It is important to emphasize that while the European Commission seeks to apply a consistent approach to all banks, it does not follow that the measures it requires are identical in all cases.

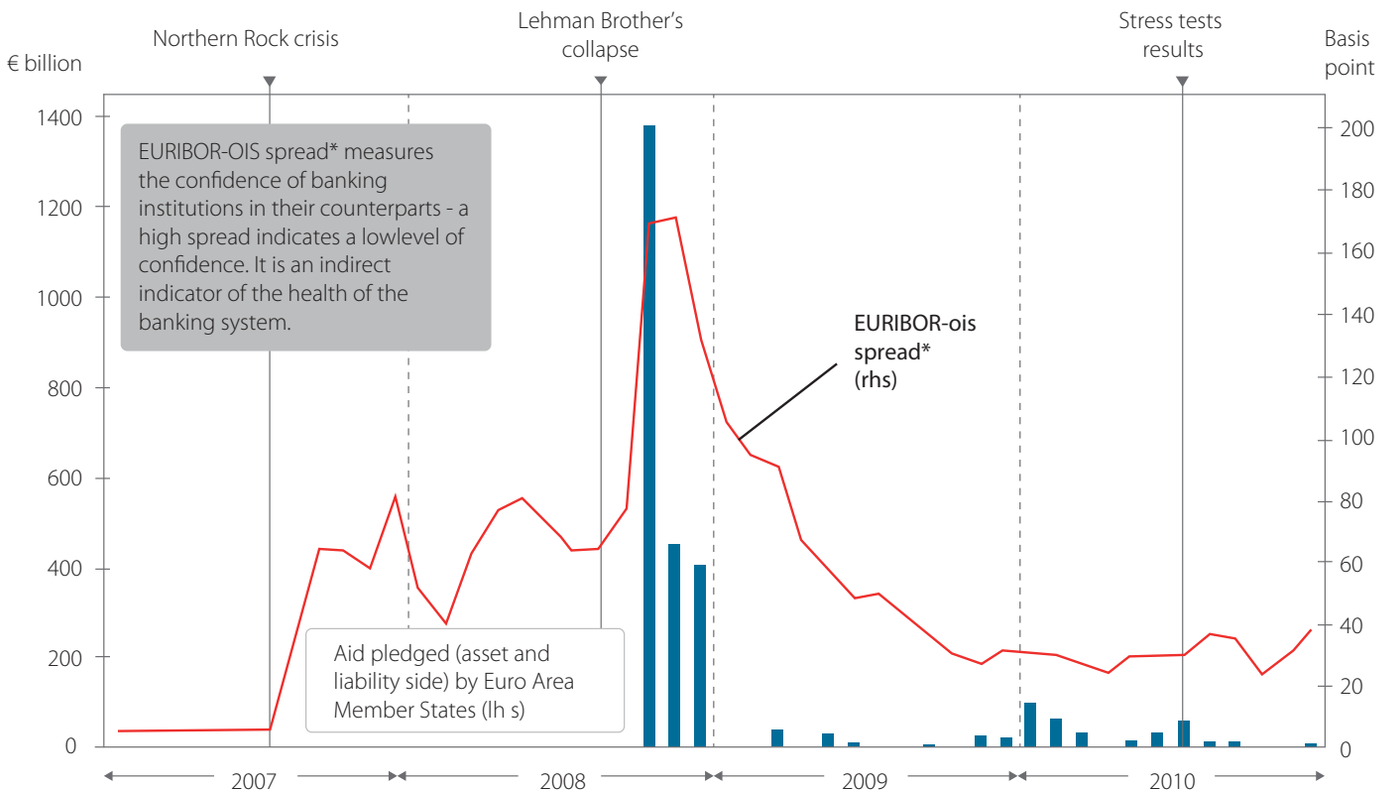
The restructuring requirements take the differences between banks into account, precisely in order to ensure equal treatment across all institutions concerned. The set up of the Task Force and the checks and balances in the European Commission all serve to ensure that this objective is met. Some Member States have taken action before Community Courts against crisis decisions by the

European Commission, but only a small number of complaints has been lodged,[25] and to date, no EC decision has been overruled by the Courts.

IV. OVERALL EFFECTS OF STATE AID AND STATE AID CONTROL DURING THE CRISIS

The significant volumes of aid to the EU financial sector, together with the intervention of the European Central Bank and the national banks, have helped mitigate the stability-eroding effects of the crisis. As Figure 3 shows, during the 2008-2010 period, the injections of aid are correlated with increases in confidence in the banking system as measured by the EURIBOR-OIS spread.

Figure 3: Evolution of EURIBOR-OIS Spread and of State Aid to the Financial Sector Pledged by Euro Area Member States



The rapid and large increases in capital in combination with the restructuring of the institutions concerned also led to improved lending conditions in the real economy as of the end of 2008 until the end of 2010.[26] A similar pattern is visible in the United States;[27] this experience stands in marked contrast with the handling of the Japanese banking crisis in the 1990s during which recapitalization, and especially restructuring, took place over the best part of the "lost decade."

Simulations using the QUEST-II macroeconomic model of the European Commission[28] also suggest that the amounts of State Aid have had a major positive effect on EU GDP. In the model, the interventions to support the financial sector mitigate the increase in equity risk premiums, thereby supporting investment that was particularly hard-hit by the crisis. Recapitalizations especially have a large GDP multiplier according to the model results.

Banks also managed to rebuild balance sheets and increase capital ratios, with the Core Tier 1 capital ratio rising by over 2 percentage points over the 2009-2010 period. The European banking sector as a whole also returned to profitability from the second quarter of 2010 onwards.

Evidence also suggests that after the initial strong tightening of credit standards and reduction in lending to the real economy, the situation began to improve again in 2010.[29] Although it is notoriously difficult to disentangle demand and supply factors, the overall evolution of the banking sector in 2010 suggests that the improvement in supply conditions played at least a supportive role.

The injections of large amounts of aid during the period 2008-2010 thus seem to have been effective in reaching their objective of strengthening macro-financial stability.

However, since the early summer of 2011, the situation of Europe's financial markets has started to deteriorate. The decline is caused by concerns with the sustainability of public finances in a number of distressed EU economies, in particular Greece, which led to steep increases in sovereign credit default swap spreads. This is illustrated in Figure 4.[30]

Figure 4: Sovereign Bond Spreads and Financial Market Stress

	10/27/2011	7/27/2011	Average 10/2008 – 3/2009	Average 10/2008 – 3/2009
Selected sovereign bond spreads over 10 year German Bunds				
Belgium	204	181	85	7
Greece	2,166	1,236	201	24
Spain	316	352	85	3
France	94	65	45	4
Ireland	625	850	178	1
Italy	370	331	123	22
Portugal	1,018	848	108	12
EU Default perceptions (Itraxx), spreads**				
High grade financials	207	173	132	n/a
Low grade financials	411	305	221	n/a

** 1 basis point equals annual cost in € 1000 for insuring against the default of € 10 million of debt for 5 years.

With markets increasingly concerned about the valuation of sovereign bonds in the hold-to-maturity accounts, the asset positions of banks, especially those located in countries with distressed sovereigns, started looking far less solid. Concerns about the consequences for the banks concerned and uncertainty about the true direct and indirect exposure of banks to weak sovereigns subsequently led to term funding drying up for many banks. On October 12, 2011, the European Commission published "Roadmap to Stability and Growth,"[31] a five-point strategy to break the vicious circle of doubts over the sustainability of sovereign debt, the stability of the banking system and the European Union's growth prospects,[32] including a plan to strengthen the resilience of the banking sector. As part of the overall support for such a comprehensive approach, ECOFIN subsequently endorsed on October 26, 2011 a proposal by the European Banking Authority ("EBA") to create temporary capital buffers after a prudential valuation of sovereign debt, and to require temporarily a 9 percent core Tier 1 level from all European banks by June 2012.[33]

It is estimated by the EBA that the largest European banks would need to reinforce their capital positions by around € 106 billion.

Although this should be accomplished from private sector sources, it is likely that further State Aid will be required. This is also likely to be the case for the effective implementation of coordinated initiative for term funding guarantees that the ECOFIN has also called for. In any event, this phase of the crisis has accentuated the strong interrelationship between the sustainability of public finances and the health of the financial sector in Europe.

The State Aid granted in 2008 and 2009 has had a positive effect on the stability of Europe's banking system (at least until the onslaught of the feedback loop from distressed sovereigns to banks), but it is difficult to isolate the effect of State Aid control during this period. The available indicators discussed below, however, suggest that the effect has been positive, both in terms of influencing stability through enhanced viability of the aided institutions, as well as with regard to its impact on the internal market.

The solvency ratio of aided institutions has increased broadly, similarly to that of non-aided institutions

over the 2008-2010 period,[34] suggesting that the restructuring and viability requirements of the former category have been successful, and that many of these banks have subsequently been able to inject private capital.

Moreover, the concentration on national and EU markets does not seem to have increased on account of the effects of the crisis and the State interventions that took place. The share of banking assets in individual Member States held by domestic institutions went up slightly in 2008, yet the trend subsequently stopped suggesting that there has not been a systematic retrenchment from cross-border activity in 2009 and 2010.[35] This is remarkable given that the State Aid framework could not substitute for public support at the European level, and cross-border banks in serious distress like Fortis and, in 2011, Dexia, had no choice but to break up into national parts as a consequence of the provision of financial support by their respective governments.

Across the European Union as a whole, the banking market does not seem to have become much more concentrated: overall, the level of concentration as measured by the Herfindahl-Hirschman Index went up by 10 percent in 2008 compared to 2005-2007, but this then decreased to 6 percent in 2009.[36] Aided banks have also not seen their overall share in the market increase. The largest aided banks typically experienced very significant balance sheet reductions as well as periods of low profitability: of the seven banks receiving aid in the Top 20 of the EU banking sector in 2008 Q1, only three still figured on the list in Q4 2010: Lloyds, Royal Bank of Scotland and ING.[37]

It is also important to emphasize that there does not seem to be much evidence that SA control would have led to a negative effect on lending to the real economy by forcing across-the-board deleveraging. Given that only banks with problematic business models were asked to divest assets, there is no indication that SA control under the crisis framework has exerted a general downward pressure on lending.

In this context, many divestments primarily lead to a reorganization of the structure of the banking sector, rather than to an impact on aggregate credit provision to the real economy.

While there is some anecdotal evidence that in the context of recently announced tighter capital standards, some banks may prefer to deleverage through reducing risk weighted assets or selling assets, rather than through accepting recapitalization aid, there is as yet no evidence that SA control has actually clashed with stability-oriented policies on account of this mechanism.

This positive assessment should be qualified in at least three ways. First, with the crisis still unfolding in the European Union, and given the short time period over which the effects of SA control have been assessed, any results at this stage are clearly preliminary and will need to be validated at a later stage by much more rigorous analysis. Second, it is clear that State Aid control cannot substitute for a reformed and revamped EU banking regulatory system, which through its design (e.g. through capital and liquidity requirements) reduces the likelihood of bank failure, and if a bank does fail, ensures that there are transparent and predictable rules in place to manage their resolution.

The Dodd-Frank Wall Street Reform and Consumer Protection Act,[38] for example, would have, according to the Federal Deposit Insurance Corporation ("FDIC"), allowed for an orderly resolution of Lehman Brothers.[39] Finally, sound macroeconomic policies, notably with regard to public finances, are a precondition for the effectiveness of all structural policies, both competition and regulatory. This latter observation is particularly relevant to the recent reemergence of the banking crisis in the European Union.

V. OUTLOOK

The European Commission is extending the SA crisis framework into 2012 to allow SA cases to be dealt with under these dedicated rules for as long as the crisis lasts. This will, therefore, apply to State Aid measures that may flow from the recapitalization and guarantee measures for European banks proposed by the EBA and endorsed by the ECOFIN Council on October 26. Given that the trigger point for these measures is linked to the EU sovereign debt crisis, it is expected that the application of the principle of proportionality will take full account of the extent to which recapitalizations occur, to offset losses resulting from prudent valuations of sovereign debt on bank balance sheets. The present framework is the appropriate tool to assess such cases, as also recognized by the European Council.

In the longer run, it is clear that SA control will need to be complemented by an appropriate regulatory framework in order to provide more stability and help de-risk the EU banking system. This principally relates to new capital and liquidity requirements for financial institutions, as well as European Union-wide rules on bank resolution. The European Commission has drawn up an ambitious work program in this respect, and many of the key proposals have already been tabled, including the CRD IV proposal made in July 2011, proposing key rules on, inter alia, capital, liquidity and leverage.[40] The creation of the European Banking Authority has strengthened centralized EU-level supervisory oversight, which is of particular relevance at the present time, given the close links between banks and the sovereign in which they are headquartered, particularly in the euro area.

An important characteristic of the proposed new capital requirements are the more demanding capital ratios required of large financial institutions: this would, to some extent, internalize the "too big to fail" advantage these institutions have in terms of funding costs. Moreover, as highlighted above, a bank resolution framework, on which the European Commission has announced that it will make a proposal,[41] would allow reducing aid to the banks in the first place or in the event of financial distress.

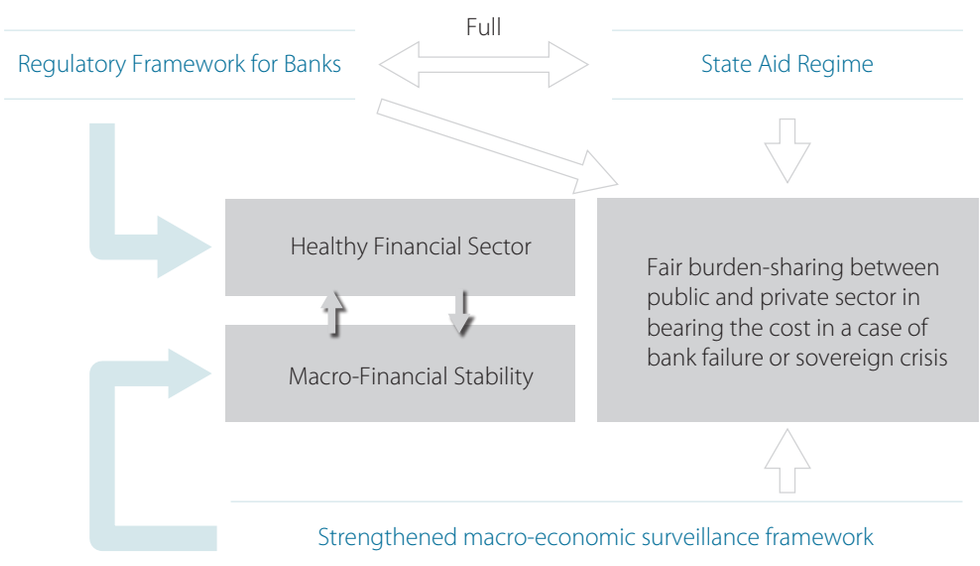
A mid-term challenge will be to ensure full consistency and compatibility between the SA rules and the regulatory framework.

This will be particularly relevant for the approach to be taken for burden-sharing. To the extent that the regulatory regime in the rules applying to all companies at all times effectively deals with moral hazard, it would fall less to the enforcement of State Aid control to achieve these objectives. Similarly, regulatory means could assist with ensuring that distressed banks would need to reform business practices and shed loss-making entities for viability reasons, even before they access public aid, if necessary and justified. It is premature to take this analysis forward at this stage, but it is already clear that the interplay between competition and regulatory policies will become more significant as the latter is further elaborated.

At the same time, a further strengthening of macroeconomic surveillance policies, including with regard to macro-prudential matters, will be of key relevance to strengthen macro-financial stability, particularly in the euro area.

Here, a more effective framework could reduce some of the pressure on State Aid control that presently attempts to integrate these concerns, inter alia, through the application of the proportionality principle. Most importantly, however, the further development of banking regulation and macroeconomic surveillance policies will lead to better overall results in terms of stability and competition in EU financial markets.

Figure 5: Complementarities Between SA Control, Banking Regulation and Macroeconomic Policies



1 This article draws heavily on European Commission, *The Effects of Temporary State Aid Rules Adopted in the Context of the Financial and Economic Crisis* (Commission Staff Working Paper, SEC 1126 final, 2011), to which many colleagues in the Competition department of the European Commission have contributed. I would also like to thank Sean Berrigan, Alexander Italianer, Stan Maes and Nicola Pesaresi for their valuable comments on a previous version of this article. Needless to say, the views expressed in this article cannot be ascribed to the European Commission and all remaining errors are mine.

2 *See* Jacques de Larosière, et al., Report of the High-Level Group on Financial Supervision in the EU (2009).

3 Treaty on European Union, 1992 O.J. (C 191), Art. 106-108.

4 For an analysis of the available options, see Gregory Nguyen, Mathias Dewatripont, Peter Praet & André Sapir, *The Role of State Aid Control in Improving Bank Resolution in Europe*, 4 Bruegel Pol'y Contribution (2010), *available at* http://www.bruegel.org/publications/publication-detail/publication/404-the-role-of-state-aid-control-in-improving-bank-resolution-in-europe.

5 *See* European Commission, *Impact of the Current Economic and Financial Crisis on Potential Output* (Directorate–General for Economic and Financial Affairs Occasional Paper No. 49, June 2009).

6 Douglas W. Diamond, *Financial Intermediation and Delegated Monitoring*, 51 Rev. Econ. Stud. 393 (1984); Franklin Allen, *The Market for Information and the Origin of Financial Intermediation*, 1 J. Fin. Intermed. 3 (1990).

7 *See* John H. Boyd & Gianni De Nicoló, *The Theory of Bank Risk-taking and Competition Revisited*, 60(3) J. Finance 60, 1329 (2005),

8 *See* Wolf Wagner, *Diversification at Financial Institutions and Systemic Crises* (Tilburg University, Center for Economic Research, Discussion Paper 2006-71, 2008).

9 Independent Commission On Banking, Final Report: Recommendations (Sept. 2011).

10 Communication from the Commission — Community Guidelines on State Aid for Rescuing and Restructuring Firms in Difficulty, 2004 O.J. (C 244/2).

11 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, 2008 O.J. (C 270/8).

12 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, 2009 O.J. (C 10/2).

13 Communication from the Commission on the Treatment of Impaired Assets in the Community Banking Sector, 2009 O.J. (C 72/1).

14 Commission Communication on the Return to Viability and the Assessment of Restructuring Measures in the Financial Sector in the Current Crisis Under the State Aid Rules, 2009 O.J. (C195/9).

15 *See, e.g.* State Aid C-11/09, Commission Decision on the Measures implemented by Dutch State for ABN AMRO Group NV (Apr. 5, 2011), *available at* http://ec.europa.eu/competition/state_aid/cases/230806/230806_1235915_338_2.pdf; State Aid N 428/09, Restructuring of Lloyds Banking Group, 2010 O.J. (C 46) 2.

16 European Commission, State Aid Scoreboard, Autumn Update (2011) (forthcoming), pp. 31-51.

17 *Id*.

18 *See* Figure 2, *infra*, for an overview of the main Institutions concerned.

19 State Aid NN 48/2008, Guarantee scheme for banks in Ireland, 2008 O.J. (C 312) 2.

20 European Commission, The Effects of Temporary State Aid Rules Adopted in the Context of the Financial and Economic Crisis (Commission Staff Working Paper, SEC 1126 final, 2011).

21 *See, e.g.*, Press Release, European Commission, State aid: Commission approves restructuring plan of Hypo Real Estate and clears the aid (July 18, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/898

22 *See, e.g.*, State Aid N 428/09, Restructuring of Lloyds Banking Group, 2010 O.J. (C 46) 2; State Aid C 18/09, Commission Decision on the State aid implemented by Belgium for KBC, 2010 O.J. (L 188) 24.

23 For a list of banks, see Federal Deposit Insurance Corporation, Failed Bank List, http://www.fdic.gov/bank/individual/failed/banklist.html (last visited Dec. 9, 2011).

24 See, e.g., State Aid N 528/2008, Participatie in het kernkapitaal van ING, 2008 O.J. (C 328) 4; Press Release, European Commission, State aid: Commission temporarily approves rescue aid for Dexia Bank Belgium; opens in-depth investigation (Oct. 17, 2011), available at http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1203.

25 See, e.g. Case T-33/10, ING Groep v. Comm'n, 2010 O.J. (C 80) 40.

26 European Commission, *The Effects of Temporary State Aid Rules Adopted in the Context of the Financial and Economic Crisis* (Commission Staff Working Paper, SEC 1126 final, 2011).

27 Troy Matheson, *Financial Conditions Indexes for the United States and Euro Area* 9-10 (International Monetary Fund, Working Paper 11/93, 2011), available at http://www.imf.org/external/pubs/ft/wp/2011/wp1193.pdf.

28 See Jan In't Veld & Werner Roeger, *The Effects of Bank Rescue Measures in the Recent Financial Crisis* (Sept. 2011) (forthcoming).

29 See Directorate-General for Economic and Financial Affairs of the European Commission, European Economic Forecast (Spring 2011).

30 *Weekly Credit Market Pulse of 10/28/2011*, internal publication of the Directorate General for Economic and Financial Affairs (on file with author).

31 Communication from the Commission — A Roadmap to Stability and Growth, COM (2011) 669 final (Oct. 12, 2011).

32 The five priorities are: (1) Give a decisive response to the problems of Greece; (2) Enhance the Euro area's backstops against the crisis; (3) Strengthen the banking system, namely through recapitalization; (4) Frontload stability and growth enhancing policies, and; (5) Build a more robust and integrated economic governance.

33 Council of the European Union, *Main Results of the Euro Summit* (Oct. 26, 2011), *available at* http://www.consilium.europa.eu/uedocs/cms_Data/docs/pressdata/en/ec/125645.pdf.

34 European Commission, *supra* note 18, at 82.

35 *Id*., at 100.

36 *Id*., at 98.

37 *Id*., at 101.

38 Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).

39 FDIC, *The Orderly Liquidation of Lehman Brothers Holdings Inc. Under the Dodd-Frank Act*, 5(2) FDIC Q. 31 (2011).

40 European Commission, *Proposal for a Directive of the European Parliament and of the Council on the access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms and amending Directive 2002/87/EC of the European Parliament and of the Council on the supplementary supervision of credit institutions, insurance undertakings and investment firms in a financial conglomerate*, COM (2011) 453 final (July 20, 2011).

41 European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the European Central Bank, regulating financial services for sustainable growth*, COM (2010) 301 final (June 2, 2010).

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

BANKING REGULATORY REFORM: "TOO BIG TO FAIL" AND WHAT STILL NEEDS TO BE DONE

Abel M. Mateus

New University of Lisbon,
University College London

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

BANKING REGULATORY REFORM: "TOO BIG TO FAIL" AND WHAT STILL NEEDS TO BE DONE

Abel M. Mateus*

ABSTRACT

Although important reforms have been undertaken in the United States and the European Union in the aftermath of the Great Financial Crisis of 2007-2009, major areas still need to be addressed. The Vickers Commission proposes a set of measures to solve the problem of too big to fail in the United Kingdom.

The proposal centers around the idea of ring fencing commercial banks and defining capital requirements separately for this compound. This paper discusses the pros and cons of the Vickers Commission proposal, comparing it with the Volcker rule, and problems of implementation. Complementary policies yet to be studied are also proposed.

* New University of Lisbon and University College London. This paper is based on a presentation at a seminar on Banking Regulatory Reform and the Vickers Report, at the Jevons Institute for Competition Law and Economics at University College of London, June 8, 2011.

It has now been about four years since the eruption of the Financial Crisis of 2007. Major reforms of the banking system have been achieved in the United States with the enactment of the Dodd-Frank Act in 2010[1], and several legislative initiatives in the European Union are to be completed by the end of 2012. Much has been achieved, but there are still areas that need further refinement and operationalization. Other areas have not yet been addressed at all. Although times of great disasters are times for major fixings of the system, we need to be aware that our present errors and omissions will seed the next financial crisis.

There are huge costs with financial crisis. The Basel Committee on Banking Supervision puts the median of the discounted cumulative costs of those crises at 63 percent of Gross Domestic Product ("GDP").[2] Andrew Haldane, Executive Director for Financial Stability at the Bank of England, estimates the costs of the 2007-2009 crises at a minimum of 90 percent of 2009 world GDP, and puts the average estimate at 220 percent of world GDP.[3]

The Independent Commission on Banking was set up in June 2010 and headed by Sir John Vickers. The main object of this paper is to comment on the Interim Report issued by the Commission in April 2011 (hereinafter "Report").[4] The Independent Commission on Banking is entrusted to formulate policy recommendations with a view to: (i) reducing systemic risk in the banking sector; (ii) mitigating moral hazard; (iii) reducing both the likelihood and impact of firm failure; and (iv) promoting competition in both retail and investment banking. In particular, the Commission is entrusted in making recommendations covering: "(a) [s]tructural measures to reform the banking system and promote stability and competition, including the complex issue of separating retail and investment banking functions; and (b) [r] elated non-structural measures to promote stability and competition in banking for the benefit of consumers and businesses."[5] The Terms of Reference explicitly state that the Commission, when making recommendations, should take into consideration the competitiveness of the UK financial and professional services sector.

The Report restates as its objective proposing reforms: (a) to reduce the probability of failure of systemically important banks by improving their resilience; and (b) to reduce the impact of failure of systemically important banks, both by providing for the orderly resolution of any institutions that fail, and by reducing levels of risk in the financial system as a whole, without disproportionately affecting the financial system's ability to provide critical financial services.[6]

There is a large consensus among the publications produced by academics and several institutions on the reforms required to strengthen financial regulation, especially in the United States and the European Union, after the 2007-2009 financial crisis. But few economists agree that those proposals have been fully translated into legislation. Although the reforms addressed in the Report are restricted to the areas indicated above, there are some reforms so interconnected that they need to be discussed in a compact. We will also use the opportunity to address some major areas related to the mission of the Basel Committee that need further work.

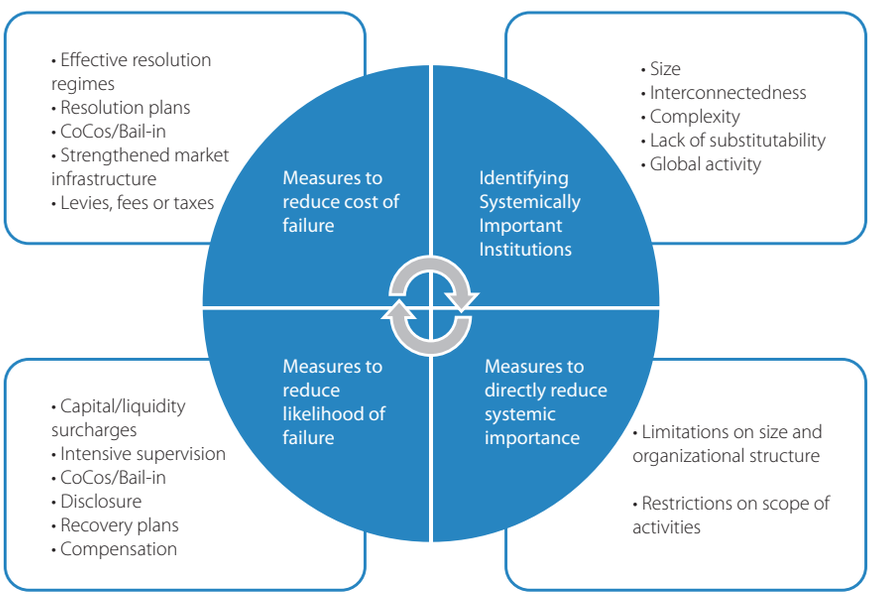
Section II confronts the problem of identifying systemic risky institutions, the basis for any discussion about this type of risk. We would not expect the Report to address a largely theoretical issue related to methodologies, but we think that without a theory to clearly identify systemic institutions, it is difficult to provide a policy addressed at them. We also discuss proposals for revising the regulation of capital and other own funds that is the most widely-known proposal for solving this risk. In contrast, both the Dodd-Frank Act and the Report make some "structural reform" proposals for solving the problem of too-big-to-fail. The proposal of the Report is discussed in Section III, along with the problems of implementation and the Volcker Rule of the Dodd-Frank Act. The Report concentrates on depository institutions.

This focus can be justified on the basis of the minimum required for bailout, both in terms of liquidity and in terms of the importance of the banks funding for the economy. However, most of the economists blame the so-called "shadow banking" not only for the source of the 2007-2009 crisis, but also for the propagation of the crisis to the overall financial system. In Section IV we discuss to what extent these institutions could undermine the banking sector, including the depository institutions, and require government intervention for preserving the stability of the financial system. In Section V we also raise the issue of governance in general, both of regulatory institutions and the regulated firms, an issue that has been completely ignored by the Report and most of the current reforms under way, and yet is at the core of the functioning of the financial system. Section VI refers to some competition issues covered by the Report, and Section VII concludes our discussion on banking regulatory reform.

II. SYSTEMIC RISK AND CURRENT RESPONSE TO REGULATION AFTER THE CRISIS: A HUGE MORAL-HAZARD PROBLEM

The large bailout programs undertaken by the U.S. and European governments in the aftermath of the 2007 crisis, either through recapitalization of banks, nationalization, blank deposit or credit guarantees in the order of trillions of dollars, have created a huge moral hazard problem for the future of banking.[7] Recent financial history has clearly established that if a large or systemically important institution is in trouble the government will come to her rescue. Taxpayers have footed the bill, leaving very little cost for shareholders, investors and bondholders to bear. Thus, we have four important problems to solve to lower the enormous moral hazard created by bailouts: (i) identifying systemic institutions, (ii) reducing structurally systemic risk by putting limits on size or building ring-fences, (iii) putting in place regimes and incentives so those institutions do not take inordinate risk, and (iv) putting in place resolution mechanisms that have a more equal burden sharing between taxpayer-shareholder-investor. To approach these four problems, the following diagram illustrates the broad tasks that need to be undertaken in any meaningful banking reform.

Dealing with the Risks Posed by Systemically Important Financial Institutions Beyond Basel III



The basis for any discussion of systemic risk is the characterization of what is an institution that is systemically risky (or too-big-to-fail).[9] There has been a substantial amount of theoretical work done in this field in the aftermath of the crisis. Most of the regulators that have been working with these concepts have used "stress tests" to evaluate systemic risk, although it is not clear to outsiders how the assumptions or scenarios given to banks for those tests are derived.

Most of the large banks, using Basel II, rely heavily on Internal Ratings-Based (IRB) risk models based on Value at Risk (VaR) calculations, considering only the bank or the banking group. The current regulatory regimes are still based in pro-cyclical capital requirements, haircuts and ratings. They focus on the asset side of the balance sheets of banks without taking into consideration the liability side and mismatches between liabilities and assets, with large implicit subsidies to short-term funding. Finally, as we will see below, the current regime has largely ignored the shadow banking system. As a result, the response by banks to current regulation is: "take positions that drag others down when you are in trouble (i.e. maximize bailout probability), become big, interconnected and/or hold similar positions."

Any analysis of systemic risk focuses on the contribution of externalities. The analysis should internalize externalities, and in terms of policy, build a fire protection wall. This requires that the analysis be translated into precise and rigorous capital requirements Only recently have there been rigorous theoretical characterizations of systemic risk. One of the major contributions is by Tobias Adrian and Markus Brunnermeier and their concept of CoVaR,[10] which is the covariance between Value at Risk of each institution vis-à-vis all the other institutions. CoVaR captures the institutions that are so large and interconnected that they can cause a negative spillover effect on the system. It also captures a subset of similar institutions that acting together can cause that negative spillover ("systemic as part of a herd").

Darrell Duffie proposes an alternative with his "10-by-10-by-10 Rule," which analyzes the results of stress tests among financial institutions.[11] A regulator would collect and analyze information concerning the exposures of N significant entities to M defined stress tests. For each stress, an entity would report its gain or loss, in total, and with respect to its contractual positions with each of the K entities for which the exposure, for that scenario, is among the K greatest in magnitude relative to all counterparties. Systemic counterparties would then be identified, stress by stress.

In addition to measuring the conditional CoVaRs, we have to eliminate the pro-cyclicality of the present ratios and build up a cushion to prevent a crisis in the future. Adrian and Brunnermeier propose to eliminate the pro-cyclicality by estimating the impact of state variables like the slope of the yield curve, the aggregate credit spread and the implied equity market volatility on tail risk. Then these time-varying CoVaRs are related to specific measures of each institution like maturity mismatch, leverage, market-to-book, size and market beta. The regression coefficients indicate how one should weigh the different firm characteristics in determining a systemic capital surcharge or Pigouvian tax.

The regulator can then establish a capital surcharge based on (forward) systemic risk contribution. It clearly changes ex-ante incentives to conduct activities that generate systemic risk. In addition, it increases the capital buffer of systemically important financial institutions, thus protecting the financial system against the risk spillovers and externalities from systemic institutions. This proposed methodology may sound complicated, but the authors have illustrated its application to major banking institutions in the United States and generated reasonable results. It shows a more complex web of interconnections than just a simple division between depository banks and the rest of the system, which is a warning sign for proposals based on the fault line proposed by the Report.

The rules proposed by the Basel Committee for Basel III have always aimed to establish minimum levels for solvability ratios, but those minimum levels are uniform. The present round of negotiations by Basel establishes a buffer capital of up to 2 percent for systemically important institutions. However, these proposed methodologies indicate that those ratios should not be uniform, but should be computed for each institution by the regulator. The same argument can be used against the proposal of the Report for what seems again a uniform solvability ratio by a depository institution.

Yet most of the studies eschew a phenomenon that deserves closer scrutiny: the fallacy of the composition. These situations may be more critical at the time of requiring an institution to improve its capital ratios. What is micro-prudent may not be macro-prudent. For example, suppose the regulator requires fire sales for resolving the problems of some large banks. It makes perfect sense at the level of the institution, but in the aggregate it will depress prices of the assets and deteriorate the balance sheets of even more banks. Other policies like ordering troubled institutions to stop giving more credit or take additional assets may force others to fire-sell or cause a credit crunch, again deteriorating the macro situation. The only policy where there is no clear conflict is when a bank is required to raise more equity.

It is quite clear that large banks that are individually systemic should be subject to both micro- and macro-regulation. However, the CoVaRs indicate that other sets of institutions also need to be regulated, even if they do not require both micro- and macro-regulation. Institutions that are systemic as part of a herd, such as leveraged hedge funds, should be subject to macro-prudential regulation but do not need micro-regulation. On the other hand, non-systemic but large institutions, like pension funds, need to be subject to micro-prudential regulation but not to macro-regulation.

Still related to the solvability ratios are two other problems. The first problem is that several rules established by Basel need an urgent revision.

Public securities continue to have a zero weight for Organisation for Economic Co-operation and Development ("OECD") countries when we have witnessed several of those states falling into unsustainable debt paths. In general, Basel has to grapple with a major conflict of interest by banks. Ratios computed based on VaRs and internal models, without proper regulatory supervision, constitute self-regulation and self-assessment of risk that has already led to major financial crisis. The problem of the ratings used in the calculation of solvability ratios has not yet been solved. Self-assessment is not an option, as some proposals have advanced, and rating agencies are still plagued by the problem of conflict of interest derived from the rule that the issuer pays.

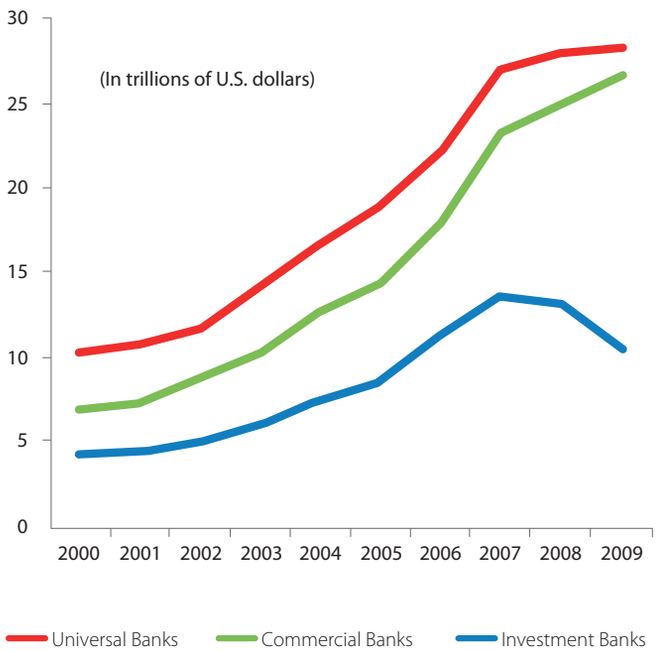
The second problem is crucial for macro-regulation. In the past, a number of asset bubbles have accumulated in the stock market. To prevent the build-up beyond a certain level of such bubbles it does not make sense to increase the capital requirements for banks. Moreover, increasing the interest rate by the central bank can precipitate a recession.[12] A solution that has not yet been implemented is to use another policy instrument that could influence the stock market more directly, namely margin requirements by all institutions trading in stocks, an instrument only rarely used in the past.[13]

II. THE TOO-BIG-TO-FAIL PROBLEM: THE VICKERS REPORT AND CONFLICTS OF INTEREST BETWEEN DEPOSITORY VERSUS INVESTMENT INSTITUTIONS

Despite the added risks they pose to financial stability, large financial institutions have important competitive advantages compared to systemically less important institutions. Large institutions possess the funding advantage of implicit or explicit government backing. Given their size and importance to their domestic economies, these institutions may enjoy strong political ties and hence may be in a position to influence policy via regulatory capture. In fact, logit analysis shows that the higher the probability of a rescue, the higher the share of the bank's assets to GDP and the higher the interconnectedness (and if it is a retail-oriented bank).

The relevance of these arguments has only increased over the past decade, as the institutions that could be considered as potentially systemic doubled their market share (see Figure 2).

Figure 2[6]: Growth in Assets (Sample of 84 Banks)



The main recommendation of the Report relates to the problem of the too-big-to-fail. It starts by focusing the analysis on depository banks or commercial banks. But are these the institutions on which to concentrate and restrict the analysis for financial stability? There are certainly very good arguments for answering in the affirmative. Deposit insurance is restricted to these institutions. Public insurance creates moral hazard problems. Bank runs are usually concentrated on depository institutions. Access to central banks is usually restricted to these institutions in order to provide funding as a lender of last resort, and they themselves have been a major provider of liquidity to the rest of the financial system. However, shadow banking cannot be ignored when dealing with financial stability today (see infra Section IV).

The Report starts to study two structural measures that are alternatives to solving the too-big-to-fail problem: break up banking groups in a depository bank and the rest of the bank, or build a ring-fence among them. These measures intend to solve three problems: (a) high impact of failure, (b) increased risk of system failure, and (c) increased risk taking.

The United Kingdom clearly opts for ring-fencing retail banking businesses from wholesale/investment banking activities through firewalls in a banking group. The Report makes a persuasive case for this solution by presenting a detailed cost-benefit analysis of each alternative. A retail ring-fence would allow for the continuation of universal banking, a form assumed by a large number of banks in Europe and the United Kingdom, with its attendant efficiency benefits of making the system more capable of absorbing shocks and reducing the perceived government guarantees. The operations to be ring-fenced are the provision of deposit-taking, payment and lending services to households and small and medium enterprises ("SMEs"). For the U.K. major banks it represents grossly 30 to 40 percent of their balance sheets. The ring-fencing serves the purposes of assigning a specific solvability ratio to the retail operations of the banking group and facilitating resolution in case of crisis.

In the United States the question of too-big-to-fail has been dealt with in several ways that diverge from the U.K. approach.

The Report considers the steps taken by the United States and assesses their feasibility in the United Kingdom. First, the United States abolished the Glass-Steagall Act of 1933[15] to separate commercial from investment banking.[16] The high costs and London's possible loss of competitiveness militate against such a measure in the United Kingdom, according to the Report. Second, the Volcker Rule contained in the Dodd-Frank Act restricts (with exceptions) banks' proprietary trading and investment in, or sponsorship of, hedge and private equity funds. The Report argues that these activities are small within the U.K. large banks and that it is difficult to separate proprietary trading from client-based trade. Furthermore, these activities in the ring-fence would be outside of the protected retail operations. Third, the Swap Pushout Rule in the Dodd-Frank Act requires certain entities relying on federal assistance and with significant swap business to move such activity to separately-capitalized nonbank affiliates.

But, as the Report recognizes, there are still important issues to be further clarified regarding (i) the implementation of the borderline between commercial and other activities, (ii) how to create stand-alone entities, and (iii) how to avoid cross-funding and funds transfer.

To protect a bank holding company seeking riskier assets to compensate for higher capital requirements,

it is necessary to have rules on what bets a retail subsidiary can make.

Such rules prevent it from ultimately behaving like an investment bank in retail clothing. The example of Lehman Brothers' failure shows a major increase in overall systemic risk that started in an investment bank and then spread to retail banking.

The Report recognizes the need to study some problems of implementation of the ring-fencing. We think that there are important technicalities and legal definitions. First, there is a need to define carefully the bail-in mechanisms—in particular, contingent capital—and the mechanisms to reinforce capital should clearly subordinate the claims of other senior unsecured creditors to those of depositors. Second, the 10 percent Core Tier I ratio requirement for retail banks by the Report should only be a benchmark. Relying on the theory expressed in Section II, supra, there is a need to use forward CoVaRs to establish the required amount by a regulator. Third, provided universal banks maintain minimum capital ratios and loss-absorbing debt for their U.K. retail operations, capital could be switched from the U.K. retail subsidiaries to other banking activities, which raises other concerns. Fourth, the current lack of a robust cross-border resolution mechanism, even within the European Union, is problematic. Fifth, assuming all the reforms are implemented,

there nevertheless continues to be a need for complementary micro-regulatory measures.

We know that at the origin of the recent financial crisis there were certain practices in the mortgage lending market.

Two important measures that should be enacted are prudential ratios in mortgage lending—like limiting loan-to-value ratio (70 to 80 percent), especially when a real estate bubble is on the making—and a reform in the governance of real estate valuations. Valuations need to be done by independent appraisers, avoiding the conflict of interest with the lending institutions. Other micro-regulatory reforms are proposed below regarding securitization.

IV. THE RISE OF SHADOW BANKING: TOO-SPARSE REFORM

Since the 1970s, there has been a major shift in the source of transaction media away from demand deposits toward money market mutual funds ("MMMFs"). MMMFs reached a peak of $3.8 trillion in 2008. Money market funds are registered investment companies that are regulated by the Securities and Exchange Commission (SEC) in accordance with Rule 2a-7 adopted pursuant to the Investment Company Act of 1940. [17]

Securitization also experienced a tremendous expansion. Securitization is the process by which traditionally illiquid loans are sold into capital markets. They are sold as large portfolios of loans to special purpose vehicles ("SPVs"), legal entities that issue rated securities in the capital markets. Total non-agency asset-backed security issuance reached $1.65 trillion on the eve of 2007.

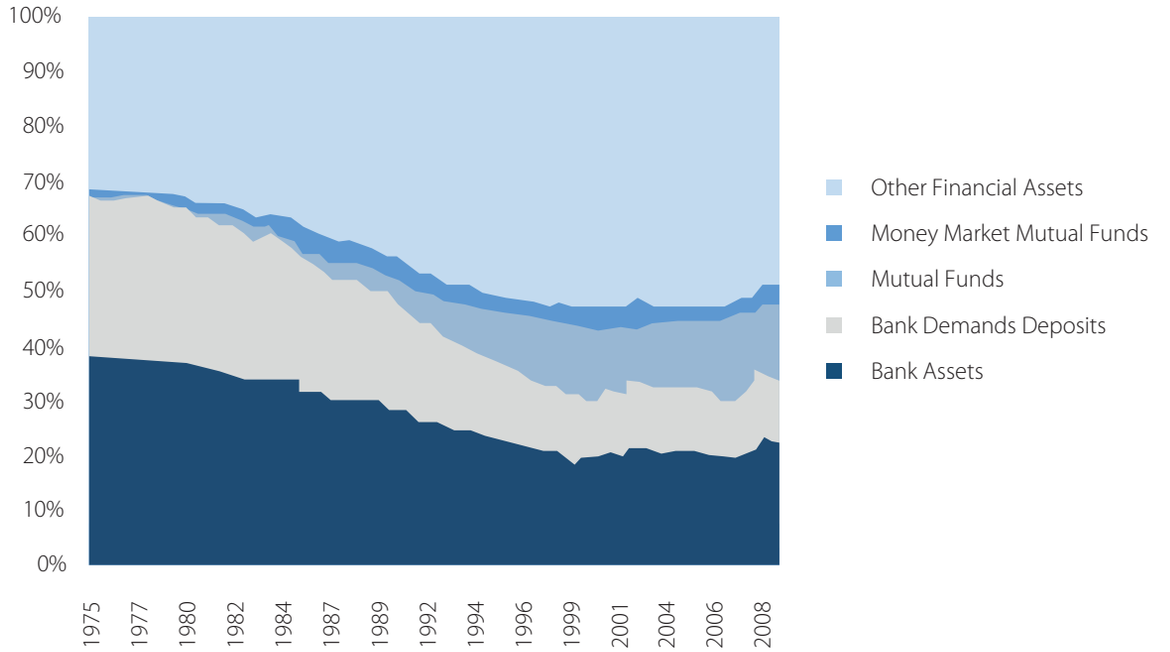
Large use of repurchase agreements ("repos"), as money under management by institutional investors (pension funds, mutual funds, states and municipalities, and nonfinancial firms) also expanded. Today they handle as many assets as banks: the repo market is about $5 trillion in the United States and $5 trillion in Europe.

Figure 3 shows the dramatic fall in the share of banks and the rise of shadow banking.

While banks had a share of 70 percent of total financial assets in the mid-1970s, it has dropped today to about 40 percent,

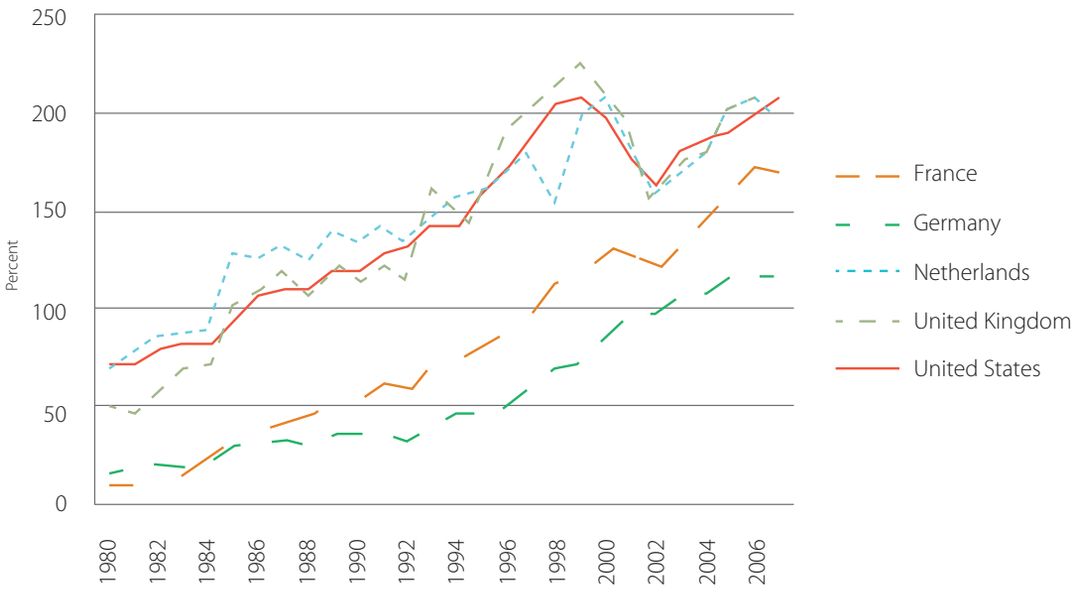
with a large part of this share being taken by mutual funds and other financial assets.

Figure 3: Money Market Mutual Funds, Mutual Funds, Demand Deposits, and Total Bank Assets as Percentages of Total Financial Assets



And, as the figure below reveals, this rise has occurred in all large developed countries.[18]

Financial Assets of Institutional Investors as a Percent of GDP



In the financial crisis of 2007-2009, problems arose in investment (shadow) banking and spread to retail banking. Investment banks transformed themselves into bank holding companies in order to have access to Federal Reserve funding and deposit insurance. The lender-of-last-resort role played by central banks saved depository banks around the world. These are simple facts usually forgotten.

A full analysis of the problems of shadow banking, including derivatives, is beyond the scope of the Report. We are going to mention just two issues more closely related with retail banking: the problem of money market funds (that are in fact quasi-banks), and securitization that has been widely used by banks to pass-on risk and acquire further liquidity.

The Group of Thirty ("G30") puts forth interesting proposals for the regulation of MMMFs.[19] G30 recommends the partition of MMMFs into two categories:

Type 1	**Type 2**
"Narrow Savings Banks" with a stable net asset values	Conservative investment funds with floating net asset values and no guaranteed return

Under this system, Type 1 funds are clearly within the safety net of explicit insurance and should be regulated as banks, while Type 2 funds should be clearly advertised as non-insured funds. G30 also proposes chartering narrow funding banks as vehicles to control and monitor securitization, combined with regulatory oversight of acceptable collateral and minimum haircuts for repos.

Regulation of securitization is certainly a major topic for reducing risk creation and subsequent spread. Securitization of mortgages lay at the center of the 2007-2009 crises, but securitization is now moving into SMEs, consumer credit, and additional areas. One of the problems of securitization oversight was the originator-to-distribution model. It is now recognized that the originator needs to retain a larger share of the risk (mainly equity risk) to avoid lax monitoring of debtors; the 5 percent imposed by the Dodd-Frank Act is insufficient. Moreover, slicing of packages should not dilute the incentive to monitor and enforce lending. Regulation of covered bonds, a new trend in securitization, is also inadequate, and forms yet another reminder that regulation usually lags behind market innovation. It is our opinion that covered bonds have a low weight (20 percent) accounting for the risk of the underwriter. Lastly, the problem of ratings of these packages has not yet been solved, see supra Section II.

V. A BASIC MISSING FRAMEWORK: GOVERNANCE REFORM

One of the most neglected areas in the reforms being discussed globally regards the governance of both the regulators and the regulated. No number of detailed new rules will succeed if the incentives on both sides are not properly aligned with the public interest of a stable and efficient financial system. To begin, regulators and supervision authorities need clear objectives and accountability to some democratic institution,

regulators and supervision authorities need clear objectives and accountability to some democratic institution

whether it is to the Executive or Parliament. If the bodies are remiss in their responsibilities, they should face serious consequences. Another major issue is to enact protections against regulatory capture. The firewalls erected between the different areas of regulatory bodies, and the activities conducted within them, need to be better defined. So do the firewalls between regulators and the government. Similarly, conflicts of interest that arise in the nominations for the regulators need to be avoided.

Regulatory bodies also need to identify and shape the incentives of their staff to maximize efficiency and productivity. A final problem is that of the thorny dilemma between transparency and confidentiality, in order to prevent false rumours or panic, those situations that hinder orderly resolutions of an institution. Publishing reports on failed institutions ex post, as audits, is only a partial solution, for such reports do not fully address consequences and responsibilities. More contentious is the publication of reports on troubled institutions in order to exert market discipline.

Turning now to governance of financial institutions, one of the most important issues is establishing rules for board nomination as controls on competence. There have been few recommendations on incentive mechanisms for board members. Bank executive pay remains substantially linked to an inappropriate metric of return on equity, which encourages executives to increase leverage.

Management has not been held fully responsible in more than a few cases of bailed-out institutions.

Beyond mechanisms within the financial institution, prompt court action should always be required in cases of fraud, which has not been the case in several European countries. Most of the recommendations in the area of governance of regulated firms address staff compensation, which, despite the Basel rules on micro-supervision, should be left largely to the institution.[20]

V. SOME COMPETITION ISSUES

The Report concludes that any limitation on the market shares of financial institutions is a blunt instrument, and that competition authorities are well-equipped to understand that it relates to abuses of dominance and mergers. We are much less confident.

Market shares may be blunt instruments, but they establish bright lines that are easy to implement.

The Dodd-Frank Act finally has it right after a hundred years of antitrust law in the United States. A simple limit of 10 percent market share in the European Union overall market should be established by European legislation, even if there is not yet any institution threatened by that restriction, which is not the case in the United States.

Methodologies for assessing bank mergers and intensity of competition are well-developed by the various E.U. competition authorities, but they seldom intervene and there have been instances where they have been overruled–most notably, by the U.K. Office of Fair Trading ("OFT") for the Lloyds TSB-HBOS merger.

The Report recognizes that banking markets are complex and subject to switching costs in current accounts for households and SMEs. The Report also recognizes that the OFT has done an excellent work in identifying those costs and taken some measures to improve competition.

We merely note that consumer protection is not enough: lowering barriers to switch may entail additional regulation, like imposing mandated reductions in those costs.

VI. CONCLUSIONS

We have surveyed in a previous paper the major reforms needed in the aftermath of the 2007-2009 financial crisis.[21]

Among the reforms required at the macro-level, the main ones are: (i) a systemic risk regulator with "teeth" that can control and reduce the systemic risk and the associated moral hazard, in particular the problem of too-big-to-fail, (ii) rationalization and coordination among regulators that are especially geared toward major financial institutions, to conduct consolidated analysis and regulatory measures, (iii) new instruments of the central bank to fight speculative bubbles, (iv) systems to resolve and maintain financial stability, including liquidity provision, and (v) regulation of over-the-counter derivative exchange markets.

The reforms required at the micro-level are mainly: (vi) strengthening the capital requirements of banks, correcting its cyclicality and its prudential role, with mark-to-market accounting systems, (vii) correcting the incentive problem of rating agencies, (viii) preventing problems of predatory lending and non-transparency of consumer products, (ix) establishing a speedy and effective resolution system for troubled institutions, (x) reducing the problem of originating and distributing in the process of securitization, and (xi) correcting the remuneration system in financial systems that gives an incentive to accumulate large risks.

The Report mainly addresses points (i), (iv) and (vi), explicitly leaving the other areas of reform to other national and international working groups, such as the Financial Stability Forum and the European Union Institutions. We are less optimistic in this appraisal.

Despite the limitations we refer to, the Report is excellent; its main proposals are well-grounded and can hardly been improved. Our suggestions address some of the points left open in the Report, in terms of implementation, and complementary policies and measures that need further analysis, either by the Independent Commission on Banking or other institutions.

1 Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub.L. 111-203, 124 Stat. 1376 (2010) (to be codified in scattered sections of the U.S. Code).

2 Bank for International Settlements, Basel Committee on Banking Supervision, *An Assessment of the Long-Term Economic Impact of the New Regulatory Framework* (August 2010), *available at* http://www.bis.org/publ/bcbs173.pdf.

3 Andrew Haldane, Executive Director for Financial Stability at the Bank of England, *The $100 Billion Question, Speech Before the Institute of Regulation & Risk, North Asia* (March 30, 2010), *available at* http://www.bankofengland.co.uk/publications/news/2010/036.htm.

4 Independent Commission on Banking, *Interim Report* (Apr. 2011), *available at* http://bankingcommission.independent.gov.uk/.

5 Press Release, Her Majesty's Treasury, Sir John Vickers to Chair the Independent Commission on Banking (June 16, 2010), *available at* http://www.hm-treasury.gov.uk/press_11_10.htm.

6 Interim Report, *supra* note 4, at 26.

7 Despite the large amounts committed for bailouts, the final fiscal cost is estimated at only a fraction of those amounts. Haldane, *supra* note 3, estimates the wealth transfer from the government to the banks as a result of the banking crisis in the United States at around $100 billion, less than 1 percent of GDP, and in the United Kingdom at £20 billion, slightly above 1 percent of GDP.

8 İnci Ötker-Robe et al., *The Too-Important-to-Fail Conundrum: Too Important to Ignore and Difficult to Solve* (IMF Staff Discussion Notes No. 11/12), available at http://www.imf.org/external/pubs/cat/longres.aspx?sk=24873.0.

9 Or similarly "too-big-to-bail" in the sense that it would cost a large amount of money to taxpayers.

10 Tobias Adrian & Markus K. Brunnermeier, *CoVaR* (Federal Reserve Bank of New York Research Paper Series - Staff Report, August 27, 2009), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1269446. According to the authors, the prefix "Co" refers to conditional, co-movement, contagion and contributing.

11 Darrell Duffie, *Systemic Risk Exposures: A 10-by-10-by-10 Approach* (National Bureau of Economic Research Working Paper No. 17281, Aug. 2011), available at www.nber.org/papers/w17281.pdf.

12 Besides, there might not be product market inflation.

13 Recall that margin requirements were used by the Chicago Mercantile Exchange in the wake of Black Monday, October 1987, when the margins were increased from 4 to 12 percent for S&P 500 futures. They were also used for the Long Term Capital Management (LTCM) crisis.

14 İnci Ötker-Robe et al., *supra* note 5.

15 Banking Act of 1933, Pub.L. 73-66, 48 Stat. 162 (1933).

16 Gramm–Leach–Bliley Act, Pub.L. 106-102, 113 Stat. 1338 (1989) (codified as amended in scattered sections of 12 and 15 U.S.C.).

17 The Investment Company Act of 1940, Pub.L. 76-768, 15 U.S.C. § 80a-1 to 80a-52 (1940).

18 İnci Ötker-Robe et al., *supra* note 5.

19 Working Group on Financial Stability, *Financial Reform: A Framework for Financial Stability* (Group of Thirty, Special Report, Jan. 15, 2009), *available at* http://www.group30.org/rpt_03.shtml.

20 Consultative Document, Basel Committee on Banking Supervision, International Framework for Liquidity Risk Measurement, Standards and Monitoring (Dec. 2009).

21 Abel M. Mateus, *After the Crisis: Reforming Financial Regulation* (Nov. 12, 2009), (unpublished manuscript, *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1504895).

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

U.S. AND EU ANTITRUST ENFORCEMENT: WHAT ROLE IN A MORE HEAVILY REGULATED FINANCIAL SECTOR?

Todd Fishman, Olivier Fréget &
David Gabathuler

Allen & Overy

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

U.S. AND EU ANTITRUST ENFORCEMENT: WHAT ROLE IN A MORE HEAVILY REGULATED FINANCIAL SECTOR?

Todd Fishman, Olivier Fréget & David Gabathuler*

ABSTRACT

The global financial crisis has led regulators and legislators in the United States and in the European Union to introduce a number of rules and regulations aimed at addressing market failures and improving regulatory enforcement in the banking and finance industry. The increasing convergence and complementarity of competition law and regulation across many regulated sectors, and the perceived commonality in interest, should mean that the antitrust authorities are strongly positioned to play an active and wide-ranging role alongside the financial regulators. Yet there is no consensus on whether unfettered competition in the banking sector will produce an optimal outcome in terms of financial stability.

Some believe that intense competition may be detrimental to stability by causing excessive risk taking, while others argue that too much oversight into the financial industries will chill investment activities and stifle the markets.

The apparent conflict between competition policy and a fundamental aim of financial regulation may explain, in part, why there has historically been a resistance to allowing competition policy to intervene heavily in the financial services sector. In particular, there are concerns regarding the ability of antitrust rules to address, quickly and effectively, conduct connected with deficiencies in market structure and transparency. This paper takes a comparative approach and examines how the enforcement of the competition rules in the United States and in the European Union could be constrained—on conflict grounds—by broadly-based rules and regulations addressing perceived market failures in the financial sector. It then briefly details the enforcement action taken by the U.S. and EU antitrust authorities in the financial sector following the advent of the economic crisis.

Finally, the paper concludes by discussing whether the apparent differences between the two systems may lead to divergent enforcement outcomes, particularly in terms of the level of scrutiny by the respective antitrust authorities. This discussion also highlights the risk of conflicts arising from the divergent interests of financial regulators and antitrust authorities.

* Todd Fishman is Partner in the New York office of Allen & Overy. Olivier Fréget is Partner in the Paris office. David Gabathuler is Senior Associate in the Brussels office. The authors are grateful for the research assistance of Marika Harjula, a trainee in the Brussels office, and Laena Keyashian, formerly a summer associate at the firm in New York. The views expressed are those of the authors alone and do not necessarily reflect those of Allen & Overy LLP or any of its clients. An earlier version of the article appeared in CPI Antitrust Chronicle, Summer 2011 (July-11(2)).

I. INTRODUCTION

The 2008 global financial crisis has given rise to a new set of supervisory and prudential rules and regulations governing the banking and finance industry. Regulators and legislators in the United States[1] and in the European Union[2], in particular, have been proposing and introducing a raft of legislative and regulatory measures to address apparent market failures and to improve regulatory enforcement.

The increasing convergence and complementarity of competition law and regulation across many regulated sectors, and the perceived commonality in interest, should mean that the antitrust authorities in the United States and the European Union are strongly positioned to play an active and wide-ranging role alongside the financial regulators.

Yet there is no consensus on whether unfettered competition in the banking sector will produce an optimal outcome in terms of financial stability. Some believe that intense competition may be detrimental to stability[3] by causing excessive risk taking, while others argue that too much oversight into the financial industries will chill investment activities and stifle the markets.

The apparent conflict between competition policy and a fundamental aim of financial regulation may explain, in part, why there has historically been a resistance to allowing competition policy to intervene heavily in the financial services sector. In particular, there are concerns regarding the ability of antitrust rules to address, quickly and effectively, conduct connected with deficiencies in market structure and transparency.

In the United States, the application of antitrust laws to regulated industries such as the financial services industry has sometimes been expressly precluded by statute, or implicitly by the courts. U.S. courts, for instance, give strong deference to traditional securities market regulators.[4] At the EU level, the exclusion of the competition rules is generally not foreseen, but the EU Merger Regulation[5] specifically provides for the competition assessment to be overruled by the need to protect other legitimate interests, in particular, "prudential rules."[6] Also, at the national level, a number of EU Member States appeared slow to grant the competition authorities unrestricted access to the banking sector.[7]

There seems to be a renewed appetite on the part of the antitrust authorities, both in the United States and the European Union, to use competition law instruments to challenge, in particular, suspected abuses of market power.

The European Union, for example, is currently examining whether the control and dissemination of financial market information by alleged dominant players unlawfully forecloses the market and distorts competition.[10]

In the United States, President Barack Obama and his administration pledged early in the presidency to increase antitrust enforcement in regulated industries and to maintain enforcement during the economic crisis. Christine Varney, Assistant Attorney General of the Department of Justice's ("DOJ") Antitrust Division, emphasized in May 2009 that "[f]irst there is no adequate substitute for a competitive market, particularly during times of economic distress. Second, vigorous antitrust enforcement must play a significant role in the Government's response to economic crises to ensure that markets remain competitive."[11]

It can be questioned, however, whether the introduction of a more robust financial regulatory scheme and the apparent resurgence of concerns about competition potentially weakening financial stability, and even possibly impeding effective regulation, will not have damaging consequences for competition law enforcement in the financial sector, and the banking industry in particular.

This paper takes a comparative approach and examines how the enforcement of the competition rules in the United States and in the European Union could be constrained—on conflict grounds—by broadly-based rules and regulations addressing perceived market failures in the financial sector. It then briefly details the enforcement action taken by the U.S. and EU antitrust authorities in the financial sector following the advent of the economic crisis. Finally, the paper concludes by discussing whether the apparent differences between the two systems may lead to divergent enforcement outcomes, particularly in terms of the level of scrutiny by the respective antitrust authorities. This discussion also highlights the risk of conflicts arising from the divergent interests of financial regulators and antitrust authorities.

II. THE U.S. POSITION

The application of U.S. antitrust laws to regulated industries, such as the banking and financial services industry, may be precluded in several ways. First, a regulatory statute may explicitly state that it precludes the application of antitrust laws. Second, when a regulatory statute is silent with respect to the application of antitrust laws, a court may find that the regulatory regime implicitly precludes the application of the antitrust laws. Congress may preserve the simultaneous operation of antitrust and regulation by the inclusion of a statutory savings clause specific to antitrust.

A) FIRST PRINCIPLES: THE U.S. SUPREME COURT'S BILLING DECISION

The Supreme Court's latest position on the application of antitrust laws to a regulated industry came in 2007 with *Credit Suisse Securities (USA) LLC v. Billing*.[12] The plaintiffs alleged that securities underwriters conspired to increase compensation for initial public offerings by inflating commissions and aftermarket prices under the pretext of the accepted practice of syndication. The Supreme Court ruled that the securities laws displaced the antitrust laws for the underwriters' activities and identified four factors to determine if "the securities laws are 'clearly incompatible' with the application of the antitrust laws,"[13] namely: (1) whether the underlying market activity is "an area of conduct squarely within the heartland of securities regulation"; (2) whether there is "clear and adequate Securities and Exchange Commission (SEC) authority to regulate" the conduct; (3) whether the conduct has been subject to "active and ongoing agency regulation," and; (4) whether a "serious conflict," or even a potential future conflict, exists between the antitrust and regulatory regimes.[14] As regards the fourth factor, the Supreme Court recognized that evidence of a "potential future conflict" might suffice for the securities laws to preclude antitrust liability "even in respect to a practice that both antitrust law and securities law might forbid."[15]

Billing left unanswered the question of how to apply the four factors and whether all four must weigh in favor of the regulated entity. This ambiguity has been reflected in the lower courts' subsequent treatment of the *Billing* test, but *the emerging consensus is that the conflict factor is decisive.*

The U.S. Court of Appeals for the Second Circuit addressed the issue in *Electronic Trading Group, LLC v. Banc of America Securities, LLC*, where it found that all four factors weighed in favor of implied immunity.[16] The short-seller plaintiff claimed that prime brokers charged "artificially inflated" borrowing fees to customers short-selling securities. The defendants allegedly designated securities arbitrarily as hard-to-borrow and fixed minimum borrowing fees for those securities. In applying *Billing*, the Second Circuit explained that, for cases involving regulated bodies, "[m]uch depends on the level of particularity or generality at which each *Billing* consideration is evaluated."[17] The court concluded that the first three *Billing* factors are to be "evaluated at the level most useful to the court in achieving the overarching goal of avoiding conflict between the securities and antitrust regimes" and that the fourth factor "is evaluated at the level of the alleged anticompetitive conduct."[18] It therefore appears that the critical factor for implied immunity is the conflict prong: where there is a conflict, or the prospect of a conflict, *the court is likely to find implied immunity to avoid a clash* between the two federal statutory regimes.

In at least one significant case since Billing, a court has determined that the antitrust laws and securities regulation are not incompatible. In *Dahl v. Bain Capital Partners, LLC*,[19] the trial court denied an effort to dismiss claims that private equity firms violated antitrust laws through the use of "club deals" (arrangements where groups of private equity funds sponsor leveraged buyouts ("LBOs"). The plaintiffs, a class of shareholders of companies that were taken private, alleged that the private equity firms conspired to allocate the LBO market in order to pay less than fair value of the target companies. Rejecting the private equity firms' argument that the U.S. Securities and Exchange Commission ("SEC") supervised the transactions in issue, the court held that "pre-emption does not apply here as the private nature of the LBOs at issue prevents the SEC from regulating these transactions."[20] Significantly, the trial court granted the plaintiff-shareholders' motion to expand the scope of their antitrust case to include ten additional transactions.[21]

While the U.S. courts wrestle with the implications of *Billing* in civil antitrust actions challenging conduct in the financial markets, the impact of the decision may be felt more acutely in two different contexts. First,

Billing is certain to be relevant to the legislative provisions of the Dodd–Frank Act

and the role of antitrust considerations in the rulemaking process within its new statutory scheme. Second, the decision is likely to reverberate throughout the investigations and other initiatives undertaken by the DOJ's Antitrust Division and its self-perceived role as an important participant in the evolution of the emerging derivative trading platforms that will define the financial markets for years to come.[22]

B) THE DODD-FRANK ACT

A notable recent example of an antitrust savings clause can be found in the influential Dodd-Frank Act, which aims to reduce risk, increase transparency, and promote market integrity within the financial system.[23] The Act enhances oversight and control in the financial sector by creating new recordkeeping, reporting, and execution requirements, and by giving regulatory bodies more power to make and enforce rules.

Billing suggests that the expansion of agency power would make activities under the Dodd-Frank Act prime candidates for implied antitrust immunity. However, § 6 of the Dodd-Frank Act contains a general antitrust savings clause[24] patterned on one that the Supreme Court found overcame implied preclusion of antitrust laws in *Verizon Communications Inc. v. Law Offices of Curtis v. Trinko, LLP*.[25] The Supreme Court upheld the effect of the savings clause, even though the enforcement scheme set up by a telecommunications regulatory regime was "a good candidate for implication of antitrust immunity."[26]

Modelling the Dodd-Frank Act's antitrust savings clause on the Trinko clause indicates a legislative attempt to combat the effects of Billing by precluding immunity.

Antitrust considerations are addressed elsewhere in the Dodd-Frank Act. The Insurance Bill contains its own antitrust savings clause that expressly mandates application of the antitrust laws even where there is a conflict.[27]

Moreover, regulators must consider antitrust where the Dodd-Frank Act requires that actions conform with provisions from other Acts containing restrictions on anticompetitive behavior, such as § 17A of the Securities Exchange Act of 1934.[28] By contrast, Title VII of the Dodd-Frank Act, which regulates the over-the-counter derivatives market and gives broad rulemaking powers to agencies, contains eight "Antitrust Consideration" provisions that place antitrust concerns behind those of the Dodd-Frank Act by allowing regulated entities[29] to engage in anticompetitive activities where "necessary or appropriate to achieve the purposes of [Dodd-Frank]...." [30]

These antitrust considerations operate, in effect, as a codification of Billing's fourth factor, consistent with Electronic Trading Group's interpretive gloss. Because Congress is capable of both specifying that conflicts should be resolved in favor of antitrust laws (as with the Insurance Bill), and delegating to regulators the responsibility of determining when antitrust laws should be pre-empted (as with Title VII), the antitrust considerations may be invoked to allow for antitrust immunity notwithstanding the general savings clause.

C) UNRESOLVED QUESTIONS

Notwithstanding its antitrust savings clause, it is an open question whether the U.S. courts will find that the Dodd-Frank Act precludes the application of antitrust laws. First, would a court apply the *Trinko* analysis in the financial context to find that the Dodd–Frank Act's broad antitrust savings clause completely bars implied preclusion of the antitrust laws? As Justice Clarence Thomas noted in his dissent in *Billing* (decided after *Trinko*), it is arguable that the antitrust savings clause contained in the Securities Exchange Act should have been given the same weight as that considered in *Trinko*.[31] The majority, however, rejected this argument. This distinction between the two savings clauses, as well as lower court decisions applying *Billing*, suggest that the *courts may view the financial industry as a special area where deference to federal regulators is especially important.*

It remains to be seen, however, whether deference to agencies will survive the perceived regulatory failures that are blamed for the credit crisis.

Second, in light of the credit crisis, will the DOJ respond by increasing its oversight of financial markets? Given the Obama administration's intensification of antitrust enforcement, coupled with the Dodd–Frank Act's general antitrust savings clause indicating the legislative intent of greater oversight and liability, the DOJ might modify its current approach.

Third, *Trinko* requires that, even if a statute contains a broad antitrust savings clause, a court "must always be attuned to the particular structure and circumstances of the industry at issue" and weigh the costs and benefits of antitrust intervention accordingly.[32] This leaves open the possibility that antitrust claims asserted in the context of a regulated industry may not survive, even in the face of a broad antitrust savings clause; indeed, the *Trinko* court ultimately found that the plaintiffs failed to state a valid antitrust claim. The inclusion of the Insurance Bill's savings clause also suggests that had Congress wanted to completely preclude antitrust immunity, it could have done so by using similar strong language as it did for the general savings clause.

The extensive new regulations (and attendant uncertainty) that the Dodd–Frank Act imposes on the banking and financial services industry, combined with the flurry of litigation arising out of the credit crisis and the possibility of treble damages for antitrust claims, strongly suggest that the intersection between antitrust law and the regulated financial market will be the subject of important litigation in the near future.

III. THE E.U. POSITION

The position taken by the Supreme Court in the *Billing* case is very different from the approach adopted by the EU institutions, including the Court of Justice of the European Union (the "ECJ"). The ECJ has consistently tried to ensure the broadest application of the competition rules in the EU Treaty[33] and has considerably limited the opportunity for parties to invoke a "regulatory defense" on the grounds of concurrent and conflicting application of sector-specific regulations and competition rules.

A) THE GENERAL APPLICABILITY OF THE EU ANTITRUST RULES

The ECJ summarily dismissed initial attempts in the 1980s to argue that the EU competition rules did not apply to the financial sector. In *Züchner v. Bayerische Vereinsbank AG*,[34] the defendant bank unsuccessfully argued that the EU Treaty provisions on competition did not generally apply to banks due to "the special nature of the services provided by such undertakings and the vital role which they play in transfers of capital."[35] In particular, the bank claimed that the financial activity (transfer of funds between Member States) should be treated as a service of general economic interest ("SGEI"[36]) falling outside the scope of the EU competition rules.

The court firmly rejected this broad assertion and stated that it would need to be established that the bank(s) had been specifically entrusted by an act of a public authority with such an SGEI.[37]

A different challenge was equally rejected by the court in *Verband der Sachversicherer v. Commission*.[38] The property insurers' association claimed that the EU competition rules could not be applied to the industry since the EU Council had yet to adopt special rules making them applicable to the insurance industry.[39] The association considered that there was an "obligation on the Council to temper the rigour of the prohibitions contained in the Treaty in so far as is necessary to ensure the survival of certain areas of economic activity."[40] It sought to highlight that "unlimited competition would result precisely in an increased risk of some insurance companies going out of business in view of the special characteristics of the industry."[41] Nevertheless, the ECJ emphasized that the Treaty contained no express derogation for the insurance industry and that the EU competition rules applied without restriction.

B) E.U. ANTITRUST RULES IN A "PRIVILEGED" POSITION

The presence of extensive (and increasing) EU and national rules and regulations addressing the financial sector creates, nonetheless, the opportunity for conflicts between regulatory provisions dealing with transparency and market conduct and EU antitrust rules which require free and open competition.

The hierarchy of norms within the EU legal system—with Treaty provisions and general principles of law at the pinnacle, above secondary legislation and implementing measures—places the competition rules enshrined in

Articles 101 and 102 of the Treaty on the Functioning of the EU ("TFEU") in a privileged provision. Nonetheless, it is difficult to envisage EU legislative acts in the financial services area being readily challenged[42] before the General Court (formerly the CFI) or the ECJ on grounds of their lack of conformity with the competition rules in the TFEU.[43] In any event, internal screening[44] within the EU institutions, and shared policy goals, including promotion of undistorted competition[45] within the Internal Market, are likely to reduce substantially the scope for conflicts between EU laws.

With regard to national laws and regulations, *the ECJ has largely limited the options for invoking a regulatory defense* to exclude the application of the EU competition rules. It has repeatedly stated that the EU competition rules are only inapplicable "if anti-competitive conduct is required of undertakings by national legislation, or if the latter creates a legal framework which itself eliminates any possibility of competitive activity on their part."[46] The EU antitrust rules would apply, however, if the national rules left open the possibility for competition, and if competition could be harmed by the autonomous conduct of the companies.[47] This would especially be the case if the national rules encouraged or made it easier for the companies to engage in anticompetitive conduct.

The EU legal order also places strict limits on the ability of Member States and national authorities to introduce or maintain legislation and regulations that could render EU laws ineffective. It is established case law that *the primacy of EU law requires any provision of national law that contravenes EU law, including the EU antitrust rules, to be disapplied by national courts and administrative bodies*, regardless of whether the provision in question was adopted before or after the EU provision. In circumstances where national rules and regulations conflict with the EU competition rules, the EU rules are given preeminence, although penalties cannot be imposed by the antitrust authorities in respect of past conduct required by national law.[48] To reduce further the scope of divergence, and ensure unity of interpretation of EU law, the ECJ will also give rulings on provisions of national law (outside the EU sphere) that refer to the content of provisions of EU law or adopt the same solutions as those found in EU law.[49]

C) A "REGULATED CONDUCT" DEFENSE?[50]

Direct conflicts between national rules and regulations and related provisions in EU law are becoming less common due to the greater convergence between European legal systems and the increasing harmonization of legal norms in the European Union, especially in relation to the Internal Market. However, opportunity for conflict in interpretation and application remains, especially in heavily regulated sectors.

In recent years, the ECJ and the General Court have considered the extent to which intervention by national regulators in the telecommunications sector could be used by companies as a defense to findings of antitrust infringement.

In the *Deutsche Telekom* ("DT") case,[51] the company argued on appeal before the General Court, and subsequently before the ECJ, that there could not be abusive pricing in the form of a margin squeeze because the charges were imposed by the German regulator ("RegTP"). However, the General Court ruled that "the fact that the applicant's charges had to be approved by RegTP does not absolve it from responsibility under Article 82 EC [now Article 102 TFEU]."[52] Both the General Court and the ECJ noted that the attribution of any infringement to DT depended on whether it had sufficient scope to fix its charges at a level that would have enabled it to end or reduce the margin squeeze. The courts found that DT had responsibility under Article 102 TFEU, despite national regulatory approval, as it had sufficient scope to end or reduce the margin squeeze within the limits imposed by regulation (i.e. in this instance, by increasing the retail prices within the price cap). The ECJ upheld the General Court's finding that DT had failed to exercise this discretion by not increasing its retail access prices.

A similar question has arisen in relation to the European Commission's ("Commission") 2007 margin squeeze decision concerning the Spanish incumbent telecoms operator Telefónica. Surprisingly, the Spanish government has itself appealed the decision on a number of grounds, including: that the decision impinged on the regulatory framework in force in Spain

(a framework grounded in EU law and supervised by the Commission); that it resulted in an ex post change to the regulatory framework, and; that the matter had already been addressed by the Spanish regulator.[53]

The pending appeal provides the courts with the opportunity to add to the jurisprudence on the interface between competition and regulation. It would, nonetheless, be unexpected for the General Court to depart from the ECJ's (and its own) previous case law and allow greater latitude for regulatory regimes to displace the EU competition rules.

IV. ANTITRUST ENFORCEMENT IN THE FINANCIAL SECTOR AFTER THE ONSET OF THE 2008 CRISIS

Parallel activity of financial regulators and antitrust authorities will not always raise questions of conflicts; there are areas where dual enforcement can be beneficial without giving rise to dispute. The complementarity of the two instruments has been highlighted by the EU Commissioner for Competition, Vice President Joaquín Almunia. He emphasized that "regulation tackles broad structural market failures" and "you need competition policy to tackle the harmful behaviour of individual market participants." [54]

The Commission has thus been very active in the financial services sector, notwithstanding the introduction of many new legislative and regulatory measures. Similarly, the Antitrust Division of the U.S. DOJ has been actively participating in the Financial Fraud Enforcement Task Force, which, for instance, has pursued a wide-ranging investigation into price-fixing in the municipal bonds investment market.

A) THE DOJ'S ANTITRUST INVESTIGATIONS AND ADVOCACY

The DOJ's activities have been marked by four recent investigations into the financial markets. In 2010, KeySpan Corp. admitted to violating antitrust laws by entering into a swap agreement with its largest competitor, thereby eliminating its incentive to sell electricity at lower prices.[55] Investigations into the municipal bonds investment market, credit derivative markets and the London Interbank Offer Rates ("LIBOR") are still ongoing. The municipal bonds investigation resulted in restitution and other financial penalties imposed on Bank of America in December 2010 and UBS in May 2011, amounting to $137 million and $160 million, respectively. In July 2011, the DOJ announced that JP Morgan Chase had agreed to pay a total of $228 million in restitution, penalties and disgorgement to federal and state agencies. This investigation also resulted in nine guilty pleas to criminal offenses and pending criminal charges against nine other individuals.

For the credit derivatives and LIBOR investigations, no public action has yet been taken and the DOJ has yet to clearly or directly target the activities of "Too Big To Fail" banks.

However, the DOJ has taken a more active role in the context of the Dodd-Frank Act rulemaking process.

It pointedly commented on the U.S. Commodity Futures Trading Commission's ("CFTC") proposed rules for derivatives clearing organizations, designated contract markets and swap execution facilities.[56] Citing its "significant experience in issues relating to the derivatives industry,"[57] the DOJ expressed its strong support for the CFTC's plan "to create meaningful limits on ownership of [derivative trading platforms], as well as its use of governance restrictions as a safeguard against conflicts of interest."[58] The DOJ explained, for example, that "major dealers might use their control of a dominant trading platform to disadvantage rivals by refusing to trade their products or to continue trading over the counter in instances where exchange trading is feasible."[59]

B) THE COMMISSION'S ANTITRUST INVESTIGATIONS

The Commission has increased the number of investigations in the financial sector following the onset of the economic crisis.

These high-profile investigations have often been targeted at areas of the financial services sector that have been viewed in some European political circles as lacking appropriate regulatory oversight and transparency.[60]

In *Standard & Poor's* (*S&P*),[61] the Commission recently investigated whether the ratings agency had been charging abusive prices in violation of Article 102 TFEU with regard to its legal monopoly over the distribution of International Securities Identification Numbers developed by ISO, the International Organization for Standardization. S&P offered commitments to change its pricing policy to address competition concerns identified by the Commission in the Statement of Objections and, following revisions made in response to observations received in the course of a market test, the Commission adopted a decision on November 15, 2011, making the commitments binding on S&P.[62]

In *Thomson Reuters*,[63] the Commission has been investigating whether Thomson Reuters is infringing Article 102 TFEU by imposing certain restrictions on the use of Reuters Instrument Codes, which prevent customers or competitors from translating these codes to alternative identification codes of other datafeed suppliers. It is reported by the Commission that, without the possibility of such mapping, customers may potentially be "locked into" working with Thomson Reuters because the procedure to replace the codes by reconfiguring or by rewriting software applications is long and costly.

The Commission is also carrying out investigations into the credit default swaps ("CDS") sector.[64] The Commission has reported that it is examining whether sixteen investment banks and Markit (a provider of financial information in the CDS sector) have been foreclosing access to raw data to other information service providers. It has also reported that it is separately investigating nine of the sixteen banks in relation to the tariffs granted by ICE Clear Europe (a clearing house for CDS) that allegedly create an incentive for the banks to use only ICE, thereby preventing entry by other clearing houses.

More recently, the Commission commenced an investigation into the sector of financial derivative products linked to Euro interest rates (Euribor) with a series of high-profile on-site inspections. The Commission reported that it was seeking evidence of possible illicit arrangements.[65]

V. CONCLUSION

The emergence of a broad set of new rules and regulations governing market behavior by banks and financial institutions, as well as the greater oversight of the sector by (in some cases) recently-created supervisory agencies, heightens uncertainty and increases the risk of substantive and jurisdictional conflicts between antitrust and financial regulation, both in the United States and in the European Union.

The mechanisms and prospects of resolving these concerns in the United States and in the European Union seem very different. The U.S. system appears to be prepared to show greater deference to regulation. It also provides the possibility for the legislature or the courts to disapply the antitrust rules in the overarching interest of avoiding conflict between financial regulations and antitrust rules. In the European Union, *the incorporation of the competition rules in the EU Treaty and their role as instruments of market integration lends them a quasi-constitutional aura,* thereby limiting the options for them to be overridden. This may explain why the DOJ's efforts indicate a cautious interventionist approach to the financial sector, while the Commission appears to be increasingly willing to launch high-profile antitrust investigations into the financial markets.

There are, however, a number of built-in safety valves in the EU system that can reduce the potential for conflicts. First, enforcement is primarily led by competition authorities, and these administrative bodies are likely to be more attuned to the risks associated with conflicting legal regimes than private litigants enforcing their rights through the courts. Second, it can be argued that the EU competition rules, and in particular Article 101(3) TFEU,[66] provide for public policy considerations to be factored into the antitrust assessment. Therefore, at least in terms of enforcement outcome, *the difference between the U.S. system and the EU system is probably less pronounced than it appears* from the underlying legal instruments and court precedents, especially as there is increasing coordination and commonality between antitrust authorities.

Conflicts in the financial sector may arise not only from a difference in antitrust enforcement by the U.S. and EU

competition authorities, but could also flow from the diverging interests of financial regulators and competition authorities. In particular, financial regulators might not share the competition priorities of antitrust authorities and might view antitrust instruments as too blunt and unwieldy to be effective in the highly complex area of banking and finance.

One can also imagine that antitrust authorities' concerns about heightened entry barriers or increasing market transparency in certain highly concentrated financial markets may sit oddly with financial regulators' aims of strengthening prudential safeguards across the industry. In this regard, it is worth highlighting, as an example, that the European Union has been substantially increasing the regulatory oversight of credit rating agencies ("CRAs").[67] In the European Union, CRAs will be subject to extensive centralized regulation by the recently created European Securities and Markets Authority.[68] There is, however, a general perception of a lack of competition[69] in the sector, due to the unrivaled position of the three leading CRAs, and it remains to be seen whether the increased regulatory burden may not further weaken competition by considerably increasing the cost of market entry.[70]

The increasing forays of antitrust into an ever more heavily regulated financial services sector bring the possibility of conflict to the fore. Given the importance of the sector to the wider economy and the concerns about stability, contagion, and systemic risk, *measures may need to be taken to ensure proper transparency of the role or authority of antitrust agencies* with regard to their sphere of influence in the banking and financial services area.

1 *See* Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).

2 The European Union created three new European Supervisory Authorities: the European Banking Authority ("EBA"), the European Insurance and Occupational Pensions Authority ("EIOPA"), and the European Securities and Markets Authority ("ESMA"). The European Union is also reviewing and revising a number of Directives and Regulations to, among other things, strengthen prudential requirements, improve internal risk management, and increase the level of available information. Measures are also being taken to improve transparency and adapt regulation to the innovation occurring in the financial markets. *See* European Commission, Regulating Financial Services For Sustainable Growth (Feb. 2011); *see generally* European Commission, Financial Services – General Policy, http://ec.europa.eu/internal_market/finances/index_en.htm (last visited Nov. 15, 2011).

3 *See, e.g.*, Barbara Casu, Claudia Girardone & Philip Molyneux, *Is There a Conflict Between Competition and Financial Stability?*, *in* Research Handbook For Banking And Governance (James Barth, Clas Wihlborg & Chen Lin eds., 2011).

4 *See Gordon v. New York Stock Exchange*, 422 U.S. 659 (1975); *United States v. National Ass'n of Securities Dealers*, 422 U.S. 694 (1975).

5 Council Regulation 139/2004, art. 21(4), 2004 O.J. (L 24) (EC) (providing that Member States may take appropriate measures to protect legitimate interests other than those taken into consideration by the Merger Regulation).

6 *See* Stéphane Kerjean, *The Legal Implications of the Prudential Supervisory Assessment of Bank Mergers and Acquisitions Under EU Law* (European Central Bank, Legal Working Paper Series No. 6, Jun. 2008), *available at* http://www.ecb.int/pub/pdf/scplps/ecblwp6.pdf (discussing prudential interests and European merger control rules).

7 For example, until the end of 2005, the Italian central bank, not the competition authority, applied the competition rules. In addition, the Dutch banking sector was excluded from the application of the merger control regime in the national Competition Act for two years following its entry into force in 1998, since mergers between banking and insurance institutions were already regulated by sector-specific legislation on the basis of a wider test applied by the Minister of Finance (or, in specific situations, the Dutch Central Bank).

8 *See, e.g.*, John Fingleton, Chief Executive, Office of Fair Trading, Speech at the Charles River Associates Conference: Competition Policy in Troubled Times (Jan. 20, 2009); Neelie Kroes, former European Commissioner for Competition Policy, Speech at the Bundeskartellamt conference Dominant Companies – The Thin Line between Regulation and Competition Law: The Interface Between Regulation and Competition (Apr. 28, 2009).

9 *See* Joaquín Almunia, Vice President of the European Comm'n responsible for Competition Policy, Speech at the Conference on Competition in Sensitive Sectors of the Romanian Economy: Fuelling Growth in Romania: The Role of Competition Policy (Oct. 21, 2011). *See also* DG Competition, How Competition Policy Is Helping Economic Recovery, http://ec.europa.eu/competition/recovery/index.html (last visited Nov. 28, 2011).

10 The Commission has recently investigated the activities of Standard & Poor, and is currently conducting investigations into (i) Thomson Reuters, and (ii) 16 investment banks and Markit (CDS market). See also Joaquín Almunia, Vice President of the European Comm'n responsible for Competition Policy, Speech at CASS Business School: Competition Policy Issues in Financial Markets (May 16, 2011). *See also* Section IV, infra.

11 *See* Christine Varney, Assistant Attorney General, Antitrust Division, U.S. DOJ, Remarks as Prepared for the Center for American Progress: Vigorous Antitrust Enforcement in this Challenging Era (May 11, 2009).

12 *Credit Suisse Securities LLC v. Billing*, 551 U.S. 264 (2007).

13 *Id*. at 285.

14 *Id*.

15 *Id*. at 273.

16 *Electronic Trading Group, LLC v. Bank of America Secs., LLC*, 588 F.3d 128 (2d Cir. 2009).

17 *Id*. at 131.

18 *Id*. at 132.

19 *Dahl v. Bain Capital Partners, LLC*, 589 F. Supp. 2d 112 (D. Mass. 2008).

20 *Id*. at 117.

21 *Dahl v. Bain Capital Partners, LLC*, No. 07-1288 (EFH), slip op. (D. Mass. Sept. 7, 2011).

22 *See*, *infra*, Section IV.

23 Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).

24 *Id.* § 6 ("Nothing in this Act, or any amendment made by this Act, shall be construed to modify, impair or supersede the operation of any of the antitrust laws, unless otherwise specified.").

25 *Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398 (2004).

26 *Id*. at 406.

27 Dodd-Frank Act § 541 ("Nothing in this subtitle or the amendments made by this subtitle shall be construed to modify, impair or supersede the application of the antitrust laws. Any implied or actual conflict between this subtitle and any amendments to this subtitle and the antitrust laws shall be resolved in favor of the operation of the antitrust laws.").

28 Securities Exchange Act of 1934, §17A, 15 U.S.C. § 78a et seq. (2006). *See, e.g.*, Dodd-Frank Act § 763 ("In reviewing a submission . . . , the Commission shall review whether the submission is consistent with section 17A.").

29 The Antitrust Considerations apply to derivatives clearing organizations, swap data repositories, swap dealers, major swap participants, swap execution facilities, boards of trade, security-based swap execution facilities, swap data repositories, security-based swap dealers, and major security-based swap participants.

30 *See, e.g.*, Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).

31 *Credit Suisse Securities LLC v. Billing*, 551 U.S. 264, 288-289 (2007) (Thomas, J., dissenting).

32 *Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398, 411 (2004).

33 EU Treaty, Title VII, Chapter 1.

34 Case C-172/80, Gerhard Züchner v. Bayerische Vereinsbank AG, 1981 E.C.R. 2021.

35 *Id*. at ¶ 6.

36 Services of general economic interest ("SGEI") are economic activities that public authorities identify as being of particular importance to citizens and that would not generally be supplied (or would be supplied under different conditions) if there were no public intervention (e.g. transport networks).

37 Züchner, 1981 E.C.R. at ¶¶ 6-9.

38 Case-45/85, Verband der Sachversicherer e.V. v Commission, 1987 E.C.R. 405.

39 Article 87(2)(c) of the EEC Treaty [now Article 103(2)(c) TFEU] allows the EU Council to define, if need be, in the various branches of the economy the scope of the provisions of Articles 101 and 102 TFEU.

40 Verband der Sachversicherer, 1987 E.C.R. at ¶ 7.

41 *Id*.

42 The legality of EU acts—producing legal effects vis-à-vis third parties—can be challenged directly before the General Court pursuant to Article 263 TFEU (and on appeal to the ECJ). They can also be indirectly challenged via a reference from a national court for a preliminary ruling by the ECJ (Article 267 TFEU). The ECJ and the General Court have exclusive jurisdiction to determine acts of EU institutions invalid (see Case 314/85, Foto-Frost v. Hauptzollamt Lübeck-Ost, 1987 E.C.R. 4199).

43 Articles 101 and 102 TFEU are directed at the conduct of private undertakings and the duty of "sincere cooperation" in Article 4(3) of the TFEU (formerly Article 10 EC Treaty) is principally addressed to the Member States. Article 4(3) of the TFEU acts as the catalyst to challenge the legality of national measures on grounds that they undermine the effectiveness of EU law which can include the application of the EU competition rules. The duty of sincere cooperation does not appear as far-reaching in relation to actions of the EU Institutions, and it has been held that it does not apply to legislative measures adopted by the EU Council (see Joined Cases C-63/90 and C-67/90, Portugal and Spain v. Council, 1992 E.C.R. I-5073, ¶ 53). Article 7 TFEU, which provides that "[t]he Union shall ensure consistency between its policies and activities, taking all of its objectives into account . . . " does not seem a sufficiently precise alternative catalyst to challenge EU legislation on ground of lack of conformity with the EU competition rules.

44 There is extensive consultation, including inter-service consultation within the Commission, whenever the European Union proposes to introduce new laws and regulations. EU legislative and policy proposals are subject to an impact assessment, which includes an assessment of the possible competition impacts. See DG Competition, Better Regulation: A Guide To Competition Screening (2005), *available at* http://ec.europa.eu/competition/publications/legis_test.pdf.

45 The reference to "ensuring the competition is not distorted" is now included in a Protocol to the TFEU (No. 27) rather than in the Preamble to the Treaty. Nonetheless, this change in position is not expected to fundamentally alter the importance of achieving free competition in the European Union, since a protocol has equal force as the rest of the Treaty.

46 Joined Cases C-359/95 P and C-379/95 P, Commission and France v. Ladbroke Racing, 1997 E.C.R. I-6265, ¶ 33.

47 *Id*. at ¶ 34.

48 Case C-198/01, Consorzio Industrie Fiammiferi (CIF) v. Autorita Garante della Concorrenza e del Mercato, 2003 E.C.R. I-8055.

49 *See* Joined Cases C-297/88 and C-197/89, Dzodzi v. Belgium, 1990 E.C.R. I-3763.

50 For a broad and in-depth discussion of the concept, see Organisation for Economic Co-operation and Development [OECD], Regulated Conduct Defence, DAF/COMP(2011)3, Sept. 1, 2011.

51 Case T-271/03, Deutsche Telekom v. Commission, 2008 E.C.R. II-47; upheld on appeal in Case C-280/08 Deutsche Telekom AG v. Commission, 5 C.M.L.R. 27 (2010). These related to appeals from the Commission's decision to fine Deutsche Telekom EUR 12.6 million for abusing its dominant position with respect to its local loop access pricing (see COMP/C-1/37.451, 37.578, 37.579, Deutsche Telekom AG, 2003 O.J. L 263/9).

52 Case T-271/03, Deutsche Telekom v. Commission, at ¶ 107.

53 Case T-398/07, Spain v. Commission, 2008 O.J. (C 8) 17. This case relates to an appeal from the Commission's decision to fine Telefónica EUR 151 million for abusing its dominant position in the Spanish broadband market (see Case COMP/38.784, Wanadoo España v. Telefónica, 2008 O.J. (C 83)).

54 Joaquín Almunia, *supra* note ix.

55 The Complaint alleged that KeySpan, an electricity generator, manipulated New York City electricity prices using a swap agreement (the "Swap") in violation of § 1 of the Sherman Act. Specifically, the Swap provided KeySpan with an indirect financial interest in the sale of electricity generating capacity by its largest competitor, Astoria Generating Company ("Astoria"). That financial interest obviated KeySpan's need to bid competitively during the sale of its own electricity generating capacity at auction. According to the Complaint, KeySpan's anticompetitive bidding drove up capacity prices as a whole and, in turn, increased the cost of electricity to consumers in New York City. *See United States v. KeySpan Corp*., 763 F. Supp. 2d 633 (S.D.N.Y. 2011).

56 U.S. Dep't of Justice, Comments in the Matter of Rin 3038-ADO1 (Dec. 28, 2010).

57 *Id*., at § 1.

58 *Id*., at § 3.

59 *Id*., at § 1.

60 *See, e.g., EU to Deal Severely with CDS; Leaders Seek Ban*, Reuters, May 17, 2010, *available at* http://uk.reuters.com/article/2010/05/17/uk-eu-regulations-barnier-idUKTRE64G3NW20100517.

61 Case COMP/39592, European Fund and Asset Management Association (EFAMA) and others vs. Standard and Poor's, *available at* http://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_39592.

62 For further background on the case, see Press Release, European Commission, Commission Market Tests Standard & Poor's Commitments on International Securities Identification Numbers (May 16, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/571&type=HTML; Press Release, European Commission, Commission Makes Standard & Poor's Commitments to Abolish Fees for Use of US International Securities Identification Numbers Binding (Nov. 15, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1354&type=HTML.

63 Case COMP/39654, Reuters Instrument Codes. For further background on the case, see Press Release, European Commission, Commission Opens Formal Proceedings Against Thomson Reuters Concerning Use of Reuters Instrument Codes (Nov. 10, 2009), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/1692&format=HTML&aged=0&language=EN&guiLanguage=en.

64 Case COMP/39745 CDS – Information market and Case COMP/39730 CDS – Clearing. For further background on the case, see Press Release, European Commission, Commission Probes Credit Default Swaps Market (Apr. 29, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/509&type=HTML.

65 Commission officials carried out unannounced inspections, starting October 18, 2011, at the premises of companies active in the sector of financial derivative products linked to the Euro Interbank Offered Rate ("Euribor") in certain Member States. See Press Release, European Commission, Commission Confirms Inspections in Suspected Cartel in the Sector of Euro Interest Rate Derivatives (Oct. 19, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/11/711&format=HTML&aged=0&language=EN&guiLanguage=en.

66 Arrangements restricting competition can be exempted, provided they meet the cumulative conditions in Article 101(3) TFEU. The conditions are arguably broad enough to extend beyond pure economic efficiencies.

67 *See, e.g.*, Regulation (EC) No 1060/2009 on credit rating agencies, as amended. On November 15, 2011, the Commission adopted proposals for a new Regulation and a new Directive on credit rating agencies, aimed at further strengthening the supervisory framework (see Press Release, European Commission, Commission Wants Better Quality Credit Ratings (Nov. 15, 2011), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1355).

68 The rules will include (i) a registration requirement for a subsidiary (or certification if the CRA has no physical presence in the EU); (ii) rules of conduct for registered CRAs (enhanced transparency and the prevention of conflicts of interest between CRAs and their most important clients), and' (iii) the supervision of registered CRAs. For a general overview of the supervision of CRAs by ESMA, see Steven Maijoor, Chair of ESMA, Keynote Address at the Bundestag Panel Discussion (May 16, 2011).

69 For a general discussion of the topic, see OECD, Competition and Credit Rating Agencies 2010.

70 See, e.g., the UK authorities' response to the Commission internal market and services consultation document on credit rating agencies ("CRAs"), *available at* https://circabc.europa.eu/d/d/workspace/SpacesStore/0d1ea101-b6d0-470b-b8b3-e28e4c1cba30/BoE-FSA-Treasury_EN.pdf (last visited Nov. 23, 2011) and, in particular, the comments in Section III - Enhancing Competition in the Credit Rating Industry (response submitted on Jan. 31, 2011). It should also be noted that the European Parliament proposed the creation of a new independent, preferably European, CRA, but this would in all likelihood distort competition by giving the EU public entity an advantage relative to its competitors.

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

ANTICOMPETITIVE REGULATION IN THE PAYMENT CARD INDUSTRY

Ronald J. Mann

Columbia Law School

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

ANTICOMPETITIVE REGULATION IN THE PAYMENT CARD INDUSTRY

Ronald Mann*

ABSTRACT

The payment card industry in the United States has come under increasing scrutiny in recent years. The Bankruptcy Abuse Prevention and Consumer Protection Act of 2005 reflects a high-water mark of congressional influence for the industry, altering bankruptcy procedures largely for the benefit of card issuers. Since that point, Congress has turned repeatedly to rein in perceived abuses in the industry. The most substantial and direct response to the perception of abuse is the Credit Card Accountability Responsibility and Disclosure Act of 2009. That statute was focused directly on the card industry and outlawed a wide variety of industry practices. More recently, in § 1075 (the "Durbin Amendment") of the Dodd-Frank Wall Street Reform and Consumer Protection Act, Congress cut permissible interchange fees for debit card transactions to amounts that approximate the costs of processing those transactions; the Federal Reserve's implementing regulation apparently will lead to a more than 50 percent decline in those fees.

So why is it at all noteworthy that Congress, in the course of reining in an industry targeted for excessive behavior, should require substantial changes in the industry's operations? My hypothesis is a simple one. Both provisions make it more challenging to operate profitably in the payment card market. Because both provisions will pose greater challenges for smaller firms than they do for larger firms, both statutes will make it harder for smaller banks to compete in the payment card market. It may not be easy to evaluate the consequences of greater concentration in the industry. But it is clear that industry concentration is not what drove Congress to action: whatever else Congress was trying to do, it certainly was not trying to drive small banks from the payment card market

* Albert E. Cinelli Enterprise Professor of Law at Columbia Law School. Co-Director of the Charles E. Gerber Program in Transactional Studies.

I. THE CCA AND THE CREDIT CARD INDUSTRY

A) INFORMATION TECHNOLOGY AND CREDIT CARD LENDING

To understand the competitive structure of the credit card industry, *it is crucial to understand the shift in industry emphasis over the last few decades from finance to information technology.*

Specifically, I argue that the profitability of firms in the industry—the growth and decline of their market shares, the success of their new products, and their vulnerability to competitors—depends much less on the cost of funds or any measure of care or "prudence" in underwriting than it does on the technological sophistication with which the firms design and manage their interactions with their customers. To explain this point, I start with a brief summary of the business of credit card issuers and how it has developed over time.

1. The Proliferation and Specialization of Credit Card Products

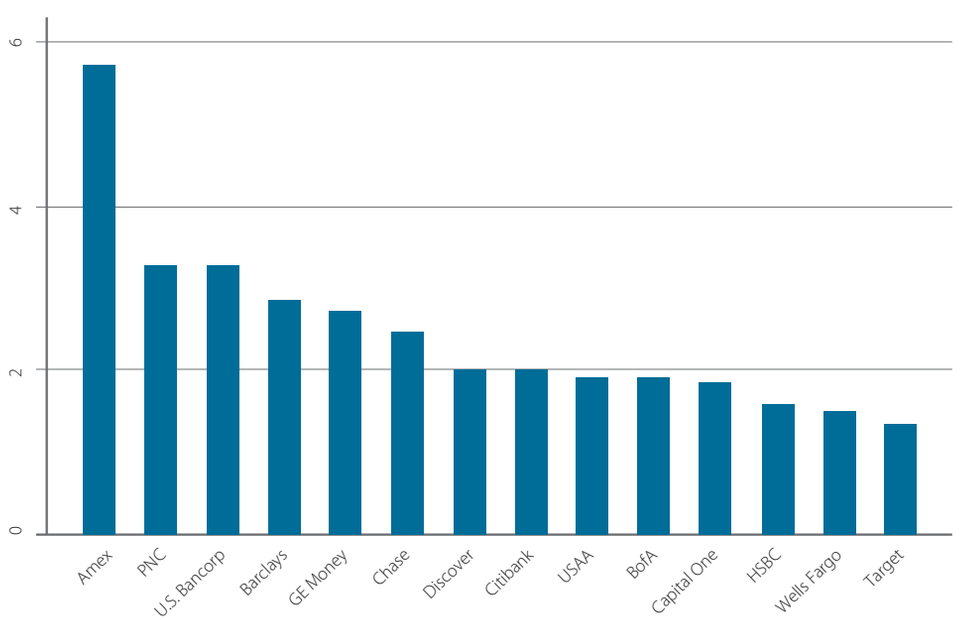
As recently as twenty years ago, the business of credit card issuing was a relatively simple one, with a more or less standard interest rate of 18 percent and three broad classes of potential customers: transactors, revolvers, and those too risky to receive cards.

Essentially, the goal of credit card issuers in those early days was to maximize the share of revolvers and minimize the share of transactors, while keeping chargeoffs at a relatively low level by excluding the risky. Even at that level of simplicity, the product was a risky one, as many issuers lost money, largely because of rampant fraud on the part of cardholders.[5] But during the intervening years, the market has changed in several important ways, primarily because

advances in information technology have improved the ability of credit card issuers to distinguish among their customers and thus segment their product offerings.

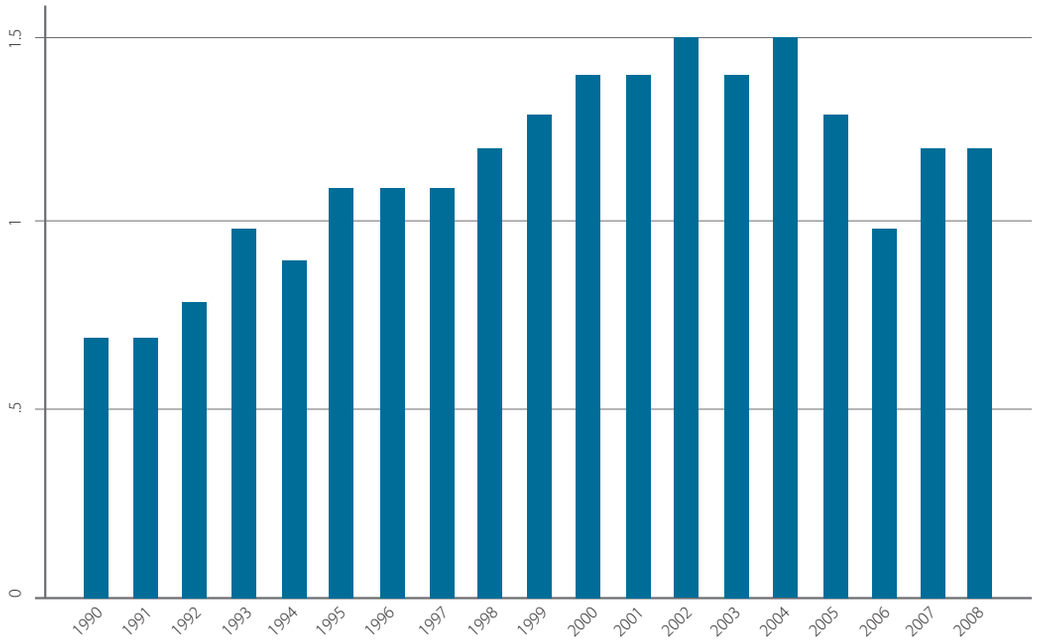
Most importantly, issuers now offer a wide variety of products, which can be placed along a spectrum from transactor-based to revolver-based. As Figure 1 shows, the ratio of purchase volume to outstanding receivables differs remarkably even among the largest issuers. Some issuers, most notably American Express, focus primarily on transactors; with a transaction volume almost six times the size of its receivables, it stands apart from all of the other substantial issuers. Issuers of that product try to earn interchange fees that exceed the cost of funds, transaction costs, and any chargeoffs. Thus, those issuers attempt to maximize the number of cardholders that use their cards frequently for high-value purchases. The products directed to those customers are likely to include substantial affinity rewards and low interest rates.[6]

Figure 1: Turnover Rates of Major Credit Card Issuers



Source: Author's calculations from Nilson Report.
Figure shows ratio of expenditures on cards during 2010 to receivables at end of 2010.

Figure 2: Late and Overlimit Fees (1990-2008)



Source: Cards Profitability Survey.
Figure shows ratio of late and overlimit fees to annual expenditures on cards.

In contrast, a revolver-based product focuses primarily on revenues from interest and late or overlimit fees. Putting American Express to the side, most of the other large issuers emphasize a revolver-based model; as Figure One shows, Chase, Bank of America, Citibank, Capitol One, and Discover all have purchase volumes about twice their outstanding receivables. That product is less likely to have an annual fee and much more likely to have a high interest rate. The central task for the issuer of that product is to identify and attract customers who will carry substantial balances, without repaying them in full each month, and without defaulting (at least before they have paid on the balances for a period long enough to amortize the card issuer's investment in the customer). That model also depends, at least in part, on late and overlimit fees. [7]

Figure 2 traces the development of that model. Several points are salutary. First, it documents the doubling of the rate of those fees during the 1990s, as issuers swarmed to the model. After a peak lasting until about 2004, however, the level of those fees began to decline. A number of possible explanations are apparent. The first is simply that consumers began to avoid these fees by altering their conduct to avoid late and overlimit transactions; as the fees became more common, consumers learned of their costs and used greater efforts to ensure that they did not accidentally charge beyond their credit limit or pay their bills late. To the extent late and overlimit fees resulted from accidental errors, rather than liquidity constraints, this would make sense. A broader, and not entirely unrelated, explanation is that more sophisticated contracting structures developed to increase interest revenues unrelated to the interest rate – double-cycle billing, minimum finance charges, and the like. As sophisticated issuers introduced those product attributes, the pressure to generate revenues from late and overlimit fees diminished, and their share of industry revenues similarly declined.[8]

One additional trend of importance, along a spectrum distinct from the transactor/revolver distinction, is the rise of affinity and rewards products. Because there is a cognizable cost to acquire and maintain each credit card account, all issuers focus on ensuring that those to whom they issue their cards use them as frequently as possible. Industry executives designing products frequently emphasize their desire that their cards will be "top of the wallet." The more the cards are used, the more profitable the issuer's operations. Because issuers cannot compete on acceptance (in the United States there is, with the exception of Discover, little variation in acceptance among the major brands), affinity and rewards cards play a particular role in the competition for "top of the wallet status." Traditionally, specialized monoline issuers like MBNA dominated that business,

but through acquisitions, that business has for the most part fallen into the hands of Bank of America and JPMorgan Chase.

The combination of those trends produces a mind-boggling potential for variation in product design. Driven both by consolidation (as the larger issuers acquire the portfolios of smaller issuers) and by market pressures, most of the large issuers now have large suites of products, including dozens of distinct credit card products, all targeting particular niches along the spectrum from transactors to revolvers, and particular pockets of affinity (specified sports teams, universities, social causes, and the like). For example, as of the fall of 2011, Bank of America displayed 72 distinct credit card products at its Web site. Although other issuers can't compete with that diversity, the number of distinct products at other major issuers is still impressive: U.S. Bank's web site advertises 29 different cards, CitiBank's 27, Chase's 20, and Capitol One's 14. Even once-stodgy American Express advertises 22 different products (15 of which are credit cards and 7 of which are charge cards). time.

2. Proprietary Predictive Models

The complexity and heterogeneity of modern credit card products presents numerous challenges to businesses that attempt to issue them profitably. For one thing, the issuance of cards involves a substantial expenditure—the process of sending solicitations, responding to applications, and issuing cards—that will produce no revenues at all unless the cardholder in fact begins to use the card for purchases. And if the cardholder maintains an unpaid balance, the consequent lending is rife with risks that are unusual for the typical bank lender. Unlike the lender on a home or a car, the lender has no collateral. The lender has no control over the uses to which the money is put. The debt is to be repaid over an extended period of time, on a payment schedule for the most part selected by the customer.

Thus, a successful credit card lender must have expertise at surveying the potential customers available to it;

predicting which ones are likely to use the cards, whether they are likely to default, and how long it is likely to be before they default; and managing accounts capably in light of those projections. There is a great deal of room for increased (or decreased) profitability based on the level of sophistication applied to those activities.

The difficulties issuers faced in the early days of the credit card industry arose directly from the primitive information technology then available to the issuers and networks. Thus, it was a bold development in the early years of the 1970s when Visa for the first time could introduce electronic processing to clear transactions among the various credit card companies—something that now is a simple and routine matter.

For the most part today, what distinguishes those who are successful and profitable from their competitors is skill at collecting, manipulating, and analyzing information. The historical example of Providian is conspicuous. At one time, it was a major player in the subprime market, but its inability to understand the risks inherent in its portfolio led to unsustainably high levels of chargeoffs, which eventually forced it to withdraw from that sector.

Issuers do not simply guess what customers will do with the cards that are offered or issued to them. Nor, like mortgage lenders, do they rely on third-party scoring systems (like Fair Isaacs) that are readily available to all in the industry. Rather, at all points along the lifecycle of each account (from the universe of potential customers through the group of existing customers at any given time), issuers access and gather immense databases, which they analyze in an effort to understand the likely patterns of use and risk associated with each customer profile. The more information issuers can use in their models, the better those models can predict future card use and the risks associated with each individual.

The better models predict future use and chargeoff risks, the better the issuer's ability to acquire (and retain) profitable customers and to avoid (or shed) unprofitable customers. The benefits drop straight to the issuer's bottom line. Models that more precisely and accurately predict the likelihood and timing of chargeoffs will allow the issuer to design a more profitable mix of product solicitations and to manage existing accounts in ways that attract or repel customers that are less (or more) likely to generate chargeoffs. Together, those will allow the issuer to keep lower reserves against future chargeoffs. Lower chargeoffs and lower reserve requirements lead directly to increased profitability.

Improving predictive models benefits issuers at several stages of the life cycle of a particular customer. First, the issuer with a better model of consumer behavior will be able to do a better job of targeting solicitations to the customers. The process of sending solicitations is extremely expensive, largely because the response rate has fallen significantly even as the number of solicitations has increased: CitiBank alone sent more than 350 million solicitations in the third quarter of 2011, expending about a quarter of a billion dollars.[9] The goal of each solicitation is to get as high a response rate as possible from the most desirable group of customers.

Thus, a solicitation can fail either because too few people respond, or because the group that responds is a surprisingly unprofitable group of customers. Given the amount of money at stake, it should be no surprise that the issuers sending such a blizzard of solicitations invest heavily in technology to predict and improve the responses they receive.

Improved predictive models also benefit issuers when they set the terms of the cards that are issued when cardholders respond to the solicitation. As individual cardholders respond to a single solicitation, issuers allocate different terms (interest rates, grace periods, credit limits) based on the issuers' assessment of the likely future behavior of the responding customers. Again, issuers can err by issuing too few cards (and thus losing desirable customers to other issuers) or by issuing too many cards (and thus issuing cards that are under-used or lead to chargeoffs).

Perhaps the most important use of these kinds of predictive models involves the ongoing management of existing cardholder accounts. Relying on those models, issuers use predictions about future cardholder behavior to make instantaneous and precisely targeted decisions about such things as increases or decreases in credit limits, alterations in interest rates, and responses to overlimit transactions or late payments. For example, sophisticated issuers customarily use predictive tools widely for such purposes as updating credit limits, issuing balance transfer offers, setting prices, and identifying likely future chargeoffs.

In sum, although it is an exaggeration to say that lending expertise is no longer important in the credit card industry, it is just as true that lending expertise and caution standing alone are not enough to compete successfully.

B) THE CCA AND CREDIT CARD LENDING

Against that backdrop, it is useful to consider the CCA. For present purposes, the principal substantive provisions of the CCA fall into two categories. The first category includes prohibitions on conduct reasonably characterized as sharp dealing, by which I mean contractual attributes and business practices that are substantially more costly to the customer than any efficiency or cost saving they might produce for the issuer. In this category, for example, I would include the prohibition on double-cycle billing,[10] the requirement that cardholders opt in to over-the-limit fees,[11] the rules requiring prompt crediting of payments,[12] and the strict limits on "fee harvester" cards.[13]

None of those provisions should substantially affect competition among the major players in the credit card industry by which I mean, loosely speaking, the large issuers identified in Figure 1, who increasingly control the market for credit card lending. In some cases, including fee harvesting, the provisions outlaw activity in which none of those issuers ever engaged.[14] In others, they outlaw arguably fraudulent behavior that was already within the control of federal regulators, such as unreasonable limitations on crediting payments.[15] In still others, they outlaw contract terms that major issuers had already stopped using before Congress enacted the CCA, like the practice of double-cycle billing.[16]

Those provisions probably are beneficial, because they outlaw conduct that serves no useful purpose. But they will not individually or collectively have any important affect on the way in which issuers design products and compete against each other.

The limitations on interest-rate increases in § 101 of the CCA (adding § 171 to the Truth in Lending Act[17]) stand out as qualitatively different. Among other things, that statute generally prohibited "retroactive" interest rate increases: interest rate increases that apply to funds already borrowed from the lender.

The only exception is for a variable interest rate that changes because of an index, rather than the borrower's individual characteristics or because of a failure of the borrower to make required minimum payments on the card account. This requires a major shift in business practices, amounting to a fundamental recasting of the basics of credit card underwriting.

Even with the predictions of future behavior drawn from their sophisticated modeling, credit card issuers traditionally have relied on product attributes that let them respond in real time to shifts in the perceived riskiness of their customer base. This is at least in part because so many of the adverse events that increase the riskiness of a particular customer are random events that have so little to do with an individual's past history that even the best modeling can do little to predict them. Thus, credit card issuers traditionally have reserved in their contracts the ability to increase interest rates on individual customers at any time or from time to time, for almost any reason that motivates the issuer to think this prudent.[18]

It always was common, of course, to increase interest rates in response to a failure of the borrower to make the required payments on the credit card account. But many lenders used "universal default" provisions, under which they increased interest rates on a credit card whenever they learned (through credit bureaus and the like) of a default by their customer on any other account: so the credit card interest rate went up, even if the cardholder was keeping that account current, solely because of a default on an electric bill.

Even more aggressively, some lenders took the opportunity of repricing interest rates before the cardholder defaulted on any payment, solely because of a shift in attributes that, in the judgment of the lender, increased the borrower's risk profile.[19]

This is related to the practice, central to the revolving-credit business model, of issuing cards on the expectation that cardholders will build balances on them relatively quickly and then pay them off slowly, over a long number of years.[20] The balance-transfer method of acquiring customers epitomizes this: instead of waiting for your own customers to charge up their balances, you acquire customers that have already built up balances on the cards of other issuers, gambling that if you do your underwriting properly they will take so long to pay off the balances that you will profit even after paying whatever enticement you offered to acquire them.

The market-oriented premise of this regime is that if the issuer increases the rate excessively, the cardholder can avoid the excessive charges by repaying the credit card lender. By hypothesis, the cardholder would simply shift its outstanding balance to any other lender willing to lend at a lower rate; if the cardholder is borrowing at any given time from the lender offering the lowest rate, then the cardholder has little about which to complain.

That market-oriented perspective overlooks a great deal of the reality that confronts the borrower. Most obviously, the borrower's ability to repay the lender is likely to be most limited at the moment the lender raises interest rates: if interest rates are raised when the borrower experiences financial distress, the borrower likely will find it hard to repay its credit card lender out of ready cash or to find a new lender willing to lend at a modest rate.

At the same time, by increasing the interest rate on the outstanding credit card debt, the lender increases the borrower's financial distress by increasing the borrower's monthly obligations.

Thus, whatever its premise, in practice the real-time interest-rate adjustment is likely to complicate the borrower's efforts to respond to financial distress, if not tip the borrower over the edge entirely.

Seen against that business model, the prohibition on retroactive interest rate increases is a major challenge. If credit card lenders cannot shift interest rates in response to changes in the borrower's risk profile as they occur, the lender instead must set an interest rate before advancing funds to the borrower in the first instance—a rate which must be adequate to compensate for all anticipated shifts in riskiness that can be foreseen as likely to occur before the debt will be repaid.

This is particularly complicated for borrowers that are likely to build up a substantial balance early in their relationship and then carry that balance for many years into the future, slowly making payments on it but not completely paying off the balance.

For those customers, the interest rate established at the beginning of the relationship, when the lender has little or no experience of the borrower's repayment behavior, will be the interest rate that must stick with the account for what well might be an extended time period. It is easy to see that this increases by an order of magnitude the difficulty of the underwriting and risk-modeling task that faces the credit card lender. It is safe to say that only the most sophisticated of credit card lenders will have adjusted to that challenge without difficulty.

C) CONCENTRATION IN THE CREDIT CARD INDUSTRY

The natural question to ask is why anybody should be concerned that Congress has made the task of credit card underwriting harder. After all, the avowed purpose of the CCA was to rein in the industry, and making the task harder should lower the profits of those lenders and thus lower the absolute or relative amount of that lending in the economy. If credit card lending imposes a negative external cost on society, then anything that lessens credit card lending is beneficial.[21]

The truth, I believe, is considerably more complex. The central problem is the industrial organization of the credit card industry, which is fissured not only along the lines of differing strategies of credit card lending as discussed in Part I, supra, but also into lending and non-lending sectors. Thus, although there are several thousand general-purpose credit card issuers, the number of significant debt issuers is much smaller.

As of 2010, the share of receivables held by the top ten issuers (those that appear in Figure 1) was about 87 percent; the top four issuers alone (JPMorgan Chase, Bank of America, CitiBank, and American Express) held 60 percent.[22]

The heavy concentration of credit card lending in such a small group of issuers is not a coincidence. The profits from "ordinary" credit card issuance, reliant on interchange fees, involve many of the typical attributes of expertise broadly distributed throughout the banking industry: building customer loyalty, attraction to the bank's brand, and the like.

Thus, it is much easier for "ordinary" banks to compete in the business of having their customers use their credit cards for ordinary day-to-day transactions. This is especially true when the credit cards are issued as one part of a broader relationship, and thus need not generate significant profit on a standalone basis. It is much harder, though, for the reasons discussed, supra Part I, and infra Part II, to issue credit card debt profitably.

Thus, the rapidly increasing sophistication of that business brought on by the use of information technology in the last two decades has made it increasingly hard for smaller issuers to compete, steadily driving them from that business and driving lending portfolios ineluctably into the hands of the largest and most technologically sophisticated issuers.

Seen through that lens, *the dead weight dropped on the industry by § 171 has a different meaning: it is yet another challenge to the "ordinary" banks trying to compete* against the few largest technologically sophisticated credit card lenders. For the largest banks, § 171 should pose a challenge, but not an insuperable one, as they presumably have been able to modify their products to predict future repayment behaviors relatively well. For smaller banks, however, with less cutting-edge modeling expertise, this should accelerate their move out of the lending market.

To be sure, those banks could invest in the modeling technology necessary to price their products as effectively as the large banks, but for several reasons this is likely to be quite difficult. The most obvious is that because their portfolios are smaller, they will have a smaller asset base over which to amortize the costs of their investment.

This problem is exacerbated by the rapid segmentation of products, infra Section I.A.1.

Where the portfolios of the larger issuers can be split into dozens of separate pieces, each with separate underwriting and pricing criteria, yet still large enough for effective modeling, the much smaller portfolios of the smaller issuers make it quite difficult for them to compete in specific segments.

A small issuer with a portfolio a fraction the size of the ones held by JPMorgan Chase and Bank of America will find it much more difficult to obtain enough customers in any particular segment to compete effectively against the precisely targeted products of those issuers. They will have many fewer customers in any particular segment, and thus much less information on which to form predictions about the likely use and repayment behavior of those customers if they do issue a competitive card. If they respond to the uncertainty by higher pricing, then almost by definition their products will not be competitive.

It is, then, difficult to see how the smaller issuers can hope, in the longer run, to compete on card product definition and management standing alone. They must, if they are to remain in the market, compete on a "whole relationship" basis.

II. THE DURBIN AMENDMENT AND THE DEBIT CARD INDUSTRY

A) THE ROLE OF DEBIT CARDS IN BANK ACCOUNT RELATIONSHIPS

Although their use at the point of sale is functionally similar to that of credit cards, the role of debit cards in financial services is completely different. Credit card markets are dominated by large national (and multinational) banks that hold gigantic portfolios unrelated to their deposit structures. Thus, the largest portfolios are constructed, for the most part, of customers that have no depositary relationship with the issuer, and often no relationship beyond the card at all.[23]

The rise of securitized financing played a major role in weakening the link to deposits, because it provided what seemed to be a low-cost and reliable source of funding that allowed banks like MBNA, Providian, and Capitol One to issue credit card loans in sizes that far exceeded the deposit base that was the traditional source of funds for credit card lending. Even now, with Capitol One the only remaining major credit card lender without a nationally significant deposit base, large-scale funding of credit card loans through securitized financing leaves the tie between deposits and credit card lending thin at best.[24]

The situation with debit cards is quite different. Debit cards are almost universally issued by a bank with which the cardholder has a deposit-account relationship.[25] Thus, debit cards and their pricing are an integral part of a larger relationship. This has several ramifications for the industry's structure. For one thing, because debit card issuance loosely parallels deposit collection, the debit card industry is much less concentrated than the credit card industry. For example, the top four debit card issuers (by purchase volume) control only 39 percent of the market; the top ten, less than half the market.[26]

Second, revenues from debit cards interact much more broadly with the account relationship; their "subsidy" is not internal to the product as it is for credit cards. Thus, debit card interchange fees essentially have funded free or low-cost checking accounts. Generally speaking, revenues from debit card interchange fees, in the range of fifty cents per transaction since settlement of the Visa and MasterCard antitrust litigation[27] in the early years of the century, have provided revenues that offset the costs of checking account services, especially for customers with relatively low average balances. Among other things, this has facilitated broader penetration of mainstream checking account services to low- and middle-income populations.[28]

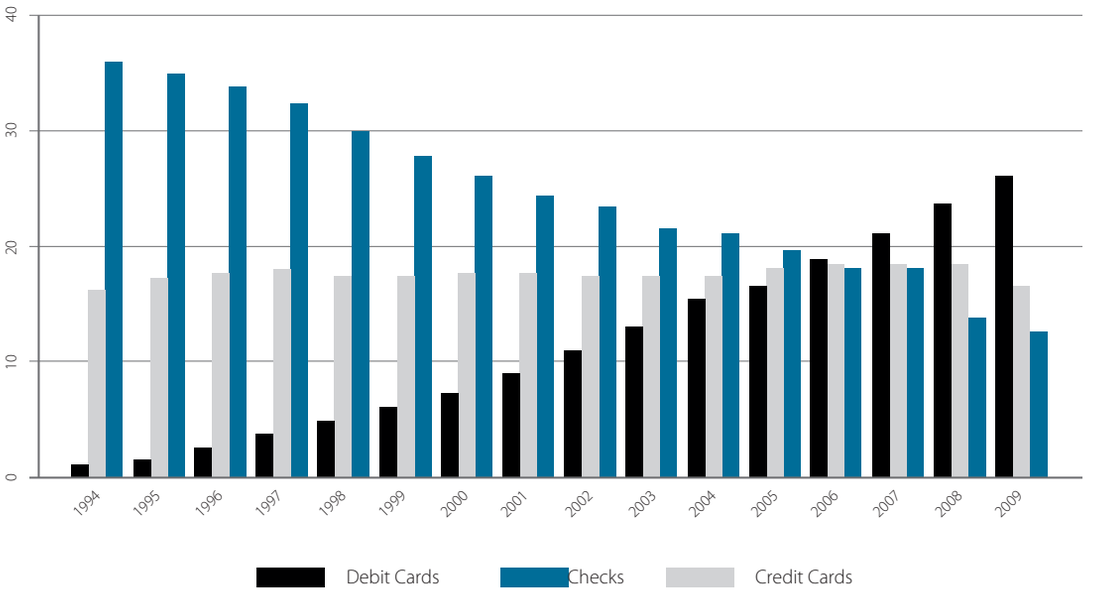
The price structure reflects the reality that debit cards, like credit cards, are a two-sided product. Thus, the manager of the relevant network must coordinate prices and terms for two distinct groups of users, managing those terms and prices to maximize the value to be derived from use of the product.

For credit cards, relatively high interchange revenues (charges imposed by the networks on the merchants) traditionally have facilitated quite generous terms for cardholders, including extensive rewards programs, which have fostered rapidly growing card usage. The parallel for debit cards has been interchange charges on merchants at levels sufficiently high to provide resources that facilitate broad provision of relatively low-cost bank accounts as a benefit to the cardholder side.[29]

Although merchants understandably have opposed cost structures in which they bear the brunt of expenses, the product borne from those cost structures has been undeniably successful.

As Figure 3 illustrates, debit card usage (as a share of consumer payment transactions) has increased from about 1 percent in 1994 to more than 25 percent in 2009, surpassing both credit cards and checks. Much of that growth has come at the expense of declining check use. Because checks are a paper-based payment system with high transaction costs, the shift to an electronically cleared system like debit cards presents a clear social benefit.[30] To the extent debit card growth also comes at the expense of credit card use, as seems to be the case since 2007, there is a parallel social gain to the extent that routine debit card use limits the unreflective borrowing associated with routine credit card use.[31]

Figure 3: Check, Credit Card, and Debit Card Transaction Shares (1994-2009)



Source: Author's calculations from Nilson Report.
Figure shows ratio of expenditures on cards during 2010 to receivables at end of 2010.

B) THE DURBIN AMENDMENT AND DEBIT CARD ISSUANCE

The Durbin Amendment to Dodd-Frank strikes at the heart of that system, requiring a drop in debit card interchange fees to a cost-justified level. Specifically (as codified in the Truth in Lending Act ("TILA")),[32] the Durbin Amendment requires the Federal Reserve to define a cap on interchange fees that is "reasonable and proportional to" the issuer's costs.[33] Also, in an effort to bolster downward pressure on network-level interchange pricing, the statute requires that each card have "bugs" from at least two non-affiliated networks, so that a merchant has two different ways to process each transaction.[34]

To implement the statute, the Board of Governors of the Federal Reserve, exercising authority delegated to it by the Durbin Amendment, originally proposed a fee cap of 12 cents per transaction, a stark drop from preexisting market levels averaging about 50 cents.

In response to voluminous comments on its proposed rule, the Federal Reserve ultimately adopted a final rule[35] that sets a cap of 21 cents plus .05 percent of the transaction amount, amounting to approximately 24 cents per transaction.[36] As compared to preexisting market levels, this amounts to a revenue drop of about 50 percent.[37]

C) CONCENTRATION IN DEBIT CARD ISSUANCE

As a matter of economic theory, the Durbin Amendment is profoundly wrong-headed. It is premised on the notion that lack of competition in the debit card industry has led to an unnaturally elevated price that banks collusively charge to merchants for the debit card product.

But there is no reason to expect a priori that a network in a fully competitive environment would set a price on either side that bears any predictable relation to the costs of providing services to that side.

Thus, the acknowledged fact that existing interchange fees are, for many banks, higher than the costs of processing debit card transactions proves nothing at all about the efficiency of the market or the "correct" debit interchange price. The relevant question is whether the networks are setting prices that maximize growth of their network. The rapid uptake in debit cards in recent years (summarized, supra Figure 3) suggests that they are.

To put it bluntly, economic theory suggests no reason to think that shifting to a cost-justified level of fees for debit card interchange will improve the efficiency of the affected markets.

To be sure, the statute rejects that understanding of the debit card market and proceeds on the supposition that existing prices reflect improper price-fixing by the major networks. Even on that basis, however, there is great reason to expect that the statute will have a substantial adverse effect on market structure.

On its face, the statute bears evidence of Congress's intention to protect small issuers. Specifically, cards issued by small issuers (those with less than $10 billion of assets) are exempted from the price-level restrictions imposed by TILA § 920(a).[38] For several reasons, however, it is likely that the statute will disadvantage the smaller institutions singled out for protection by the small-issuer exemption.

The first and practical reason is that networks have little or no incentive to establish separate, higher price levels for their smallest and least influential issuers. As discussed above, networks that control two-sided products like debit cards thrive by coordinating the prices and terms on the two sides of the network so as to maximize the growth of the network.

Among other things, they must attract transactions to their network and they can do that only by attracting issuers that issue large volumes of cards. The basic problem this creates is that networks that establish special elevated pricing for small issuers will offend their most important issuers, the large issuers on whose cards the overwhelming majority of debit card transactions occur.

Thus, the most likely response of large networks is to adopt fee structures that minimize the revenue advantages of small issuers over large issuers.[39]

The second is the ability of the merchant to steer customers away from high-cost cards. For one thing, the Durbin Amendment allows merchants to control routing, to select whatever network they prefer from the networks on a card, and small issuers are not exempt from that provision.[40]

Furthermore, although the Durbin Amendment does prohibit merchants from discriminating on the basis of an issuer's identity,[41] it does not prohibit them from discriminating on the basis of price.

Accordingly, it appears that merchants could lawfully refuse to accept small-issuer cards on any network that allows small issuers to collect substantially greater interchange fees than the Durbin Amendment caps for large issuers.

The third reason that the Durbin Amendment is likely to affect small issuers particularly harshly relates to the cost structure of the industry. Before promulgating Regulation II, the Federal Reserve collected data from issuers on their cost structures.

The data show substantial economies of scale in the costs of debit card processing. For the largest issuers, average variable costs appear to be less than four cents per debit card transaction; for those issuers, Regulation II leaves approximately twenty cents per transaction to subsidize other account services.

This is, to be sure, much less than what they had before the Durbin Amendment, but it is a substantial continuing revenue stream. For most small issuers, by contrast, average variable costs appear to exceed twenty-five cents per transaction.[42]

Thus, for those issuers, transactions processed at the cap would be processed at a loss; not only would this eliminate the subsidy of other services entirely, it would require a flow of funds from other sources to debit-card processing. For those institutions, then, maintaining revenues substantially above the Regulation II cap is not only attractive, it is crucial to the existing business model. Because continuation of those revenue levels is unlikely, small issuers face daunting challenges in the years to come.

III. ROOTING OUT COMPETITION

So what does this mean? Let us suppose I am correct in my conjecture that the CCA and the Durbin Amendment will exacerbate the market push driving relatively small banks from the payment card industry. What adverse effects can we attribute to this? The first and obvious one is lessening competition.

Although it is easy to suggest that competition between Visa and MasterCard has rarely been aggressive, competition at the bank level traditionally has been vigorous.

For credit card issuance, thousands of issuers produce a blizzard of television advertisements and billions of annual solicitations sent by mail. For the basic business of consumer banking, the medium is different—the local billboard supplementing nationwide television advertising campaigns—but the market for consumer banking accounts traditionally has been relatively robust.[43]

Yet as the number of effective players falls ever lower,

the point is coming (if it is not already here) when there are few issuers competing for the business of any particular consumer.

This is particularly salient in the credit card sector given the trend toward segmentation, which allows larger issuers to provide products that can compete nationally for smaller and smaller groups of precisely defined customers.

The consequences of such a decline of competition, at least under classic economic theory, are simple: an increase in price and a fall in supply. It is safe to assume that neither Congress nor the federal competition regulators (the Department of Justice and the Federal Trade Commission) would applaud a conspicuous decline in competition in such an important industry. Indeed, the Durbin Amendment was enacted on the stated premise that small issuers would be protected.

For several reasons, however, I doubt this simplistic take on the competitive impact of these statutes is adequate. On the one hand, it is easy to argue after the recent economic meltdown that unbridled competition by financial institutions is itself socially harmful. What we have seen in the last decade is the specter of financial institutions substantially unconstrained by regulators, driven by cutthroat competition into lending that was so plainly imprudent as to drive the world financial system to the brink of collapse.

Commentators can speculate and argue about what particular piece of the system led to such an exuberance of irrationally imprudent lending. Was it the existence of deposit insurance and related regulatory institutions that left banks too little concerned about the effects of imprudent lending?[44]

Was it the markets that allowed (or even encouraged) banks to make loans based on insupportable valuations by making it so easy for them to transfer the risks of nonpayment to others?[45]

Was it the relative asymmetry of institutions that made it easy to withdraw home equity during times of rising prices but left no similar exit strategy for times of falling prices?[46] Or was it the failure of financial analysts to understand the nature of the risks created by the financial instruments into which these loans were packaged?[47]

Whatever the reason for the problem was, it is clear that unbridled competition exacerbated the problems into which they have driven our economy. Accordingly, we should acknowledge at least one beneficial side effect to reforms that undermine vigorous competition in financial markets: they lessen the risks we all face from destructive competition in those markets.

On the other hand, a look at the particular actors affected here tells a less sanguine story of the aggregate effects of these statutes. In both cases, they accelerate shifts away from an older, more relational style of financial services toward a more information- and product-centered model based in technocratic norms.

Thus, if I am right in thinking that the CCA is effectively driving the smaller, more relational issuers from the lending sector of the credit-card industry, the market response will not be limited to a decline in competition about price. It also includes a broader eradication of a model of banking in which the bank sees a credit card as one of a suite of products issued to a particular customer, out of which the bank needs to profit in aggregate.

Because this model involves less of the highly aggressive lending characteristic of the largest, most information-intensive lending experts, it probably has less of the adverse social costs that go with that lending. If the only issuers with competitive significance are the very largest and most technologically focused lenders, we should be concerned about the potential for a shift to

a market in which all credit card lending is conducted at the harsh edge of riskiness that maximizes the adverse social cost of the product.

The Durbin Amendment's effects are likely parallel. By putting inordinate pressure on the cost structures of community banks and credit unions,

the statute is likely to accelerate the shift toward the large money-center institutions

and away from the smaller, more fragmented localized financial institutions. This seems particularly perverse, given the role money-center institutions played in the recent crisis and given the unique role the smaller institutions play in funneling capital to small businesses and employers remote from national financial centers.

It would be easy to view these statutes through a simple public-choice model as yet two more examples of the continuing political power of the largest financial institutions.[48] To me, however, it makes more sense to emphasize the particular perversity that the CCA and the Durbin Amendment share: a failure to recognize the links between product design and market structure. The central flaw in the CCA is its failure to recognize the relation between interest-rate flexibility and the ability of smaller banks to manage credit card lending effectively.

The central weakness of the Durbin Amendment is its misunderstanding of the relation between interchange fee levels and the cost structure of small institutions.

Given Congress's stated intention to protect small institutions in Durbin, I find it more accurate to view the statutes as example of ineptitude – poor craft in policymaking – than venality in intentionally favoring the interests of the largest institutions. I leave it to the reader to judge which perspective bodes better for the future of financial regulation.

1 Bankruptcy Abuse Prevention and Consumer Protection Act of 2005, Pub. L. 109-8, 119 Stat. 23 (2005).

2 Ronald Mann, *Bankruptcy Reform and the Sweat Box of Credit Card Debt*, 2007 U. Ill. L. Rev. 345.

3 The Credit Card Accountability Responsibility and Disclosure Act of 2009, Pub. L. 111-24. 123 Stat. 1734-1766 (2010).

4 Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, 124 Stat. 1376 (2010).

5 Lewis Mandell, The Credit Card Industry: A History (MacMillan Publishing Company 1990); Timothy Wolters, *"Carry your credit in your pocket": the early history of the credit card at Bank of America and Chase Manhattan*, 1 Enterprise & Soc'y 315 (2000).

6 Ronald Mann, *Patterns of Credit Card Use Among Low- and Moderate-Income Households*, *in* Insufficient Funds: Savings, Assets, Credit, And Banking Among Low-Income Households (Rebecca M. Blank & Michael S. Barr eds., 2009).

7 Mann, *supra* note 2; Mann, *supra* note 6.

8 Mann, *supra* note 6.

9 Suzanne Kapner, *Debit or Credit? Citi Places Its Bet*, Wall St. J., Sept. 20, 2001, at C1.

10 CCA § 102(a), codified at TILA § 127(j).

11 CCA § 102(a), codified at TILA § 127(k).

12 CCA § 104, codified at TILA § 164.

13 CCA § 105, codified at TILA § 127(n).

14 As defined in TILA § 127(n), "fee harvesting cards" assess fees at issuance that exceed 25 percent of the card's credit limit. Although it is not easy to obtain data about all products issued by large issuers, I have never observed such a product from a large issuer. The majority of those products were issued by relatively small issuers such as CompuCredit (the object of a major FDIC enforcement action that terminated its role in the industry), First Premier Bank, First National Bank of Pierre, First Bank of Delaware, and Applied Bank. *See* National Consumer Law Center, Fee-Harvesters: Low-Credit, High-Cost Cards Bleed Consumers (2007).

15 The most well-known incident involved Providian Bank, which was the subject of a 2000 consent decree under the Federal Trade Commission Act requiring it to pay approximately $300 million to the OCC. *See* David Leonhardt, *Credit Card Issuer Will Repay Millions to Some Customers*, N.Y. Times, June 29, 2000, at A1.

16 At least in part, issuers stopped those practices in anticipation of their likely prohibition by Congress. But a large part of the cessation also involved the phenomenon of "term cycling," in which issuers routinely change the terms of their card agreements as consumers learn how to avoid the costs of more familiar adverse terms. *See* Mann, *supra* note 2.

17 Truth in Lending Act, 15 U.S.C. §§ 1601-1665 (1968).

18 Ronald Mann, *Contracting for Credit*, 104 Mich. L. Rev. 899 (2005).

19 Although I am unaware of data about the prevalence of this phenomenon, conversations with industry professionals convince me that it was quite common in the portfolios of large issuers, especially those with exposure in the "prime" market to five- and six-figure credit lines likely to be repaid over long periods of time (during which the cardholder's riskiness and the issuer's tolerance for risk both might change substantially).

20 Mann, *supra* note 2.

21 See Robert D. Manning, Credit Card Nation: The Consequences of America's Addiction to Credit (2001); Ronald Mann, Charging Ahead: The Growth and Regulation of Payment Card Markets Around the World (2007).

22 The Nilson Report 966 (Feb. 2011).

23 David S. Evans & Richard Schmalensee, Paying with Plastic: The Digital Revolution in Buying and Borrowing (2nd ed. 2004).

24 Although that market obviously slowed with the economic downturn, most of the largest issuers have successfully securitized credit card receivables in 2011.

25 The only exception is a trivial number of "decoupled" debit cards, generally issued by nondepositary banks. Capitol One pioneered the only general-purpose decoupled debit card in 2007, but stopped issuing it after a one-year trial, so that at the present time there is no substantial portfolio of decoupled debit cards.

26 The Nilson Report 971 (May 2011).

27 *In re Visa Check/MasterMoney Antitrust Litigation*, No. CV-96-5238 (E.D.N.Y. Apr. 1, 2003).

28 David S. Evans, Robert E. Litan, & Richard Schmalensee, *Economic Analysis of the Effects of the Federal Reserve Board's Proposed Debit Card Interchange Fee Regulations on Consumers and Small Businesses* (Feb. 22, 2011) (unpublished, *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1769887).

29 Evans & Schmalensee, *supra* note 23; David S. Evans & Richard Schmalensee, Catalyst Code: The Strategies Behind The World's Most Dynamic Companies (2007); Evans, Litan & Schmalensee, *supra* note 28.

30 Mann, *supra* note 21.

31 Ronald Mann, *Adopting, Using, and Discarding Paper and Electronic Payment Instruments: Variation by Age and Race* (Federal Reserve Bank of Boston, Public Policy Discussion Paper No. 11-2, 2011), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1862169; Mann, *supra* note 21.

32 TILA, *supra* note 17.

33 TILA § 920(a)(2).

34 TILA § 920(b).

35 Regulation II, 12 C.F.R. pt. 235 (2011).

36 Debit Card Interchange Fees and Routing, 76 Fed. Reg. 43394 (July 20, 2011) (describing interim rule and justifying shift to final rule).

37 Evans, Litan & Schmalensee, *supra* note 28.

38 TILA § 920(a)(6).

39 Although it is far to soon to see how this will play out in practice, as I write it appears that none of the major networks have yet implemented fee schedules that advantage small issuers over large issuers. Indeed, perversely enough, it appears that Visa and MasterCard's new schedules will increase fees on small-ticket transactions for large issuers (up to the regulatory maximum, which is higher than the previous market level fees for those transactions), leaving small issuers with **lower** revenues for those transactions. *See* Digital Transactions News, *Applying the Durbin Maximum, Visa and MasterCard Could Squash Small Tickets*, Sept. 27, 2011, http://digitaltransactions.net/news/story/3217. If it seems strange that the Durbin fee cap actually raised rates for small-ticket transactions, the key is in the way the rate is defined. Market-set rates traditionally have had a substantial variable component, increasing with the size of the transaction. The Regulation II rate, by contrast, is almost entirely fixed. Thus, although the overall effect of Regulation II is to cut rates by about 50 percent, the Regulation II cap is above the preexisting market rate for small-ticket transactions.

40 TILA § 920(b)(1)(B).

41 TILA § 920(b)(4).

42 Letter from Independent Community Bankers of America, to Jennifer J. Johnson, Sec'y of the Board of Governors of the Federal Reserve System (Feb. 22, 2011) (on file with author), *available at* www.icba.org/files/ICBASites/PDFs/cl022211a.pdf.

43 Evans, Litan & Schmalensee, *supra* note 28 (presenting industry concentration data for retail banking in large SMSAs).

44 Daniel Immergluck, Foreclosed: High-Risk Lending, Deregulation, and the Undermining of America's Mortgage Market (2009).

45 Robert J. Shiller, The Subprime Solution: How Today's Global Financial Crisis Happened, And What To Do About It (2008).

46 Amir Khandani, Andrew W. Lo, & Robert C. Merton, *Systemic Risk and the Refinancing Ratchet Effect* (MIT Sloan Research Paper No. 4750-09, Harvard Business School Finance Working Paper No. 1472892, Sept. 15, 2009), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1472892.

47 Donald MacKenzie, *The Credit Crisis as a Problem in the Sociology of Knowledge*, 116(6) Am. J. Soc. 1778 (May 2011); Gillian Tett, Fool's Gold: The Inside Story Of J.P. Morgan And How Wall St. Greed Corrupted Its Bold Dream And Created A Financial Catastrophe (reprint ed., 2010).

48 *See* David A. Skeel, The New Financial Deal: Understanding the Dodd-Frank Act and Its (Unintended) Consequences (2010) (arguing that the entirety of Dodd-Frank reflects a similar failure to grapple with the social costs of financial centralization, and that the statute inevitably will lead to increased power by the largest financial institutions).

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

DURBIN'S FOLLY: THE ERRATIC COURSE OF DEBIT CARD MARKETS

Richard A. Epstein

NYU School of Law
Hoover Institution
University of Chicago Law School

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

DURBIN'S FOLLY: THE ERRATIC COURSE OF DEBIT CARD MARKETS?

Richard A. Epstein*

ABSTRACT

The passage of the Durbin Amendment in July, 2010 followed extensive claims by Senator Durbin and retailers that the only consequence of the law would be to bleed out the excessive debit interchange charges that platform operators and issuing banks collected from retailers. In their view, the proper source of revenues was from debit card holders themselves, as in the Canadian system. Events have not played out that way. After the Federal Reserve authorized a $0.21 base interchange fee, which was generous given the narrow statutory language, the major banks found it impossible to raise rates in the face of sustained political and market pressure, goaded on in part by Senator Durbin himself. At the same time, there is no evidence that the reduction in debit card fees have been passed through by merchants to their customers .

The reason this adventure into regulation has failed is that Senator Durbin and his allies did not understand the operation of the fast-moving two-sided debit card market. In their view, platform operators like Visa and MasterCard operated a duopoly that afforded them the market power to extract rents from merchants while feeding oversized fees to issuing banks in order to attract new streams of customers. That analysis ignores two brute facts. First, the only contest between platform operators, banks and merchants is over the considerable surplus generated by a debit card interchange system. Those fees are constrained because merchants always have the option to pull out of the system if interchange rates are set too high. Second, the interchange fees paid to the issuing banks are not just kept in a vault, but are spent in maintaining the fixed costs of running the system and recruiting new customers, so that all rents are dissipated by these competitive forces. There is, therefore, no unearned surplus, and issuing banks are now forced to adopt inefficient systems of fee collection to offset the nearly $8 billion in lost interchange fees.

Full awareness of the competitive nature of the debit interchange market should have led the courts to declare the current regulatory system a confiscatory form of ratemaking. The combination of higher administrative fees under the Durbin Amendment and lower returns necessarily pushes banks below a competitive rate of return on key debit card services, especially since subsequent events have made clear that there will be zero recoupment in revenues from charges to debit card holders. The level of confiscation is still greater because smaller banks—those with under $10 billion in assets—may continue to collect their full interchange fees in ways that tilt the market even further. Nonetheless, by adopting an all-too-forgiving rational basis test, the courts sustained the statute by showing excessive deference to Congress.

Given the situation today, tinkering will not fix the inherent structural defects of the Durbin Amendment, which should be repealed forthwith before it does any greater damage to debit card transactions.

---

* Laurence A. Tisch Professor of Law, New York University School of Law; Peter and Kirsten Bedford Senior Fellow, the Hoover Institution; James Parker Hall Distinguished Service Professor of Law Emeritus and Senior Lecturer, University of Chicago. For the record, I worked as a lawyer for TCF in the initial stages of the case through the preliminary hearing before Judge Lawrence L. Piersol of the South Dakota District Court of April 4, 2011. I would like to thank Samuel Eckman, University of Chicago Law School Class of 2013, for his usual excellent research assistance.

I. INTRODUCTION: THE COMING OF AGE OF THE DURBIN AMENDMENT

News coverage on debit cards has increased exponentially since Senator Richard Durbin proposed his amendment to the Dodd-Frank financial reform legislation in March 2010. The Amendment gained a lot of initial traction in the Congress, and, with no Congressional hearings in either house, was signed into law on July 21, 2010.[1] The key feature of the Durbin Amendment is that it sets a hard cap on the level of interchange fees that may be charged by big banks (defined by statute as those whose assets exceed $10 billion).

These fees are not set by the Amendment itself, which only contains an explicit instruction limiting these fees to the "incremental costs" associated with the "authorization, clearance, or settlement" of a discrete debit card transaction.[2] The actual numerical rules for calculating these fees were set under the Amendment by the Federal Reserve Board, by capping the amounts charged to about $0.21 cents per transaction with small allowances for additional payments to cover the costs of fraud prevention and fraud loss.[3] That basic $0.21 figure was challenged in an unsuccessful lawsuit brought by TCF Bank. Once the challenge was decisively rejected by the Eighth Circuit,[4] the program went into effect on October 1, 2011.

During the litigation, the claim was commonly made that the regulated banks could offset any revenue loss from the Durbin price caps by raising the direct fees that they charged to their own customers for debit card use. That offsetting fee could take place either on a monthly basis or on a per transaction (or per-swipe) basis. Bank of America,[5] Wells Fargo,[6] and several other banks sought to make good on that option by setting debit card fees at between $3 and $5 per month.[7]

But the huge level of popular resistance by debit card customers, spurred on by Senator Durbin,[8] the media,[9] and various activist groups, turned this recoupment effort into a nonstarter.[10]

Consumers who have long gotten debit cards for free are in no mood to pay a dime.[11] All the proposed fees were eliminated, leaving big banks to scramble for other ways to either reduce costs or increase revenues to control the near-$8 billion gap that the imposition of the Durbin Amendment left on bank balance sheets.[12]

The banks were already in perilous condition because of the general downturn in the market, the glut of real estate in foreclosure, and the various restrictions that the CARD Act[13] had imposed on credit card fees in 2009.

At the same time, the Durbin Amendment has caused dislocations for small merchants.[14] Prior to the passage of the Amendment, debit interchange fees were commonly a percentage of the particular transaction, such that banks made up any losses on small transactions by the higher fees on the larger ones. That system meant that merchants were willing to keep all transactions in the system because the debit card fees did not eat up the profits. But once the Durbin Amendment capped maximum fees, the banks raised the fees on smaller transactions to the level of the cap to make up for those lost fees on larger transactions, which in turn chipped away at profit margins for retailers whose business consisted of large numbers of small transactions.

Moreover, a recent study by the Electronic Payments Association finds that merchants have not passed through debit interchange savings to consumers. One does not have to accept that extreme conclusion, for competition may result in some degree of price reduction on the merchant side.[15] To be sure, representatives of retailers have consistently claimed that they would pass their savings through to customers.[16] As a matter of basic political economy, however, it is highly unlikely that those pass-throughs would be dollar-for-dollar: why would merchants push so hard for the Durbin Amendment if they could not keep a large chunk of the gains for themselves?

In the end, therefore, the new situation is likely to prove unstable, so that the entire system could partially unravel as some merchants opt out of the system, which in turn means that there are fewer transactions to cover its fixed costs. As is always the case, price controls have unintended, and unwelcome, consequences.

In this article, I review the various economic and legal issues that arose from the time that Senator Durbin first proposed his Amendment in May of 2009 until the present.[17]

Until the Durbin Amendment, the good news for debit cards was that it had sparked an expanding market with relatively little legislation or administrative action–a sign that it was in good health. Left to its own devices, that well-functioning market demonstrated its continued ability to process billions of transactions in an apparently effortless fashion.

It did so because all market participants had strong incentives to extract virtually all potential gains through their repeated transactions.

Everyone stayed with the system because everyone profited from it; each player had at least some piece of the overall gains.

The activities that meet this mutual gain condition are, in turn, capable of generating sufficient profits to insure the continued quiet expansion of the market sparked by strong innovation and powerful consumer acceptance.

Once public cries of discontent gain traction they usually translate into ill-considered regulation that, caught in the vise of the rule of unintended consequences, only makes operations on the ground worse.

At present, the final stage of that cycle is now being played out in the debit card market. To show the trajectory of debit card regulation, I proceed as follows: in Section I, I recount the institutional arrangements that have made the debit card system a continued success story; Section II recanvasses the arguments that were invoked successfully to justify the major form of regulation contained in the Durbin Amendment; I examine the legal arguments that surround the unsuccessful, but sound, constitutional challenge that was raised against the Durbin Amendment in Section III. I conclude with a broad look at the consequences of the Amendment on banking institutions and the broader economy.

II. THE DEBIT CARD IN GOOD TIMES

For many years, the most notable feature of debit cards was their ability to gain an ever-larger fraction of payment transactions. In 2009, debit cards became the most common form of payment, whether measured by dollars or by number of transactions.[18] Debit cards outpaced their more expensive credit card rivals; they lapped the track with clunky paper checks and made major inroads into cash purchases. During the years 2005 to 2009, the volume of transactions increased, yet the average interchange fell. Thus in 2005, average debit interchange fees were about 1.83 percent on a purchase volume of $2.651 billion. By 2009, the rate had dropped to 1.69 percent on a volume of $3.663 trillion purchase, which translates into a 7.7 percent drop in rates on a volume increase of about 39 percent.[19]

Success on that order of magnitude did not occur by chance. All the relevant players in the payments market enter into hundreds of billions of transactions each year. If a system's overall design contains a serious glitch, players will discover it and thereafter will alter their behavior to mitigate their losses. Yet by the same token, when the new payment system produces systematic net benefits, the same parties will gravitate toward its use, even if they do not understand its precise mechanics.

This path of development marked the rise of the debit card. The debit-card universe involves, in its simplest iteration, five discrete players. At one end of the transaction lies the bank customer (doubling as a retail consumer) who receives a debit card from an (issuing) bank. The bank earns its revenues not by any direct charges against that customer, but by collecting an interchange fee from the retailer or merchant at whose establishment the customer uses that debit card to complete a transaction. The merchant requests through its own (merchant) bank verification that the customer is in fact able to pay the charge. This information is forwarded via a network platform—typically, but not always, Visa or MasterCard—back to the customer's bank, which can then authorize or decline the payment. Because that bank has up-to-date information about the status of the customer's bank balance, it can use complex algorithms to decide whether to authorize or decline payment. When the transaction is approved, the issuing bank keeps part of the proceeds (typically around 1.35 percent of the transaction amount)[20] to cover

its own costs, and forwards the rest of the money to the network platform. The network platform then takes its (smaller) slice for routing the transaction to the merchant bank, who takes its own slice for servicing its business client. When these three slices are removed from the sales price, the merchant receives about 98 percent of the face amount of the transaction.[21]

A two-percent take from gross sales is a considerable expense in a low margin business.

But owing to the repeat nature of these transactions, retail merchants, faced with fierce competition, continue to pay this fee for one and only one reason: they receive in return benefits in excess of the costs imposed.

Defenders of the Durbin Amendment say that merchants only pay their fees because they need to keep up with the competition. They write: *"Because the RLC's [Retail Litigation Center] members must accept debit cards to remain competitive, they have had no choice but to pay these fees."*[22] But that is precisely the point. Merchants compete by supplying the same low-cost services as their competitors. Keeping up with competition is good, not bad. Indeed, if it were bad, these keen competitors would opt out by reverting back to credit cards (which carry a higher interchange fee because of the greater credit risk assumed by the bank) or checks and cash, each of which have their own problems as payment mediums. Thus the system endures because it generates benefits to all players, including those who have groused about it most. At this point, it is possible to identify just what those benefits are.[23]

1) Speed. The rapid pace of debit card transactions on check-out lines reduces the collateral costs of servicing these accounts. Check-out clerks can process more transactions per hour, and fewer customers walk away because they do not wish to stand in long lines for small purchases. Furthermore, in those settings where debit card transactions don't make sense, retailers can set up cash-only lines so its operations move smoothly, or they can just decline to accept the cards. These multiple options lead to advantages in the hands of a skilled professional.

2) Ticket uplift. Debit card use typically increases the size of a particular purchase because the customer is no longer constrained by the amount of cash in his or her wallet. A virtuous circle is at work in these situations: the knowledge that merchants accept debit cards offers yet another reason for consumers to carry less cash, thereby reducing the personal risk of theft or loss.

3) Information. The debit card gives all parties an accurate and instantaneous record of each individual transaction that can be used for multiple purposes. To the merchant, the collection of this information assists with better inventory control, cash management, and marketing. For customers, the transaction record allows them to know how much money remains in their deposit accounts. For banks, it allows better recordkeeping of their customers' balances. Good information in these cases is a clear advantage across the board.

4) Guaranteed payments. The debit card improves risk management by allowing the issuing bank to make up-to-date credit checks that reduce the risk of default. In this regard, it is critical to correct the common misconception that debit card holders must have sufficient funds in their accounts for their bank to authorize the transaction. It does wonders for customer relations to allow clients to overdraw their accounts toward the end of each pay period, and to recoup those lost revenues when the next pay check comes in a day or two later. Since the banks have the superior information, they can take the credit risk away from the merchants by guaranteeing payment on authorized transactions (only), whether or not the customer defaults. That risk of loss can also be shifted with the use of debit cards. But that arrangement will work only if issuing banks in the long run can cover two costs: those of running the system, and those of covering the losses that still occur from assuming these risks. At this point, however, the existence of these debit-card losses is not a sign of social dislocation, for the banks have every incentive to make the right trade-offs at the margin, by refusing to extend credit when the risk of loss appears too great. This risk-shifting operation is far more costly when payments are made by check, since the issuing bank has no better information about the proposed transaction than the party who accepts the check.

In the abstract, this bundle of benefits may not be sufficient to justify the interchange fees that banks exact for debit cards. But there is no need to speculate as to the relative benefits and tradeoffs, given the system's growth.

At this point the correct inquiry is: when a system produces net benefits to all participants, is it nonetheless possible to identify some market imperfection that justifies regulation?

Once that imperfection is identified, the next question is whether it is possible to develop a regulatory scheme that corrects that imperfection at an acceptable cost. Under this inquiry, it is not enough to show some deviation from the standards of perfect competition. It is necessary to show that the deviation is large enough to justify the particular regulatory response that is imposed.

III. RATE REGULATION FOR DEBIT CARDS: IMPERFECT INFORMATION AND MONOPOLY POWER

In the context of debit cards, there are only two plausible justifications for rate regulation of debit cards. The first is to find some information asymmetry. The second is to find some exercise of monopoly power. Both justifications are addressed below.

A) IMPERFECT INFORMATION

The Durbin system of rate regulation cannot be justified as a means to counter supposed informational deficits in the debit card market, especially on the part of merchants who are so intimately familiar with its operation. Likewise, the customers who continually revert to debit card use learn debit card operations, if not as a matter of abstract logic, then through experience. Customers prize debit cards for the want of direct fees, for their convenience, and for the various extra bonuses that banks dispense to lure customers to sign up. It has been suggested that consumers should be told of the debit card fees that are hidden in the total purchase price,[24] yet retailers are not required to disclose all the other cost components of their business, such as electricity, rent, salaries and merchandise. The information that consumers use to make decisions concerns the price and quality of goods that are under consideration for purchase.[25]

They buy if the price is less than the net benefit, and decline to purchase if it is not. Information about component costs is just a distraction, for few consumers will switch from the lower to higher cost identical good merely because the cheaper good embeds a higher cost of electricity. Moreover, whatever disclosures might be required under this logic deal not with the debit fee itself but with the bank charges for administration and bad debt losses. Yet neither of these charges bears any relation to the severe price caps now authorized under the Durbin Amendment.

B) MONOPOLY PROFITS

The question of potential monopoly profits requires more analysis than that devoted to information asymmetry. The proponents of the Durbin Amendment, including the Senator himself, have long insisted that the credit card companies, most especially Visa and MasterCard, exert market power over the industry in virtue of their "duopoly" controlling the two major platforms for debit card transactions. Senator Durbin insisted that the Amendment "will prevent the giant credit card companies from using anti-competitive practices."[26]

On this account, this precarious situation did not come about through competition or through a healthy market for debit cards. To the contrary, debit card networks such as Visa and MasterCard, and the banks that issue their debit cards, have imposed this system on merchants through collusion, with banks agreeing not to compete over these fees, and through the market power that the banks exercise through the networks. Because retailers must accept debit cards to remain competitive, they had no choice but to pay these fees.[27]

The clear implication of this position is that collectively, the banks and the debit card networks are able to extract some additional revenue by virtue of their control over the key middle step in the debit card system. The claim of "collusion" is not wholly correct, for it were, all parties to the collusive arrangement would have organized a system of horizontal price restrictions

that would be in violation of Section 1 of the Sherman Act. There have been cases involving tie-in arrangements that have raised these antitrust concerns,[28] but

there is no case in which anyone has attempted to show collusion between rivals at either the network or the bank level.

The claim that such collusion exists therefore has to rest on the claim that both Visa and MasterCard has chosen to raise their rates, knowing that the other is likely to follow suit. That form of "parallel pricing" is difficult to prove (or disprove), since there is no independent evidence of what the pure competitive rate would otherwise be.[29]

The level of gains that can be reached through any form of tacit collusion are always smaller than those that can be achieved by direct linkage between the parties; under a Cournot duopoly, each party ignores the gains that the other receives from raising prices, so that in equilibrium, the prices set are somewhere between the monopoly and competitive price. That possible gap shrinks still further for two other reasons. First, Visa and MasterCard do not have 100 percent of the market, but 83 percent,[30] which means that additional competitive forces are at work. Second, the market in payments generally is highly dynamic so that the prospect of further entry through new forms of payment, e.g. mobile phones, will induce the incumbents to lower their prices still further, at least in the long run, and perhaps sooner. Hence, even if there were some supracompetitive profits in this industry, they are likely to be small relative to the overall gain. In addition, *the alleged ability of Visa and MasterCard to extract these supposed rents is still limited by the key constraint that the rates charged to merchants must be low enough to keep them in the system.*

The estimated size of these rents matters, because the smaller their size relative to the gains generated by the system, the weaker the target for constructive interference, given that all forms of regulation are prey to two systematic risks: error costs in administration, and flaws in design, induced by the price restrictions imposed through the political process.

These concerns with regulation are very much in play in the context of the strict rate caps imposed by the Durbin Amendment. The first, and most obvious, point is that the Durbin Amendment does not impose any restrictions at all on the rates that the debit card companies can charge for their services. That figure is small, in the neighborhood of 0.20 percent of the transaction, and far smaller than the fee charged by the issuing bank (around 1.35 percent) or the merchant bank (around 0.5 percent). It would be difficult for the merchants to claim that their own banks are involved in any collusion, so the only target that they have is the issuing banks, who also earn the bulk of the fee in question. The issue then arises, however, about market power, of which not even the largest bank—Bank of America—has.

What is needed is some theory that can explain why debit card companies are prepared to use their supposed monopoly clout in order to benefit the issuing banks with whom they have only an arm's-length contractual position.

During the litigation, the retailers relied on a report by Steven C. Salop,[31] written on behalf of the Merchants Payment Center, which purported to explain that connection. Under Salop's view, the optimal system is the Canadian system, which uses no interchange fees at all, but has each side pay its own costs for running the system. As Salop explains, "The most economically reasonable way to satisfy the mandate for debit interchange fees that are reasonable and not disproportional to issuers' costs is to adopt a presumptive standard of at-par interchange ("API"). Under this standard, there would be a strong regulatory presumption that interchange should be at-par for all debit card networks."[32] In effect, consistent with the logic of the Durbin Amendment, the model that is used for checks is carried over here.

The initial question to ask is whether there is any need for regulation at all, given the successful growth and evolution of the debit card system. One way to insure "minimal market intrusion,"[33] as Salop wishes, is not to regulate at all and let evolution take its course.

Cutting interchange rates by 100 percent (as the API model suggests), or reducing them to the incremental costs of authorization, clearance and settlement costs specified in the Durbin Amendment (Salop's fall-back position), are far cries from minimal market intrusions.[34] "Minimal" takes on a different coloration if there is a market failure. In this instance, however, Salop finds a sufficient source of market power, because "Visa and MasterCard have the ability to exercise significant market power over merchants with respect to the acceptance of debit cards by raising their interchange fees, which then are passed on to retailers by the acquiring banks."[35] The first half of the sentence is subject, of course, to the limitation that they cannot raise these fees to a level that drives merchants to other debit card providers or other forms of payment. Salop implicitly acknowledges the point when he observes that merchants accept the charges because "losing the sale would be costlier to the merchant than accepting debit and paying the high interchange fee."[36] Salop's remark is another way of saying that the debit interchange fee makes sense for merchants, such that the only remaining issue is to determine the division of gains from operating the system.

The Durbin Amendment therefore hits the wrong target by attacking issuing banks for market power which is said to lie with platform operators.

The question is why these operators would gratuitously divide monopoly rents, if any, with a group of competitive banks.

Salop's explanation for the transfer is that banks are in a stronger position vis-à-vis the platform operator than the merchants. Every merchant has to accept all debit cards, but each bank needs only one platform operator to run its business. The banks therefore can play the card companies off against each other in order to increase receive higher interchange fees.

This analysis is incomplete because it never asks what banks do with the added interchange revenues. They cannot just pocket the funds, for they are in aggressive competition with each other, forcing them to spend money to recruit and retain customers. Those activities work to the benefit of merchants and platform owners. By lowering price and raising service, interbank competition brings additional debit card users into the system.

At no point does Salop explain why competition for customers does not drive bank returns down to competitive levels. His report offers no evidence of any extraordinary returns for the banking industry. Nor does it ask the vital question of whether the use of interchange fees supplies an efficiency advantage that cannot be duplicated by forcing issuing banks to cover all their debit card cost through fees collected solely from their own customers, in order to let debit card transactions clear at par.

To see what is missing, it is necessary to discuss more deeply the operation of these two-sided markets.[37] In his expert testimony, Salop refers repeatedly to these markets,[38] but does so solely to show their monopoly tendencies: "In two-sided markets, networks with market power may be able to exercise substantial market power over one side of the market but be able to exercise less (if any) market power over the other side of the market."[39] But at no point does he address the classic efficiency explanations for the rise of voluntary two-sided markets, which date back to the classic paper by William Baxter in 1983,[40] and which were elaborated in the expert testimony prepared for TCF National Bank by Kevin Murphy.[41]

The ability of the platform operator to coordinate cross payments from one side of the market to the other opens up a potential gain from trade that cannot be captured in the at-par system that Salop champions. The simplest version of the story is that these payments take advantage of the different levels of elasticity on the two sides of the market. The merchants, whose demand for debit payments is highly inelastic, pay something to customers, whose demand is highly elastic, and who thus are more willing to leave the system. These merchant payments are not made directly, but through the interchange system to the issuer, in order to help them bring in customers to the system who might otherwise stay out.

Those payments are not specific to any merchant, and hence when spent by the issuing banks, they improve the entire operation of the system. The individual merchants therefore need not worry about free-riding by rivals because they know that the standardized interchange fees reduce the ability of any given merchant to foist its costs off on other parties. The greater number of customers who are brought into the system further increases the willingness of consumers and merchants alike to remain in the system.

The fixed rate schedule for interchange fees, when set by the platform operators in competition with each other, eliminates the expensive costs of negotiating individual transactions and thus improves the overall efficiency of the system.

This increased efficiency undermines the view that fixed fees are only of use in setting cartel-like prices for issuing banks that are in heavy competition with each other.

The American debit card system has proved more innovative than its Canadian rival because it allows for both foreign and online transactions (at the time of the litigation, these two types of transactions could not be performed over the Canadian system,[42] which also featured very high first-party interchange fees running between $0.50 and $0.60 per transaction in 2004).[43]

Since all consumers are involved on both sides of the deal—that is, with both the merchant and the issuing bank—their preference is for a set of arrangements that minimizes the sum of their costs on both sides of the transaction.

In this environment, the broad scale acceptance of the system suggests that it has done that. No one can argue that a network industry achieves a perfect competitive solution, for that result is impossible no matter what interconnection rules are adopted. But it should remind us how difficult it is to construct a regulatory framework that works better than these voluntary deals.

In this institutional environment, the Durbin Amendment constitutes a classic case of regulatory overkill.

The rate restrictions that are set exceed those necessary to combat any risk of market power by Visa or MasterCard, and they bear no relationship to the tiny monopoly rents, if any, that issuing banks can extract from the system. Indeed, the Durbin Amendment makes no effort to calibrate its price caps to offset any supposed level of market power.

Instead, it treats the entire debit interchange system as if it were some kind of public utility that is allowed to recover its costs, but not a risk-adjusted competitive rate of return. The rate base for the regulation is tied to the notion of incremental cost, which does not allow for any recovery of the extensive fixed costs incurred to operate the system.

In addition to the transaction-specific costs of authorizing, clearing and settling a transaction, a bank must design, construct, maintain and upgrade the basic system, supply support services for existing customers, and invest in soliciting new customers, which is no mean task given the high rate of debit card turnover. TCF, for example, "has been required to open 500,000 to 600,000 accounts each year just to maintain its customer base,"[44] all at, in the pre-Durbin days, no cost to the customer. The Durbin scheme therefore contemplates that all these costs should now be switched from the interchange system to the customers in ways that approach the Canadian system. For that to happen, all these additional costs should be recouped directly from customers, which has turned out to be institutionally impossible once the Durbin Amendment has gone into effect. The question is how these various elements line up in connection with the legal challenges to the Durbin Amendment that were turned aside in the Eighth Circuit.

IV. CONSTITUTIONAL CHALLENGES TO THE DURBIN AMENDMENT

A) PROSPECTIVE AND RETROSPECTIVE REGULATION

The basic challenge to the Durbin Amendment rests on the view that the rate regulation imposed under this system has to be tested by the same constitutional rules that apply to other forms of rate regulation.[45] In this instance, there are two approaches to the question, one of which deals with prospective regulation generally, and the other which deals with the greater protection that is afforded for public utilities that have made specific investments in plant and equipment.

The difference between these prospective and retroactive forms of regulation is reflected in the standard of constitutional review that is applied. For prospective regulation, the current standard supplies a low level of protection under the "rational basis" test, and the challenger faces a steep uphill climb.

This was in fact the test adopted in the TCF case, where the Eighth Circuit opined:

"Parties making substantive due-process claims concerning economic regulations generally face a highly deferential rational basis test, whereby the burden is on the one complaining of a due process violation to establish that the legislature has acted in an arbitrary and irrational way. Similarly, the standard for determining whether a state price-control regulation is constitutional under the Due Process Clause is well established: Price control is unconstitutional if arbitrary, discriminatory, or demonstrably irrelevant to the policy the legislature is free to adopt."[46]

Perhaps the most famous rate-making case of this sort is *Nebbia v. New York*, where the Supreme Court upheld **minimum** prices for milk in a competitive industry against a charge that they violated the economic liberties of the milk producers, who wanted to sell milk at below the regulated prices.[47]

Similar arguments have been used to uphold rent control statutes, which set maximum rentals, against charges of confiscation, at least if the rentals allowed were sufficient to cover the cost of providing services to the tenant even if they did not allow the landlord to any rent increases to reflect the appreciation of the underlying asset.[48] Indeed the most influential formulation of the rules that are associated with regulatory takings are those found in *Penn Central Transportation Co. v. City of New York*,[49] under which the local government was allowed to prevent the use of air rights for new construction of a landmarked building on the ground that the revenues received from the operation of the existing facility were sufficient to cover its costs. Finally, in *Yakus v. United States*, the Supreme Court sustained a general system of prospective price controls put forward for all goods and services under loose guidelines that left a fair level of administrative discretion in the joints.[50]

Faced with these precedents, it may be easy to conclude that virtually any system of rate control passes constitutional muster. But the issues are far more subtle than this initial analysis suggests. *Yakus*, in particular, was decided only two months after the Supreme Court handed down its public utility rate regulation case in *Hope Natural Gas v. Federal Power Commission*,[51] which took a very different approach toward public utility regulation with respect to invested capital that had already been committed to a particular venture.

B) TRADITIONAL PUBLIC UTILITY RATE REGULATION

At issue in *Hope* was the form of rate protection for public utilities that must incur huge sunk costs **before** they can begin the operations that allow them to recoup their initial investment and operating costs. In these situations, two warring concerns require reconciliation.[52] The first is that, traditionally, the public utility has a natural monopoly in the geographical region in which it operates. The high fixed costs of building a plant are such that no second company can enter the market at a cost below that which the incumbent can charge for its services, even if allowed to do so as a matter of law.

The key assumption that supports this view is that over the relevant range of output, the incumbent has declining marginal costs that allow it to price additional units of service below those which the new entrant must charge in order to cover the heavy costs to set up his initial system. Put otherwise, the industry operates at a lower cost with one firm than it does with two. The level of monopoly power is only entrenched further if the public utility commission is vested with the power to deny a license to any new entrant that might decide to brave entry. Rate regulation is one permissible means to combat the use of this monopoly power.[53]

The problem is that the system of rate regulation cannot operate in a fashion that makes it impossible for the utility to recover its sunk costs over the useful life of its capital investments. Thus, in a world devoid of constitutional protection, the public utility commission could trap the utility once it has gone into operation by setting rates that allow it to recover revenues beyond its marginal costs, but that do not allow it to recover the fixed costs plus a suitable rate of return over the useful life of the regulated facility. So long as these rates are above marginal cost, the utility loses more money if it withdraws from the market than if it remains.[54] As a result, unless there is protection against that misbehavior once the utility is in operation, no one will set up a plant in the first place, given the ever-present risk of confiscation.

The utility therefore needs ironclad guarantees that it will be able to recover not only its fixed costs, but also a reasonable profit that is needed to attract and retain capital.

The system of rate regulation under Hope Natural Gas must make some provision that the rate structure will allow for that return.

The question then arises how the courts supervise this constitutional standard given the difficulty of its administration. Courts have adopted a two-part approach. The first is setting the ideal standard, on which the command is categorical. The legislature cannot drive the regulated firm below that risk-adjusted rate of return, where the adjustments in question take into account that a natural monopoly in a stable geographical market faces lower risks than the ordinary competitive return. The regulator is then given wide discretion on the way in which various items of revenue and expense are taken into account, on the ground that intermediate errors along the way should not attract judicial attention, so long as the "bottom line" meets the appropriate standard.[55]

The key point is that this process under Hope Natural Gas applies to any firm that has made fixed investments in its own facilities.

The state could avoid any and all obligations of this sort if it announced in advance that it will only allow a firm to enter this market if it accepts the risk of confiscation, at which point the firm can protect its position by declining the opportunity. Thus under current law, if (before the onset of the debit card business) Congress passed a statute that forbade all debit interchange fees, the regulation would stick, and the business as it emerged might well follow along Canadian lines. In this environment, however, regulators are unlikely to impose these restrictions, because they understand the brutal truth that these prohibitions could easily discourage or block the needed investment in the first place. No regulator therefore imposes confiscatory rates that operate in futuro only.

At this point, it must be stressed that all banks made their initial investments in debit cards in an unregulated, competitive market, in which their ability to work out in advance the details of the debit interchange system through long-term contracts protected them against merchant expropriation. Since these investments were all made in depreciable assets that are typically not sold, there is no possibility that they will appreciate over time in the manner of residential real estate.

Hence it is perfectly sensible to set the rate of return in ways that allow for the recovery of the initial costs over the useful life of the assets. Rate regulation of industrial facilities does not pose the serious danger of abuse present in rent control, where the appreciation of rental property is in effect transferred to the tenant through the statutory right to remain on the premises long after the expiration of the original lease. Instructively, however, the rent control rules also provide that the rates cannot be set so low as to deny recovery on the original investment, which is all that is claimed in this case.ations are addressed below.

C) DEBIT CARD REGISTRATION

Once rate regulation is imposed on banks that have invested capital in their debit card systems, the same consideration applies: the revenues that the firm receives over the useful life of that equipment must be sufficient to allow for the recovery of all relevant costs. These costs are, as in the public utility cases, much more extensive than the incremental costs associated with the supply of individual units of service, and each variation on the public utility rules requires the rate base to include those elements.[56] The same situation is required here, for otherwise the government is allowed to engage in a bait and switch, whereby it encourages investment under one legal regime, only thereafter to deny the firm its needed recovery once the investment is made under a second. The only question is how the analysis plays out as the discussion moves from traditional public utilities to debit card transactions.

In dealing with that issue *there are two major differences between the debit interchange market and standard public utility regulation,* there are two major differences between the debit interchange market and standard public utility regulation, one of which strengthens the TCF challenge to the Durbin Amendment and the other which cuts against it. The first difference is that rate regulation here is imposed on what is a virtual competitive industry, where any pocket of monopoly power is tiny relative to the systematic long-term territorial monopoly of the standard public utility. That is, the analysis above makes it clear that there are no supracompetitive profits for government regulation to

bleed out of the system. As a matter of simple math, if the current rate of return in this industry is **already** at the risk-adjusted competitive rate, any government effort to reduce that rate of return and to add administrative costs into the system **necessarily** pushes the regulated issuing bank below the competitive rate of return. The monopoly cushion that is available to regulated industries that have monopoly power, either because of their economic position or because of some legal privilege, is **never** available for a firm that is already at the competitive position that is the end point for any sound system of rate regulation.

The reduction of revenues under the Durbin Amendment thus leads to a confiscatory rate structure for all invested capital unless the revenues lost through regulation can be recovered from the other side of the market.

In dealing with this issue, the Eighth Circuit found that the opportunity to recoup lost revenues from customers, in the manner of the Canadian system, was the Achilles' heel to TCF's case. The Eighth Circuit wrote:

The Durbin Amendment only restricts how much certain financial institutions issuing a debit card may charge for processing a transaction; it does not restrict how much those institutions may charge their customers for the privilege of using their debit-card services. Since TCF is free under the Durbin Amendment to assess fees on its customers to offset any losses under the Durbin Amendment, it is unlikely that the Durbin Amendment has created a sufficient price control on TCF's debit-card business so as to trigger a confiscatory-rate analysis, or that the law could, in fact, produce a confiscatory rate. Indeed, the heart of any confiscatory-rate claim is the ability to show that the government has set a maximum price for a good or service and that the rate is below the cost of production (factoring in a reasonable rate of return), which TCF has simply not shown on this record.[57]

In making this argument, the Eighth Circuit is conscious of the procedural posture of the case, under which TCF must meet a heavy burden of proof in order to enjoin the statute before it is put into effect.[58] That claim in turns stands or falls on the question of how best to valuate that right. In these circumstances the analysis can occur in three separate ways.

The first involves the analytics of the matter, without taking into account the statutory exemption for banks with less than $10 billion in assets. The second takes that exemption into account. The third considers the political fallout from the Durbin Amendment.

Under the first scenario, there is clearly no direct information about the various strategies that banks will use to recoup their losses. For the purposes of this analysis, it must be assumed that once the banks are faced with the rate regulation under the Durbin Amendment, they will take all steps within their power to mitigate the losses imposed on them. We can assume for the sake of analysis that they will engage in error-free strategy, so that they will cut back on benefits and increase their fees in ways that maximize their profit position, conditional on the passage of the Durbin Amendment.

These changes do not matter, so long as the attack is directed to the investment that the regulated banks have made in the system.

Given the operation of two-sided markets, it follows that any system that bans interchange fees forces banks to get all their income from the consumer side of the market.

The chances that even the best alternative strategy can put the banks back to their pre-Durbin state of earnings (recall that this is the competitive rate of return, which eliminates any margin of error) by making that 100 percent recoupment are zero. Two sources of revenue are always greater than one, especially when the theory of two-sided markets holds unambiguously that payments across the platform generate additional efficiencies that the Canadian at-par system cannot hope to match. If the Canadian system were as efficient, the banks would have no reason to object to the Durbin Amendment, and indeed no reason to set up the debit interchange system in the first place. The only disputed question therefore is the extent of the loss, not its existence. If litigation involved efforts to determine the amount of money that the government owed for imposing these restraints, the question could not be resolved before the systems were put into effect. But given that the government has made it clear that no compensation is in the cards, the size of the shortfall is utterly immaterial to the outcome of the case.

There is no state of the world where the compensation derived from customers could, even conceivably, provide the perfect offset needed to restore the competitive rate of return.

It might be said that this point ignores the possibility of market power, which I criticized above. Ironically, that objection was disposed of on appeal by the District Court's finding that the debit card industry was in fact competitive among the issuing banks.[59] As Justice Piersol noted, "there is no monopoly power assumed to be associated with issuing debit cards. Plaintiff is not a public utility under rate case jurisprudence. The case law relied upon by Plaintiff is therefore inapplicable to its due process claim."[60] His point gets it exactly backwards. As the case law has long recognized, firms in competitive markets are entitled to the opportunity to run their business at a profit.[61] The want of market power strips the government of any reason to regulate debit card rates in the first place. Accordingly, the level of scrutiny to rate regulation should be **higher** when the government seeks to regulate the rates of a competitive firm.[62] The mathematics show that the government should lose under any and all circumstances. Thus in a competitive market, the relationship between revenues and costs sets up the risk-adjusted rate of return as follows:

$$\frac{R - C}{R} = \pi$$

In this simple equation, R equals revenues, C equals costs, and π equals profit under competitive conditions. In the new environment R* (= 0.5R) is less than R, and C* (= 1.5C) is greater than C, such that π^* is necessarily less than π. To see why, take the case of a regulation that cuts revenues in half—as under the Durbin Amendment—and increases compliance costs by the same amount. To this point,

$$\frac{R^* - C^*}{R^*} = \frac{0.5R - 1.5C}{0.5R} = 2(0.5)\frac{R - 3C}{R} = \frac{R - 3C}{R} = \pi^* < \pi$$

The results do not depend on the choice of coefficients for R and C after regulation. So long as the revenues are less than one, and the costs are greater than one, the coefficient for C will be greater than one while those for R will be one, so that the inequality holds in all states of the world. The case against Durbin on the assumption that markets are competitive rises to the level of a truth.

The second state of the world is the current one, where the banks whose assets are below $10 billion are exempt from the restrictions on interchange fees. At this point, the case against the Durbin Amendment is stronger than it was before, because the differential form of regulation necessarily reduces the pricing and marketing strategies available to the big banks. The only question that is worth asking is how significant the cost differential would turn out to be. In work done for TCF Bank, Anne Layne-Farrar estimated that there would be high slippage rates if the banks sought to recoup the estimated $10 in lost interchange fees through direct monthly charges.[63] In dealing with this issue, the Eighth Circuit held that these concerns did not matter because it looked at the rate differential only in connection with an asserted equal protection claim, and not as part of the larger rate-making position. Under that view, it was easy to note that there was an understanding that Durbin was "protecting smaller banks, which do not enjoy the competitive advantage of their larger counterparts and which provide valuable diversity in the financial industry."[64]

The Eighth Circuit did not explain how the smaller banks managed to compete and retain market share prior to the passage of the Durbin Amendment, or why their presence in the market did not already contribute to some needed diversity in the financial industry.

By partitioning the small banking exemption from the taking claim, it ignored the close connection between them. Confiscation in the guise of protection of small (but certainly not infant) industry is a convenient intellectual crutch that avoids all serious analysis.

At this point, however, there is no longer any need for speculation, so that the heavy burden of proof that is needed for a preliminary injunction is not in place. Everyone who worked on the TCF litigation was of the view that some recoupment of debit card fees was at least possible after the imposition of the Amendment. Today, we know better. There were efforts by Bank of America and Wells Fargo to impose fees, but the onslaught of negative publicity and thinly veiled threats by Senator Durbin[65] led to their withdrawal.

V. CONCLUSION

The history of the Durbin Amendment offers powerful evidence for the sources of economic decay in the United States. In the Senate, a strong populist appeal by Senator Durbin drives forward a system of rate regulation that never received any scrutiny before it was added into the Durbin Amendment. The economic analysis that supported his position rested on untenable claims of market power that overlooked the efficiency dynamics of interchange fees, even in a competitive market.

The stubborn unwillingness of courts to look critically at how these markets operate leads them to ignore the inexorable reasons why the Durbin Amendment, under existing constitutional standards, should have been Dead On Arrival.

The stubborn unwillingness of courts to look critically at how these markets operate leads them to ignore the inexorable reasons why the Durbin Amendment, under existing constitutional standards, should have been Dead On Arrival. The litigation is over for the moment, but the bad consequences remain. The disruption of the mechanics of the debit card system show a nation committed to a converging point of view where large sums are invested in regulation that will reduce the operating efficiency of the market. Banks are at this point under siege in virtually all their operations. The scope of modern regulation seems to have a new message, which is to help secure the failure of big banks that are exposed to serious risks of failure.

1 Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, 124 Stat. 1376 (2010) (to be codified in scattered sections of the U.S. Code).

2 More precisely, the Durbin Amendment commands that interchange fees be "reasonable and proportional to the cost incurred by the issuer with respect to the transaction." The terms "reasonable" and "proportional" cannot to be read in isolation, but must be understood in light of their own statutory definition, which requires Federal Reserve in issuing regulation to differentiate between:

> (i) the incremental cost incurred by an issuer for the role of the issuer in the authorization, clearance, or settlement of a particular electronic debit transaction . . . ; and (ii) other costs incurred by an issuer which are not specific to a particular electronic debit transaction . . .

Dodd-Frank Act, sec. 1075, § 920(a)(4)(B)(i)–(ii), 124 Stat. at 2068–69 (to be codified at 15 U.S.C.A. § 1693o-2(a)(4)(B)).
In addition, the Durbin Amendment instructs the Federal Reserve to "consider the functional similarity" to "checking transactions that are required within the Federal Reserve bank system to clear at par." Dodd-Frank Act, sec. 1075, § 920(a)(4)(A)(ii), 124 Stat. 1376, 2068 (2010) (to be codified at 15 U.S.C.A. § 1693o-2(a)(4)(A) (West 2010)).

3 12 C.F.R. pt. 235, *available at* http://www.gpo.gov/fdsys/pkg/FR-2011-07-20/pdf/2011-16861.pdf.

4 *TCF Nat'l Bank v. Bernanke* (TCF I), No. CIV 10-4149, 2011 U.S. Dist. LEXIS 45059, at *14 (D.S.D. Apr. 25, 2011), *aff'd*, 643 F.3d 1158 (8th Cir. June 29, 2011).

5 Linette Lopez, *After Dropping Debit Card Fees, Bank of America Is Quietly Cranking Them Up Elsewhere*, Business Insider, Nov. 14, 2011, http://articles.businessinsider.com/2011-11-14/wall_street/30396471_1_debit-card-swipe-fees-overdraft-fees-costs-banks. The introduction of new, and more inefficient fees, is what one should expect.

6 Vivek Shankar, *Wells Fargo Cancels Pilot of Monthly Fee for Debit Cards*, Bloomberg, Oct. 29, 2011, http://www.businessweek.com/news/2011-10-28/wells-fargo-cancels-pilot-of-3-monthly-fee-for-debit-cards.html.

7 Robin Sidel, *Big Banks Blink on New Card Fees*, Wall St. J., Oct. 28, 2011, *available at* http://online.wsj.com/article/SB10001424052970204505304577002041853240850.html.

8 *See* Letter from Senator Richard Durbin, to John Stumpf, CEO of Wells Fargo (Oct. 29, 2011), *available at* http://durbin.senate.gov/public/index.cfm/statementscommentary?ContentRecord_id=b4fe225f-fc62-455f-abbb-6f9f4c1c60d5:

> *But you did not earn these fees by bettering your competitors in a free market, which is how Main Street businesses have to make their money. Rather, you made this lucrative revenue stream because the Visa and MasterCard duopoly fixed the same high swipe fee rates for your bank that they did for every other bank—thus immunizing this revenue stream from competitive pressure and enabling fees to keep going up even as processing costs went down. It is disingenuous for banks to claim they are somehow entitled to make up reductions to a revenue stream that they never would have received in the first place in a transparent and competitive market.*

For a discussion of these claims, see *infra* at Section III, B (page 18).

9 *See* Ann Carrns, *Fees Help Drive Working Poor From Banks*, N.Y. Times, Oct. 21, 2011, *available at* http://bucks.blogs.nytimes.com/2011/10/21/fees-drive-working-poor-from-banking-system (explaining the fees without stopping to ask how banks were able to secure debit cards under the supposedly defective system that the Durbin Amendment had just dismantled).

10 For my account, see Richard A. Epstein, *The Debit Card Stealth Tax*, Defining Ideas, Oct. 4, 2011, http://www.hoover.org/publications/defining-ideas/article/95011.

11 *Four in Five Consumers Won't Tolerate a Monthly Debit Card Fee*, PYMNTS.com, Dec. 7, 2011, http://pymnts.com/regulations/debit-checking/four-in-five-consumers-won-t-tolerate-a-monthly-debit-card-fee.

12 "The limits, mandated by the Dodd-Frank Act, may cut annual revenue by $8 billion at the biggest U.S. banks, according to data compiled by Bloomberg Government." Hugh Son, *Debit-Fee 'Flop' Leaves Banks Seeking $8 Billion in Revenue*, Bloomberg, Nov. 2, 2011, http://mobile.bloomberg.com/news/2011-11-02/debit-fee-flop-leaves-u-s-banks-looking-for-8-billion-in-lost-revenue.

13 The Credit Card Accountability Responsibility and Disclosure Act of 2009, Pub. L. 111-24. 123 Stat. 1734-1766 (2010).

14 Eamon Murphy, *How Durbin's Debit Card Fee Cut Backfired on Small Merchants*, Daily Finance, Dec. 8, 2011, http://www.dailyfinance.com/2011/12/08/how-durbins-debit-card-fee-cut-backfired-on-small-merchants.

15 **"There is no evidence that American consumers are benefitting from the Durbin amendment, despite overwhelming evidence that the retail industry is experiencing significant savings."** (bold in original), Electronic Payments Coalition, *Where's the Debit Discount: Durbin Price Controls Fail to Ring Up Savings for Consumers* 3, Electronic Payments Coalition, Dec. 12, 2011, http://wheresmydebitdiscount.com/wp-content/themes/epc/media/Where's%20My%20Debit%20Discount%20-%20Durbin%20Price%20Controls%20Fail%20to%20Ring%20Up%20Savings%20for%20Consumers.pdf, [hereinafter EPC, *Debit Discount*].

16 *See, e.g.*, Statement of Mallory Duncan Senior Vice President and General Counsel for the National Retail Federation: "Merchants are ready to pass lower swipe fees along to consumers in the form of discounts and other benefits as soon as reform goes into effect." Press Release, National Retail Federation Press Release, Legislation Introduced to Delay Swipe Fee Reform up to Two Years (Mar. 15, 2011), *available at* http://www.nrf.com/modules.php?name=Newsletter&op=viewlive&sp_id=323&id=51.

17 For a more extensive treatment of many of these issues, see Richard A. Epstein, *The Constitutional Paradox of the Durbin Amendment: How Monopolies are Offered Constitutional Protection Denied to Competitive Firms*, 63 Fla. L. Rev. 1307 (2011), [hereinafter Epstein, *Durbin Paradox*]; *see also* Richard A. Epstein, *The Dangerous Experiment of the Durbin Amendment*, 34 Regulation 24 (Spring 2011) [hereinafter Epstein, *Durbin Experiment*].

18 The Wall Street Journal reported the cycle as follows: as late as 1995, debit card use was minuscule; by 2000, debit transactions were still only a small fraction of credit card transactions; yet, by the end of 2008, Visa debit card volume had overtaken credit card volume by number of transactions, but not by value. In the next year, 2009, total outstanding credit card debt dropped by over $100 billion, from $957 billion to $866 billion, but debit usage continued to grow. Simultaneously, consumers cut back sharply on the use of checks. Total check volume fell five percent per year each year from 2000 to 2006. By 2005, aggregate debit card transaction value exceeded the sum of aggregate cash and check transaction value for retailers.

19 EPC, *Debit Discount*, *supra* note 14, at 4 n. 6.

20 For data, see Amended Complaint, *TCF Nat'l Bank v. Bernanke* (TCF I), No. CIV 10-4149, 2011 U.S. Dist. LEXIS 45059, at *14 (D.S.D. Apr. 25, 2011), aff'd, 643 F.3d 1158 (8th Cir. June 29, 2011); Anne Layne-Farrar, LECG, Assessing Retailers' Costs and Benefits from Accepting Debit Cards 3 (2011).

21 Brief for Retail Litigation Center, Inc. as Amicus Curiae Supporting Appellees at 1, *TCF Nat'l Bank v. Bernanke*, 643 F.3d 1158 (8th Cir. June 29, 2011) (No. 11-1805) [hereinafter RLC Brief].

22 For a more exhaustive statement, see Layne-Farrar, *supra* note 20, at 11-14.

23 *See* Layne-Farrar, *supra* note 20, at 3.

24 *See supra* note 19.

25 The only cases in which these inputs matter, and often greatly, are in connection with campaigns that address child labor or trafficking in endangered species. Such campaigns present issues far removed from those involved with debit cards.

26 Press Release, Durbin Statement on His Debit Card Swipe Fee Amendment (May 13, 2010), *available at* http://durbin.senate.gov/public/index.cfm/pressreleases?ID=506e66c9-13bd-455c-ba21-d749148b5d5e.

27 RLC brief, *supra* note 21, at 4.

28 *In re Visa Check/Mastermoney Antitrust Litig.*, 297 F. Supp. 2d 503, 506–07, (E.D.N.Y. 2003), *aff'd* sub nom. *Wal-Mart Stores, Inc. v. Visa U.S.A., Inc.*, 396 F.3d 96 (2d Cir. 2005), *cert. denied* sub nom. *Leonardo's Pizza by the Slice, Inc. v. Wal-Mart Stores, Inc.*, 544 U.S. 1044 (2005).

29 Although it is clear that these cases do not allow for antitrust liability, it does not follow that some administrative remedy beyond antitrust law is necessarily out of order.

30 RLC brief, *supra* note 21, at 10.

31 Steven C. Salop et al., Merchs. Payments Coal., Economic Analysis Of Debit Card Regulation Under Section 920 10 (2010), *available at* http://www.federalreserve.gov/newsevents/files/merchants_payment_coalition_meeting_20101102.pdf.

32 *Id*. at 3.

33 *Id.* at 6.

34 *Id*. at 3, 17, 24 and 38-42.

35 *Id*. at 3, 17, 24 and 38-42.

36 *Id*. at 14.

37 For my longer account, see Richard A. Epstein, *The Dangerous Experiment of the Durbin Amendment*, Reg.: Cato Rev. Bus. & Gov't, Spring 2011.

38 *Id*. at 1, 13 and 29.

39 *Id*. at 1.

40 William F. Baxter, *Bank Interchange of Transactional Paper: Legal and Economic Perspectives*, 26 J.L. & ECON. 541 (1983).

41 For an extensive account, see Declaration of Kevin M. Murphy at 2, *TCF Nat'l Bank v. Bernanke*, 643 F.3d 1158 (8th Cir. June 29, 2011) (concluding emphatically, "[P]roponents of debit regulation have not identified any market failure that justifies intervention, because there are none"). For further discussion, see Epstein, *Durbin Paradox*, *supra* note 17, at 17-18.

42 Visa in fact pushes these two advantages in advertising its Visa Debit Card, which "lets you shop in more places online and around the world." What is Visa Debit?, http://www.visa.ca/en/personal/visa-debit-card/index.jsp (last visited Dec. 8, 2011).

43 John Bulmer, Payment Systems: The Debit Card Market in Canada (2009), *available at* http://www.parl.gc.ca/Content/LOP/researchpublications/prb0909-e.htm. "In 2004, the typical transaction fee paid by debit cardholders to card issuers in Canada ranged from $0.50 to $0.60 per transaction; however, a number of cardholders paid neither a transaction nor a monthly fee for debit transactions."

The Canadian system, moreover, is run exclusively through an operation known as the Interac Association, which was created by its member banks, and which has its own antitrust issues. The American fees through the debit interchange system are, if anything, slightly lower. *See TCF Nat'l Bank v. Bernanke* (TCF I), No. CIV 10-4149, 2011 U.S. Dist. LEXIS 45059, at *10 (D.S.D. Apr. 25, 2011) (quoting Debit Card Interchange Fees and Routing, 75 Fed. Reg. 81,722, 81,725 (proposed Dec. 28, 2010) (to be codified at 12 C.F.R. pt. 235)); Epstein, *Durbin's Paradox*, *supra* note 17, at 1315 (reviewing how fees on a $100 debit transaction are allocated).

44 Amended Complaint at ¶ 30, *TCF Nat'l Bank v. Bernanke* (TCF I), No. CIV 10-4149, 2011 U.S. Dist. LEXIS 45059, at *14 (D.S.D. Apr. 25, 2011), *aff'd*, 643 F.3d 1158 (8th Cir. June 29, 2011).

45 For more detail, see Epstein, *Durbin's Paradox*, *supra* note 17.

46 *TCF Nat'l Bank v. Bernanke* (TCF II), 643 F.3d 1158, 1163 (8th Cir. June 29, 2011).

47 *Nebbia v. New York*, 291 U.S. 502 (1933).

48 *Block v. Hirsh*, 256 U.S. 135, 153–54 (1921).

49 *Penn Central Transportation Co. v. City of New York*, 438 U.S. 104 (1978).

50 *Yakus v. United States*, 321 U.S. 414, 422–23 (1944).

51 *Hope Natural Gas v. Federal Power Commission*, 320 U.S. 591 (1944).

52 For a useful summary of the difficulties, see *Duquesne Light Co. v. Barasch,* 488 U.S. 299 (1988). For discussion, see Michael W. McConnell, *Public Utilities' Private Rights: Paying for Failed Nuclear Power Projects*, 12(2) Regulation 35 (1988), *available at* http://www.cato.org/pubs/regulation/regv12n2/reg12n2-mcconnell.html.

53 Richard A. Posner, Natural Monopoly and Its Regulation (1999).

54 There are additional complications if the utility is not permitted to withdraw from the market at all, at which point the rates could be set so low that it can only escape bankruptcy. At this point the risk of confiscation is even greater. But it is wrong to claim, as the government did in TCF, that there is no risk from government regulation unless the government explicitly blocks the exit right. For further discussion, see Epstein, *Durbin Paradox*, *supra* note 17, at 1339–41.

55 *See Hope Natural Gas*, 320 U.S. at 600-602.

56 For further discussion, see *Duquesne Light Co v. Barasch*, 488 U.S. 299 (1988).

57 *TCF Nat'l Bank v. Bernanke* (TCF II), 643 F.3d 1158, 1164 (8th Cir. June 29, 2011).

58 *Id*. at 1162.

59 *TCF II*, 643 F.3d at 1164.

60 *TCF Nat'l Bank v. Bernanke* (TCF I), No. CIV 10-4149, 2011 U.S. Dist. LEXIS 45059, at *12–13 (D.S.D. Apr. 25, 2011) (citing *Minn. Ass'n of Health Care Facilities v. Minn. Dep't of Pub. Welfare*, 742 F.2d 442 (8th Cir. 1984)). A similar point was made on appeal, TCF II, 643 F.3d at 1164, note 2.

61 While it is not constitutionally required to fix rates that will guarantee a profit to all insurers, it may not constitutionally fix rates that are so low that if the insurers engage in business they may do so only at a loss. *See Aetna Casualty & Surety Co. v. Commissioner of Insurance*, 263 N.E.2d 698, 703 (Mass. 1970).

62 Epstein, *Durbin Paradox*, *supra* note 17, at 1345–47.

63 Anne Layne-Farrar, *supra* note 20, at ¶ 13. Her key finding is:

> *I conclude that TCF's DDA [debit deposit accounts] customers are highly price sensitive. The statistical analysis suggests that should TCF impose a monthly fee of roughly $8.33 on its DDA customers that use a debit card, it would likely see its rate of account closings rise by 58% to 81%.*

64 TCF II, 643 F.3d at 1165.

65 Dan Fitzpatrick and Robin Sidel, *BofA Retreats on Debit Fee, Citing Uproar*, Wall St. J., Nov. 1, 2011, available at http://online.wsj.com/article/SB10001424052970204528204577011813902843218.html.

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

PAYMENTS INNOVATION AND INTERCHANGE FEES REGULATION: HOW INVERTING THE MERCHANT-PAYS BUSINESS MODEL WOULD AFFECT THE EXTENT AND DIRECTION OF INNOVATION

David S. Evans

Global Economics Group, University College London, University of Chicago Law School

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

PAYMENTS INNOVATION AND INTERCHANGE FEES REGULATION: HOW INVERTING THE MERCHANT-PAYS BUSINESS MODEL WOULD AFFECT THE EXTENT AND DIRECTION OF INNOVATION

David S. Evans*

ABSTRACT

This paper examines the possible impact on innovation involving payment cards as a result of price caps that lead to a the significant drastic reduction in interchange fees. Such reductions invert the traditional business model for the payments card industry from a merchant-pays model to a consumer-pays model.

The paper argues that this inversion is likely to reduce the overall level of innovation in the industry, divert innovation away from the role of payments in transactions and towards improvements for which consumers can be charged non-transaction related fees, and discourage the entry of new payment systems.

* Chairman of Global Economics Group, Lecturer at University of Chicago Law School, and Visiting Professor at University College London. I would like thank Howard Chang, Steven Joyce, Scott Walster and Margaret Weichert for helpful comments and The Monnet Project for financial support. I retain all responsibility for opinions and errors in this paper and none of the above necessarily agrees with anything or everything in his paper.

I. INTRODUCTION

In most parts of the world, when a person pays a merchant with a card the bank that issued that card receives a payment from the acquirer that processes transactions for that merchant. These "interchange fees" have come under increasing scrutiny by governments around the world. Antitrust authorities, central bank regulators, and legislatures in various jurisdictions have imposed price caps on these fees. Usually the fees decline—sometimes by more than 80 percent—following the regulations.1

Most of the work on interchange fees has focused on static models that examine how the payment system sets the profit-maximizing interchange fee, whether the interchange fee deviates from the interchange fee that would maximize social welfare, and how to regulate prices.2 Little work has considered the relationship between interchange fees and the level and type of innovation. Yet getting innovation right is likely to be far more important than getting prices right. Innovation generates new products that provide considerable improvements in social welfare while changing prices for existing products typically leads to marginal improvements in social welfare.[3]

This topic is especially important given the recent experience of ISIS. ISIS is a joint venture of the three largest mobile operators in the United States (AT&T, T-Mobile and Verizon). It said on its formation last year that it was going to develop a mobile payments system in United States working with the Discover Network and with Barclaycard US as its first issuer.[4]

Recent reports indicate that ISIS has abandoned this plan because the sharp reductions in debit-card interchange fees proposed by the United States. Federal Reserve Board made its original business model untenable.[5] It was going to distinguish itself by having a low merchant fee model but the proposed price caps would eliminate that source of differentiation. There are similar concerns in Europe over the impact of interchange fee caps on the incentives for starting new payment schemes. Although some banks are interested in starting a new EU card scheme to challenge MasterCard and Visa Europe, it is unclear whether these schemes would be viable if the European Commission required them to adopt the same low interchange fees as MasterCard and Visa have agreed to.[6]

Any economist who opines on innovation must be humble. Innovation is an extraordinarily complex process. After years of research economists have not found that it is possible to make many definitive statements either as a matter of theory or empirical evidence. Moreover, there has been no significant work concerning innovation involving multi-sided platforms. Nor have economists conducted much research on innovation in the payments industry.[7]

The aims of this paper are correspondingly humble. The focus is on examining how the interchange fee model—and is referred to as the "merchant pays model" more generally for reasons explained below—has influenced innovation in the payments industry and conjecturing how flipping it to a consumers pay model, as a result of low price caps on interchange fees, would alter innovation. A driving observation for the analysis is that interchange fee regulation that caps these fees a low level does not simply regulate prices but inverts the business model from one in which merchants bear most of the cost of the system (a merchant-pays model) to one in which consumers do (a consumers-pay model). It is like telling ad-supported media companies such as newspaper and television networks that they have to reduce their advertising rates by 80 percent and make up the difference by charging for content.

The paper argues that the merchant-pays model has resulted in drastic innovation that has resulted in considerable benefits to merchants and consumers and has been behind significant incremental innovation as well. While it is not possible to prove that these benefits could not have come without interchange fees, or with much lower ones, one should be at least mindful of these benefits in considering a radical change to the business model that was relied on by the entrepreneurs who created these benefits.

The paper also considers how adopting a consumer-pays model would alter the direction and pace of innovation. It would go much too far to suggest that sharply reducing interchange fees would eliminate innovation. Entrepreneurs will adapt to the new regime and adjust the types of payments innovation they develop accordingly. In fact, there will likely be a flurry of innovation resulting from such radical change in business models. Nevertheless, *the amount of innovation and investment in payments could decline if there was switch to a consumer pays model*

for the simple reason that the amount of profits that payments systems can obtain from the consumer side is less than what it can obtain from the merchant side. It is simply less interesting to invest in innovation in an industry that is smaller and less profitable all else equal.

It is also likely that adopting the consumer-pays model would hinder new payment systems, such as ISIS in the United States and some of the new proposed schemes in the European Union, from starting or reaching critical mass, and shift the direction of innovation away from increasing payment card transactions and towards other types of improvements for which it is possible to charge and earn profits.

These considerations go beyond the usual concern that government regulation—and price caps in particular—deter innovation.[8]

The next section explains the merchant-pays model and describes how most payment systems have adopted this model from the beginning of the general-purpose payment card industry. Section III documents the social welfare that has resulted from the merchant-pays systems. Section IV describes how inverting the business model from merchant to consumer pays would affect the amount and direction of innovation. Section V concludes.

II. THE MERCHANT-PAYS MODEL

The merchant-pays model has been the basis for general-purpose payment card networks since these systems were first introduced in the 1950s. Before the invention of these networks consumers could pay with "store cards" that merchants issued. Consumers used those cards to identify themselves to the merchant who would put charges on a house account.

Consumers could then pay those charges off at the end of the month or finance them. Some groups of merchants developed standard identification cards that could be used at any of the merchants in that group. The merchants bore the costs of running their payment and financing programs and managing the risk associated with those activities. Many merchants did not offer payment cards, which were, at that time, largely confined to department stores.

Diners Club introduced the first general-purpose payment card in 1950 in the United States. Unlike the store cards it was possible for cardholders to use these cards to pay at any merchant that had joined the Diners Club network. Initially, Diners Club signed up restaurants but then expanded to hotels, airlines, car rentals, and other parts of what was called "travel and entertainment." The new network also quickly expanded internationally. American Express and Carte Blanche entered eight years later and also became internationally used cards primarily for travel and entertainment.[9]

These three-party[10] systems all adopted the merchant-pays model to cover the costs of operating this network and earn a profit. They charged merchants a fee—this was initially 7 percent of the transaction but declined to about 5 percent by the end of the 1950s. Cardholders did not bear much of the direct cost of these systems. They paid a modest annual fee but that roughly covered value of the float they received as a result of delaying their payments until the end of the month. Moreover, they did not have to pay any transactions fees—fees associated with using the card. As is well known, these card systems were examples of two-sided platforms that helped facilitate exchange between two groups that needed each other—in this case merchants and customers.[11]

> *Like many two-sided platforms they charged a low price to one side (the "subsidy" side) and a higher price to the other side (the "money" side).*[12]

A number of banks tried to enter the payment card business in the 1950s in the US. Bank of America introduced a credit card in 1958 in California that was particularly successful in part because it could promote this card to merchants and consumers statewide in a state with a large population. The credit card provided a personal line of credit that enabled consumers to finance their purchases. The finance charges to consumers who used it provided an additional stream of income to the issuer beyond merchant fees.

Interstate banking regulation prevented Bank of America and most banks from operating nationally while state regulation sometimes prevented them from operating even beyond a single location. These government-imposed restrictions therefore limited their ability to scale.

Banks formed two national associations in 1966 that evolved into MasterCard and Visa in response to these restrictions. Many of the members were initially banks that had their own local card programs. Like American Express, they signed up merchants and cardholders and charged both sides. As part of becoming associations, the banks agreed to allow consumers to pay with the card of any bank that belonged to the association at any merchant that had been signed up by any bank that belonged to the association. Eventually, the card associations adopted "interchange fees" to pay the bank that issued the card a fee when the card was used at a participating merchant.

The interchange fee determines in large part how much of the overall revenue (and profits) for the system come from the consumer versus the merchant side. It does this by influencing the prices merchant acquirers—the companies that sign up merchants and process merchant transactions—charge to merchants and card issuers charge to consumers for using the card.

The card association—or four-party system[13]—model was adopted around the world. In some countries MasterCard and Visa organized bank associations.[14] In many countries domestic schemes emerged which affiliated with MasterCard or Visa for the purpose of international card acceptance. Banks in these four-party systems issued credit cards, debit cards, or both. Countries quickly diverged, however, on the relative issuance of credit versus debit cards. Credit cards became the leading card type in the United States initially while debit cards became the leading card type in most of continental Europe. Debit cards started taking off in the United States in the mid 1990s and today account for 45 percent of payment card volume.[15] Credit cards have grown slowly in most other parts of the world with the exception of the Commonwealth and some of the Nordic countries.

PayPal provided another significant innovation by serving as an intermediary between consumers and merchants who wanted to transact online. Buyers provided PayPal with a means of payment (a payment card or their bank account number), which PayPal billed; sellers did the same and PayPal credited their cards or their bank accounts. Following its early acquisition by eBay, it mainly provided this service to buyers and sellers on eBay. Later it promoted its service more broadly to merchants off of eBay so that consumers could pay anyplace that took PayPal. PayPal is free to payers and it makes its money from charges to recipients of funds.

While it is not possible to obtain precise figures, it would appear most payment card systems are based on a merchant pays model in which the preponderance of the cost of the provision of payment transaction services is borne by merchants.[16] On the merchant side, almost all countries have interchange fees in which the bank that issued the card to a consumer receives a fee—often a percent of the transaction amount—from the merchant's acquirer when the consumer pays with her card.[17] Merchant acquirers pass on some or all of these fees to merchants either as a separate interchange fee assessment or as part of the overall merchant service fee. The three-party systems collect these charges directly from merchants usually. Therefore, merchants almost always pay some percent of the transaction amount. Merchants incur other costs as well to accept cards including obtaining terminals, training staff, and paying merchant processing fees on top of interchange fees. On the cardholder side, people pay little directly for using payment cards. Debit cards account for the preponderance of card transactions around the world. The bank usually provides these cards to customers as part of their checking account. Banks normally do not impose transaction fees for using these cards.[18] In some countries, credit cards account for a significant share of card transactions. Credit card customers do not pay transaction charges (and in fact sometimes receive rewards for using their cards). They do pay annual fees but the cost of these is offset in part by the free float that they receive as a result of not having to pay charges until the end of the month. About half of the people who use these cards, at least in the United States, pay off their charges in full every month and do not finance. For them the annual fee is the only cost of using credit cards. The other half finances their charges; the finance fees cover at least in part the cost of providing risky lending to customers.[19]

Payment card systems act as intermediaries between consumers and merchants. As it turns out, the merchant-pays business model appears to be common not just for payment card systems, but also for most businesses that serve as intermediaries between consumers and merchants.

The three leading examples of well-developed industries that provide intermediation services between consumers and merchants are shopping malls, e-commerce sites, and advertising-supported media.

1) Shopping mall owners usually charge merchants store rental fees and sometimes a percent of transaction volume; they usually provide consumers with free access to the malls.

2) e-Commerce sites such as amazon.com and ebay.com charge merchants fees for access to their sites and a "referral fee" or "final value fee" that are typically a percentage of the transaction price of the goods sold.

3) Advertising-supported media usually attracts viewers or listeners by providing them with valuable media content for free or for a fee that usually would not be sufficient to cover the cost of developing and delivering the content. They then sell access to these viewers to advertisers. Variants of the advertising-media model include search engines, social networking, and yellow pages.

Two recent innovative businesses that were started in the United States represent new variants of the merchant pays model.

OpenTable has a web-based platform that provides reviews and information on participating restaurants and enables consumers to make reservations at those restaurants. Consumers do not pay anything for the service. However, restaurants pay $1 per patron they get in addition to a monthly fee for reservation management software and a one-time set up fee.[20] TopTable, which OpenTable acquired in September 2010, provided similar services to restaurants in a number of European countries.[21]

Groupon helps businesses obtain traffic to their stores by providing coupons to people at heavily discounted prices for the products or services offered by the business. Groupon does not charge consumers anything for access to its discounting platform. It collects all of its revenues from merchants who pay 50 percent of the face value of the coupon as a commission to Groupon.[22] Groupon has expanded into 43 countries.[23] A number of other companies have started similar businesses in the United States or other countries.

It would appear, then, that over long periods of time and in diverse countries, payment cards have been using the merchant-pays model, and the same is true for other businesses that provide intermediation services between merchants and consumers.[24]

The merchant-pays model was also adopted by new businesses that had no market power at all. It is possible that a different pricing structure—one more balanced or tilted towards consumers—could enable the consumer-merchant intermediary businesses, including payment cards, to start, grow and sustain themselves profitably. But it would seem more likely that there is some fundamental market dynamic about the demand and costs for these businesses that has led them to structure themselves this way.

III. THE ROLE OF THE MERCHANT-PAYS MODEL IN INNOVATION

Over the last 60 years consumers and merchants have been able to participate in a number of innovative payment systems that were based on business models in which the merchant paid for most of the cost of the system. This section describes this innovation and the social welfare that they provided.

New businesses fail in part because it is very difficult to persuade customers to change their existing behavior. When a new venture succeeds there is a strong presumption that it is providing significant value to its customers. This statement is a strong version of the revealed preference theorem in economics:

the best way to determine what consumers value, and by how much, is to observe what they choose relative to the alternatives.

Over the last 60 years individuals and merchants (the customers of the two-sided payment systems) have flocked to new payments methods that they have determined provide them value.[25] The focus here is in explaining the sources of that value.

Generally there is an opportunity for the creation of a multi-sided platform when the provision of intermediation services to the different customers of the platform generates enough value to cover the cost of the platform itself as well as any subsidies that need to be paid by one side or the other.

For example, for advertising-supported media, merchants obtain enough value from advertising that the media entity can charge enough money to cover the costs of operating the platform as well as to cover the cost of the content that is used to lure consumers to come to the platform where they will, in turn, be exposed to advertisements.[26]

When Diners Club started in 1950, consumers and merchants both faced imperfections in transactions. Merchants incurred expenses from maintaining their own charge programs. They had to issue cards, manage their books, collect money, and so forth. The cards they issued were mainly relevant for repeat customers since occasional customers would probably not spend the time applying for a card and giving an occasional customer even temporary credit was likely risky. The merchant cards were also not relevant for travelers. At the same time, many merchants obviously found that, despite the availability of cash and checks for payment, it was profitable to establish a charge card program. It was presumably a valuable service to their customers and increased sales even though it must have been more costly than accepting cash or checks. Cash and checks were inconvenient in some cases for consumers. Especially in the days before ATM machines, it was inconvenient to carry cash for payment especially for occasional large purchases. Check books were more convenient but because they were not a secure method of payment for merchants not all merchants accepted them and did not accept them from all people.

Diners Club and subsequent entrants created three-party payment systems to solve these transaction problems by adopting a merchant pays model as described above. Diners Club charged a 7 percent commission on transactions to the merchant; it charged cardholders an annual fee that roughly compensated it for the cost of the float and did not charge cardholders any transaction fees. Although consumers clearly obtained value from the charge cards, Diners Club chose a strategy that did not seek to extract a significant payment for that value. Diners Club grew quickly in the United States and around the world.

Having demonstrated that there was merchant and consumer demand for a general-purpose card system that enabled multiple merchants and consumers to transact with each other, Diners Club soon faced competition from other firms, including American Express.

By the early 1960s, eighteen thousand merchants including most travel and entertainment businesses accepted cards from the three-party systems and a million consumers had and used these cards.[27]

In the United States, MasterCard and Visa were particularly important for solving another problem for merchants and consumers: the provision of credit. Before the advent of credit cards, merchants—especially large ones and ones that sold consumer durables—offered financing to their customers.[28] Often, these merchants allowed consumers to buy on an installment plan that enabled them to spread the cost of their purchases, and therefore finance them, over time. Consumers sometimes availed themselves of these plans or took out personal loans from their banks.

This, of course, was an extremely cumbersome system. The scale of lending operations was limited by the size of the merchant's customer base. Consumers faced high implicit interest charges from installment loans and had to apply separately at each store they patronized. They could obtain better rates from their banks, but securing a personal loan each time a new purchase was desired was a time consuming and inconvenient process Credit cards provided a more efficient method of financing for both merchants and cardholders. Not surprisingly, over time these programs displaced merchant lending programs including store cards and enabled consumers to avoid applying to their banks for personal loans when they wanted to make large purchases.

The four-party system itself was a major innovation. Banks had obvious skills in facilitating payments and lending money. However, no single bank had the scale in most countries to start its own card system.

By standardizing on a single brand and having interoperable cards, they made it possible to generate network effects quickly as a result of pooling merchants and cardholders and making it possible for them to transact with each other, regardless of which bank had issued their card.

The four-party system created by MasterCard and Visa provided a business model that banks around the world could imitate.

Most of these payment systems appear to have adopted an interchange model that required merchant acquirers to pay a percent of the transaction amount to the card issuer. That resulted in these four-party systems having a merchant-pays model that was similar to what the three-party systems had. These four-party systems then helped spread the use of debit and credit cards around the world.

The introduction of debit cards outside of the United States starting in the 1970s, and in the United States starting in the late 1990s, was another major innovation. In many countries, these cards helped merchants, consumers, and banks reduce the use of checks that, of course, are cumbersome on many dimensions. Data for the United States and the European Union indicates that debit cards have become the preferred non-cash method of payment for consumers. In the United States debit cards accounted for 35 percent of all non-cash transactions in 2009 and were the most commonly used non-cash payment method.[29] In Europe, cards with a debit function made up over 28 percent of all non-cash payment transactions and were second only to credit transfers in terms of the most commonly used form of payment.[30]

The merchant-pays model and the interchange-fee based four-party system model were therefore behind the development of an industry that, sixty years after its start, provides one of the leading payment methods in the world. Millions of merchants around the world have chosen to accept cards for payment and hundreds of million consumers use these cards to make purchases. The theory of revealed preference implies that merchants and consumers are obtaining value from using these cards. Otherwise, merchants would not accept these cards and consumers would not use them. There also does not seem to be any serious question about the overall value of payment cards. It is generally acknowledged that they have reduced the use of paper-based methods of payment and therefore moved society to the use of more efficient payment mechanisms.[31]

It is possible as a matter of theory that society could have gotten the benefits of these innovations if the entrepreneurs behind the payment card industry had chosen the consumer-pays model that would result with drastically lower interchange fees. That seems quite implausible though. It is hard to imagine that most entrepreneurs in the payments industry, over extended periods of time, in varying market circumstances, and in most countries, stumbled upon the wrong model to starting payments systems.

If the inverted consumer-pays model could have lead to the innovations described above, then we would have expected that more than a handful of entrepreneurs in a few countries would have adopted it.

This is not to say that the particular pricing adopted by the merchant-pays model is the socially efficient pricing that an all-knowing social planner would adopt. The two-sided markets literature has identified a variety of reasons why interchange fees, for example, could be set too high or too low relative to the socially efficient benchmark. It would be quite extreme, and inconsistent with the evidence, however, to assert that almost every payment system in almost every country over six decades is upside down in having a merchant-pays rather than a consumer-pays model.

IV. THE IMPACT OF A CONSUMER PAYS MODEL ON INNOVATION AND INVESTMENT

Competition authorities and regulators have imposed reductions in interchange fees of around 50 percent thus far. The Reserve Bank of Australia, for example, reduced the credit card interchange fee from .95 percent to .55 percent (a 42 percent reduction) during the 2000s.[32] The European Commission, in settlements with MasterCard and Visa Europe, reduced the interchange fee by about 60 percent.[33] The Federal Reserve Board originally proposed a 73 to 84 percent reduction in debit card interchange fees but ultimately reduced it by about 45 percent.[34] Some commentators in the United States and Europe have argued that interchange fees should be zero, which would largely eliminate the costs of payment cards for the merchant side of the business.[35]

Such regulation is much more radical than the price regulation that governments usually impose on public utilities or former state-owned enterprises.

Traditional regulation typically results in marginal adjustments in prices within the confines of a well-established business model. Interchange fee regulation results in an inversion of the business model. The two-sided market literature has recognized that interchange-fee regulation results in determining the "pricing structure"—the relative prices for the two sides—rather than the overall pricing level. But it has not focused on the inversion issue and the radical departure it would result in from existing ways of doing business.[36]

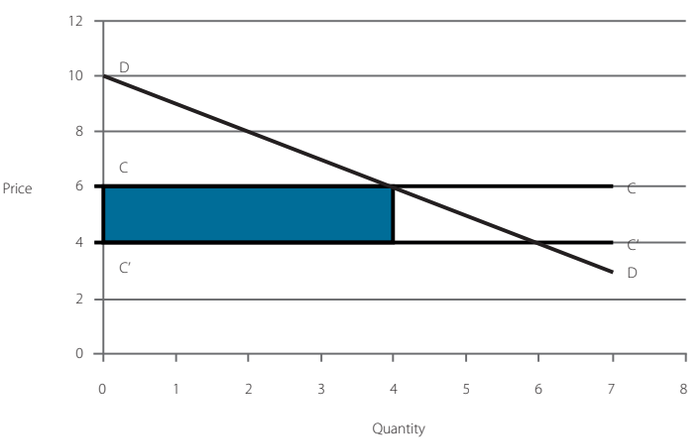
One would expect that such an inversion would have consequential results including on innovation as this section describes in more detail.

A) IMPACT ON PROFITS AND RETURN ON INVESTMENT

The theory of two-sided platforms finds that the relative prices for the two sides of the platform depend, in part, on the elasticities of demand.[37] The platform charges a higher price to the side with a more inelastic demand and a lower price to the side with a more elastic demand, all else equal. It seems plausible in the case of payment cards that consumers have a relatively elastic demand since they can use free payment methods such as cash for many transactions or other relatively low-cost substitutes such as checks. It likewise seems plausible that merchants have a relatively inelastic demand conditional on a modest fraction of customers carrying cards. The merchant stands to lose a sale—and the margin on that sale—if a consumer cannot pay or decides they do not want to pay unless they can do with their preferred method. Indeed, some of the economics literature that finds that there may be a market failure in the setting of interchange fees argues that merchants do not have any choice but to accept the card.[38]

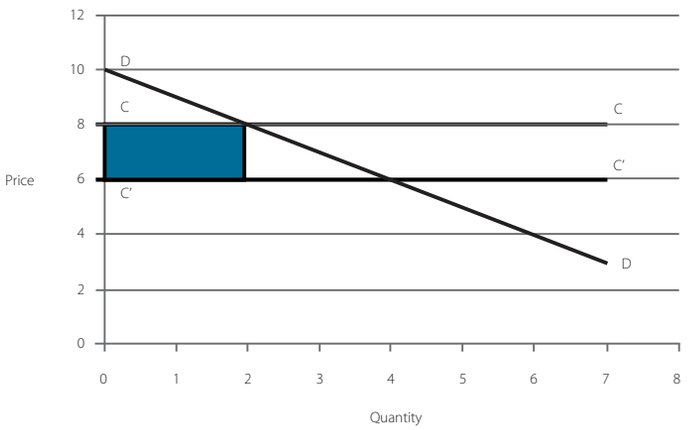
If consumers have a more elastic demand than merchants then it would not be possible for payment systems overall to earn as much revenue or profit if the price to merchants were, indirectly through interchange fee regulation, regulated to zero or a very low level. We can reasonably assume that the payments system would have been maximizing private profits before government intervention to lower interchange fees. After price caps are imposed on the merchant side of the business we would expect that there would be an attempt to increase fees to the consumer side of the business. However, since consumers have relatively elastic demand we would not expect that the payments systems overall would be able to fully replace revenue and profit after increasing prices and reducing service offering. Total profits would tend to decline since the revenue base would fall and because average profits are likely to be lower as well. The reduction in revenue and profits would tend to reduce the overall level of investment in innovation in payment card systems, and ultimately, the amount of innovation that would take place. Most economic models of investment in research and development find that the optimal investment depends on sales. For example, all else equal a business that is considering investing in process improvements will obtain greater returns if it can average the fixed costs of its research and development efforts across a larger business. An entrepreneur, and its venture backers, would, to take another example, realize a greater return if the sales and profit potential is greater. Those sales and profits would be smaller after imposing the constraint that it is not possible to earn significant revenues and profits from the side of the market with more inelastic demand. This process can be illustrated with a simple example based on a textbook model of innovation.[39] Consider a situation that is initially competitive, with a large number of issuers setting price equal to marginal cost and earning zero economic profit. Suppose one of these firms is considering investing in an innovation that would lower its costs. If it makes the investment, it will gain a temporary cost advantage over the other firms. While its advantage lasts, the innovative firm charges a price slightly below the old price (because the competitive threat of the other firms prevents it from charging any higher price), captures the entire market, and earns profits indicated by the shaded rectangle in the graph below.[40] The firm will make the investment if the net present value of these profits (taken over the expected duration of its cost advantage) is greater than the cost of the investment.

Figure 1: Incentive to Innovate - Before Cost Increase



Now suppose government regulations reduce issuers' interchange revenue, raising both the pre-innovation marginal cost and the post-innovation marginal cost (but with the same difference between the cost levels). This shifts the rectangle upward, as shown in the second graph below. Since demand slopes down, this reduces the incentive to innovate. The magnitude of the reduction is determined by the elasticity of demand. The more elastic the demand curve, the greater the reduction in the size of the rectangle.[41]

Figure 2: Incentive to Innovate - After Cost Increase



Although we can be confident that investment in innovation would decline as a result of switching from the merchant-pays to the consumer-pays model it is difficult to forecast the degree of the decline. That depends on how elastic the demand by consumers is and how clever banks, networks, and other members of the payment card systems are in raising fees for consumers and mitigating the losses from the merchant side. However, two sources of evidence should make us concerned that depressing effects of regulation on innovation could be significant.

First, empirical studies have found regulated industries tend to be relatively less innovative.

An early survey of the effects of regulation found mixed evidence of the effect of regulation on innovation.[42] Some heavily regulated industries had high productivity growth (electric power, telecommunications, airlines, and trucking), whereas others had low productivity growth (railroads, and pharmaceuticals).

One study estimated that 15 percent of the productivity slowdown of the 1970s in the United States could be explained by increased regulation.[43] More recent research has found more substantial evidence of the negative effects of regulation on productivity growth.[44] In particular, price regulation in the pharmaceutical industry has been found to deter the launch of new drugs.[45] It is difficult to be separate out cause and effect for these studies—perhaps industries that are regulated are ones that would have less innovation anyone. Nevertheless, the studies are consistent with the view that there is a negative effect of regulation on innovation.

The experience of the check-based payments system that has been subject to price regulation, for all intents and purposes, in the United States since 1914[46] provides a second source of evidence and also raises some concerns. As a result of a combination of common law and Federal Reserve Board regulation, there are significant constraints on the ability of financial institutions to charge individuals who cash checks—there is on par payment so banks have to pay the face value of the check.[47]

While there are apparently no systematic studies of innovation in the checking business, two tendencies are apparent in the United States. First, there has been a great deal of process innovation to reduce the cost of handling paper checks. This was born of necessity given the exponential growth in the use of checks over time. Second, there seems to have been little innovation that has benefited merchants or consumers. For most of the last century, there was little progress in how consumers wrote checks and managed their checkbooks; only recently have they benefited from online banking which has made it easier to use funds in a checking account. For most of the last century, there was little progress in how merchants authenticated and handled checks. Merchants today are able to use electronic capture, and some third-party check verification systems have arisen. For many consumers paying with a check at a store in the United States in 2011 would not appear to be much different than paying with a check at a store in 1911.

B) IMPACT ON STARTING A NEW SYSTEM

A price cap on interchange fees would tend to have two implications for entrepreneurs seeking to start a new four-party system.

First, for the reasons just discussed, the regulation would reduce the expected overall profitability of the new system. The system would not be able to earn as much profits under the constraint that it cannot charge the side of the market that has inelastic demand. Therefore entrepreneurs would be less motivated to start a system under these circumstances. Suppose, for example, that American Express was told in 1957 that, as a result of government regulation imposed following complaints from merchants, it was not possible to have a merchant discount of more than 50 basis points at a time when Diners Club was charging more than 500 basis points. We would expect that even if American Express recognized that Diners Club and other systems would face the same price cap, American Express would forecast a smaller revenue and profit for its business. That is because it, as well as the other systems, would have to charge the more elastic consumer side of the business. As it was, American Express almost did not survive—it tried to sell itself to Diners Club and also considered shutting down by the early 1960s—even under the merchant-pays model.[48]

Second, the price cap would interfere with the ability of the system to use the relative prices to merchants and cardholders to generate enough interest on the part of consumers and merchants to create critical mass. Putting aside the issue of how much money the system would make at maturity, most card systems appear to have started by providing incentives to consumers to get and want to use cards and then using the consumers amassed to motivate merchants to accept those cards for payment. Low prices to merchants as a result of low or zero interchange fees would increase merchant interest. But merchants would still need to incur costs to accept cards and would not do so unless the system had enough consumers. The system would therefore not have significant numbers of merchants to entice cardholders to join. Of course, the entrepreneur behind the system could seek other sources of funding for providing consumers with incentives to join. However, that could be very expensive and risky.[49]

The experience of ISIS illustrates the impact of inverting the business model from merchant to consumer pays.[50]

ISIS announced in November 2010 its intention to create a new mobile payments network that would allow consumers to pay at physical points of sale using their mobile phones. As noted earlier, ISIS was a joint venture between three mobile carriers: AT&T, Verizon and T-Mobile. ISIS also planned to use the Discover network to process transactions across its network, and Barclaycard US to issue its cards at launch. Consumer phones would have NFC-chips that would interact with merchant terminals to process these transactions, across the ISIS network.

The ISIS value proposition to consumers was the ability to transact at physical retail locations with a mobile phone and to use those phones to receive offers from merchants as inducements to shop in their stores, using cards that ran over the ISIS network. The proposition to merchants was lower acceptance fees since ISIS was planning to process transactions at a lower cost to that merchant than Visa or MasterCard was charging, presumably by using Discover's PULSE network[51] and by presumably persuading consumers to use a debit-like product. The combination of lower "swipe fees" and merchant offers was thought to be attractive enough for merchants to sign on, in spite of Discover's low market share.[52]

The ISIS business model was going to be funded in several ways: it was going to receive a commission on sales driven to merchants as a result of offers that were served to customers and from fees charged to merchants for processing payments across its network, even though those fees were said to be lower than those charged by MasterCard or Visa.

In May of 2011, ISIS abruptly announced a change in strategy, abandoning its ambition to be, in effect, the fifth payment network. It announced that it would reposition itself as a NFC-wallet, open to all issuers and networks. ISIS' spokesperson, Jaymee Johnson, stated that, "ISIS was forced to re-evaluate its strategy after financial reform legislation made it more difficult for companies like itself to make money off payment networks."[53] Johnson went on say that merchants were interested in the ISIS mobile network initially because it could deliver a mobile payments experience at a lower fee, but since Durbin was likely to so significantly reduce the fees associated with accepting cards, there was no future to the business model and the business the way it was initially conceived.

ISIS was planning to enter, therefore, by differentiating itself from existing system by charging lower merchant fees.

The government-imposed price caps largely eliminated that source of differentiation by forcing the four-party debit card systems to have low interchange fees and therefore likely low merchant fees. One could argue that ISIS provided value only because it was bypassing systems with inefficiently high interchange fees. However, by restricting competition on an important dimension government imposed price caps likely reduce the prospects for entry and differentiated-product competition.

The possible introduction of new card schemes in Europe also illustrates how low interchange fee caps could affect the decision to invest in new possibly innovative card schemes. Monnet, Payfair, and EAPS[54] have been considering starting pan-European card systems partly in response to European regulations that mandate the development of a single European payments area (SEPA). The SEPA initiatives are designed to encourage the development of an integrated European payments system. In payment cards, Europe has multiple schemes in most countries and these schemes do not interoperate well across borders. A possible result of SEPA, however, is the erosion of the domestic schemes and their replacement with cross-border schemes. That provides a business opportunity for new entry especially given that the only cross-border schemes are MasterCard and Visa.

At least two considerations come to bear on launching a new scheme. The first is the long run question of whether the new system could earn enough profits overall (which would then need to be paid to issuers, acquirers, the network and other participants) to warrant the investment and risk. To the extent that reduced interchange fees, for the reasons discussed above, reduce revenue and profits, they would likely also reduce the return on investment for a new system. The second is the shorter run question of whether it is possible for a new system to achieve the critical mass necessary for ignition.[55] This presents a practical business problem. Interchange fee setting by a pan-European system would likely be viewed by the Commission in the same way as it viewed price setting by MasterCard and Visa. If so, that would mean it would be faced possibly with a similar price cap in order to have an acceptable regime. However, in order to persuade banks that currently issue cards with domestic schemes to shift some or all of their volume to a new scheme the new scheme would, in many countries, be competing with domestic schemes that offer a higher interchange fee. It would therefore be difficult to attract cardholders and as a result hard to obtain merchant acceptance.

Part of the problem with a new scheme is that it would be required to compete with incumbent systems that have been able to use interchange fee revenues to recruit bank issuers and consumers over many decades. Even if all schemes were subject to the same price cap—zero for example—the new scheme would be at a competitive disadvantage. It would lack a major tool for getting consumers on board but at the same time would not have a better price to offer merchants.[56]

C) IMPACT ON THE DIRECTION OF INNOVATION

Although the reduced profitability of four-party payment systems would likely reduce overall innovation, there is no reason to believe that innovation would stop. In fact, the disruption in the existing business model would provide the opportunity and incentives to do things differently. However, interchange fee regulation would likely alter the direction of innovation.

Consider the following plausible scenario. Bank issuers do not impose transaction or other fees on cardholders because consumers have elastic demand; instead banks try to recover their losses through other fees related to the consumer's current account or through reduction in service. That seems like the most likely outcome in the United States.

As a result, for banks and for the system overall, not much revenue is based directly on transactions taking place. In addition, there is much less revenue coming from merchants directly. Getting an additional merchant or merchant location on board does not result in any direct increase in revenue since neither the merchant nor the cardholder would be paying transaction fees. The value only comes indirectly from increasing the value of the card brand to the consumer. In these circumstances we would expect that innovation will be directed towards products and services that can earn revenue as a result of consumers being more likely to take out a checking account, and purchasing complementary products, and possibly paying annual fees for the use of a debit or credit card. That is more or less what has happened in checking in the United States. There has been little consumer or merchant innovation surrounding checking account transactions, as noted above. The innovation has occurred in the overall checking account services provided to merchants and consumers such as online banking and online bill pay as a way to lock in consumers to those services, and ultimately the checking accounts that they underpin.

Eliminating monthly fees, being able to deposit checks at ATMS without putting them in envelopes, mobile banking and transactions alerts are just a few examples of how innovation is happening on top of checking accounts in the United States.

D) COULD LESS CARD INNOVATION BE A GOOD THING?

Of course, one might argue that this redirection of innovative effort is a good thing. At least one theory of payment cards is that they are a clever way to extract money from merchants: card systems bribe consumers to sign up and use the card and then charge merchants who do not want to lose sales from these consumers. Others have argued that payment card systems provide a subsidy to the wealthy that is paid for by a tax to the poor.[57]

Assessing the social value of payment cards versus other payment methods is beyond the scope of this paper. However, the view that we have too much use of payment cards and too much investment in payment card innovation has a couple of implications that would appear implausible on their face. The first implication is that we should have more cash and check transactions. Much of the information in the world has moved from physical to digital media in the last 15 years. We would expect that the same would be true for payments, which is information all of which can be expressed digitally. In part it has. Check use has declined in a number of countries and cash use in some. Much of the growth of electronic payments has come from the use of debit and credit cards. Debit cards are the most popular non-cash electronic payment method in the United States and the second-most popular method in Europe. Nevertheless, even in developed countries a large fraction—in many cases the majority—of consumer payments transactions are based on exchanging paper money, coins, or paper checks. It is hard to imagine that countries should have moved even more slowly from paper-based methods to electronic methods of payment than they actually have.

The second implication of objecting to the growth of debit and card cards is that given the government's reservations over the private-sector payments systems, perhaps, we should count more on the government for payments innovation.

When Diners Club was created in 1950, general-purpose payments instruments were tightly controlled by the U.S. government, which controlled the cash and coins and largely controlled the checking account system through the Federal Reserve Board. Although the Federal Reserve Board is widely credited with making an intrinsically inefficient paper-based check system more efficient, one would be hard pressed to look at the history of cash and checks—and more recently the ACH system—and argue that it has been a fountain of innovation. Looking around the world, whether it is M-Pesa in Kenya (a mobile phone based payments and banking system), PayPal's online wallet and recently introduced applications platform, DoCoMo's contactless mobile payments system in Japan, or Greendot's prepaid card products in the United States, one does not typically see governments behind payments innovation. The inexorable rise in the use of debit and credit throughout the world after the introduction of Diners Club in the United States and especially after the creation of the four-party system model, and the innovation surround those payment products, is best seen as a response to a lack of innovation by government-controlled payments systems. These private payments systems obtained traction with consumers and merchants because of the existence of transaction-cost problems that the government payment systems were not solving.

V. CONCLUSION

Consumers and merchants around the world have benefited over the last 60 years as a result payments innovation largely driven by for-profit payment card systems. There is no way to prove how much of this innovation—or alternative innovation—would have been possible under a consumer-pays model rather than the merchant-pays model that was actually used. However, given that the merchant-pays model is the one that entrepreneurs gravitated towards and that a consumer-pays model would have faced elastic demand from consumer it appears likely that society would have had considerably less innovation with the consumer pays model.

Interchange fee regulation has, or has proposed, forcing payment card systems to drop the merchant-pays model which would necessarily resulting in requiring them to flip their business models to consumers-pay. Such a radical change in business models, combined with the fact that it would impose price caps on the side of the

market with inelastic demand and require recovery of costs and profits from the side of the market with elastic demand, must have material effects including on innovation. Forecasting innovation is difficult in the best of worlds but more so in the case of two-sided markets where theory is undeveloped.

Nevertheless, the most likely scenario is that investment in payments card innovation will decline overall and will shift towards the creation of value-added services for accounts that include payment cards as a feature. As we have already seen with the decision by the U.S. joint venture of the three largest mobile carriers to drop its ambitious plans to start a new mobile-phone based payments system given the expected drop in debit-card interchange fees, it is likely that the inversion of the business model will result in the discouragement of the formation of new payment card systems, or other systems for which payments is an essential attribute.

1 The European Commission filed complaints that MasterCard and Visa violated the European Union's antitrust laws by setting interchange fees and entered into agreements with both card systems to lower those fees for the cross-country transactions as a result. *See* Matthew Dalton, *EU Says MasterCard Won't Face Antitrust Penalties*, Dow Jones Int'l News, Apr. 1, 2009, http://www.advfn.com/news_EU-Says-MasterCard-Wont-Face-Antitrust-Penalties-Over-Fees_37122940.html, and Foo Yun Chee, *EU Accepts Visa Europe Fee Cuts, Drops Probe*, Reuters, Dec. 8, 2010, http://www.reuters.com/article/2010/12/08/eu-visaeurope-idUSLDE6B70VH20101208. The U.S. Congress enacted legislation that in 2010 that requires the Federal Reserve Board to regulate debit interchange fees. See Dodd–Frank Wall Street Reform and Consumer Protection Act, Pub. L. 111-203, § 1075, 124 Stat. 1376, (2010). The Reserve Bank of Australia has imposed price caps on debit and credit card interchange fees. *See* Reserve Bank of Australia, Payment System Board Annual Report, 2004 (2004), available at http://www.rba.gov.au/publications/annual-reports/psb/2004/pdf/2004-psb-ann-report.pdf and Reserve Bank of Australia, Reform of Australia's Payments System: Issues for the 2007/08 Review ¶ 28 (May 2007), *available at* http://www.rba.gov.au/payments-system/reforms/review-card-reforms/review-0708-issues/index.html. Other countries have imposed price caps on interchange fees or have started inquiries concerning these fees.

2 For a summary, see Marianne Verdier, *Interchange Fees in Payment Card Systems: A Survey of the Literature*, 25(2) J. Econ. Surveys 273 (2011).

3 For the classic study on new products see Jerry A. Hausman, *Valuation of New Goods under Perfect and Imperfect Competition*, in The Economics of New Goods (1997).

4 *See* Troy McCombs, *AT&T, T-Mobile and Verizon Wireless Announce Joint Venture to Build National Mobile Commerce Network* (Verizon Wireless News Center) Nov. 16, 2010, *available at* http://news.vzw.com/news/2010/11/pr2010-11-16.html.

5 *See* Robin Sidel & Shayndi Raice, *Pay-by-Phone Dialed Back*, Wall St. J., May 4, 2009, *available at* http://online.wsj.com/article/SB10001424052748704740604576301482470575092.html; Maria Aspan, *Dodd-Frank Hurt Mobile Payment System Plans: AT&T*, Reuters, May 13, 2011, http://www.reuters.com/article/2011/05/13/us-summit-att-isis-idUSTRE74C61V20110513.

6 John B. Frank, *Monnet Could Challenge V/MC with Introduction of European Debit System*, Epayment News, July 10, 2009, *available at* http://epaymentnews.blogspot.com/2009/07/monnet-could-challenge-vmc-with.html#axzz1OXiNepsF.

7 For a descriptive review of what is happening in payments innovation and why, see David S. Evans & Richard Schmalensee, *Innovation in Payments*, *in* Moving Money: The Future of Consumer Payments (2009).

8 *See* Paul L. Joskow & Nancy L. Rose, *The Effects of Economic Regulation*, *in* 2 Handbook of Industrial Organization, 1449 (1982), and the extensive literature they discuss; *see also* W. Kip Viscusi, John M. Vernon & Joseph E. Harrington Jr., Economics of Regulation and Antitrust (2005).

9 Bank of America also entered in 1958 with a card program in California. Interstate banking restrictions in the United States prevented in from operating outside of California.

10 They are call three party because they involve the merchant, network, and consumer.

11 Jean-Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, 1(4) J. Eur. Econ. Ass'n 990 (2003); David S. Evans & Richard Schmalensee, Catalyst Code: The Strategies Behind the World's Most Dynamic Companies (2007).

12 This distinction between the "money side" and the "subsidy side" is used in the business strategy literature. Given joint costs and indirect network effects it is often not strictly correct to say that one side provides a "subsidy" to the other. Rather one side is more important for generating profit than the other. For evidence on the prevalence of low and zero prices though see David S. Evans, *Some Empirical Aspects of Multi-sided Platform Industries*, 2(3) Rev. Network Econ. 191 (2003).

13 They are called four-party systems because they have merchant, acquirer, issuer, and cardholder although strictly speaking they are five-party systems that include the network.

14 Some multinational card schemes also emerged, such as Eurocard which eventually was merged into MasterCard. *See MasterCard Completes Europay Merger*, Electronic Payments International (July 26, 2002).

15 The Nilson Report 948 (May 2010).

16 Two other payment systems are notable. Cash is provided by the government and is financed in effect from payers and payees from seigniorage and general tax funds. Checks are provided through a complex set of institutions and regulations at least in the United States. We discuss this more below.

17 A small number of card schemes have zero interchange fees or have systems in which the merchant acquirer is paid by the card issuer. The European Commission's first interim report listed four EU countries in which banks participated in a debit payment network without interchange fees. In all the report listed 16 major domestic debit networks. Also, the primary credit card networks in Europe operate in a non-zero interchange fee structure. *See* European Commission, Interim Report I Payment Cards (Apr. 12, 2006). Card networks in other markets such as China, Singapore, and the United States operate with a non-zero interchange fee.

18 Of course, banks could charge indirectly for debit cards as part of the overall banking relationship.

19 In the four-party systems the network also charges the acquirers and the issuers directly and the acquirers may pass on some of these costs to merchants. In the United States, debit networks collected 48 percent of network fees from acquirers and 52 percent from issuers; these fees are small relative to interchange fees, however, and therefore the interchange fees largely determine the overall cost to the merchant versus the consumer side of the business.

20 Randall Stross, *The Online Reservations That Restaurants Love to Hate*, N.Y. Times, Dec. 11, 2010, *available at* http://www.nytimes.com/2010/12/12/business/12digi.html.

21 Interestingly, OpenTable has attracted the same sort of complaints from restaurants that the payment card systems received early in their existence. *Compare* Randall Stross, *supra* note 20, with David S. Evans & Richard Schmalensee, *System Wars*, *in* Paying With Plastic: The Digital Revolution In Buying And Borrowing (2005).

22 Bari Weiss, *Groupon's $6 Billion Gambler*, Wall St. J., Dec. 20, 2010, *available at* http://online.wsj.com/article/SB10001424052748704828104576021481410635432.html.

23 Groupon, Registration of Securities (Form S-1) (June 2, 2011), *available at* http://sec.gov/Archives/edgar/data/1490281/000104746911005613/a2203913zs-1.htm.

24 I have not conducted any systematic surveys of business models for advertising, shopping malls, or e-commerce businesses around the world but my impression from the countries that I am familiar with is that it is generally the case that the merchant pays.

25 Some of the advocates of interchange fee regulation claim that merchants do not have a choice. But all merchants need to enter into contracts to accept cards and then must install equipment and train staff to take payment cards. In the United States MasterCard and Visa have lowered interchange fees to various segments that did not accept cards. As prices declined merchants changed from making the business decision of not accepting cards to accepting them. In some countries that have high merchant discounts many merchants choose not to accept cards.

26 Simon Anderson & Jean Gabszewicz, *The Media and Advertising: A Tale of Two-Sided Markets*, *in* Handbook On The Economics of Art and Culture 567 (Victor A. Ginsburgh & David Throsby eds., 2006).

27 David S. Evans & Richard Schmalensee, *More Than Money*, *in* Paying with Plastic: The Digital Revolution in Buying and Borrowing (2005).

28 Lendol Calder, Financing the American Dream: A Cultural History of Consumer Credit (2001).

29 The Federal Reserve System, The 2010 Federal Reserve Payments Study: Noncash Payment Trends in the United States: 2006 – 2009 (Apr. 5, 2011), *available at* http://www.frbservices.org/files/communications/pdf/press/2010_payments_study.pdf.

30 European Central Bank, Payment Statistics (Sept. 2010), *available at* http://sdw.ecb.europa.eu/reports.do?node=1000001440. The reported percentages excludes France for which there was no subtotals provided for the categories credit, debit, and delayed debit.

31 *See, e.g.*, William Poole, *President's Message: Checks Lose Market Share to Electronic Payments – and the Economy Gains*, The Regional Economist, Jan. 2002, *available at* www.stlouisfed.org/publications/re/articles/?id=451 ("Replacing checks with electronic payments is good for the economy; electronic payments are just plain more efficient."); Press Release, Federal Reserve Financial Services Policy Committee, Federal Reserve Study Shows More Than Three-Quarters of Noncash Payments Are Now Electronic (Dec. 8, 2010), *available at* http://www.federalreserve.gov/newsevents/press/other/20101208a.htm ("The results of the study clearly underscore this nation's efforts to move toward a more efficient electronic clearing system for all types of retail payments.").

32 Reserve Bank of Australia, Payment System Board Annual Report, 2004 (2004), *available at* http://www.rba.gov.au/publications/annual-reports/psb/2004/pdf/2004-psb-ann-report.pdf.

33 *See* Press Release, European Commission, Antitrust: Commission Makes Visa Europe's Commitments to Cut Interbank Fees for Debit Cards Legally Binding (Dec. 8, 2010), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/10/1684 and Press Release, European Commission, Commissioner Kroes Takes Note of MasterCard's Decision to Cut Cross-Border Multilateral Interchange Fees (MIFs) and to Repeal Recent Scheme Fee Increases (Apr. 1, 2009), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/09/515.

34 The debit card interchange fee was reduced from 44 cents for an average $38.58 transaction to 21 cents, plus 1 cent for fraud losses, a 5 basis points for fraud prevention efforts. This works out to 24 cents for an average $38.58 transaction. *See* Debit Card Interchange Fees and Routing, 12 C.F.R. pt. 235 (2011).

35 *See* Alan Frankel, *Towards a Competitive Card Payments Marketplace*, Reserve Bank of Australia, Payments System Review Conference, Proceedings of a Conference held in Sydney (Nov. 29, 2007). *See also* Alan Frankel & Allan Shampine, *Economic Effects of Interchange Fees,* 3 Antitrust L. J. 627 (2006).

36 An exception is Calvano who, in a submission to the Federal Reserve Board, noted that it was unlikely that a drastic reduction in interchange fees was optimal. *See* Emilio Calvano, *Note on the Economic Theory of Interchange* (submitted to the Federal Reserve Board regarding the implementation of the Durbin Amendment, Feb. 22, 2011), *available at* http://www.federalreserve.gov/SECRS/2011/March/20110308/R-1404/R-1404_030811_69122_621890579792_1.pdf.

37 E. Glen Weyl, *A Price Theory of Multi-sided Platforms*, 100(4) Am. Econ. Rev. 1642 (2010).

38 Özlem Bedre-Defolie & Emilio Calvano, *Pricing Payment Cards* (European Central Bank, Working Paper No. 1139, 2011), *available at* http://dl.dropbox.com/u/123685/Website/ppc.pdf.

39 Based on Jean Tirole, The Theory Of Industrial Organization (1988); a similar argument applies to many of the other models of innovation presented in this chapter. This particular model was chosen for simplicity.

40 This assumes that the cost reduction is not too large (called a non-drastic innovation). If the cost reduction is large enough that the post-innovation monopoly price is lower than the pre-innovation marginal cost (called a drastic innovation), then the innovating firm will charge the monopoly price. This does change the conclusion that the reduction of interchange will reduce the incentive to innovate, but does complicate the graphical presentation.

41 This discussion is based on the issuer incentive to innovation. However, the point applies more generally to the payments ecosystem. By constraining the price on the inelastic side of the ecosystem the overall prospects for revenue and profit must decline overall.

42 Paul L. Joskow & Nancy L. Rose, *The Effects of Economic Regulation*, *in* Handbook of Industrial Organization (Richard Schmalensee, ed., 1989).

43 Gregory B. Christainsen & Robert H. Haveman, *Public Regulations and the Slowdown in Productivity Growth*, 71 AM. ECON. REV. 320 (1981).

44 John W. Dawson & John J. Seater, *Federal Regulation and Aggregate Economic Growth* (Department of Economics, Appalachian State University, Working Paper No. 09-02, Dec. 2008), *available at* http://econ.appstate.edu/RePEc/pdf/wp0902.pdf; Simeon Djankov, Caralee McLiesh, & Rita Maria Ramalho, *Regulation and Growth*, 92(3) Econ. Letters 395 (2006); Giuseppe Nicoletti, Stefano Scarpetta & Philip R. Lane, *Regulation, Productivity, and Growth: OECD Evidence*, 18(36) Econ. Pol'y 9 (2003).

45 Margaret K. Kyle, *Pharmaceutical Price Controls and Entry Strategies*, 89(1) R. Econ. Stat. 88 (2007); Abdulkadir Civan & Michael T. Maloney, *The Effect of Price on Pharmaceutical R&D*, 9(1) B. E. J. Econ. Anal. & Pol'y 15 (2009).

46 Stephen Quinn & William Roberds, *The Evolution of the Check as a Means of Payment: An Historical Survey*, 93 Fed. Reserve Bank Atlanta Econ. Rev. 1 (2008).

47 Howard H. Chang, David S. Evans & Daniel D. Garcia Swartz, *An Assessment of the Reserve Bank of Australia's Interchange Fee Regulation* (2005) (unpublished manuscript, *available at* http://www.ny.frb.org/research/conference/2005/antitrust/chang_evans_garcia.pdf).

48 David S. Evans & Richard Schmalensee, *supra* note 11.

49 Some of the advocates of interchange fee regulation recognize that interchange fees may be needed early on to solve this "chicken and egg problem." *See* Steven C. Salop et al., *Economic Analysis of Debit Card Regulation Under Section 920* (submitted to the Board of Governors of the Federal Reserve System Concerning Its Rulemaking Pursuant to Section 920 of the Electronic Fund Transfer Act, Oct. 27, 2010).

50 For some background, see Gene Retske, *ISIS Changing Course? Conflicting Reports Cloud Water* (June 1, 2011), http://www.prepaid-press.com/wordpress/?page_id=4397.

51 PULSE was purchased by Discover in 2005. PULSE is one of the leading EFT/ATM networks, processing debit transactions.

52 Discover controls about 3.5 percent of all transactions, MasterCard and Visa together account for roughly 90 percent of transactions.

53 *See* Maisie Ramsay, *Isis Explains Strategy Shift*, Wireless Week (May 9, 2011), http://www.wirelessweek.com/news/2011/05/Isis-Explains-Strategy-Shift.

54 The Euro Alliance of Payment Schemes (EAPS) was formally announced in 2007 and is an international alliance of European bank and interbank networks designed to create a pan-European debit card system.

55 See David S. Evans, *How Catalysts Ignite: The Economics of Platform-Based Start-Ups*, *in* Platforms, Markets and Innovation (Annabelle Gawer ed., 2009); David S. Evans, *Launching New Payments Businesses: The Role of Critical Mass and Ignition Strategies*, *in* Ignition Series eBook (2010), http://pymnts.com/briefingroom/commentary-2/ignition-series.

56 Similar interchange fee caps on credit cards could have reduced the incentives of Discover to enter the U.S. market in the mid 1980s. One part of its strategy was to secure merchant acceptance by offering lower merchant fees than the rates being charged by MasterCard, Visa, and American Express brands. A price cap would have limited the strategies available to it for obtaining merchant acceptance in competition with the established brands.

57 Scott Schuh, Oz Shy, & Joanna Stavins, *Who Gains and Who Loses from Credit Card Payments? Theory and Calibrations* (Federal Reserve Bank of Boston, Public Policy Discussion Paper No. 10-03, Aug. 31, 2010), *available at* www.bos.frb.org/economic/ppdp/2010/ppdp1003.pdf.

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

COMPETITION AND VERTICAL INTEGRATION IN FINANCIAL EXCHANGES

Craig Pirrong

C.T. Bauer College of Business, University of Houston

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

COMPETITION AND VERTICAL INTEGRATION IN FINANCIAL EXCHANGES

Craig Pirrong*

ABSTRACT

Financial exchanges have come under increasing antitrust scrutiny of late. Competition authorities—especially those in Europe—have focused critical attention on the integration of trade execution and post-trade services in a single "silo." This hostility is predicated on a belief that integrated exchanges are immune to competitive entry. The conditions in financial trading markets do not match those that the "post-Chicago" literature has shown can make integration anti-competitive.

Moreover, the cost and demand conditions in trade execution and post-trading services make integration efficient as a means of reducing double marginalization problems and transactions costs. In particular, the liquidity network effects tend to lead to consolidation of trading on a single venue, and risk sharing considerations give rise to extensive economies of scale and scope in post-trade services like clearing.

Integration reduces the double marginalization and opportunism problems that would arise if dominant trading and post-trade venues were operated as separate firms. Liquidity network effects can be mitigated by order handling rules like RegNMS in the United States, but the issues with post-trade services are far less amenable to regulatory remediation.

Thus, the hostility to vertically integrated exchanges is misguided. Moreover, even if order handling rules that reduce market power in execution are adopted, post-trade services are likely to present chronic competitive concerns.

* New University of Lisbon and University College London. This paper is based on a presentation at a seminar on Banking Regulatory Reform and the Vickers Report, at the Jevons Institute for Competition Law and Economics at University College of London, June 8, 2011.

I. INTRODUCTION

Historically, antitrust authorities paid little attention to financial exchanges— like stock exchanges and exchanges where derivatives like futures are traded—despite the fact that they are often monopolies or near-monopolies.1 This has changed of late. In 2000, the Antitrust Division of the United States Department of Justice ("DOJ") sued options exchanges for not competing in the listing of options contracts.2 More recently, the DOJ released a letter arguing that vertical integration between exchanges and clearinghouses was anticompetitive.3The merger between Deutsche Börse and NYSE Euronext has come under antitrust scrutiny in Europe over these same vertical integration issues, as has the sale of the Toronto Stock Exchange in Canada. The European Commission's recently proposed Markets in Financial Instruments Directive ("Mifid") and Markets in Financial Instruments Regulation ("Mifir") regulations would require open access to vertically integrated clearinghouses. One exchange CEO warned that exchanges must "rethink their global strategies" due to increased antitrust scrutiny."4

One can speculate as to the reasons for this change in the antitrust posture toward exchanges, but regardless of the explanation, the shift has been profound. Moreover, as my brief sketch of developments suggests,

the focus of antitrust scrutiny has been directed primarily at vertical relationships in the financial marketplace.

In particular, competition authorities—and those in industry advocating a more aggressive competition policy towards exchanges—have expressed suspicion of vertical integration between the actual execution of stock or derivatives trades on the one hand, and post-trade services like clearing and settlement on the other. That is, vertical "silos"—exchanges like the CME Group and Deutsche Börse that operate systems for executing trades and clearinghouses—have been the main subject of antitrust scrutiny. Indeed, even exclusive contracts between exchanges and the operators of data centers providing services to exchanges have been the subject of antitrust investigation.[5] The basic concern underlying this scrutiny is that post-trade services are a natural monopoly or nearly so, but execution is competitive or potentially competitive.

By integrating into post-trade services, exchanges foreclose competition in execution and extend a post-trade monopoly into an execution monopoly.

There are reasons to suspect the validity of these concerns. The standard Chicago School "one monopoly rent" view implies that they are, in fact, invalid. There are, of course, post-Chicago theories that identify conditions under which integration or exclusive contracts can foreclose competition, but as shown in detail below, those theories are inapposite in this context.

Furthermore, vertical integration (or exclusive contracting) is an economizing response to the characteristics of both the trading and post-trade segments of the value chain.

Trading and post-trading services are highly complementary, and are consumed in near-constant proportions.

Moreover, under the laws and regulations governing securities and derivatives trading in most jurisdictions, there are strong natural monopoly elements in both trading and post-trade services. The economics of risk create strong scale and scope economies in clearing, for instance. In execution, when exchanges have no obligation to route orders to other exchanges offering better prices, network effects associated with liquidity tend to cause trading to gravitate to a single exchange that can exercise market power.

Thus, absent integration, back-to-back trading and post-trade monopolies (or near monopolies) would be the likely outcome in financial markets. This results in double marginalization problems. It also raises the potential for opportunism problems that can preclude efficient responses to market crises like a stock market crash and impede innovations that require coordinated investments in trading and post-trade functionalities. Vertical integration therefore makes economic sense because it mitigates both ex ante and ex post contracting hazards, and is likely welfare enhancing.

There are some policies that can encourage competition in the execution of transactions. In particular, the "socialization" of order flow through the creation of an open access limit order book, or by requiring competing exchanges to direct orders they receive to other exchanges offering better prices, can break the order

flow network effect that induces tipping to a single exchange. Such policies would reduce the benefits of vertical integration.

But there are no comparable policies that can mitigate or eliminate the competition-reducing effects of powerful scale and scope economies in post-trade services. Therefore,

it is likely that the coming decades will see chronic antitrust disputes involving trading services, post-trading services, or both.

II. THE U.S. POSITION

The completion of a financial transaction typically involves a variety of complementary activities.

The first function is the execution of a transaction. In exchange markets, orders to buy and sell are directed to a central marketplace, that is, the exchange. In a traditional floor-based, open outcry exchange, orders to buy or sell are represented by agents (floor brokers) on the exchange floor, or by exchange members physically present on the exchange dealing on their own account. The terms of a transaction are determined in a two-sided auction process. In newer, computerized exchanges, orders are routed electronically to a central computer that matches buy and sell orders based on priority algorithms.

Once the buyer and seller agree on the terms, a transaction must be cleared. The clearer first establishes that all terms submitted by the buyer and seller match. In most centralized markets, the clearing entity is then substituted as a principal to the transaction, becoming the buyer to the seller, and the seller to the buyer. That is, the clearer becomes the central counterparty ("CCP") that bears the risk of default by those with whom it transacts, and the original buyer and seller have no contractual obligation to one other. As a result of this "novation" process, CCPs bear the risk that one of the parties to a derivatives deal fails to perform on her obligations. CCPs attempt to protect themselves against losses from default by collecting collateral (margins) from traders. To the extent that margins are insufficient to cover a defaulter's losses, the remaining losses are shared among the CCP's members, who are usually banks or brokerage firms. Thus, CCPs mutualize default risk.

CCPs—often referred to as "clearinghouses"—engage in a variety of activities, including: calculation and collection of collateral (margin); determination of settlement obligations; determination of default; collection from defaulting parties, and; remuneration of participants in the event of a default. The CCP usually nets the obligations of those for whom it clears by determining the net amount each part owes or is owed. Since a party may owe money on some transactions, and be owed money on others,

netting typically reduces the flows of cash (and securities) between transacting parties.

As will be seen, this netting function is economically very important.

Clearers service the financial intermediaries who broker customer orders, and who sometimes trade on their own account. That is, clearinghouses serve as a central counterparty only to so-called "clearing members," and collect margins, collect and disburse variation payments, and charge fees from/to these members. They typically do not deal directly with the ultimate buyers or sellers for whom the brokerage firms serve as agents

Settlement is the process whereby parties discharge their contractual obligations to pay cash or deliver securities. At one time, settlement agents facilitated the physical delivery of stock certificates, bonds, or other delivery instruments. Today, delivery is performed by debiting or crediting the securities and cash accounts of the counterparties to transactions. This typically involves the maintenance of a central register that records ultimate ownership of securities.

A securities or derivatives transaction involves all three functions. Thus, *these functions are complementary, and the demand for each service is a derived demand.*

This has important implications for the organization of exchanges, and the role of vertical integration and exclusive contracting.

III. SCALE AND SCOPE ECONOMIES IN TRADING AND POST-TRADING SERVICES

The efficient organization of the firms providing the highly complementary execution, clearing, and settlement services depends crucially on the costs of providing them. Importantly, each function is subject to strong scale and scope economies.

The execution of transactions in securities and derivatives is subject to substantial economies of scale due to the nature of liquidity. It is typically cheaper to execute transactions in markets where large numbers of other transactors congregate. There are a variety of formal models that demonstrate that trading of financial instruments is subject to network economies that cause average trading costs to decline with the number of traders.[6] These trading costs include the bid-ask spread and the price impact of trades. The extant empirical evidence is consistent with these predictions.[7]

The crucial source of these network economies is informed trading.

Informed trading imposes adverse selection costs on those who do not possess private information. The uninformed mitigate their exposure to adverse selection by congregating on a single trading venue.

These models imply that the trading of financial instruments is "tippy" when uninformed market participants decide where to direct their orders based on expected execution costs, because in the presence of adverse selection, expected costs are decreasing in the number of uninformed traders. That is, trading activity in a particular instrument should gravitate to a single platform or exchange. With multiple exchanges, the exchange with the larger number of participants exhibits lower expected trading costs. This attracts traders from the smaller exchanges, which exacerbates the cost disparities, attracting yet more defections to the larger venue. Absent strong clientele effects, in equilibrium this process results in the survival of a single exchange.[8]

Empirical evidence is consistent with this tipping hypothesis.[9]

In practice, it is known that sometimes trading in financial instruments (notably, equities) fragments, with a given security being traded on several venues. Theoretically, however, this fragmentation is a form of "cream skimming" whereby orders submitted by those who are verifiably uninformed are executed off-exchange, while all orders that are not verifiably uninformed are submitted to a dominant exchange.[10] Off-exchange block trading mechanisms attempt to screen out the informed traders and limit participation to those who are unlikely to have private information about valuations. Trades executed away from the primary exchange typically have less information content than those executed on the primary exchange.[11] Both theory and empirical evidence suggest that trading activity that is not verifiably uninformed tips to a single venue. Put differently, price discovery is a natural monopoly.

This natural monopoly is unlikely to be contestable.

Exchanges must incur sunk costs in specific assets to enter. A traditional open outcry (floor) exchange must construct a specialized trading facility that has no use other than that for which it is designed. Moreover, floor traders invest in specific human capital that is of little use in other professions. Modern electronic exchanges create specialized trading systems involving investments in hardware and specialized software that has little to no value in other uses. In addition, the customers of electronic exchanges invest in linkages customized to a particular exchange to connect it. Thus both open outcry and computerized trading exchanges incur sunk costs, and customers incur costs to switch exchanges. Finally, to compete on liquidity in open outcry and electronic exchanges, an entrant must attract the near-simultaneous defection of a large number of traders on an incumbent exchange. Coordinating this movement is costly, and these coordination costs are sunk once incurred.[12] Sunk costs in physical trading infrastructure and human capital, switching costs, and coordination costs all impair the contestability of the trade execution venue.[13]

The foregoing analysis depends critically on the assumption that

uninformed traders choose where to trade based on expected execution costs.

This occurs when exchanges are under no obligation to direct orders to another exchange at which better prices are available, and indeed is the case in most markets around the world. In 2005, however, the U.S. Securities and Exchange Commission ("SEC") promulgated Regulation National Market System ("Reg NMS"),[14] which required an exchange to direct orders to another venue if the latter offered better prices.[15] This effectively socialized order flow, and undermined the liquidity network effect. Consistent with the theory outlined above, the NYSE had a market share of approximately 85 percent prior to Reg NMS, and accounted for virtually all of the price discovery. After Reg NMS, the NYSE's market share plunged into the 30 percent range. This reveals how the nature of competition in financial instruments turns on whether or not exchanges are under any obligation to direct orders to markets offering superior prices.[16] In the case of an obligation, order flows go to where the best price is; when there is no obligation, order flows go to where the best price is expected to be. This difference is crucial.

Clearing and settlement are also subject to strong scale and scope economies.[17] These economies arise primarily from the economics of risk bearing. Several factors are at work here.

First, recall that CCPs absorb default risk. Default risk is like an option: the best thing that can happen to the CCP is that it does not have to pay out on the default option. However, if a member firm defaults on its obligations, the amount that the CCP must pay out is positive and depends on the price of the defaulted instrument. Aggregate default losses equal the sum of these option payoffs across all customers. The average expected option payoff is declining in the number of members because the cost of an option on a portfolio (such as a portfolio of members) is smaller than the cost of a portfolio of options.[18] This is a source of scale economies.

This option-like nature of the CCP's exposure also leads to economies of scope. A CCP can net gains and losses on the different instruments in a defaulter's portfolio that it clears. These netting opportunities (diversification effects) are greater, the larger the number, and more diverse, the instruments cleared. Again, the option on the portfolio is less costly than the portfolio of options on the individual components.[19] Average clearing costs therefore tend to be lower when the risks cleared by a CCP are more diverse.

Diversification reduces costs in another way as well. CCPs collect margins to protect against default losses: the CCP can seize a defaulter's margins to cover losses. Due to diversification effects, the amount of margin required to provide a given level of protection on a diverse portfolio is smaller than the sum of the margin amounts that would be required to provide the same level of protection on the individual positions. This again reflects the ability to net gains and losses. It means that a CCP clearing a portfolio of risks can charge lower margins to achieve a given level of protection than would CCPs clearing the individual risks. Since margins are costly (as they must be met using low-yielding government securities or cash), portfolio margining reduces the costs of trading. This is another source of scope economies.

Netting provides a further source of scale economies. Some firms buy and sell the same instrument. For instance, A may sell to B, who may sell to C. Here B has both bought and sold, and in a clearing arrangement his positions can be eliminated, which also eliminates the risk that B will default. These risk-reducing multilateral netting possibilities increase with the number of traders that participate.[20]

IV. SCALE AND SCOPE ECONOMIES AND INTEGRATION

The foregoing analysis in Section III, supra, indicates that there are strong economies of scale and scope in both execution and clearing; similar economies exist for settlement as well. Indeed, these economies are so strong that execution,[21] clearing, and settlement are plausibly natural monopolies. Virtually every major derivatives contract traded around the world is traded on a single exchange. There are few examples of an entrant competing successfully against an incumbent. Indeed, the most prominent example demonstrates the power of the liquidity network effect: trading in German

government bond futures tipped from LIFFE to Eurex in a period of months.[22]

Furthermore, there are few examples of the survival of multiple clearers for a particular financial instrument, and the pursuit of scope economies in clearing has been a driving force in the consolidation of derivatives exchanges that has occurred in the 2000s. These extensive scope and scale economies would pose serious difficulties if execution, clearing, and settlement were provided by separate firms.

Avoiding the difficulties provides a motive for vertical integration of execution, clearing and settlement, or exclusive contracts between the suppliers of these services.

First, there is the potential for double marginalization. The sum of prices chosen by profit-maximizing back-to-back (or back-to-back-to-back) monopolists exceeds the price for the bundle of trading and post-trading services that an integrated monopolist would charge. The integrated monopolist's price generates both larger producer rent and larger consumer surplus than the unintegrated monopolists prices.

Double marginalization can occur even if a not-for-profit "utility" supplies clearing services to an execution venue.[23] For example, a group of banks or brokers can form a CCP that clears for an exchange. In fact, this CCP can provide clearing services for multiple exchanges, thereby permitting it to exploit greater scope economies. This "horizontal" model is epitomized by the London Clearinghouse (LCH) and LCH.Clearnet.[24] Even if this CCP is formally organized as a non-profit, it can exercise market power. In particular, it can restrict membership to a suboptimally small number of firms that supply clearing services. Even if the CCP itself does not earn a profit, its members can earn rents due to the limitation of the supply of clearing services. The scale and scope economies imply that it is possible to choose a membership that is suboptimally small, but just large enough to permit this CCP to have lower costs than any potential competitor.[25]

Execution venues can avoid this potential double marginalization problem by integrating into clearing. They can then set requirements for clearing membership based on prudential risk management criteria, thereby preventing a coalition of brokers and banks from exercising market power by limiting clearinghouse membership. Similar results can be obtained by contract. For instance, although the Board of Trade Clearing Corporation ("BOTCC"), which cleared for the Chicago Board of Trade ("CBT") from 1925 to 2008, was set up as a separate corporation, all Board of Trade members had the right to become BOTCC members. This prevented BOTCC from extracting rents from CBT members by restricting access to the clearinghouse.

Second, arm's-length contracting between an upstream clearing (or settlement) monopolist and a downstream execution monopolist can increase transactions costs. That is, whereas double marginalization from back-to-back monopoly creates ex ante contracting inefficiencies, successive monopoly can also create ex post contracting costs.

Specifically, even if the exchange, clearer, and settlement agent enter into a contract (or set of contracts) that prices each firm's services in a way that avoids multiple-marginalization and ensures that the ultimate customer of financial transaction services pays the monopoly price (which maximizes the rent to be divided between the three entities), wasteful rent-seeking and opportunism can arise. Each employs specific capital, and such capital is likely to be quite durable. These considerations lock the (putatively separate) suppliers of execution, clearing, and settlement services into long-term, trilateral relationships. Due to the long-term nature of the relationships, the parties are likely to rely on long-term contracts to govern their interactions. However, the specific assets of the clearer, exchange, and settlement firm give rise to quasi-rents, and each firm has the incentive to engage in ex post opportunism to expropriate them. As a result, even if the parties sign long-term contracts, they have an incentive to violate the contract or evade performance in order to expropriate these quasi-rents. Unpredictability in the economic environment makes complete contracts impossible, and parties can exploit this incompleteness in an attempt to profit at the expense of their contracting partners. This rent-seeking utilizes real resources.

Some specific examples are illuminating. To begin, the putatively separate clearer cannot necessarily internalize all benefits from investments to improve productivity or improve service quality because some of these benefits accrue to the monopoly supplier of execution services. If the cooperative invests in technology that reduces costs,

and this investment is non-contractible, the exchange's derived demand rises. In response, the exchange raises the price of execution, thereby capturing some of the cost reduction. This reduces, at the margin, the cooperative's incentives to invest, leading to underinvestment.

As another example, separation of trade execution and post-trade services can impede coordination. A change in a trading or clearing system, such as the addition of a new product for trading, or the offering of a new clearing or trading functionality such as straight-through processing, often requires changes to both the clearing and trading systems.

The incentives to adopt efficient changes may not be well-aligned when trade execution and post-trade services are carried out by different firms.

Similarly, sometimes there is a need to coordinate responses to market shocks or regulatory changes. Implementation of such changes requires negotiation across firm boundaries, which can provide an opportunity for hold up to extract the quasi-rents that arise from specific investments. This impairs incentives to introduce efficiency-enhancing innovations or to respond efficiently to shocks.

These coordination problems can be particularly acute during market crashes. The experience of the Hong Kong Futures Exchange ("HKFE") in the 1987 Crash is illustrative. HKFE secured some clearing services (e.g., trade matching) from ICCH (Hong Kong) Ltd., but this latter firm did not guarantee futures trades. That clearing function was performed by the Hong Kong Futures Guarantee Corporation ("FGC"). During the Crash, many brokers defaulted, and the FGC did not have adequate financial resources to cover the default losses. The exchange closed for a time, and the FGC was bailed out by the Hong Kong government and three large banks. A post-mortem determined that "the tripartite structure . . . confused lines of responsibility and effectively obstructed the development of an adequate risk-management system . . . all three agencies should have acted to contain the dangers in the expansion of the business and buildup of large positions by a few investors."[26]

Another review determined:
The clearing house [ICCH HK] was responsible for monitoring positions, but was not exposed to losses in the event of default, whereas the guarantee fund was exposed to losses but dependent on the clearing house for its risk monitoring. This meant not only that the guarantee fund was exposed if information was not effectively shared, but that traders, who were not exposed to the losses of the guarantee fund, had little incentive either to monitor the clearing house's risk management or to follow prudent trading strategies.[27]
Thus, given the successive monopoly problem driven by scale and scope economies, vertical integration (or various forms of exclusive contracts) can mitigate ex ante and ex post contracting hazards.[28] This is not to say that integration is free. Integration usually requires the use of low-powered incentives.

However, high-powered incentives can be extremely problematic for a risk sharing entity like a CCP because it can give rise to moral hazard. Moreover, integration can be expensive when there is a mismatch between the scope economies in execution and clearing (or settlement). As noted above, diversification effects create pervasive scope economies in clearing. The scope economies in execution historically have not been as pronounced. An integrated exchange that executes and clears trades on a narrow product range foregoes the clearing scope economies that could be realized by obtaining clearing services from a horizontal entity that clears for several specialty exchanges.

This model has existed, most notably in London, where the London Clearinghouse and its successor, LCH. Clearnet, cleared for several narrowly focused exchanges, like the London Metal Exchange and the London Commodity Exchange.[29] However, several exchanges that obtained clearing from LCH.Clearnet (including the LME, the Intercontinental Exchange, and EuronextLIFFE) have recently integrated into clearing, or are considering doing so. The Swiss Stock Exchange also integrated into clearing in 2007, and the London Stock Exchange has a deal to purchase LCH.Clearnet. The publicly stated rationales for these changes comports with the transaction cost rationale given above.

In particular, exchanges have stated that they can adopt new trading and clearing technologies more rapidly and efficiently when clearing and execution are performed within a single firm. The development of computerized trading has made the execution business much more technologically dynamic; prior to computerization, the

technology of trading had remained nearly static for well over a century. This technological dynamism has increased the need to coordinate the development of trading and post-trade systems, which the foregoing analysis implies should lead to more integration. The movement towards integration by even narrowly-focused exchanges suggests that this is indeed the case, and that transactions cost-related efficiencies now outweigh the loss of diversification-driven scope economies in clearing.

The shift in execution technology from face-to-face auctions on trading floors to computerized trading systems has increased scope economies in execution.

Traders around the world can use a computerized system like the CME Group's GLOBEX II to trade a dizzying array of products. The system is scalable because the same algorithms and software can be used to trade any product. The technology-driven expansion of scope economies in execution has driven the consolidation of the derivatives exchange industry into two huge exchanges, CME Group (which purchased the Chicago Board of Trade in 2008 and the New York Mercantile Exchange in 2009) and Deutsche Börse-EuronextLIFFE-NYSE (which also trades stocks). These groups can exploit scope economies in both trading and execution. Significantly, however, they do not compete head-to-head in any major product: each group has a near-monopoly on the products it trades.

In sum, vertical integration between trading and post-trade services can reduce costs arising from market power (double-marginalization) and transactions costs (from ex post opportunism and coordination problems). Moreover, the computerization of trading has made the execution business much more technologically dynamic, which has increased the benefits of integration. These technological developments have led to a closer match between scope economies between trading and post-trade services, which has reduced the opportunity cost of integration, and led to the formation of large, vertically integrated global exchanges.

This analysis provides an efficiency-based explanation for vertical integration between trading and post-trade services. It can also explain some of the changes in organization observed over the last decade, in particular, the move toward integration even by narrowly specialized exchanges.

The alternative view, which motivates much of the skepticism of integration among antitrust authorities in Europe and the US, is that integration is instead anticompetitive, and driven by a desire to extend monopoly. The next section evaluates the plausibility of this view.

V. THE PLAUSIBILITY OF MONOPOLY LEVERAGING THROUGH INTEGRATION

The efficiency explanation for integration hinges on the claim that both execution and post-trade services are natural monopolies, or nearly so. The alternative view agrees that clearing is a natural monopoly, but is predicated on the belief that execution is potentially competitive. In this view, an operator of a clearing monopoly can thwart competition in execution by creating a vertical silo, and providing clearing services exclusively to its integrated execution arm. The clearing monopolist can thereby leverage his market power into execution, which would otherwise be competitive.

As Sam Peltzman notes, and as Aaron Director argued well over a half-century ago, this fear of leveraging one monopoly into two is commonsensical, but more often than not, wrong.[30] The essence of the Chicago critique is that the monopolist (in this case, the operator of the clearing service) can extract all of the monopoly rent by choosing the monopoly price for his product. Keeping out potentially more efficient suppliers of complementary services (execution, in this instance) merely reduces the profit the monopolist could extract. The monopolist wants complements sold for the lowest price possible, in order to push out the demand curve for the monopoly good as far as possible. Thus, keeping out a more efficient supplier of the complementary good, or reducing competition in the sale of the complementary good, is counterproductive.[31]

Chicagoans starting with Director explained vertical restrictions as a form of price discrimination (which has ambiguous welfare consequences); a means of addressing free rider problems[32]; or as a way to eliminate double-marginalization problems. Transaction costs economists devised other efficiency-related explanations for vertical integration. Yet the suspicion of vertical integration, ties and exclusive dealing, and other vertical restraints lives on, as exhibited by the fighting over "silos" in the exchange space.

Post-Chicago, there have been several attempts to produce models which lead to anti-Chicago implications, i.e., to show that monopoly leveraging is possible. An examination of these models shows that they do not apply to the facts of the exchange case.

The most prominent post-Chicago leveraging model is by Michael Whinston.[33] In his model, there is a monopoly good, M. Some customers want to consume that good along with another good, C, that could be produced by competitive firms. But some customers don't want to buy M; they wish to consume C alone. The M monopolist may want to tie or vertically integrate into C (and not sell to other producers of C) if entry into C production requires payment of a fixed cost. By tying/integrating, those who want to buy M have to buy C from the M producer, too. Thus, potential entrants into the C market can sell only to those who want to buy C alone. If there are too few of those customers, or if fixed costs are too high, it will be unprofitable to enter into the production of C. Then the monopolist can sell C to the stand-alone customers at a monopoly price.

This model clearly does not fit the facts in the clearing-execution case.

Those products are highly complementary. Indeed, they are consumed in nearly fixed proportions—if you want to trade, you need to clear, and if you clear, you need to trade. The whole point of the Whinston model is monopolization of a product some customers do not find complementary to M. The monopolist uses his power over the customers who have strong complementarity to gain a monopoly over customers who do not experience any complementarity with M. This is clearly at odds with the assertions of those who assert that clearing monopolies use their power to achieve execution monopolies, because those assertions rely heavily on the notion that clearing is an essential service—i.e., highly complementary to execution, and a service that all traders consume. That is completely at odds with the Whinston story, so it is of no help to the silo opponents.

Dennis Carlton & Michael Waldman have an interesting model that embeds complementarity,[34] but arrives at similar conclusions to Whinston's model. Yet whereas Whinston argues that ties/integration can be used to extend a monopoly to a non-complementary good, Carlton & Waldman devise a two-period model in which a monopolist ties a complementary good to protect his M monopoly. A firm has a monopoly over M. It is guaranteed this monopoly for one period, but in the second period, a competitor can enter. The M monopolist can also produce a good C, and a firm that can enter the M market in the second period can produce C in the first period.

In one model, the rival incurs a fixed cost to enter the C market. By tying the complementary good in the first period, the M monopolist deprives the entrant of any sales in the first period. The profits from producing C and M in the second period may not be sufficient to cover entry costs, meaning that with the tie, entry may not occur in either market, thereby preserving the M monopoly. In contrast, without a tie, the entrant can produce C in the first period and make a profit that contributes towards covering fixed costs: he can make a profit because his C good is superior to that of the monopoly producer of M. The profit from entering C production in the first period may cover fixed costs of entering the C market. Then, in the second period, it may be profitable to enter the M market as well. In this case, tying protects the M monopoly.

In the second model, there is customer lock-in due to network effects. By tying in the first period, the monopolist of M locks in many consumers of C, and deprives the entrant of any sales in the first period. The customer lock-in reduces the profitability of entry into M and C production in the second period, likely by enough to make such entry unprofitable. Again, the tie protects the M monopoly.

These models work best to explain ties in highly technologically dynamic industries where monopolies are likely to be short-lived in any event.

Such a description does not fit the exchange-clearing case. Moreover, there is no legal or economic bar on entry into clearing and execution simultaneously, and the necessity of sequential entry is the key driver of the Carlton-Waldman results. Indeed, integrated exchanges have entered in competition with incumbents, and execution platforms have secured clearing services by contract, so simultaneous entry has occurred.

A third type of model relies on contracting externalities to explain how exclusive dealing and integration can

impair competition. One example of this is a model by Oliver Hart & Jean Tirole.[35] In the Hart-Tirole model, an upstream monopolist can sell to multiple downstream retailers (in the exchange case, the upstream firm would be the clearing monopoly, and the retailers execution venues).

The upstream monopolist in the Hart-Tirole model negotiates with the downstream firms individually and secretly. In a key assumption, the firms negotiate over output—the quantity sold. Hart & Tirole show that under these conditions, the monopolist cannot credibly commit to selling the monopoly output Qm. By way of illustration, if he sells .5Qm to one firm, he has an incentive to sell more than .5Qm to the other: he cannot credibly commit to selling .5Qm to the second firm once he has sold that amount to the first firm. Total output exceeds the monopoly output and the monopolist's profit is smaller. Indeed, he can only achieve the Cournot duopoly profit. If he sells to N retailers, he can get only the N-firm Cournot profit.

By integrating, or selling to only a single retailer, the monopolist effectively commits to the monopoly output. This may come at a cost. There may be diseconomies of scale in retailing, or retailers may be differentiated and service different customer clienteles. But the gains from eliminating the commitment problem may exceed the costs arising from diseconomies of scale or underproduction of variety/customization.
The monopolist obviously has incentives to avoid the commitment problem that drives the exclusionary result.

He could charge the monopoly price, post that price publicly, and let the downstream firms buy as much as they want—which would be .5Qm. This would require the avoidance of secret price discounts. Reputation may ensure this in a repeated game. The retailers could monitor competitors' sales to see if the monopolist were cheating.

Moreover, this doesn't seem to match up well with the mechanics of the exchange case. "Output" is not the choice variable; prices are. And trading volumes are readily observable, making it possible to detect whether a clearing monopolist were offering secret price cuts.

A similar model is one in which a downstream monopolist buys from two upstream suppliers who compete in an input market in which the supply curve for the input slopes up. Similar commitment problems preclude achievement of the monopsony outcome in the input market. This model has the same choice variable problem as the Hart-Tirole model, and furthermore, it is difficult to imagine what the relevant input with the upward-sloping supply would be—computer programmers, or, servers? Again, the model is inapposite to the exchange case.

Another model of anti-competitive integration is by Janusz Ordover, Garth Saloner, & Steven Salop.[36] In that model, two downstream firms D1 and D2 compete, as do two upstream firms U1 and U2. If D1 and U1 integrate, and the integrated firm refuses to sell to D2, D2 now has to buy an input from a monopoly supplier U1. D2 pays a higher price for the input, making it a less formidable competitor for the integrated firm who therefore becomes more profitable.

This model is quite fragile. What's more, an example posed in a related paper by Michael Riordan & Steve Salop makes it seem nearly trivial.[37] Their example of how the Ordover-Saloner-Salop story could work is that the purchase of Autolite—a spark plug maker—by Ford could raise the price of spark plugs to GM and Chrysler, thereby allowing Ford to raise the price of cars. Richard Posner dismisses the applicability of this theory by pointing out the complete absence of credible examples.[38]

Finally, exchange silos do not add to the (non-existent) stock of credible examples. The premise behind criticism of integration between clearing and execution is that clearing is a natural monopoly. But the Ordover-Saloner-Salop model relies heavily on integration reducing competition upstream (i.e., in clearing). That cannot happen if clearing is already a monopoly. Ordover-Saloner-Salop is not a theory of monopoly leveraging.

VI. NATURAL EXPERIMENTS HELP DISCRIMINATE BETWEEN EXPLANATIONS

There is a powerful natural experiment that makes it possible to test the back-to-back monopoly hypothesis against the monopoly-leveraging alternative. Prior to 1973, each U.S. exchange had its own clearing operation. Then the paperwork crisis of the 1960s led to the creation of an industry settlement utility, the Depository Trust Corporation ("DTC"), and an industry clearing utility, the National Securities Clearing Corporation ("NSC"), in 1973.

The two facilities were combined in 1999 to form the Depository Trust Clearing Corporation ("DTCC"). DTCC (and its predecessors) operates as a not-for-profit, member-governed utility that provides services to members at cost.

Under the monopoly leveraging theory of integration, the formation of horizontal, open access CCP and settlement entities should have led to entry of new exchanges providing execution services, and a decline in the market share of the dominant NYSE. Under the back-to-back theory, the NYSE's large market share reflected liquidity network effects, and the change to a horizontal structure should have had no effect on its market share.

In fact, after the formation of NSC and DTC, NYSE remained the dominant exchange in the United States. Until 2006, its market share of the shares it listed was approximately 85 percent, and even this understates is dominance of price-discovery (the implication of the liquidity network theory). Most non-NYSE trades of NYSE-listed shares were executed under various sorts of screening/preferencing arrangements that skimmed verifiably uninformed orders. The liquidity network theory implies that this is the only kind of orders that satellite execution venues can attract.[39] Thus,

the result of the natural experiment of the creation of the DTC and NSCC supports the liquidity network theory

which implies that clearing and execution should be back-to-back monopolies—and is inconsistent with the monopoly leveraging theory.

A subsequent natural experiment bolsters the point. In 2005, the SEC issued Reg NMS. This regulation dramatically tightened the obligation of an exchange to route orders sent to it to other markets displaying better prices.[40] Prior to Reg NMS, orders would be sent to the market where market participants expected to get the best price, which was typically the biggest market: this created the self-reinforcing liquidity network effect.

After Reg NMS, orders were directed to the market actually posting the best price. This broke the network liquidity effect. Within months, the market share of the NYSE plunged, and upwards of 65 percent of trades in NYSE-listed stocks are now executed on other exchanges.

The two natural experiments support the view that absent some rule like Reg NMS, back-to-back monopoly between execution and post-trade services is the most likely outcome. Moreover, it contradicts the claim that preventing integration is sufficient to achieve vigorous competition between execution venues, thereby undercutting the monopoly leveraging view of exchange silos.

VII. EXCLUSIVITY PUZZLES

Not only do vertically integrated exchanges combine trading and clearing (and sometimes settlement, where relevant), they also typically are exclusive. For instance, integrated exchanges typically refuse to clear for execution venues they do not own.

This exclusivity is not immediately consistent with the one monopoly rent view, which would predict that absent some other cost, a putative clearing monopolist would be willing to sell at the monopoly price to all comers in order to maximize profit; turning away potential customers to favor an affiliate is not profit maximizing. Although some of the models just discussed can explain exclusivity, and, as in the Hart-Tirole model, turning away business from some potential customers, these models are not plausible for the reasons shown above.

There are plausible reasons why dealing with multiple execution venues, some not owned by the clearing firm, creates costs that can be avoided through exclusivity.

Most notable of these costs are those arising from integrating trading and post-trade systems,[41] and coordinating changes and innovations across firm boundaries. Relatedly, there are potential spillovers between the execution venue and the clearer. For instance, a system failure or programming error can cause a problem at the execution venue that disrupts the clearer's operations. The clearer's ability to influence the likelihood of such an event is more limited across firm boundaries than inside them, and charging the execution venue a price that reflects the potential spillover cost it imposes on the clearer is greatly impeded by the difficulty of obtaining information about the technology and operations of a separate firm, especially inasmuch as that information is likely to be highly sensitive. Ex post "pricing" through legal liability is expensive, and many actions are almost certainly

non-contractible due to the difficulty of courts in adjudicating disputes involving the operations of technologically complex firms.

The issue of "open access" to clearing facilities, a regulatory response to exclusivity,[42] raises another complication. In Europe particularly, this is viewed as facilitating competition not just in execution, but in clearing as well. Under open access, clearer C1 would have to provide clearing services to execution venue E even if E were a separate firm. But as envisioned by some European regulators and legislators, there would be two or more clearinghouses. Under open access, E could demand access not just to C1, but to another clearer (if one were to enter), C2. If a buyer and a seller who execute on E can choose individually where to clear their sides of a trade (as would likely be necessary in an anonymous market), the buyer might choose C1, and the seller C2. This would create a contract between C1 and C2: the clearinghouses would have to interoperate.

Interoperability is highly problematic, not least because a CCP is highly reluctant to take on risk exposure from another CCP due to its inability to monitor effectively the other's risk management. Interoperability also increases collateral costs because CCPs are almost certain to require collateral on inter-CCP exposures, meaning that whereas with a single clearer only the buyer and seller post collateral, now each CCP must as well. In addition, it will be necessary to coordinate systems and interfaces across independent clearers, a process rife with potential for opportunism and coordination failures. Lastly, interoperability raises difficult competition issues because ostensible competitors need to contract with one another, and price the services and risks they exchange.

Open access also raises the issues of "at what price?" and "on what terms?" Open access is likely to trigger efforts to regulate the prices and terms of service of the dominant clearer (or clearers). This is the rule in network industries, and the rule is likely to apply in clearing.

Banning exclusivity by mandating open access is therefore highly dubious policy, predicated on a faulty understanding of the economics of clearing and execution. The scale and scope economies discussed throughout make it improbable that multiple CCPs clearing a particular product will survive in equilibrium. In this case, open access will likely result in excessive transactions costs associated with coordinating and integrating clearing and execution functions across firm boundaries. If, alternatively, multiple clearers do survive, interoperability creates costs and risks.

VIII. THE FUTURE OF ANTITRUST AND FINANCIAL EXCHANGES

The natural experiments, plus the analysis above, cast serious doubt on the monopoly leveraging theory, and hence on antitrust authorities' suspicions of integrated exchanges. Integration is far more plausibly an economizing response to liquidity-driven scale economies in execution, and risk-driven scale and scope economies in clearing, than an anticompetitive attempt to exercise market power.

This means that vertical silos should not be a major antitrust concern. But it does not imply that competition issues will be absent in markets for stocks and derivatives in the years to come. Indeed, the strong scale and scope economies will likely continue to ensure that market power and monopoly or near-monopoly will be the rule for financial exchanges in years to come. Competition policy involving financial trading and clearing is difficult primarily because the fundamental cost and demand conditions are not conducive to the survival of even a handful of highly rival firms.

There are policies that can reduce some sources of market power in financial markets. The risk-driven scale and scope economies are inherent in the nature of clearing, and not amenable to policy intervention. As the Reg NMS experience demonstrates, however, it is possible to increase competition in execution through order handling rules. Yet it must be recognized that these rules would face tremendous political opposition, especially in derivatives markets because of the political power of major exchanges such as the CME Group.

If that is done, regulatory policy will need to focus on clearing and settlement, as rigorous competition between CCPs or settlement agents is unlikely due to the oft-mentioned scale and scope economies. Here, a utility-type model along the lines of DTCC would have some advantages, although (a) access/membership standards would still have to be determined, and (b) this model would likely raise the costs innovation due to the difficulties of coordinating between the clearing (or

settlement) utility and execution venues, especially inasmuch as execution venues would attempt to gain competitive advantages by influencing the utility.

Regardless, the historical indifference of competition authorities to the organized trading of financial instruments will not continue in the future. The fundamental characteristics of trading and post-trading make market power an inherent condition in this industry. Some policy prescriptions—such as unbundling execution and post-trade services, or mandating open access to post-trade services—are defective because they ignore these fundamental characteristics. Vertical integration is a response to scale and scope economies and market power, rather than a cause of market power. Some sources of scale economies and market power, most notably the network effect, are amenable to policy changes. Others, particulary those in post-trade services, are not.

Given the extreme complementarity between trading and post-trade services, moreover, policymaking must deal with both simultaneously in a coordinated fashion.

Going forward, competition policy in organized financial markets is likely to resemble that in telecommunications markets, a discouraging prospect indeed. But as in telecommunications, fundamental technological considerations defy easy fixes to improve competition.

It is therefore essential that antitrust and competition policymakers dramatically improve their understanding of these fundamental considerations. Scholarship in finance, particularly market microstructure, has insights that are essential for competition policy in financial markets, but this scholarship is terra incognita for most antitrust and industrial organization scholars and policymakers. Similarly, scholarship in industrial organization sheds light on crucial issues in financial markets, but it has had only limited impact on finance scholars and financial regulators. Devising sensible competition policies will require an integration between these different and largely distinct branches of economics.

1 There are exceptions of course, such as the famous case *Chicago Board of Trade vs. United States*, 246 U.S. 231 (1918), which established the Rule of Reason principle.

2 *United States v. American Stock Exchange, LLC*, No. 00CV02174 (D.C. Cir. Dec. 6, 2000).

3 U.S. Department of Justice, Comments to the Review of the Regulatory Structure Associated with Financial Institutions, 72 Fed. Reg. 58,939 (Oct. 17, 2007), Jan. 31, 2008, *available at* http://www.justice.gov/atr/public/comments/229911.htm

4 Jeremy Grant, *Exchanges "Need Rethink" Over Failed Tie-Ups*, Fin. Times, July 17, 2011, *available at* http://www.ft.com/intl/cms/s/0/2f97b0fe-b07e-11e0-a5a7-00144feab49a.html.

5 Janina Pfalzer & Adam Ewing, *NASDAQ-OMX Raided by Swedish Regulator Over Data-Center Probe*, Bloomberg News, July 14, 2011, *available at* http://www.bloomberg.com/news/2011-07-14/nasdaq-omx-raided-by-swedish-regulator-over-data-center-probe.html.

6 *See* Craig Pirrong, *The Self-Regulation of Commodity Exchanges: The Case of Market Manipulation*, 38 J.L. & Econ. 141 (1995); Marco Pagano, *Trading Volume and Asset Liquidity*, 104 Q. J. Econ 255 (1989); Anat Admati & Paul Pfleiderer, *A Theory of Intraday Patterns: Volume and Price Variability*, 1 Rev. Fin. Stud. 3 (1988).

7 Craig Pirrong, *The Industrial Organization of Trading, Clearing, and Settlement in Financial Markets* (2010) (unpublished manuscript, on file with author).

8 Cross-border trading restrictions are one potential source of clientele effects. Geographic proximity gave rise to clienteles prior to the advent of telegraphic and telephonic communications.

9 *See* Craig Pirrong, *Securities Market Macrostructure: Property Rights and the Efficiency of Securities Trading*, 18 J.L. Econ. & Org. 385 (October 2002) [hereinafter Pirrong, *Securities Market Macrostructure*]; Craig Pirrong, *The Industrial Organization of Financial Markets: Theory and Evidence*, 2 J. Fin. Markets 329 (1999) [hereinafter Pirrong, *The Industrial Organization of Financial Markets*].

10 Pirrong, *Securities Market Macrostructure*, *supra* note 9.

11 *See* Hendrik Bessembinder & Harold Kaufman, *A Cross-Exchange examination of Trading Costs and Information Flow for NYSE-Listed Stocks*, 46 J. Fin. Econ. 293 (1997); David Easley, Nicholas Kiefer, & Maureen O'Hara, *Cream Skimming or Profit-Sharing? The Curious Role of Purchased Order Flow*, 51 J. Fin. 811 (1996); Bruce Smith, Alasdair Turnbull, & Robert White, *Upstairs Market for Principal and Agency Trades: Analysis of Adverse Information and Price Effects*, 56 J. Fin. 1723 (2001).

12 Pirrong, *supra* note 6; Pirrong, *Securities Market Macrostructure*, *supra* note 9.

13 It is well known, moreover, that contestability requires some constraint on the incumbent's ability to cut prices in response to entry. *See* Jean Tirole, The Theory of Industrial Organization (MIT Press 1988). However, in the exchange market, the incumbent can cut prices in response to entry. When Eurex attempted to enter the market for U.S. Treasury futures in competition with the Chicago Board of Trade ("CBT"), the CBT cut trading fees dramatically, only to raise them again when Eurex gave up its attempt. One of the few examples of an exchange wresting a market from an incumbent illustrates this point. The London International Financial Futures and Options Exchange ("LIFFE") did not cut fees in response to a price cut by Eurex. This caused the volume in trading of German government bond futures to tip from LIFFE to Eurex in a period of months.

14 Regulation NMS, 17 C.F.R. pt. 242 (2005).

15 For instance, if a customer sends a buy order to the NYSE when the lowest sell price posted for that stock there is 10, but there is an order to sell at another exchange at 9.95, the NYSE is obligated to route the order to the exchange posting the better price.

16 Under Reg NMS, the SEC opted for a "information and linkages" approach rather than mandating a single, open access, central limit order book ("CLOB"). One weakness of this approach is that linkages can break down, particularly during periods of market stress like those observed during the "Flash Crash" of 2010.

17 In the interest of space, I will focus only on clearing.

18 *See* Robert C. Merton, *Theory of Rational Option Pricing*, 4(1) Bell J. Econ. Manag. Sci. 141 (1973); Craig Pirrong, *Rocket Science, Default Risk, and the Organization of Derivatives Markets* (2008) (unpublished manuscript, on file with author).

19 This was important in one major recent default. The CME Group suffered no loss as a result of the default of Lehman Brothers because of this netting/diversification effect.

20 Darrell Duffie & Haoxiang Zhu, *Does a Central Clearing Counterparty Reduce Counterparty Risk?* (Rock Center for Corporate Governance at Stanford University Working Paper No. 46, Stanford University Graduate School of Business Research Paper No. 2022, 2009). It is not strictly correct to say that netting reduces risk. Although it does reduce default losses among derivatives counterparties, it actually shifts that risk to other claimants on a bankrupt firm.

21 Pirrong, *The Industrial Organization of Financial Markets*, *supra* note 9.

22 *See* Craig Pirrong, *Bund for Glory: Or, It's a Long Way to Tip a Market* (2009) (unpublished manuscript, on file with author). Although stock trading has always been fragmented to some degree, and this fragmentation has increased in recent years, this does not contradict the natural monopoly argument. In particular, fragmentation in the form of "third markets," "dark pools," internalization, and block markets, largely reflects efforts of verifiably uninformed traders to reduce their execution costs. In a traditional anonymous exchange market, informed and uninformed traders cannot be distinguished. Informed traders impose costs on market makers, but since the latter cannot distinguish informed from uninformed traders in an anonymous market, even uninformed traders pay a cost related to the market makers' adverse selection problem. This provides an incentive to devise trading mechanisms that screen out informed traders. Dark pools and third markets represent various ways of separating the informed from the uninformed, thereby reducing the adverse selection costs imposed on the latter. Since less informed trading takes place in dark pools and the like, trades there are less informative than trades on an anonymous market where the bulk of informed trading occurs. "Tipping" minimizes the adverse selection costs incurred by uninformed traders who cannot avail themselves of the alternative venues that screen out the informed because the cost of signaling or screening is too high for these traders. Thus, fragmentation reflects segmentation by information, and "price discovery" is a natural monopoly. *See generally* Pirrong, *Securities Market Macrostructure*, *supra* note 9.

23 Cooperatives are sometimes employed to mitigate market power. *See* Henry Hansmann, The Ownership of Enterprise (Harvard University Press 1996); Tomas Philipson & Richard A. Posner, *Antitrust and the Not-For-Profit Sector* (Nat'l Bureau of Econ. Research, Working Paper No. 8126, 2001).

24 Peter Norman, The Risk Controllers (Wiley 2011).

25 See Pirrong, *Securities Market Macrostructure*, *supra* note 9; Pirrong, *supra* note 6. Historically, non-profit exchanges exercised market power, and generated rents for their members, by restricting membership. See Pirrong, *The Industrial Organization of Financial Markets*, *supra* note 9.

26 Securities Review Committee, The Operation and Regulation of the Hong Kong Securities Industry (1988).

27 Bob Hills et al., *Central Counterparty Clearing Houses and Financial Stability*, 6 Fin. Stab. Rev. 122 (June 1999).

28 For a general treatment of these issues, see Dennis Carlton & Jeffrey Perloff, Modern Industrial Organization (Addison Wesley 2004).

29 LCH.Clearnet also offers clearing for over-the-counter derivatives and repo transactions, thereby exploiting scope economies across a broader variety of contracts.

30 Sam Peltzman, *Aaron Director's Influence on Antitrust Policy*, *in* Pioneers of Law and Economics (Lloyd Cohen & Joshua Wright eds., Edward Elgar 2011).

31 *See* Einer Elhauge, *Tying, Bundled Discounts, and the Death of the Single Monopoly Profit Theory*, 123 Harv. L. Rev. 397 (2009) (declaring that the Single Monopoly Profit theory is dead). *But see* Paul Seabright, *The Undead? A Comment on Professor Elhauge's Paper*, 5 Comp. Pol'y Int'l 243 (November 2009) (demonstrating that Elhauge's report of the theory's death are greatly exaggerated).

32 *See, e.g.*, Lester Telser, *Why Should Manufacturers Want Fair Trade*, 3 J.L. & Econ. 86 (1960) (providing a model of resale price maintenance).

33 Michael Whinston, Lectures on Antitrust Economics (MIT Press 2006).

34 Dennis Carlton & Michael Waldman, *The Strategic Use of Tying to Preserve Market Power in Evolving Industries*, 33 Rand J. Econ. 194 (2002).

35 Oliver Hart & Jean Tirole, *Vertical Integration and Market Foreclosure*, in Brookings Papers on Economic Activity, Microeconomics 205-286 (Clifford Winston &Martin N. Baily, eds., Brookings Institution Press 1990).

36 Janusz Ordover, Garth Saloner, & Steven Salop, *Equilibrium Vertical Foreclosure*, 80 Am. Econ. Rev. 1263 (1990).

37 Michael Riordan & Steven Salop, *Evaluating Vertical Mergers: A Post-Chicago Approach*, 63 Antitrust L.J. 513 (1995).

38 Richard Posner, Antitrust Law 226 (University of Chicago Press 2001).

39 Pirrong, *Securities Market Macrostructure*, *supra* note 9.

40 *See also* discussion *supra* Section III.

41 *See id*.

42 The recently developed European proposal on regulation of financial instruments mandates "non-discriminatory access to a CCP." In the United States, some—including senior policymakers—have advocated "fungibility" of trades conducted on different execution venues, which would require open access.

CPI COMPETITION POLICY INTERNATIONAL
Volume 7 | Number 2 | Fall 2011

REVISITING WILLIAM BAXTER'S PERSPECTIVES ON BANK INTERCHANGE OF TRANSACTIONAL PAPER

Thomas Brown

O'Melveny & Myers

Published in Competition Policy International (ISSN 1554-6853)
Fall 2011, Vol. 7. No. 2. For more articles and information, visit
www.competitionpolicyinternational.com

A TRIBUTE, OF SORTS, TO WILLIAM F. BAXTER'S "BANK INTERCHANGE OF TRANSACTIONAL PAPER"

Thomas P. Brown*

ABSTRACT

In 1983, the Assistant Attorney General of the Antitrust Division in the United States Department of Justice, Bill Baxter, did something that would be unfathomable today. He published an academic paper in a scholarly journal that related directly to a piece of antitrust litigation then pending in federal court in which he had served as an expert. The paper did not ignite a storm of controversy. Indeed, outside of the court presiding over the litigation to which Baxter's article related, Baxter's paper attracted little immediate attention. Even twelve years ago, when a group of friends and colleagues gathered to celebrate Baxter's work, this paper took a distant back seat to his tenure at the Department of Justice, his monograph on environmental law, and his one article on choice of law. Today, the paper is recognized as the seminal work on a topic that has attracted considerable attention for the last several years and seems likely to remain on the public agenda in the United States and elsewhere for the indefinite future: interchange.

The consensus on Baxter's paper ends there. There is considerable disagreement about what Baxter's paper actually says. For example, Jean Charles-Rochet & Jean Tirole credit Baxter for observing (1) that the decision to use a payment type requires coordination between the consumer and the merchant, (2) that the merchant and consumer in a four-party payment system may be served by different payment institutions, and (3) that maximization of output frequently requires a transfer from one side of the system to the other.[1] Dennis Carlton extracts a different lesson from Baxter. According to Carlton, Baxter's paper demonstrates that interchange can be used to enable merchants to charge two sets of prices: a higher price for cash customers and a lower price for credit customers.[2]

The various interpretations of "Bank Interchange of Transactional Paper" flow from two omissions in the paper that a contemporary reader will notice—a formal model and discussion of the work of other economists. Baxter's article has none of the former and very little of the later. It precedes by a few years the modeling revolution of Industrial Organization, and like other famous and roughly contemporaneous articles,[3] it makes little effort to explain where it stands in relation to the contributions of other economists. Baxter's article limits its discussion of the work of other economists to two short footnote discussions of an article by Bowen entitled The Interpretation of Voting in the Allocation of Economic Resources and the classic article by Landes and Posner, Market Power in Antitrust Cases.

With some trepidation,[4] this essay attempts to make the going easier. It provides a short map of the paper. It also fills in some of the obvious holes in Baxter's article and flags portions where an unwary reader might get trapped. Baxter's article, like a proverbial Michelin-starred restaurant, is worth the trip. But it is also worth attempting to smooth an otherwise bumpy journey.

* O'Melveny & Myers, U.C. Berkeley Law School. I want to thank Richard Schmalensee and Thomas Hubbard for comments on an earlier draft of this introduction. This paper does not represent the views of O'Melveny & Myers or any of its clients, and the errors and omissions are entirely my own.

I. A TRUNCATED ROAD MAP TO BAXTER'S BANK INTERCHANGE OF TRANSACTIONAL PAPER

Baxter's paper follows a simple outline. It contains three sections labeled as follows: "I. The Theoretical Viewpoint;" "II. The History of Four-Party Transaction Vehicles;" and "III. Conclusion." Like other features of the paper, the apparent simplicity is deceiving. The first and second sections each contain subsections. The first has two—"A. The Demand for Transaction Paper" and "B. The Supply of Transactional Paper." The second has three—"A. The Practice of Paying Checks 'At Par,'" "B. Bank Credit Cards and the Interchange Fee," and "C. Modern Developments." None of the subsections has sub-subsections, though the subsections devoted to "at par" checking and interchange would greatly benefit from them, as they cover quite a bit of ground.

In the interest of brevity, this essay devotes most of its attention to the sections central to Baxter's discussion of interchange—i.e., "I. The Theoretical Viewpoint" and "II. B. Bank Credit Cards and the Interchange Fee." It skips entirely the discussion of "The Practice of Paying Checks 'At Par'" and offers only limited observations about the "Modern Developments."

A) A BRIEF GUIDE TO BAXTER'S "THE THEORETICAL VIEWPOINT"

Baxter's paper does not, at least at the outset, waste any time. After a brief two-paragraph introduction, it jumps into a discussion of four-party payment systems by offering a generic vocabulary to describe those systems. The introduction of this vocabulary plays two important roles for the discussion that follows. First, it literally defines away the obvious differences between checks, credit cards and other forms of non-cash payments that might otherwise complicate the narrative. Second—and this was more important for the case to which this article related than any overt goal of the paper itself—the common vocabulary tends to suggest some degree of interchangeability or substitution among the instruments.

Baxter's vocabulary for four-party payments is quite simple. He posits the following participants:

1) a "merchant (M)" who receives transactional paper in exchange for goods or services;

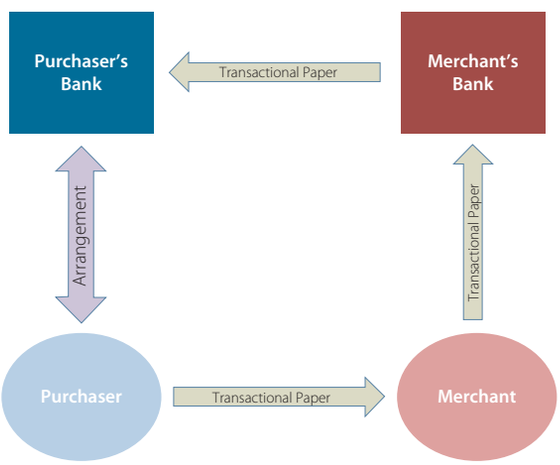
2) a "merchant's bank (M bank)" where M deposits its transactional paper;

3) a "purchaser (P)" who gives M transactional paper in exchange for goods or services; and

4) a "purchaser's bank (P bank)" where P has established "an arrangement that contemplates acceptance of and payment against" the transactional paper presented by P to M.

Baxter's vocabulary, although useful for advancing the points noted above, has one significant drawback. It omits any role for the administrator of the system. In other words, most of the systems that Baxter labels four-party systems are actually five-party systems. This is not as obvious when the system is introduced verbally as Baxter does, but the omission is striking when Baxter's instructions are illustrated.

BAXTER'S FOUR-PARTY TRANSACTION WHITER VISA, MASTERCARD OR THE FED?



Baxter's vocabulary provides the foundation for the paper's first major insight: the selection of a medium of exchange, unlike the decision about whether to purchase a traditional product, is contingent on the choices made by the counter-party to the transaction. Baxter draws the distinction with aid from a pedestrian example. He asserts that when a consumer contemplates purchasing a pair of shoes, the consumer's evaluation of the benefit from those shoes "is usually independent of other consumers' evaluations."

Payments, Baxter claims, are different. In order for a purchaser and merchant to use a particular payment instrument, both have to agree to it:

Rather than considering the demands of P and M as demands for separate products, define one unit of product to consist of the bundle of transactional services that banks must supply to P and M in order to facilitate the execution of one exchange of goods or services between P and M.

Baxter's insight that demand for payments requires coordination among payers and recipients is, as others have observed, profound. But if anything, the paper underplays the significance of the observation by failing to distinguish it from the work of other economists. Long before Baxter wrote his paper, economists had devised tools to model the impact that one person's decision might have on another. Both Alfred Marshall and Arthur Pigou had examined and debated the importance of externalities, and positive as well as negative externalities had appeared in models of everything from pollution to proliferation of intellectual property.[5]

Similarly, the challenge of reaching optimal outcomes through independent action had been a topic of conversation in economic circles at least since John Nash had helped introduce the world to game theory.[6] Even though the works of Marshall, Pigou and Nash do not directly anticipate Baxter's insight, the paper would surely be easier to understand had it taken the time to explain why.

After introducing the vocabulary, Baxter launches into a description of joint demand for transactional services that accompanies Figure 1, a graphical representation of that demand. The graph depicts two crossed demand curves—one for merchants (denoted dM) and one for purchasers (denoted dP)—that are summed "vertically" into an aggregate demand curve denoted d'.

The paper explains that the diagram should be understood to show the relationship between price and quantity for transactions conditioned on P and M coordinating their relative contributions to pay for the jointly consumed service.

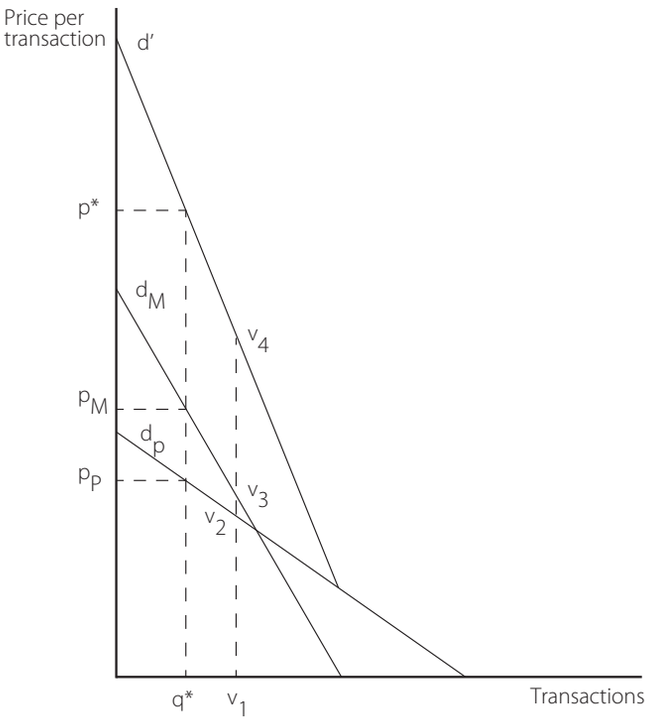


Figure 1

Baxter's paper then takes the discussion to the supply-side. Here, the coordinating parties are P's bank and M's bank. Baxter assumes that the costs to support the service that Purchasers and Merchants jointly consume are distributed over their respective institutions. Based on this assumption, he concludes, "the geometry of aggregate supply is analogous to that of aggregate demand."

And as with joint demand, Baxter offers a depiction of the independent supply curves as well as the joint whole.

He then combines the separate geometric depictions of supply and demand into a figure that "depicts the resulting demand-supply equilibrium."

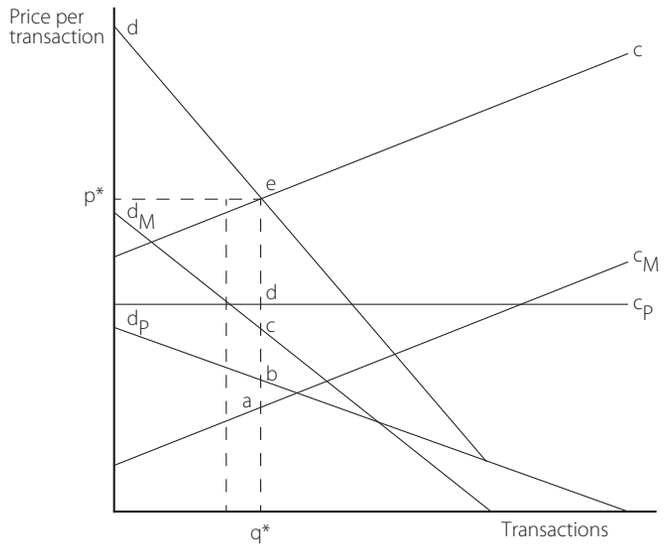


Figure 4: Merchant makes sales of amount S; M bank discounts q*c; merchant gets S – q*c; P bank collects S + q*b from purchaser; together banks retain (S + q*b) P + (–S + q*c) M = q*b + q*c = q*e; P bank remits S + q*b – q*d to M bank

This diagram sets up the paper's second critical insight:

*What is of critical importance is that the marginal cost q*d of the activities performed by the purchaser banks bears no necessary relation to the amount of revenue q*b forthcoming from the purchasers with whom those banks have contractual relationships.*

*Similarly, the costs q*a associated with the activities performed by merchant banks have no necessary relation to the amount of revenue q*c forthcoming from the merchants with whom they have contractual relationships.*

In other words, unless the banks on either side of the transaction are permitted to coordinate their joint supply decision through a side payment, from the side that collects too much to the side that collects too little, the four-party payment system will supply fewer transactions than is socially optimal.

Again, however, Baxter avoids presenting his conclusion in the language and form of formal economics. As Rochet & Tirole explain in a widely circulated draft of the famous paper noted above, those conclusions can be extracted from Baxter's analysis with a slight change in his notation. If benefits to purchasers and merchants are defined as marginal net benefits from the use of cards relative to other forms of payment, and if costs to purchasers and banks (or fees to their respective banks) are defined as marginal costs (or revenues), and if those banks are price-takers rather than price-setters, then as a payment system it will achieve the socially optimal level of output by setting the transfer payment between the two sides of the transaction at exactly the rate necessary to ensure that neither earns an economic profit.[7]

B) A BRIEF GUIDE TO BAXTER'S "BANK CREDIT CARDS AND THE INTERCHANGE FEE"

After theory comes history, at least in this article. Baxter takes the reader on an extended tour of the evolution of the check clearing system in the United States[8] and then returns the discussion to credit cards by way of merchant credit. "For centuries," Baxter observes, "merchants have extended short-term, interest–free credit to customers whose patronage is highly valued." And as Baxter explains, the rationale is quite intuitive. By making credit available to their best customers, merchants make it possible for consumer to (i) buy more on (ii) fewer visits and (iii) choose "higher-priced items" than they might otherwise.

Baxter then points to a shift from merchant credit to third-party credit following World War II. He posits the existence of a "frequent traveler" with "high income and high time costs" who would have access to local merchant-supplied credit at home but not on the road. This hypothetical "frequent traveler" would have poor payment options available to him—cash, traveler's checks, and personal checks. Cash, from a consumer perspective, carries a significant risk of loss. Traveler's

checks come with "high time costs." Checks require the presentation of "identification at a moment when time costs [are] greatest" (i.e., the moment of purchase) and "not infrequently" involve "humiliat[ion]" with the effort to confirm identity at the point of sale.

According to Baxter, card-based payment systems arose to meet this demand. Non-banks such Diner's Club and, later, American Express offered three-party systems. Such systems, of course, did not need an interchange mechanism. One firm both signed merchants to accept cards and issued cards to consumers. Four-party bank systems came later. Three-party bank systems that had evolved in specific geographies became four-party to achieve ubiquity that "by reason of our geographically restrictive banking laws, could not be obtained by any single banking enterprise."

Having laid out the four-party model earlier, Baxter then delivers the rhetorical coup de grâce on the need for an interchange mechanism:

[M]ultibank organizations were from their inception four-party systems having the peculiar economic characteristic previously described. Given the distribution of charges between P and M that would achieve equilibrium in their demands, it was overwhelmingly improbable that the revenue stream from M to M bank or from P to P bank would equal the costs of the subset of activities that a particular bank was required by the technology of the payment system to perform; thus some redistribution of those revenues between M bank and P bank was likely to be necessary for the payment system to compete effectively with alternative mechanisms.

Although the article--or, at least, the section--could end there, it does not. Baxter proceeds to answer three discrete questions: (1) whether individual bank negotiations might take the place of centrally set interchange; (2) whether interchange fees should be set at 0 (as in the checking system); and (3) whether interchange rates are currently set at the socially optimal level. The article does not, however, attempt to motivate the discussion, and it seems, at least without context, a bit forced.

Context is, however, available. These questions flow directly from the litigation that served as the inspiration for Baxter's paper, NaBanco v. Visa U.S.A., Inc.[9] NaBanco argued (1) that individual negotiations between counterparties to specific transactions could take the place of centrally set interchange; (2) that the court should simply set interchange at 0, effectively allowing acquirers to keep the entirety of what they collect from merchants; and (3) that interchange had been set "too high."

The article's answers to these questions are not entirely satisfying. The article marches through them as if it were following an indisputable chain of logic. But the explanations are not entirely persuasive. The problem is largely rhetorical. After asserting that interchange is necessary for a four-party payment card system, Baxter's writing becomes significantly more conditional. Key sentences throughout the discussion use words like "can," "could," and "possible." And as in the theoretical section that opens the piece, Baxter eschews external references.[10] In at least this respect, the court's discussion of these points is more satisfying. The court rejects NaBanco's efforts to replace interchange with individual bi-lateral negotiations by observing that the transaction costs in such a system would be "high and stultifying."[11] The court similarly dismisses the claim that the Sherman Act requires interchange fees to be set to $0. Using more or less the same verbal formulation that Baxter's article uses to introduce interchange, the court explains that nothing in the system "suggests, much less guarantees" that revenue streams on either side of the system will be sufficient to cover the costs unique to that side of the platform.[12] The court also has little patience for the argument that Visa arrived at its interchange rate through a flawed process. As the court explains, although the process through which Visa set interchange may not have been perfect, it "was and is careful, consistent, and within the bounds of sound business judgment."[13]

II. FINAL THOUGHTS

Baxter's paper is the first scholarly paper to discuss a tool that helped propel the rise of electronic payments around the world and that has been the subject of nearly constant legal and regulatory scrutiny since its introduction nearly forty years ago. With the passage of time, it has become difficult to separate Baxter's contribution from those who helped to formalize and extend his work.[14] But even if lawyers and economists interested in interchange and payment card networks must look beyond Baxter for answers to their questions, his article remains, even after the passage of time, the best place to start.

1 *See* Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: Some Economics of Payment Card Associations*, 33 Rand J. Econ. 549, 564 (2002).

2 Dennis W. Carlton, *Externalities in Payment Card Networks: Theories and Evidence* 126, *in* Kansas City Federal Reserve, The Changing Retail Landscape: What Role For Central Banks? (2010).

3 *See, e.g.*, Benjamin Klein et al., *Vertical Integration, Appropriable Rents, and the Competitive Process*, 21 J. L. Econ. 297 (1978).

4 The trepidation arises from the quasi-religious devotion and disdain that Baxter continues to inspire in fans and critics. *Compare* Richard A. Posner, *Introduction to Baxter Symposium*, 51 Stan. L. Rev. 1007 (1999) (recalling his impression upon meeting Baxter in 1967 when interviewing for a junior faculty post at Stanford—"I was instantly, immensely, and permanently impressed by the power of his mind and the clarity of his expression") with Lloyd Constantine, Testimony Before The Antitrust Modernization Commission (2005), *available at* http://govinfo.library.unt.edu/amc/commission_hearings/pdf/Constantine.pdf (complaining that Baxter, who as head of the Antitrust Division had reportedly defied President Reagan in pursuing the case against AT&T, had "prophe[sied]" and sought to eliminate "federal antitrust enforcement").

5 *See* Gary S. Becker, Economic Theory 85 (1971) (discussing "the Marshall-Pigou tradition" and formally describing models with externalities running between competing firms).

6 *See* John Nash, *Two-Person Cooperative Games*, 21 Econometrica 128 (1953).

7 *See* Jean-Charles Rochet & Jean Tirole, *Cooperation Among Competitors: The Economics of Payment Card Associations* 4-5 (May 16, 2000), available at http://www.wcas.northwestern.edu/csio/Conferences/CSIO-IDEI-2000/tirole.pdf. The working draft also explicitly credits Baxter for observing that in a four party system "there is no reason why both banks should break even on the transaction." *Id*. at 4.

8 As discussed above, this essay is going to skip Baxter's discussion of the check system. That section of the article is well footnoted and generally straightforward. Moreover, although some critics harbor objections to some elements of Baxter's history, see, e.g., Alan S. Frankel, *Monopoly and Competition and Exchange of Money*, 66 Antitrust L.J. 313 (1998), Baxter's discussion of the evolution of the check system is generally regarded as authoritative.

9 *NaBanco v. Visa U.S.A., Inc.*, 596 F. Supp. 1231 (S.D. Fla. 1984), *aff'd* 779 F.2d 592 (11th Cir. 1986).

10 Two very short sections follow Baxter's discussion of interchange in the credit card systems. The first is labeled "Modern Developments," and the second is simply "Conclusion." The section devoted to "Modern Developments" offers some predictions about debit cards that, at least in the wake of the cases challenging Visa's and MasterCard's respective honor all cards rules proved quite prescient—on page 585, Baxter notes, "It seems likely . . . that the two payment vehicles [debit and credit] will have to be differentiated and subjected to different patterns of distributing charges between merchants and card holders and, in all probability, to different interchange fees." The Conclusion contains very truncated discussions of two key legal issues—(1) whether arrangements setting interchange should be viewed as price fixing; and (2) whether further cooperation among the banks that make up the card networks should be condoned. According to Baxter, the answer to both questions is no.

11 *Nabanco*, 596 F. Supp at 1261.

12 *Id*. at 1260.

13 *Id*. at 1262.

14 Since Baxter's piece was published, it has been cited by 260 papers and legal opinions in English and other readily scannable languages. A full 170 of those citing works also cite the work of David Evans, Jean-Charles Rochet, Richard Schmalensee or Jean Tirole.

BANK INTERCHANGE OF TRANSACTIONAL PAPER: LEGAL AND ECONOMIC PERSPECTIVES

William F. Baxter*

Consumer purchases by means other than currency—for example, by check, credit card, or debit card-generate a paper record that must be handled by the merchant, the merchant's bank, the purchaser's bank, and the purchaser. Before coming to Washington, I was involved in several controversies over the terms on which these types of records would be created and exchanged between banks. That involvement led me to think that economics provides novel and useful insights into the process of interchange and the payment systems of which they are a part.

In this article I examine some of those lessons. I focus primarily on the economics of financial institutions in generating and exchanging accounting information essential to the operation of four-party cashless payment systems. Section I develops the economic theory of these systems, and Section II examines the evolution of four-party cashless payment systems in the light of this theory.

I. THE THEORETICAL VIEWPOINT

The payment systems I discuss all involve four parties and four consensual arrangements. For example, in the checking context, the parties are the payee of the check, the bank in which the payee deposits the check for credit to his account, the bank on which the check is drawn (typically a bank with which the maker of the check has a depository arrangement), and finally, the maker of the check, usually a depositor with the drawee bank. In the context of the credit card or the debit card, four functionally analogous parties are involved, although the labels attached to them differ.

Because I focus on what is common to these payment mechanisms rather than on the distinctions between them, I use neutral terms to describe the actors and operations inherent in these mechanisms—terms not associated with any particular payment mechanism. Each payment system generates certain accounting information, which is exchanged among the four parties in order to facilitate an exchange of goods or services between two of the parties. (Although electronic signals soon may replace much of the paper that embodies the accounting information required for cashless payment systems, this would not affect the basic economic issues addressed in this article.) For convenience, I refer to the embodiment of this accounting information as **transactional paper** regardless of its physical form, and to the generation and exchange of transactional paper as **transactional services**. I assume that the person who initially receives the transactional paper is a **merchant** (*M*) who receives it in payment for goods; I refer to the bank in which he deposits the paper for credit to his account as the **merchant's bank** (*M* bank);[1] I assume that the person who gives the paper does so in his capacity as **purchaser** (*P*) of the goods sold by the merchant; and I refer to the bank with whom the purchaser has an arrangement that contemplates acceptance of and payment against that paper as the **purchaser's bank** (*P* bank). Nothing turns on the assumption that the purchaser and the merchant are in fact playing those particular roles. What is critical to the analysis is that there are at least four parties and that their relationship to the payment mechanism is analogous to the one I have described.[2]

---

* This article was originally published in volume XXVI of the Journal of Law & Economics. At the time, William Baxter was Assistant Attorney General of the Antitrust Division in the United States Department of Justice. In his introduction, Baxter wrote, "This paper was written while I was Professor of Law at Stanford University and revised thereafter. The views expressed here are my own and are not official policy statements of the Antitrust Division or the Justice Department. I thank J. Anthony Chavez and Greg Sidak for their helpful research assistance and suggestions." The article is reprinted with the permission of the University of Chicago Press.

A) THE DEMAND FOR TRANSACTIONAL PAPER

Any bargained-for exchange requires *P* to pay *M* for goods or services received. Once an economy moves beyond barter, the concept of payment involves much abstraction. Even if *P* tenders the gold coins of the realm, M is willing to accept the coins not because *M* can use them to fashion jewelry or fill his teeth but because he expects other merchants to "honor" the coins-that is, to be willing to deliver goods and services which *M* wants in exchange for the coins. The progression from gold coins to bank notes, to negotiable paper, to credit card charge slips, to electronic impulses as acceptable forms of payment makes clear that what is involved is a mechanism for causing multiple accounting entries to be made in several different sets of books, entries that in their totality constitute the community's recognition of each person's entitlements to consume. Merchant *M*, having delivered goods to *P* at an agreed price, wishes to have his consumption credits enhanced on the books of the community by the amount of the price; and since the rules of the community require that books balance, P agrees to have the consumption credits posted to his name reduced by an equal amount. Adjustments of the community's books in crediting *M*'s account and in debiting *P*'s account on the occasion of a purchase are accounting services that facilitate the needs of both the merchant and the purchaser. In terms of supply and demand, *M* and *P* have demands for transactional services in order to effect the appropriate entities in the community's books; banks supply such services.

Although a given transactional service may have as its fundamental purpose adjustment of the accounts of M and P, it will also have a variety of other product characteristics, such as cost of supply, convenience to the consumer of service (whether M or P), speed of adjustment, and accuracy of entry. There is no a priori reason to believe that the preferences of merchants for a given transactional service would be the same as that of purchasers or even that different merchants (or purchasers) would have identical preferences. Consequently, the distribution of transactional services in terms of their product characteristics, the prices for these services, and the volume of their production are all questions remaining to be answered in the context of a market equilibrium.

At first impression transactional services appear to be private, not public, goods. Banks are able to extend such services to those who are willing to pay for them, whether merchants or purchasers, and to exclude from the services those who are not. Yet transactional services are unlike most private goods, because one cannot determine the aggregate (or industry) demand for them in the traditional way by horizontally summing the individual consumers' demands.

Demand for a private good depends on each person's evaluation of the good's marginal utility and can be described by a function indicating the amount of product the person is willing to buy at a given price. Each consumer's evaluation of the marginal utility of a private good is usually independent of other consumers' evaluations, and so aggregate demand at any price level is the sum of the individual demands at that price. For example, if the prevailing price of shoes is $30 a pair, consumer Jones will buy one, and then another, and then another pair of shoes until the marginal value he attaches to the next pair (which he does not buy) falls below $30. The same is true for consumer Smith, although there is no reason to expect that at any particular price each will demand the same number of pairs, because there is no particular reason to suppose that the marginal value that Jones attaches to the third or fifth or eighth pair of shoes is the same as the marginal value that Smith attaches. Because the evaluations of the marginal value of shoes by Jones and Smith are independent of one another, the aggregate demand of Jones and Smith for shoes at $30 a pair is simply the sum of their individual demands at that price.

In the case of transactional services, however, although consumer *P*'s marginal valuation of the additional use of a particular payment mechanism may differ markedly from consumer *M*'s marginal valuation,[3] these valuations cannot be independent of one another as in the case for shoes. The mechanics of transactional services require that for every transaction in which a purchaser becomes a maker of a check, there must be one—and precisely one—transaction in which a merchant becomes a payee; similarly, each use of a credit card by a card holder must be matched by precisely one act of acceptance of the card (or, more accurately, the paper that the card generates) by a merchant.

This identity in the type of transactional service used by the merchant and purchaser in a given exchange introduces a constraint not normally found in markets for private goods and reflects the interdependence in the marginal valuations between merchants and purchasers. Because the mechanics of transactional services require the acceptance of a particular payment mechanism by **both** the merchant and the purchaser

to effect any given purchase, the marginal valuation of a transactional service by one party to the purchase is contingent on the acceptability of this form of service by the other party. On the one hand, given that particular payment mechanism is acceptable to the other party, marginal valuation is determined in the usual manner for private goods. On the other hand, if the payment mechanism in question is unacceptable to the other party for whatever reason, the marginal valuation by the first party is zero regardless of the magnitude of its value when the mechanism is acceptable. The contingent nature of these marginal valuations of transactional services by merchants and purchasers, and hence the contingent nature of the individual demands for these services, destroys the independence necessary to permit the calculation of aggregate demand by summing the individual demands horizontally and largely renders intractable the economics of transactional paper in this particular description of the market.

Perhaps the most intuitively appealing way to resolve the difficulties posed by this market model is to redefine what we mean as one unit of the product consumed. Rather than considering the demands of *P* and *M* as demands for separate products, define one unit of product to consist of the bundle of transactional services that banks must supply jointly to *P* and *M* in order to facilitate the execution of one exchange of goods or services between *P* and *M*. Under this interpretation, the supply price of the product is the sum of the individual charges to *P* and to *M*. Furthermore, the demand for that product is a joint demand of *P* and of *M*: in combination they must make a payment of that magnitude to the banks to induce the necessary supply, but independently neither *P* nor *M* necessarily confronts any particular price as one he must pay in order to have his demand fulfilled.[4] This model preserves the excludability property of transactional services.

Figure 1 illustrates the derivation of aggregate demand for transactional services of a given type in a single-merchant, single-purchaser economy. The quantity axis is calibrated in units which represent the bundle of services that must be provided by banks to both *P* and *M* in order to facilitate one exchange. The vertical axis gives the reservation prices of the two traders for various levels of consumption of the transactional services. Line d_M represents the demand schedule of *M* for such complete units of transactional service on the assumption that *P*—*M*'s customer—is willing to use this particular service but unwilling to make any contributory payment for the units when purchased from the bank.

Line d_p represents the demand schedule of *P*, based on the assumption that *M* is unwilling to make any contributory payment for those services. Given the information shown in line d_M and line d_p, the aggregate demand schedule of *M* and *P* for these units of transactional services is line *d'*, which is obtained by summing vertically the separate demand schedules of *M* and *P*. In other words, the schedule *d'* is constructed so that if any vertical line is drawn through the figure, the distance v_1v_4 equals the sum of distances v_1v_2 and v_1v_3.

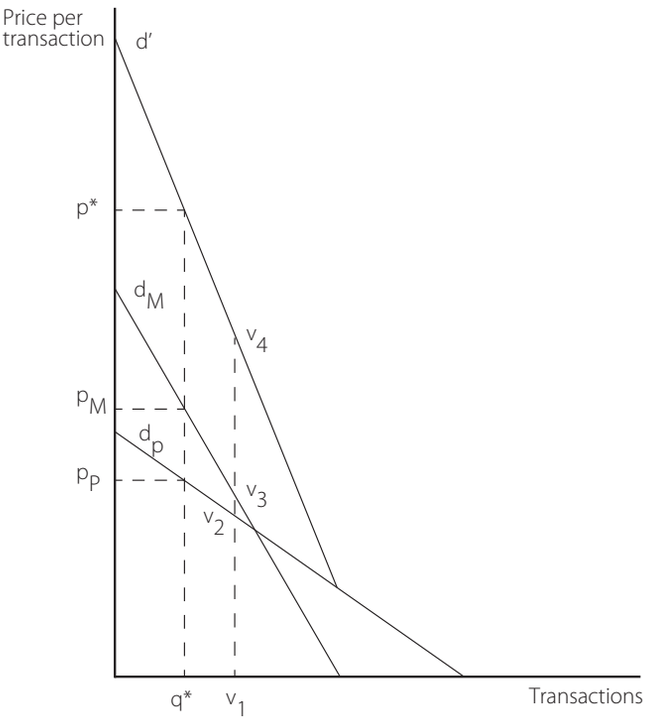


Figure 1

Figure 1 should be interpreted as follows: if the price per complete transaction—that is, the total revenue banks will demand to provide the services necessary to facilitate one exchange between *M* and *P*—is *p**, then the quantity of transactions that *M* and *P* should demand is *q**, the quantity indicated by a vertical line dropped from the intersection of *p** and *d'*. I say "should" rather than "will" be demanded because, although *q** is the quantity of transactions that maximizes the aggregate benefits of *M* and *P*, a certain amount of coordination is prerequisite to *M* and *P*'s arriving at that outcome. Specifically, this favorable outcome will result only if the aggregate price *p** is apportioned between *M* and *P* in the proportions represented by the height of their respective demand curves at output level *q**. That is, for each transaction, *P* must find a way to make some payment p_p to the banks, and *M* must find a way to

make some payment p_M to the banks; when p_P and p_M are summed they will, by construction in Figure 1, equal p^*, the price that the banks demand for providing those services. If there are no bargaining costs-that is, if *P* and *M* have perfect information and neither persists in strategic bluffing to reduce his own costs at the expense of the other-they would bargain to this particular outcome. On the other hand, if either *P* or *M* strategically insists on paying less, then, because the other can be induced to pay no more at so high a level of transaction services, both *P* and *M* will be harmed, for the sum of their contributions will be less than p^*; thus the banks will decline to provide services that *M* and *P* together value at p^*.

One must resist any impulse to say that *M* is paying too much and *P* too little in the circumstances depicted by Figure 1. Given that the banks will insist on receiving revenues per transaction in the amount p^*, and given that *P* is unwilling to pay more than p_P per transaction at output level q^* for the very good reason that he does not value the service any more highly, *M* can only worsen his position by declining to make a payment per transaction in the amount pm. For it is inescapable that *M* and *P* must agree on some specific number of transactions to be effected by the payment mechanism in question. And if that number is to be q^*, then in our hypothetical case depicted in Figure 1 agreement can only be reached if *M* is willing to pay the preponderant share of the price p^*. In the region q^*, *M* values the marginal transaction more highly than does *P*, and *M* pays accordingly.

In our example, the individual demand schedules imply that if the level of transaction prices required by banks fell substantially, *M*'s valuation of these transaction services would decline more rapidly than would *P*'s. There is a particular output level, corresponding to the intersection of the individual demand curves where equal contribution would be required for equilibrium. And there is a still higher output level at which *M* would be unwilling to pay anything for additional services: to the right of that point *P* would have to bear all bank-imposed charges in order for equilibrium to be attained.

Figure 1 depicts how the individual demand schedules of a particular merchant and purchaser must be aggregated vertically in order to obtain a well-defined expression of the aggregate demand for transaction services in this miniature economy. However, since in our model merchants trade only with purchasers and not with other merchants, as we increase the number of merchants beyond one we must sum their individual demand schedules horizontally to obtain the aggregate merchant demand schedule. Similarly, if more than one purchaser exists in the economy, we must sum their individual demand schedules horizontally to obtain the aggregate purchaser demand schedule. Then, as in our one-merchant, one-purchaser case, the total aggregate demand schedule in the multi-merchant, multi-purchaser economy is obtained by summing vertically the two partial aggregate demand schedules of the two classes of traders.

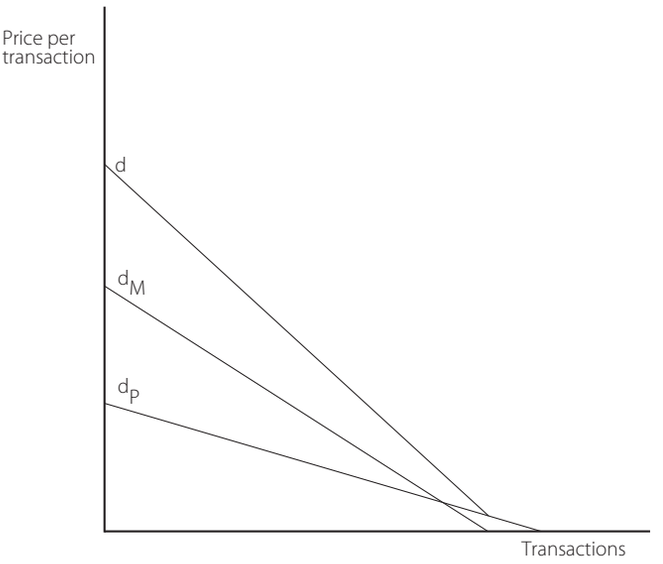


Figure 2

The multi-merchant, multi-purchaser case is illustrated in Figure 2. Although the total number of transactions demanded industry-wide will be orders of magnitude larger than that depicted in Figure 1, Figure 2 retains the basic feature of Figure 1: merchant demand and purchaser demand are each depicted individually, and the aggregate demand for transaction services that confronts all participating banks in the community consists of the vertical aggregation of these two partial aggregate demands. For it remains true in the industry context, as in the case of the individual merchant, that a transaction is a two-sided arrangement, that transaction services facilitate the needs of both merchant and purchaser, and that agreement on a common number of transactions to be effected through the particular payment mechanism will not be possible with an equal division of charges between merchants and purchasers except under the extremely unlikely coincidence that the aggregate level of charges per transaction required by the banks lies directly above the intersection of those separate demand curves.[5]

B) THE SUPPLY OF TRANSACTIONAL PAPER

A polarity corresponding to that of *M* and *P* on the demand side exists on the supply side as well: *P* has his banking relationship with one institution, *P* bank, and *M* has his banking relationship with another, *M* bank.[6] Both *M* and *P* bank will incur costs associated with establishing the payment system and providing services essential to effecting each transaction between *P* and *M*.

One can identify a set of activities that, at least in the typical case, will be performed by the employees of *M* bank, in principal part at *M*'s business premises. Such activities include soliciting, negotiating, and executing contractual agreements with merchants who do business in the geographical vicinity of *M* bank; participating in the periodic delivery by merchants to *M* bank of *M*'s records of transactions with purchasers; entering on the books of *M* bank credits to the account of *M*; capturing, in one form or another, the identity of the purchasers with whom *M* dealt and the identity of *P* bank with whom each *P* has his banking relationship; forwarding those data through some interchange or clearance mechanism to *P* bank; and bearing the cost of capital to the extent that unconditional credits are posted to *M*'s account before payment is received from *P* bank.

Analogously, there will be certain activities that typically will be performed by the employees of *P* bank, in major part at its business premises: soliciting, negotiating, and executing agreements with purchasers who wish to use the payment mechanism; receiving from a large number of *M* banks data about transactions executed by those purchasers; posting debits to the individual accounts of its various purchasers; transmitting periodic statements of those accounts to its various purchasers; and, in the case of arrangements not involving antecedent deposits by purchasers, receiving payment from those purchasers and entering credits to their account corresponding to their payments; bearing the costs of capital to the extent that unconditional credits are forwarded to *M* banks before payment from purchasers is in hand; and bearing the risk of purchaser default.

To describe the activities traditionally performed by one bank or another is not to say that the costs of these activities must be borne by the bank performing them. Just as it is true on the demand side that there must be an identity between individual purchaser transactions and individual merchant transactions, so also is it true on the supply side that there must be an identity between individual merchant bank transactions processed and individual purchaser bank transactions processed. For example, signing up merchants would be pointless if purchasers were not simultaneously being signed up. Hence, on the supply side, the costs of the activities of *M* bank and *P* bank must be regarded as joint costs with respect to each individual transaction, in the same sense that, on the demand side, demand of merchants and purchasers is strictly interdependent.

Correspondingly, the geometry of aggregate supply is analogous to that of aggregate demand. It is conventional to think of the supply curve for an industry as being constituted by the horizontal aggregation of the supply curves of the individual firms. But because the costs incurred by the banks are joint, when *P* bank participates on behalf of purchasers and *M* bank participates on behalf of merchants, the costs of the two firms must be aggregated vertically, not horizontally, in order to obtain an analytically useful representation of the full marginal cost per transaction and hence of the number of purchaser-merchant exchanges that banks will facilitate at any particular price level for transactional services.

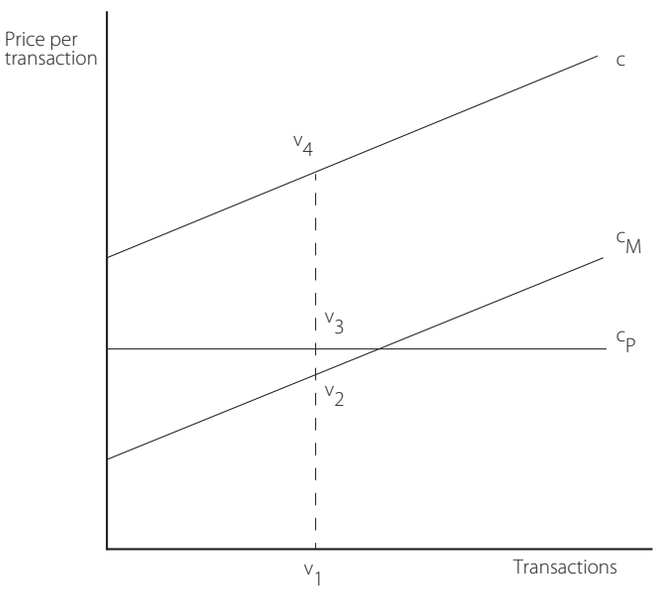


Figure 3

Figure 3 depicts possible marginal cost curves c_M for *M* bank, and c_P for *P* bank, together with their vertical aggregation *c*, which corresponds to the total marginal cost per exchange facilitated by the two participating banks. As before, the technique of vertical aggregation is such that, given any vertical line drawn through the curves, the distance v_1v_4 equals the sum of the distances $v_1v_2 + v_1v_3$.

Somewhat arbitrarily, I have drawn Figure 3 in a way that suggests that *P* bank's costs exhibit constant returns to scale whereas *M* bank's costs exhibit decreasing returns to scale, but nothing in the analysis turns on those particular assumptions.[7] Figure 3 also could be thought of as depicting industry supply, if one views c_P as a traditional horizontal summation of the marginal cost curves of all purchaser banks, and c_M as the traditional horizontal summation of marginal cost curves of all merchant banks. But in this interpretation, too, the vertical summation *c* of those two sets of costs depicts the industry supply curve, for with respect to each transaction, revenue equal to *c* must be forthcoming in order to cover all industry marginal costs

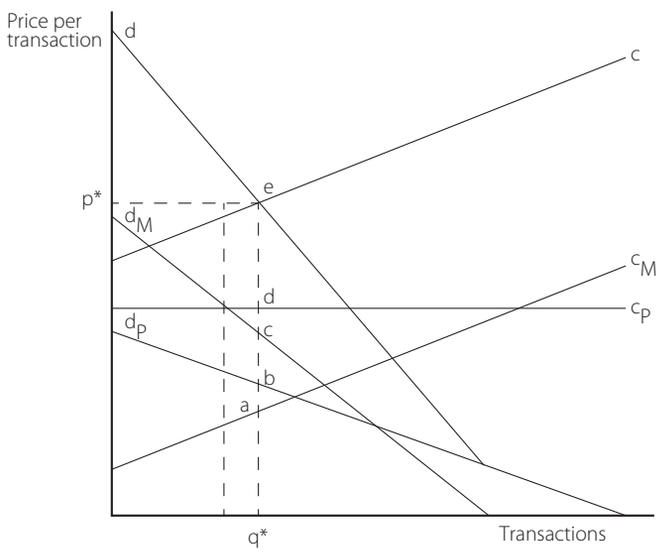


Figure 4—Merchant makes sales of amount *S*; *M* bank discounts *q*c*; merchant gets *S* – *q*c*; *P* bank collects *S* + *q*b* from purchaser; together banks retain (*S* + *q*b*) *P* + (–*S* + *q**c) *M* = *q*b* + *q*c* = *q*e*; *P* bank remits *S* + *q*b* – *q*d* to *M* bank. At close,

P's position	*–S*	*–q*b*			
P bank position	*+S*	*+q*b*	*–S*	*–q*b*	*+q*d*
M bank position	*–S*	*–q*c*	*±S*	*+q*b*	*–q*d*
M's bank position	*+S*	*–q*c*			
Totals down	0	0	0	0	0
Totals across:					
P bank	*+q*d* = cost				
M bank	*q*c + q*b – q*d = q*a* = cost				
"Interchange fee"	(*q*d – q*b*) = (*q*c – q*a*)				

Figure 4 depicts the resulting demand-supply equilibrium. In view of the total marginal cost per completed transaction, the industry is willing to supply transactions along the positively sloped marginal cost curve. These total marginal costs may be subdivided into costs incurred by merchant banks and those incurred by purchaser banks. Purchasers, on the other hand, through their pooled willingness to purchase transaction services, have effective demands along the line *d*. The intersection of *d* with *c* at point *e* implies an equilibrium price of *p** to facilitate *q** exchanges. In the process of producing an industry output of *q**, merchant banks incur marginal costs in the amount *q*a* and purchaser banks incur marginal costs in the amount *q*d*; and the sum of those two sets of costs is *q*e*. In consideration for transactional services to facilitate *q** exchanges, purchasers are willing to make expenditures in the amount of *q*b* and merchants are willing to make expenditures in the amount *q*c*; the sum of those two revenues streams is *q*e*.

What is of critical importance is that the marginal cost *q*d* of the activities performed by purchaser banks bears no necessary relation to the amount of revenue *q*b* forthcoming from the purchasers with whom those banks have contractual relationships. Similarly, the costs *q*a* associated with the activities performed by merchant banks have no necessary relation to the amount of revenue *q*c* forthcoming from the merchants with whom they have contractual relationships. Nonetheless, the sum of the two revenue streams equals the sum of the two marginal cost streams, *q*e*, and it follows that there must be some particular side payment between a merchant bank and purchaser bank with respect to any particular exchange that will bring the receipts of each bank into equality with the marginal cost it has incurred in providing transactional services to facilitate the exchange.

In Figure 4, *M* bank receives *q*c* of revenue from merchants and must pay over to *P* bank the amount *ac*; and *P* bank receives from its purchasers revenue in the amount *q*b*, which is less than it costs, *q*d*, by the amount bd. The side payment from *M* bank, *ac*, precisely equals the deficiency, *bd*.[8]

It is true, of course, that a side payment of ac per facilitated exchange from *M* bank to *P* bank is not the only conceivable institutional adjustment, but it appears to be by far the simplest and the least expensive.[9] Since any redistribution mechanism will itself involve a transaction cost which will serve to raise *C*, the

mechanism that minimizes transaction costs is in the interest of all the parties. Since remittance of funds in some amount from *P* bank to *M* bank is an inescapable feature of any payment mechanism of the type under consideration, adjustment of the magnitude of that remittance to achieve the equilibration of costs and revenue clearly appears to be the preferred mechanism.

In summary, one would expect to observe the following behavior in the operation of cashless payment systems: after the purchase transaction between *P* and *M* (1) *M* bank buys the paper from *M* at face value, minus a discount in the dollar magnitude q^*c, thus bringing revenues of q^*c into the banking system; (2) *P* bank buys the paper at face value from *M* bank, minus a discount ($q^*c - q^*a$), leaving *M* bank with net revenues q^*a; (3) *P* bank bills its customer *P* in an amount equal to the face of the paper plus the premium q^*b, thus bringing revenues in the amount q^*b into the banking system. Thus in total *P* bank has received revenues in the amount $q^*b + q^*c - q^*a$. But the first two terms in that expression are equal to q^*e; and q^*e minus the third term, q^*a, is equal to q^*d, *P* bank's costs.

One important assumption underlies the preceding paragraph: banks participating in the payment system are behaving competitively and charging prices to *P* and *M* corresponding to the bank's marginal costs and, in equilibrium, to their average total costs including the opportunity costs of invested capital. There are two quite distinct reasons why this assumption may not hold in any particular real world context. First, through collusion the banks might have acquired enough market power to be able to charge both purchasers and merchants prices that exceed the banks' cost.[10] I explore the implications of collective action among banks more fully, later in this paper.[11] For the present, I note only that the problem of cartel profit maximization will be complicated by the fact that, in order to maintain an equilibrium number of transactions, the cartel must increase prices each to merchants and to purchasers in amounts dictated by the slope of their demand curves—amounts that, in all probability, are equal neither in absolute magnitude nor in percentage markup over the competitive price. Hence cartelization of the industry would be comparatively difficult.[12]

The second reason that some degree of market failure might be observed involves the relations between the two sets of banks. Each *M* bank collects transaction paper that must be forwarded for collection to many *P* banks, including some with which that *M* bank will never before have dealt. At that time, *M* bank faces a monopsonistic buyer for each piece of paper. One can imagine a variety of institutional solutions for this problem. Conceivably, *P*'s participation in the payments system could be conditioned on his assuming an obligation to redeem his paper from any bank that presented it to him. Under that arrangement, *M* bank would face a competitive set of bidders for *P*'s paper, but such an arrangement would so increase *P*'s transaction costs that the competitive viability of the payment system, in competition with others, would be in serious doubt. Moreover, if the payment system in question involves a deposit relationship between *P* and *P* bank, accompanied by an understanding that the paper will be debited against *P*'s deposit, *P* bank would nevertheless remain in a significant monopsonistic position: it would have lower float costs and lower default costs because of the security afforded by the existence of the deposit.

In short, if *P* is to be afforded the transaction costs savings associated with having his paper returned to him through one particular *P* bank, and if deposit-based transaction systems, as opposed to pure credit systems, are to be among the set of systems available, *M* bank must have, at the time it acquires paper from its set of merchants, a preexisting understanding governing interbank discount with each bank in the set of participating *P* banks. If the number of *P* banks participating in this system is large, as it often will be, a complete set of bilaterally negotiated agreements would be excessively cumbersome and costly. Some uniform understanding between the set of *M* banks on the one hand and the set of *P* banks on the other would appear to be essential to any cost-effective payment system. As we shall see, the practical and legal difficulties of bringing into existence such a uniform understanding constitute a significant part of the history of the various payment systems.

II. THE HISTORY OF FOUR-PARTY TRANSACTION VEHICLES

Over the last 150 years, three distinct categories of four-party cashless payment systems have evolved. The check and the bank credit card are heavily used today to facilitate exchanges, and the debit card is increasingly being promoted. This section presents a brief history of the commercial environment in which each of these

developed in conjunction with each of them. By use of the economic theory developed in Section I, it is possible to uncover previously unrecognized forces in the evolution of these payment systems.

A) THE PRACTICE OF PAYING CHECKS "AT PAR"

In the early 1800s the two principal means of payment in commercial transactions were (i) bank notes issued by state banks and (ii) drafts. These two media can be thought of as corresponding to (i) currency and (ii) checks today. Although checks had an early origin,[13] they did not become common until after the Revolutionary War.[14] In the years between the demise of the Second Bank of the United States and the Civil War, checks were commonly used as a means of paying local bills only in the nation's commercial centers.[15] City banks encouraged the use of deposit currency because inferior country bank notes of uncertain value tended to drive the sounder city bank notes out of circulation.[16] For the most part, the attempts of the city banks to prevent the discounting of these notes were unsuccessful.[17] During this time, transportation outside the nation's commercial centers was slow, expensive, and often dangerous. Only infrequently did either goods or people travel very far. Markets were predominantly local, and goods consumed in any geographic area usually had been produced there.

In those commercial circumstances, *P* and *M* were almost always residents of the same area. Accordingly, payment media rarely had to be sent beyond the local area. Bank notes, issued by the local bank or banks, circulated through the area and were used in a far greater fraction of transactions than currency is used today.[18] In the larger local transaction, and also in the relatively infrequent long-distance transaction, the draft was the typical medium used.[19]

If P became indebted to *M*, who resided in a distant place, *P* would execute payment by purchasing a draft made payable to *M* as payee. His local *P* bank would prepare a draft instructing *M* bank in *M*'s geographic vicinity to make payment to *M* in the amount of the indebtedness. For this service, *P* would pay a very substantial fee in comparison with present day transaction costs. In the terminology of the day, *P* was said to "purchase exchange" from *P* bank.[20] The draft thus obtained would then be sent through the mail, usually by *P* bank but perhaps by *P* himself, ad- dressed either to *M* bank or to *M* himself. If sent to *M*, the draft would be presented by him to *M* bank for payment; or if sent to *M* bank, the draft would be held while notice was transmitted to *M* that funds were available to him at *M* bank.

This transaction satisfied the obligation of *P* to *M* but created a new indebtedness on the part of *P* bank to *M* bank. This interbank indebtedness might then be settled in any of several ways. Settlement was simplest if *P* bank customarily maintained a positive balance with the remote *M* bank; and the existence of such a correspondent relationship between *P* bank and *M* bank would have been a sufficient reason to select *M* bank as drawee of the draft in *M*'s favor. If no such balance was maintained, *P* bank might now settle its indebtedness by issuing and mailing yet another draft, payable to *M* bank, to some third bank with which it did maintain a balance, that third bank being selected because it was geographically close to *M* bank. Alternatively, if *P* bank maintained no such balance in *M* bank's vicinity, *P* bank would now be obligated physically to transport to *M* bank a mutually acceptable form of currency. In either event, the cost of the transaction was substantial: the costs of shipping bank notes or gold were high, as were the opportunity costs of maintaining non-interest-bearing balances at distant locations. It was to cover these costs that *P* paid to *P* bank a substantial service charge in addition to the face amount of the draft.[21]

In 1864 Congress passed the National Bank Act,[22] reinstituting the rivalry between state and national banking systems that had existed during the nation's first half century. Federal taxes were levied on bank notes issued by state banks in an endeavor to drive the notes, and perhaps the banks, out of existence.[23] Although the 1864 Act required that national banks maintain reserve deposits, it permitted a large fraction of those reserves to be held as deposits in designated "reserve banks" in various major cities; and, because drafts could be issued against these reserves, the national banking system became instrumental in the payments system.[24]

The era was one of rapid technological change in both transportation and communications. The railroads, waterways, and post roads expanded rapidly, frequently under the spur of government subsidies, and the telegraph was invented and deployed. These changes tend to explain the increase in use of transactional paper relative to currency, but it is less clear why the use of checks relative to drafts also increased very rapidly during this period.[25] When a check was used to pay a distant payee, *P*, having a positive balance with *P* bank,

sent the instrument (usually by mail) to *M*, who presented it to *M* bank for collection. Then *M* bank accepted the instrument for collection and might or might not credit *M*'s account with *M* bank for the amount of the check before collection had been achieved.[26] The instrument was started by *M* bank on what was often a circuitous journey from one bank to another until through some series of correspondent relationships it arrived at *P* bank.[27] The check was accepted by *P* bank and debited against *P*'s account. At this point *P* bank again faced the problem of making payment to *M* bank, just as when drafts were used. Again, its costly alternatives were the actual transport of currency or the maintenance of geographically dispersed balances against which a draft in favor of *M* bank could now be issued.

To obtain revenues, *P* bank might have levied a service charge against *P*'s account and made remittance to *M* bank in the full face amount of the check; but this was not the custom. Rather, it was customary to make remittance to *M* bank in an amount less than the face of the check, the discount being called an "exchange charge," a term that reflected the functional similarity of the charge to the prepaid service charge characteristically imposed on *P* in the earlier period when a draft was issued on his behalf. The preservation of that term, however, tended to obscure the important fact that the direct economic incidence of the service charge had been shifted-initially to *M* bank, or to some intermediate bank in the chain which might be willing to absorb the charge, but ultimately to *M*.

Early descriptions of the checking system suggest that the contemporaneous view in the banking community of this shift in incidence was that it reflected an understandable conflict of interests between *P* bank and *P* on the one hand and *M* bank and *M* on the other.[28] But that explanation fails for two reasons. First, the conflict of interests had been present no less during the earlier period when drafts were the predominant transaction vehicle; and old causes cannot explain new effects. Second, the explanation attributes a widespread and persistent pattern of behavior to an erroneous perception, for it implicitly assumes that the checking system could attain equilibrium without regard to the proportion in which banking costs were imposed on *P* and *M* so long as all costs were borne by them in combination. To the contrary, as I argued in Section I, equilibrium in the level of checking services demanded and supplied is possible only with some specific distribution of costs between *P* and *M*.

If the shift in incidence reflected rational business behavior, as I prefer to think it did, then it had to reflect either a change in the relative demands of purchasers and merchants for checking services or changes in the relative costs of *P* bank and *M* bank in providing them. Several contemporaneous developments support the inference that such shifts actually occurred.

The advent of faster and cheaper transportation and communication had two consequences for the supply costs of transactional paper. First, it reduced the banking system's aggregate direct costs of processing checks and, when necessary, transporting currency. Second, because they tended to convert local markets into regional and national markets, these cost reductions greatly increased commercial transactions between remote parties. This increase in the volume of distant transactions enabled banks to exploit scale economies in maintaining balances at distant locations; for, given the law of large numbers, higher turnover velocities in those balances could be achieved with disproportionately small increases in the magnitude of the balances. This factor, too, must have contributed to a reduction in average cost per transaction.

In addition, although under the draft system *P* contributed substantially to bank revenue by purchasing "exchange," those transactions imposed large indirect costs on *M*: the cost of the float during the slow process of paper interchange and the cost associated with the risk of default. In addition to the reductions in direct cost brought about by better transportation and communication, these indirect costs to *M* would also be significantly reduced by shortening the period of float, by providing cheaper access to credit references, and by reducing the costs of collecting delinquent obligations. Hence, even if there had been no reduction in aggregate direct costs, the redistribution of those direct costs toward *M* might well have been necessary to attain equilibrium in view of the reduction of *M*'s indirect costs.

Finally, the widespread emergence of clearinghouses also significantly reduced direct costs and accelerated the process of interchange, further reducing float costs.[29]

For some or all of these reasons it seems to have been necessary for the industry to redistribute the direct costs of the checking system away from *P* and toward *M* so that the market for transactional paper could equilibrate. That need may itself best explain the relatively sudden displacement of the draft by the check. A new and less familiar instrument, the check was accompanied by

fewer customs and fixed expectations than the more familiar draft. And the check, although very similar to the draft in most respects, passed through the hands of the four parties in a different sequence, a sequence that tended to enhance monopsonistic position of *P* bank as a buyer of paper.

As Figure 4 demonstrates, if the level of total banking costs (and therefore the values of p^* and q^*) changed significantly, then no change in the aggregate demand curve of *P* and *M* would be necessary to change the relative magnitudes of their individual demand levels for use of a payment system. It is well established that from the Civil War to the end of the nineteenth century p^* fell by a considerable amount and q^* increased enormously.[30]

The clearinghouse seems to have had consequences beyond mere reduction of costs to the banking system. With increasing urbanization of the nation, many banks found themselves in cities served by many other banks. The local clearinghouse—at which each bank in its role as *M* bank would transfer to every other bank in its role as *P* bank a bundle of checks, packaged and tallied in advance-had enormous potential for reducing the costs of the payment system by expediting both presentment and remittance. Interbank debits among clearinghouse members could be netted out on the books of the clearinghouse; and actual payment, usually made to the clearinghouse, was necessary only intermittently to the extent that an individual bank's presentment over a period of time had aggregated more or less than the aggregate, over the same period, of its remittance obligations.

Clearing arrangements were negotiated not only among banks in individual urban areas but also between banks in widely separated urban areas. These intercity arrangements were often bilateral agreements by which one large bank in the first city would accept for forwarding to all other banks there checks gathered in the second city by the other large bank from all other banks located there.

These clearing arrangements were significant because they both reduced the cost per item substantially and encouraged standardization. Because of the large number of items involved and because cost reductions depended heavily on use of routinized procedures for assembling the items in batches and tallying the totals for the items in each batch, it was highly desirable that every item be susceptible to handling in the same routinized way.[31] If different exchange charges were to be charged on different items by different *P* banks—charges not appearing on the instruments—handling procedures would be complicated.

Moreover, many banks were indifferent whether exchange charges were low or high or even made at all. The typical bank presented to other banks about the same volume of items as were presented to it; and for such a bank the aggregate of exchange charges represented a wash. The increased administrative cost of accounting for different exchange charges on different individual items constituted a useless cost for such a bank. Therefore, there was a strong incentive to standardize such charges, and fixing them at zero was an obvious and entirely acceptable form of standardization.

For these reasons, many banks agreed to handle each other's items "at par"—that is, to make no exchange charges. For similar reasons, many clearing organizations required their members to remit at par on all items sent through the clearing arrangement.[32]

An exchange charge equal to zero obviously has no unique potential for cost reduction; any uniform exchange charge would have facilitated routinized processing. Any advantage of a zero price over others is rooted less in economics than in psychology.[33]

Parties to individual items on which varying amounts of exchange would be charged when they reached *P* bank were at a disadvantage in competing with parties to items eligible for routinized clearance. Clearance mechanisms tended to get a check from *M* bank to *P* bank via quite direct paths, but items on which exchange charges were due tended to follow slow and circuitous routes.[34] Each bank would prefer to transfer the item to another bank with whom it had negotiated a bilateral arrangement to remit at par than to send to *P* bank, which would impose exchange charges. Consequently, both float and handling costs were relatively greater for items with nonstandardized exchange.

Notwithstanding the advantages of uniform (perhaps uniformly zero) exchange charges, a very large number of banks strenuously resisted remitting at par. The banks that continued to charge exchange into the twentieth century were, almost without exception, small banks in isolated agricultural communities. For the banks that adhered to this practice, revenue in 1964 from exchange charges constituted about 10 percent of total current

operating revenue, and the percentage was higher for the smaller institutions among the group.[35] It seems likely that in the late 1800s and early 1900s, when the nonpar controversy was at its height, this form of income was even more important to the small country bank.[36]

There are at least two possible explanations of how these rural banks benefited from charging exchange. One is that, even though they charged exchange in their role as *P* bank, they managed to collect at par in their role as *M* bank. No doubt this explanation is at least partly correct, for banks that did not remit at par were not, for that reason alone, prohibited from forwarding for collection items drawn on banks that did remit at par via a correspondent bank through the Federal Reserve clearing system, and the same may have been true of some earlier, private clearance systems. But because remittance at par, at least generally, was a reciprocal practice, it seems unlikely that this was the whole explanation. Moreover, although this hypothesis tends to explain why some banks clung to the practice and might, when coupled with another factor I address hereafter, tend to explain why the practice was most common for banks in isolated communities, it does not explain why the practice should have been confined so largely to isolated agricultural communities, rather than, for example, mining communities.

A different factor must have been at work. The amount of exchange charged was customarily a percentage of the face value of the item. But a minimum charge, often ten cents, was charged on all items having a face amount of $100 or less, and $100 was a large sum then. A bank benefits from charging exchange if, notwithstanding that its aggregate dollar volume of remittances roughly equals its collections, a larger number of small items are presented to it than it presents to other banks. In isolated agricultural communities, the receipts of the farmers, who constituted the local depositors, probably took the form of several large payments at harvest time. On the other hand, farmers more nearly resemble nonfarmers in their purchase patterns, for they engage in personal consumption and the purchase of farm supplies throughout the year. And, of course, the magnitude of most individual purchasers must be much smaller than the magnitude of the small number of income items. Although apparently no data exist that would constitute hard evidence for this hypothesis, it is the only explanation that enables me to make sense of the available information about the nonpar controversy.

Why nonpar practices tended to be confined to small isolated communities is more obvious. A situation in which one or more nonpar banks occupied the same market with one or more par banks is inherently unstable. It had always been an unambiguous understanding about any bank's obligation on a check that payment had to be made at full face value if the check were presented for payment at its banking premises. If there was a par bank in the same areas as *P* bank, *M* bank would forward items drawn on nonpar *P* bank to that neighboring bank so as to avoid exchange costs; and the neighboring bank would present such items at *P* bank's premises. Hence, the conversion from nonpar to par of any one bank in an area usually led to the conversion of all in the area. Nonpar banking thus survived primarily in isolated communities able to support only one, or a few, banks. However, in the early twentieth century it was Federal Reserve pressure, not competition, that reduced the practice of charging exchange to a trivial level; where the practice survived it was state legislation, not monopoly enclaves, that sheltered it.

After the monetary panic of 1907, a national monetary commission was appointed to study the American banking system.[37] Its report led to the passage of the Federal Reserve Act in 1913.[38] This legislation, its subsequent amendments, and the practices and rules of the Federal Reserve Board, which the legislation created, eventually tipped the balance in favor of par clearance in the United States. It was not obvious from the initial legislation that this outcome would result, nor is there any reason to believe that the practice of nonpar banking particularly concerned either the National Monetary Commission or the Congress of 1913.[39] The key provisions of the Federal Reserve Act were sections 13 and 16. Section 13 initially read, in part:

Any Federal reserve bank may receive from any of its member banks... deposits... or, solely for exchange purposes, may receive... checks and drafts upon solvent member or other Federal reserve banks, payable on presentation.[40]

Section 16 read, in part:

Nothing herein contained shall be construed as prohibiting a member bank from charging its actual expense incurred in collecting and remitting funds, or for exchange sold to its patrons. The Federal Reserve Board shall, by rule, fix the charges to be collected by the member banks from its patrons

whose checks are cleared through the Federal reserve bank and the charge which may be im- posed for the service of clearing or collection rendered by the Federal reserve bank...[41]

Section 16 is silent on the practices of nonmembers. It preserves the right of members to impose costs on their check-writing depositors and implies obliquely that language elsewhere in the Act might be read to curtail member *P* bank's ability to charge exchange to *M* bank; but no curtailing language is to be found elsewhere. The power vested in the Reserve Board to standardize fees for clearance or collection at a level other than zero has never been exercised.

More generally, the Act provided that the Federal Reserve Board would establish a check clearance system throughout the United States, each federal reserve bank being required to act as a clearinghouse for member banks in its region. After establishing this system, the Fed began to establish more pervasive clearing mechanisms. Funds for the clearance system were available, for the Act also required member banks to deposit substantial reserves with federal reserve banks in accounts bearing no interest.[42] Deposits, however, were invested in government securities; and the investment yield constituted a very substantial source of funds to the system. It seems clear that the clearance systems established by the Fed were largely subsidized by these earnings. Although member banks did not receive a "free" clearing system—the forgone investment yield on their reserve deposits paid for it—the Fed clearing system was available to members at a price included in the sunk cost of maintaining the required reserves. The alternatives (to continue using private clearinghouses or to establish a new, private, interregional clearinghouse) would have required that member banks bear the full system costs in addition to the cost of maintaining reserves with the Fed. Accordingly, the economic incentives for member banks to use Fed clearing mechanisms were strong.

The incentive for member banks to use the Fed's clearance system, coupled with the Fed's requirement that member banks remit at par against items presented to them through the clearance system, served as a significant direct force in the adoption of clearance at par by member banks. This same force operated, albeit indirectly, on nonmember banks. Member banks were allowed to forward through the system for collection not only checks drawn on other member banks throughout the nation but also checks drawn on such nonmember banks as had agreed to remit at par. In order to identify for member banks those nonmember banks whose checks could be sent through the Fed clearance system, the Fed began regularly to publish the "par list," a complete state-by-state list of all nonmember banks that had agreed to remit at par. In addition, from the beginning of the system nonmember banks could use the Fed clearing system by forwarding acceptable items through correspondent banks that were member banks; but in this context, too, a check drawn on a bank not on the par list was not an acceptable item. Such checks had to be cleared outside the system and were denied the benefits of subsidized clearance.

In 1916 Congress amended section 13. Because the Act initially authorized any federal reserve bank to "receive . . . for exchange purposes . . . checks and drafts upon . . . member or other Federal reserve banks," some doubt existed whether checks on nonmember banks could be received.[43] The clause was amended to read: "Any Federal reserve bank . . . solely for purposes of exchange or of collection, may receive . . . checks and drafts, **payable upon presentation within its district. . . .**"[44] Congress thereby made clear that the federal reserve banks were authorized to accept from their member banks checks drawn on nonmember banks.[45]

Notwithstanding these various enticements, many banks refused to remit at par and stayed outside the federal clearance system.[46] To entice or coerce more banks into its clearance system, the Fed in 1916 made its system mandatory for all member banks with respect to items drawn on them, but the system remained voluntary with respect to items forwarded by them.[47] And nonmember banks on the par list were permitted to ship funds for the purpose of clearance to the Fed at the Fed's expense. Thus a subsidy was employed to expand the par list of nonmembers.

In 1917 Congress further amended section 13 by adopting the "Hardwick Amendment," which added the language, "Nothing . . . in this Act shall be construed as prohibiting a member or nonmember bank from making reasonable charges, to be determined . . . by the . . . Board, but in no case to exceed 10 cents per $100 or a fraction thereof, based upon the total of checks and drafts presented at any one time, for collection or payment . . . but no such charges shall be made against the Federal reserve banks."[48] In its annual report for 1917, the Fed said of the Hard- wick Amendment and its legislative history:

An effort was made, in the interest of some member and non-member banks to amend the Act by providing for a standardized exchange charge, not to exceed one-tenth of 1 percent, to be made by member banks against Federal reserve banks for checks sent for collection. It was not successful, and the Act as finally amended provides that a member or non-member bank may make "reasonable charges to be determined... by the... Board... ; but no such charges shall be made against the Federal reserve banks." The Attorney General has been re- quested to give his opinion as to whether this proviso applies to non-member banks. An affirmative opinion will make possible the establishment of a universal par clearing system, but if, on the contrary, it should be held that the proviso applied to member banks only, the further development of the collection system will necessarily be slow, and in the absence of further legislation will depend upon the voluntary action of many small banks.[49]

This comment is noteworthy in two respects. First, it tends to support the view that standardization of exchange charges was seen as a means, alternative to par payment, to facilitate the clearance process. Second, it reveals that the Fed as early as 1917 perceived that the last twelve words of the amendment, if "favorably" interpreted by the attorney general, could be used to coerce a general abandonment of any exchange charges—making "possible the establishment of a universal par clearing system"—and thus achieving standardization of a special kind.[50]

In 1918 the Fed dropped all per item service charges for using its clearance system. It also began operating a leased telegraph system (the "Fed Wire") between all federal reserve banks, the Fed, and the Treasury. The use of the Fed Wire was made available to member and par-list banks to adjust clearing balances. Despite this additional carrot, there remained at the end of 1918 about 20,000 nonmember banks, half of which also remained off the par list.[51]

In 1918 the Fed succeeded also in obtaining from the attorney general an opinion that in effect prohibited precisely what the Hardwick Amendment seems, at first glance, to have permitted. Focusing on the last few words in the Amendment, the attorney general ruled that the federal reserve banks were prohibited by law from paying, even in the sense of passing on, exchange charges in the course of the clearance process.[52]

Since, in the period under discussion, the system would not accept items drawn on nonmember banks not on the par list, the clause, even thus interpreted, would appear to have been inconsequential. But the Fed made it of consequence in 1919, adding substantially to the number of banks on the par list by introducing a new coercive device.

It began to accept for clearance items drawn on nonpar banks and then to demand that they be paid at par. If that request was refused, as it often was, the local reserve bank gathered up the checks of the nonpar bank and presented them at the bank's premises ("at the window"), demanding payment in full in currency.[53] This tactic proved to be very powerful while it was available to the Fed. It has always been regarded as the legal obligation of P bank to P to pay in full on demand if an item was presented at the window;[54] only with respect to items presented through the mails had banks asserted the right to remit at discount. The batch presentation of checks in the manner described often required more currency than the bank had in its vault; yet if payment in full was not made, the checks could be returned to the depositor dishonored, placing the drawee bank in violation of its contractual obligation to its customer. Through this tactic the Fed succeeded in forcing many recalcitrant banks onto the par list.[55]

Commenting on its endeavors in its annual report for 1919, the Fed said:

[The] proviso in Section 13... has been constructed by the Attorney General... as meaning that a Federal reserve bank cannot legally pay any fee to a member or non-member bank for the collection and remittance of a check. It follows, therefore, that if the Federal reserve banks are to give the service required of them under the provisions of Section 13 they must, in cases where banks refuse to remit for their checks at par, use some other means of collection, no matter how expensive.

The action of the various Federal reserve banks in extending their par lists has met with the cordial approval the Federal Reserve Board, which holds the view that under the terms of existing law the Federal reserve banks must use every effort to collect all bank checks received from member banks at par. Several of the Federal reserve banks are now able to collect on all points on their respective districts at par, and new additions to the other par lists are being made every day. The board sees no objection to one bank charging another bank or a firm or

individual the full amount provided in Section 13 of the Federal Reserve (10 cents per $100) and has not undertaken to modify these charges, but the Act expressly provides that no such charge shall be made against the Federal reserve banks.[56]

The legality of this practice by the Fed was challenged in the courts. While the cases were making their way to the Supreme Court, a number of states, mostly in the rural Southeast, passed legislation providing that a state bank should not be deemed to have dishonored a check—that is, to have violated its obligation to its depositor—if it refused to accept the check merely because exchange would not be paid.[57] The constitutionality of these state statutes was also challenged on preemption grounds.[58]

The two groups of cases made their way to the Supreme Court, which in 1923 held, first, that in the absence of the state statue prohibiting its practice, the Fed was authorized to employ the tactic of making presentment at the drawee bank window[59] and, second, that the state statutes prohibiting the practice were also constitutional.[60] Thus nonpar banking continued to be sheltered in those few states that chose to adopt such statutes but substantially disappeared elsewhere. At the end of 1964, there were 1,547 nonpar banks in fourteen states, but their deposits accounted for only about 2 percent of total deposits in FDIC-insured institutions.[61] On April 1, 1980, there were only fifteen nonpar banks left in the United States.[62] All these banks were located in Louisiana. By September 1980 all but one of these had become par banks.[63]

Thus the role of the interchange fee in the process of check clearance, a commercial context in which an unregulated market solution might have been expected to work reasonably well and to yield instructive results, was aborted and continues to be suppressed by a mixture of subsidies and coercion by the Federal Reserve System.

B) BANK CREDIT CARDS AND THE INTERCHANGE FEE

About a century passed between the date the check gained common acceptance and the date another four-party payment instrument—the bank credit card—was introduced. The precursors of the bank credit card were the retail merchant's open book account and later the travel and entertainment card.

For centuries merchants have extended short-term, interest-free credit to customers whose patronage is highly valued. The shopping behavior of customers varies widely, and those behavioral differences make transactions with some customers more profitable for the merchant than transactions with others. A customer whose own time costs are high will tend to shop regularly at a particular retail outlet because of its geographic proximity to his other activities, and he will tend to shop when it is convenient for him rather than waiting for occasions when merchandise is on sale. He will tend to shop on fewer occasions and buy a larger number of items on each occasion. He will consume less time of sales personnel because he is attempting to save his own time, and he will be able to decide more quickly because he conceives his quest to be locating the items he wants rather than making closely balanced trade-offs with reference to price. Finally, he will tend to buy higher-priced items, which are likely to carry higher percentage markups and are certain to carry higher absolute dollar markups.

There is a strong although not perfect correlation between customers with high time costs, high incomes, and high wealth positions, so the default risk of extending credit to such customers is also relatively low. For all these reasons merchants have long used the selective extension of open book credit as a competitive tool by which to attract and retain the patronage of such customers.

The customer to whom open book credit was extended, having purchased on various occasions during the month, received by mail at the end of the month a bill in the face amount of his purchases; soon thereafter, he would remit payment by mail. On the average mid-month purchase, the merchant was absorbing the cost of capital for about three weeks. The merchant thus remitted to these customers in a fairly direct way part of his cost savings attributable to their shopping behavior; he also conferred minor indirect cost savings by reducing the customer's need to carry cash on his person.

Open book credit well served the parties affected while travel outside one's home community was relatively infrequent. After World War II, the frequent traveler was likely to have a high income and high time costs and therefore to have been extended open book credit in his own community; but away from home he could not readily be identified at the point of sale. He could carry large amounts of cash, but the risk of loss was substantial.

Traveler's checks were an alternative, but they involved high time costs because they required the traveler, first, to visit the bank before departing and, second, to predict with reasonable accuracy how much money would be needed during the trip or to make another journey to the bank on return to redeem the excess checks, or to leave funds tied up on a non-interest-bearing certificate until a later time when the traveler's checks might be used. A second alternative—attempting to cash personal checks at one's destination—involved tediously presenting identification at a moment when time costs were likely to be greatest; not infrequently, the attempt was humiliatingly unsuccessful. From the standpoint of the merchant located at the traveler's destination, the situation was also unsatisfactory. If the merchant could easily identify the traveler as a creditworthy consumer with high time costs, he would be only too happy to extend to the traveler the same credit facilities extended to comparable local customers.

The first commercial response, in the early 1950s, to this obvious transactional need was the travel and entertainment (T&E) card, notably the American Express card and the Diner's Club card. The issuing organization signed up merchants across the country of the type frequently patronized by travelers: hotels, resorts, restaurants, and a relatively small number of prestigious merchandise outlets. After investigating an applicant's creditworthiness, it issued a card for an annual fee that would tend to make the card attractive only to persons who traveled relatively frequently. Thus self-selection as well as the financial eligibility criteria of the issuer combined to produce the result that only persons with relatively high incomes and high time costs were likely to use the card. Thus, having a T&E card signaled to the distant merchant that the holder had the same income and consumption characteristics that induced the merchant to extend open book credit to local customers.

The issuing organization bought the transactional paper thus generated at a discount. Even though by present bank-card standards this discount was relatively large, the relation was worthwhile to the merchant: the system not only enabled the merchant to identify a new group of high-income customers and compete for their patronage but also protected him against default risk, performed billing and collection, and, perhaps most important, eliminated the capital costs of extending credit during the billing cycle.

Because the T&E card was a three-party instrument rather than a four-party instrument, the feature of jointness was present on the demand side but not on the supply side. Again, there was one particular distribution of costs between the merchants and the card holders that would bring their demands for the transactional service into equilibrium. But the card-issuing organization was a single enterprise; periodic adjustment was within its control, and there was no problem of coordinating two enterprises to determine how to distribute charges between card holders and merchants.

The national T&E cards were not the only three-party transaction cards that appeared during these years. Many major oil companies distributed similar cards, but their merchant base was limited primarily to their distributors. A number of banks also distributed three-party cards. Although these cards were accepted by a more heterogeneous set of merchants, their use was limited to the geographic region to which the banking laws limited the bank's deposit-accepting activity. One of the most successful three-party bank cards was BankAmericard. The Bank of America, enjoying the advantage of a large and populous state with relatively permissive statewide branching laws, was able to reach more card holders and merchants than most other three-party bank-card systems.

Several characteristics of the late 1950s and early 1960s set the stage for the introduction and rapid expansion of the four-party bank credit card. Those were years of relatively rapid growth in real income in the United States. The number of high-income, high-time-cost persons increased rapidly, as did the number who traveled frequently outside their own community. Simultaneously, data processing and electronic communications experienced dramatic technological advance, which enhanced the demand for transactional services and, on the supply side, significantly reduced the costs of maintaining accessible documentation on creditworthiness and of billing and collection.

Moreover, as nominal interest rates began to rise by the late 1960s, interest costs became a larger fraction of the total cost of extending consumer credit. The comparative advantage of banks and other financial institutions over all but the very largest of the retail chains became ever more decisive as interest costs predominated in the total cost of performing the retail credit function. Finally, there were scale economies from consolidating one consumer's transaction with a number of merchants into a single statement, a single billing, and a single remittance.

All these factors favored substituting bank-card systems for the traditional merchant function of extending retail credit.

The four-party bank credit card was introduced in 1966 in order to obtain for bank-card payment systems a ubiquity that, by reason of our geographically restrictive banking laws, could not be obtained by any single banking enterprise in its deposit acceptance activities. In that year the Bank of America licensed its "BankAmericard" service mark on a nationwide basis. Licensees were authorized to issue cards bearing the logo, to sign up merchants who would accept the card in the area of the licensee's operation, and to engage other banks as agents to expand the merchant base still further.

At about the same time, under the leadership of the major Chicago banks, the Midwest Bank Card system was established as a joint venture among a number of banks in the Great Lakes area. Shortly thereafter, the Interbank Card Association was formed as a nonprofit membership organization owned by its card-issuing member banks. Its initial purpose was to provide nationwide interchange facilities to a number of regional bank card systems. Among these local programs was the Western States Bank Card Association, which owned the "Master Charge" service mark. In 1969, after that card association had joined InterBank, the Master Charge mark was assigned to InterBank and then licensed to all InterBank members. Thus within three or four years, today's major bank-card systems made their appearance. In 1970 the BankAmericard system changed its structure to that of a membership corporation; in 1977 the name of the national organization changed to "Visa" and exclusive rights to the name "BankAmericard" reverted to the Bank of America.

These organizational changes did not alter the fundamental point that these multibank organizations were from their inception four-party systems having the peculiar economic characteristic previously described. Given the distribution of charges between *P* and *M* that would achieve equilibrium in their demands, it was overwhelmingly improbable that the revenue stream from *M* to *M* bank or from *P* to *P* bank would equal the costs of the subset of activities that a particular bank was required by the technology of the payment system to perform; thus some redistribution of those revenues between *M* bank and *P* bank was likely to be necessary for the payment system to compete effectively with alternative mechanisms.

Hence, half a century after Fed coercion resolved this problem of redistributing revenues in the context of four-party check clearance transactions, the bank-card systems confronted the question how to determine the appropriate magnitude of the necessary transfer payment between *M* bank and *P* bank. It makes no difference when addressing this question in the abstract whether the transfer payment is made by card-issuing banks to merchant banks or by merchant banks to card-issuing banks; I will assume, as recent cost patterns suggest, that income from card holders is too small for the average card-issuing bank to cover its costs, whereas income from merchants is, on average, more than sufficient for merchant banks to cover their costs. As shown in Section I, given the assumption about competitive equilibrium stated there, the magnitude of the deficiency must equal the magnitude of the surplus; I will refer to that magnitude as the optimum transfer fee.

The monopsonistic position of *P* bank—which is determined by the direction of the paper flow and hence would be present even if the transfer fee had to move in the opposite direction—implies that each *P* bank cannot be permitted to announce daily the price at which it will buy paper to be billed to its card holders. If a system involved very few *P* banks and *M* banks, bilateral agreements could be negotiated between each *P* bank and *M* bank, and each agreement could establish for some substantial period of time the magnitude of the transfer fee. This approach has two substantial drawbacks in practice. First, the number of agreements to be negotiated in each time period is equal to the product of the number of *P* banks and the number of *M* banks; second, and probably more important, there is a significant free-rider problem that increases with the number of participants.

Imagine a card system composed of ten *P* banks that act only as purchaser banks and ten *M* banks that act only as merchant banks. Assume that each *P* bank receives from each *M* bank 1 percent of the aggregate paper flow of the system and has 10 percent of the aggregate card-holder base. Assume, finally, that the optimum transfer fee is 1 percent of the face value of the paper and that this fee amounts to $0.30 per item. Although it is subversive of the system as a whole to demand a higher fee, each individual *P* bank faces a strong temptation to do so—let us assume a 10 percent increase in the transfer fee to 1.1 percent, or $0.33. Any individual *P* bank that so behaves, provided that it is unique in demanding an excessive fee, will increase its fee revenues by about

10 percent but will increase the effective costs confronted by each *M* bank only by 1 percent. Even assuming that the *M* banks immediately pass on this cost differential, the merchant discount would be increased by 1 percent on the paper of all *P* banks, for it is not feasible for the *M* banks to discriminate against paper en route to that particular *P* bank without creating, on the part of all the merchants, an incentive to refuse to honor cards issued by that *P* bank; moreover, any endeavor by all merchants selectively to refuse cards issued by a particular *P* bank (at least outside the context of an on-line electronic system) would substantially increase the transaction costs of all merchants and of all card holders. The utility of the system to all participants would diminish, as would the system's viability in competition with other payment systems.

Similar, although perhaps less immediately dramatic, consequences would follow if either the set of *M* banks or the set of merchants chose to absorb the percent cost increase that flows from *P* bank's 10 percent increase in the transfer fee. Some might drop out of the system entirely because of economic losses; others would alter their behavior in less drastic ways to shift from using the card system to using some other payment systems. These adverse consequences would eventually reduce the transaction volume of the individual *P* bank that raised the transfer fee, but the adverse effect would be spread across all *P* banks. The one *P* bank would realize 100 percent of the revenue gains from its fee increase but would bear only 10 percent of the adverse consequences. More generally, in a card system involving *x* number of *P* banks, any one bank can exploit the monopsonistic position it enjoys over its own paper and can realize 100 percent of the revenue gains while suffering only a fraction of the adverse consequences, that fraction being 1/*x*. Accordingly, it is essential that the participants in a four-party payment system collectively adopt some internal mechanism that prevents individual exploitation of the monopsony power endemic to such systems.

As discussed earlier, banks were prevented from exploiting their monopsonistic power in the checking system initially by collective agreements among clearinghouse members and later by the Fed's coercive tactics. But the problem was resolved for the checking system without explicit recognition of the problem's characteristics, without any inquiry into the costs of the system, at the apparently arbitrary transfer fee of zero, and largely by government coercion rather than agreement. These all make it unlikely that the resolution was optimum when first made, even less likely that the resolution could have continued to be optimum after the enormous changes in check-processing technology. Compared to the checking system, the bank credit card system has evolved so far under less government intervention with respect to the transfer fee. Perhaps for that reason, perhaps also because there are many institutions for which items transmitted in their capacity as *M* bank are unequal to items received in their capacity as *P* bank, behavioral characteristics of those payment systems more closely correspond with the behavior implied by the theoretical considerations discussed in Section I.

Before those transfer fee arrangements are examined, two important differences between the checking system and bank-card systems should be noted, differences that significantly affect the cost to the parties. First, under the checking system, *M* bears the risk of default: if funds adequate to cover the check are not on deposit at *P* bank when the instrument arrives for payment, the check is dishonored and charged back through the clearance system against *M*'s account with *M* bank. But under the bank-card system, provided that *M* complies with the prescribed authorization procedures, *P* bank guarantees payment by the card holder and thus bears the risk of default. This shifting of risk under the bank-card system obviously increases *P* bank's cost, enhances *M*'s demand for the system, and increases the amount of discount *M* is willing to pay to *M* bank. Thus, one would expect to observe larger transfer fees from *M* banks to *P* bank than those in the checking system.

The second basic difference between the checking and bank-card systems also has the effect of increasing *P* bank's costs of the bank-card system. Because a check forwarded to *P* bank is debited immediately against funds on deposit, *P* bank incurs only minor float costs. Whatever float costs remain are borne either by *M* bank (if it credits *M*'s account on deposit) or by *M* (if his account with *M* bank is not credited until funds are remitted). Float costs under the bank-card system are borne in different proportions from those under the checking system and are substantially greater. The paper generated by the card holder is not issued against any existing deposit with *P* bank; remittance is made by *P* only at the end of the monthly billing cycle. Unlike the check clearance cycle, which takes only a few days, bank-card items will on average be outstanding on *P* bank's books for two weeks before *P* is sent an accounting statement and for about three and a half weeks before *P*'s remittance is received.

Clearly, *P* bank bears the cost of this extended period of float, but the incidence of the corresponding benefit on demand is ambiguous. In comparison with use of a currency or a check method of payment, *P* is the beneficiary, and his demand for the bank-card system should increase. On the other hand, to the extent that the bank-card system is being used by *P* and *M* in lieu of open-book credit, it is *M* whose float costs have been reduced, and his demand should be enhanced.

Before turning to the messy world of reality, it is useful to ask what one would expect to find there, reasoning from the theoretical joint demand and supply model developed in Section I. Both *M* and *P* banks will be incurring activity costs, and both will be receiving a revenue stream. Because the revenue stream of each probably will not equal its cost stream, one would expect to observe some side payment that will bring the net revenue stream of each bank, after the side payment, back into the same proportion with respect to its cost stream as the proportion between total revenue and total bank costs. Obviously, any side payment that brings those ratios into equality for the two banks (or sets of banks) has the same effect. Equally obviously, the value of all these ratios will, in competitive equilibrium, equal one.

With these features in mind, one can attempt to derive by arm-chair empiricism a picture of both the demand and the supply sides of the bank-card industry as revealed by present behavior. So far as demand is concerned, there is unmistakable evidence that a positive demand exists on the part of many merchants for bank-card services; and, although the evidence is less clear, there are persuasive reasons to believe that a demand exists also on the card holder side and that it also is positive at prevailing transaction levels. No direct observation of the contours of these demand functions is possible; we catch glimpses of segments of the functions only as demand is revealed by the willingness of merchants and card holders to pay for bank-card services. Thus, in our endeavor to explore demand functions, we are led to examine the charges that banks have historically imposed on merchants and card holders.

Before nominal interest rates skyrocketed in early 1980, the bank-card industry imposed substantially all the costs or bank-card transaction services (as opposed to financing services, a distinction developed hereafter) on merchants. Since each merchant bank is free to negotiate whatever arrangement it can with its own set of merchants, enough variance exists among arrangements to make generalization difficult. Typically, though, merchant discounts have been between 2.25 and 3 percent of total transaction dollars, the discount being higher for merchants who have smaller aggregate dollar volumes or who have smaller average dollar amounts per item. To facilitate discussion I assume where precision is not essential that the typical merchant discount is 2.5 percent.

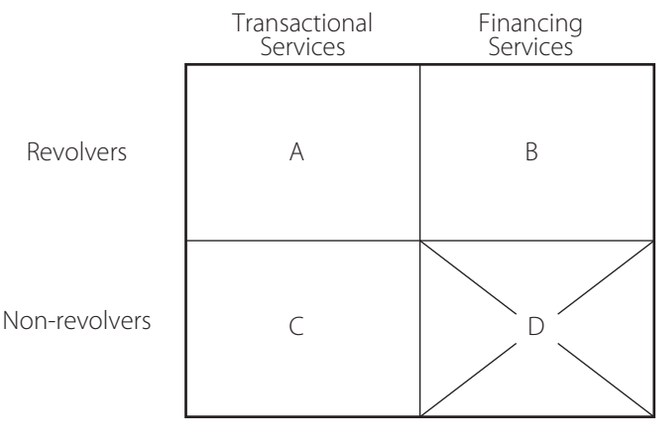
With exceptions to be discussed later, no charge has been imposed on the card holder. In this context, too, each card-issuing bank is free to negotiate such arrangements as it wishes with its card holders. Before 1980 only a few card-issuing banks had imposed either transaction fees or periodic "membership" fees on their card holders; in the overwhelming preponderance of instances, banks have been willing to play the role of *P* bank as a competitive gambit to attract the individual demand deposits of its card holder. Until recent regulatory reform permitted banks to pay interest on demand deposits, the value to the card-issuing bank of attracting incremental individual demand deposits on which no interest was or could be paid was a sufficient inducement, at least when coupled with the interchange fee received from the merchant bank, to compensate *P* bank. Thus, although revealed demand plainly exists on the merchant side, it is less clear on the card holder side.

The picture is complicated on the card-holder side by the fact that the bank credit card historically has not been merely a payment mechanism. The card holder has had the option of paying, at the end of a billing cycle, only a minor fraction of the charges incurred during that billing cycle and deferring payment of the preponderant portion of the balance. But if he does "revolve" his account in this way, interest payments become due not only on the balance deferred, but also on each new charge subsequently incurred until the balance is, at the end of some billing cycle, reduced to zero. In short, card holders who revolve their accounts not only pay interest on the deferred balances but lose the advantage, available to those who do not revolve, of about three weeks "free" float on current transactions.

Thus the card-issuing bank can be viewed as engaged in two different businesses. It sells a transaction service involving valuable float to those "nonrevolvers" who choose to pay their statement in full at the end of each billing cycle. It also sells a combination transaction service and consumer finance service to those who use their bank cards as an extended credit mechanism.

Because certain activities essential to providing the payment service—receipt of interchange items, posting to individual card holder accounts, billing, collection, posting of credits, bearing the risk of default, etc.—must be performed with respect to revolvers as well as nonrevolvers, complex accounting allocation problems arise.

Several different views of the bank-card industry can be taken. Figure 5 will aid in distinguishing the possible views and the accounting differences that seem to follow from taking one view rather than another. The alternative views present the industry as engaged in only one business or in two different businesses. If the industry is thought to be in two businesses, there are alternate ways of defining those two businesses. If two or more business segments are truly joint (in the sense that one set of services cannot be rendered economically without simultaneously performing the other), it is pointless and potentially misleading to regard them as separate businesses. Equalization of both *P* bank and *M* bank revenue-to-cost ratios throughout all such segments is our theoretical expectation. If jointness in that sense between any two segments is not present, then one should expect to observe an endeavor, first, to engage in cost allocation and revenue allocation as between such disjoint segments and, second, to observe an endeavor to equalize, within each of those segments, the revenue-to-cost ratios of the two sets of banks. The significance of disjointness is that, should the system-wide revenue-to-cost ratio for one such segment consistently fall below the value of one while the ratio for the other segment exceeded one, the former activities would be abandoned as a commercial failure and the latter activities would be continued.



As the matrix in Figure 5 illustrates, the industry provides three distinct services: transaction services to revolvers (cell A), financing services to revolvers (cell B), and transaction services to nonrevolvers (cell C).

One possible "two-business" view separates activities according to the type of service so that the provision of transaction services to revolvers and nonrevolvers is one business, the provision of financing services to revolvers a second. From an accounting standpoint, this view suggests a cost allocation to cell B of (1) the interest cost of the outstanding balances of revolvers; (2) the incremental billing and collection costs, if any, associated with the extended credit function (as distinguished from those associated with the payment mechanism function); and (3) the incremental costs, if any, of risk of default or fraud associated with the extended credit function (as opposed to the payment mechanism function). Under this view, the periodic interest charge to revolvers would be set at a level just sufficient to cover that set of incremental costs. The costs associated with the payment system features of the card, for those transactions engaged in by card holders who regularly took advantage of the extended credit feature and for those transactions by nonrevolvers, would be regarded as payment system costs that would be covered by some other revenue stream, which might consist of the merchant discount or a separately identifiable charge imposed upon all card holders, such as a periodic membership charge or a per-item charge or a per-dollar volume charge. This first view involves the difficult problem of deciding the extent to which bookkeeping costs and risk costs are incrementally associated with the extended credit function.

Alternatively, one could view the industry as being engaged in two businesses but, rather than linking cell A with cell C and defining cell B to be the separate business, this second view links cell A with cell B and defines cell C to be a separate business. This view defines the two businesses with reference to card holder payment practices. One business consists of providing transaction and financing services to revolvers; another consists of providing transaction services to nonrevolvers. The implied accounting allocation problem is to allocate each category of banks' activity costs either to revolvers as a group or to nonrevolvers as a group. Under this view, the cost allocation problem is to associate some fraction of total bookkeeping costs and total fraud and default costs with habitual revolvers and the remaining fraction with habitual nonrevolvers.

For habitual revolvers, there are three possible revenue sources: periodic interest charges on outstanding balances, the merchant discount, and other card holder charges such as membership or per dollar fees. For nonrevolvers, only the two latter revenue sources are available.

A third view is that the industry engages in a single business. No cost allocation is attempted; three possible revenue sources previously identified are seen as being available to cover all costs.

From a theoretical standpoint it seems clear that cells B and C are disjoint. One can readily conceive of a bank-card service that did not offer the extended payment feature. Although nothing resembling the financing service that is provided to revolvers would be possible unless a transaction service was being rendered as well, it would be possible for banks to render transaction services without providing financing services. The T&E cards typically do just this. Accordingly, sensible business practice requires that the avoidable costs of the extended credit activity be ascertained and compared with the incremental revenues to assure that a revenue-to-cost ratio of not less than one exists. But if incremental revenues equal or exceed incremental costs, the extended credit function is commercially viable so long as transaction services continue to be provided: no more stringent test—for example, a requirement that total revenue equal or exceed total cost—is appropriate.

C) MODERN DEVELOPMENTS

Several events since 1980 require significant adjustments by the bank- card industry. Among the most important are the changes introduced by the Depository Institutions Deregulation and Monetary Control Act of 1980.[64] This legislation, and the regulations that implement it, require the Fed to impose cost-based fees on banking institutions to which it renders services, including check-clearing and collection service; authorize the Federal Home Loan Bank Board to render clearing and collection services, again on a cost-based fee basis, to savings and loan institutions (S&Ls); authorize a significantly broadened scope of activities by S&Ls, including nonbusiness demand deposits (NOW accounts), broadened lending authority, and credit card services; and authorize both banks and S&Ls to pay interest on demand deposits.

The second significant development was the unprecedented escalation in 1980 of nominal interest rates on debt instruments of all maturities and, in particular, the sharp increase in both nominal and real interest rates on short-term paper.

The third development is the decline of usury laws. The Deregulation Act preempts some state usury laws, and some states are moving quickly to raise or remove other usury limits. These several developments comprise a set of diverse and substantial shocks that will require both a short-run and long-run industry adjustment. Some of the short-run adjustments are already quite visible.

The most significant of these recent developments is likely to be the elimination of the prohibition against paying interest on demand deposits. Heretofore, in most urban areas, and some rural areas as well where the structure of the retail banking industry was conducive to rivalry, commercial banks have engaged in vigorous nonprice competition to attract demand deposits. In significant part, this rivalry took the form of a geographic proliferation of retail bank establishments: multiple branches where branching was freely permitted and small independent establishments where it was not. Thus, banks competed for demand deposits by offering potential depositors geographic convenience. Unless one assumes that the interest prohibition had no effect on the industry at all, one must conclude that, at least to some extent, depositors would have preferred interest payments to incremental geographic proximity and that they will now avail themselves of that possibility. Some fraction of existing banking establishments will prove to be uneconomic, but their disappearance will require a long-run adjustment. Bank payment of interest on deposits will be and is being made in the short run. Profitability will be adversely affected until long-run adjustments have occurred.

The other important dimensions on which banks competed for demand deposits included the provision of checking services without the imposition of transaction charges and the "free" provision of collateral services such as safety deposit boxes and bank card issuance. In these dimensions, short-run adjustments are feasible, and the introduction of charges for such collateral services has been widespread. Since 1980 a large fraction of card-issuing banks have imposed either periodic fees or per transaction fees on card holders. Periodic interest charges on the outstanding balances of extended credit users have also been increased by a number of banks. Both of these changes were facilitated by the removal or escalation of usury limits.

It is clear that these various developments have had and will have a substantial effect on the credit card industry. In the past, users of checks have faced artificially low marginal prices for incremental check transactions. Uncompensated demand balances have yielded adequate bank revenues to cover those costs. The widespread introduction of NOW accounts by S&Ls will erode any remaining supracompetitive profitability associated with demand deposits, increasing pressure to impose transaction charges. And the payment of interest by banks on demand deposits will both add to that effect and alter competitive strategies for attracting demand deposits. The introduction of cost-based fees for federal collection and clearance services also will increase the cost of using checks. All these factors will work together to dissuade the providers of demand deposit services from providing those services without imposing explicit transaction charges. Many depositors who previously received free checking services will now face per item transaction charges, and the level of charges demanded of other depositors will increase. These in-creases in the marginal cost of using checks will shift out the demand curve for credit cards.

Simultaneously, however, the supply curve for credit card transactions will also be shifting to the right because of the high cost of funds. Not only the height of these functions but also their shapes over the relevant range will undoubtedly change in ways we do not yet know. As I emphasized in Section I, the shifting cost function under consideration cannot usefully be viewed as reflecting the cost of dealing with card holders; it reflects the joint cost of providing transaction services to both card holders and merchants. Nevertheless, substantially all of the recent price changes are in the charges imposed on card holders rather than in the merchant discount.

It would be an astounding coincidence if at the end of this first round of price changes the distribution of charges between card holders and merchants happened to equilibrate the individual demand functions of those two sets of parties so that each set wished to engage in the same number of transactions at the prevailing price. It seems more probable that a lengthy process of adjustment will ensue, during which financial institutions will gravitate by trial and error to some new equilibrium. And it seems equally probable that the new equilibrium will involve either a higher or a lower interchange fee than that presently in effect. As previously explained, the interchange fee for any one card system must be determined collectively by the system's members: any attempt to set that fee bank by bank, to reflect each bank's individual costs (rather than the system's average costs), would invite each bank to free-ride on the others and set inappropriately high fees.

In addition to the present perturbations in the industry, the "debit card" is for the first time being distributed widely. Apparently many institutions in the industry believe that the debit card and the credit card can be combined and embodied in a single set of plastic cards. Transactions using the cards would be subject to the same merchant discount and the same interchange fee notwithstanding that the card-issuing bank would handle the two types of transactions quite differently. This outcome seems most unlikely unless the contractual terms that have traditionally accompanied the credit card are materially altered. From the standpoint of the card-issuing bank, debit card transactions will be substantially cheaper than credit card transactions, for debit card transactions will not be authorized unless they are for amounts less than the card holder's deposit balance, in which case the default risks are relatively low. Moreover, since the transaction amount is immediately debited against the card holder's deposit balance, the float costs of the debit card are substantially less. These considerations alone seem to dictate quite a different distribution of fees between card holder and merchant and a different interchange fee, as well. In addition to these cost factors, demand factors suggest a similar conclusion. From the card holder's standpoint, the debit card is less attractive than the credit card. The float costs that the bank saves when a debit card is used are precisely the float benefits that the card holder forgoes when he uses a debit card. One would expect therefore that any card holder entitled to use a credit card will always use it rather than a debit card. It follows that the only frequent users of debit cards will be people whose incomes and other indicators of creditworthiness do not enable them to obtain and use credit cards.

The characteristics that distinguish credit card users from debit card users will substantially affect the demand curve of merchants for transactions with these two different types of card holders. The holder of a credit card will continue to be identified as a customer for whose patronage the merchant wishes to compete by extending a free float period; but that will not be true of the holder of a debit card, and one would expect merchants to be unwilling to accept discounts on debit card paper as large as the discounts traditionally accepted on credit card paper.

It seems likely, therefore, that the two payment vehicles will have to be differentiated and subjected to different patterns of distributing charges between merchants and card holders and, in all probability, to different interchange fees. Hence I believe that card-issuing institutions will be engaged in not one but two different learning processes in the period immediately ahead; and both processes will be retarded if these institutions are reluctant to recognize the sharply different cost and demand characteristics of the two payment vehicles.

III. CONCLUSION

Four-party payment vehicles such as the check, the credit card, and the debit card are characterized by joint costs and also by interdependent demand on the part of their users, which, despite the antiquity of such mechanisms, neither the economic literature nor the institutions that provide their services have fully recognized. Those characteristics, in my judgment, were an important contributing cause to the controversy over "clearance at par" that troubled the banking industry for more than half a century and was quieted at last only by means of federal coercion and subsidy. A repetition of the same basic controversy in the context of new payment mechanisms—credit cards and debit cards—is likely to occur in the next few years. Because of sharp cost and demand changes attributable to legislative amendments, because of the effect of inflation on nominal interest rates, and because of governmental responses to inflation that have taken the form of restrictive monetary policies that increase the real interest rates on short-term obligations, those years are likely to be characterized by disequilibrium, confusion, and controversy. In such a period, reliance on governmental intervention to reduce uncertainty is likely to appeal to at least some of the disputants. Such intervention should be resisted.

Once the economic peculiarities that underlie such payment mechanisms are recognized, one can conclude that legal mechanisms already in place are entirely adequate for the task of equilibrating the market. The courts should recognize that collective institutional determination of the interchange fee is both appropriate and desirable. To an unsophisticated observer this collective process of equilibration resembles horizontal price fixing, but, for the reasons set forth in this paper, it should not be so treated. Because of the potential for free-rider behavior, individual establishment of interchange fees will almost certainly produce chaotic results, such as higher fees and instability within card systems.

On the other hand, the fee that is collectively set should not be binding prospectively on any pair of banks within the system. Any pair of banks in the system should be free to negotiate a different bilateral arrangement by higher or lower fees for paper interchanged between them. The collectively determined interchange fee should be merely a guarantee that no card-issuing bank will demand a higher fee on paper presented to it in the absence of such a bilateral arrangement. Of course, the fee should be regarded as binding retroactively for transactions already executed. Sen- sible administration of section 1 of the Sherman Act, applied in a rule of reason context, is sufficient to arrive at this result.[65]

It seems equally clear that the movement toward a competitive equilibrium requires no other collaborative action between participants in such payment systems. It is entirely compatible with that competitive equilibrium that individual *P* banks compete with respect to the charges imposed on cardholders and *M* banks with respect to the magnitude of the merchant discount.

Although collaboration among competing banks with respect to the interchange fee should be permitted under the antitrust laws, any expansion of the range of cooperative action should be viewed with healthy skepticism. Thus antitrust and banking authorities should be alert to ensure that the number of payment systems is as large as the attainment of scale economies permits. Though unbridled autonomy within a system cannot be attained, unbridled rivalry between a multiplicity of systems should be encouraged.

In this regard it is regrettable that the Antitrust Division did not give a less qualified response in 1975 to Visa's request for a business review letter pertaining to its then-effective prohibition against dual membership. Visa sought advice with respect to a by-law that prohibited any card-issuing bank or any merchant bank in the Visa system from serving simultaneously either as a card-issuing bank or a merchant bank in any other system. In a business review letter dated October 7, 1975, to outside counsel for Visa from the assistant attorney general, the Division gave a blessing so limited and so carefully hedged as to leave unresolved the legal permissibility of an effective prohibition against dual membership. Visa responded by withdrawing all restrictions on dual membership, even the limited

In the last five years dual membership in the Visa system and the MasterCard system has become the rule. This widespread pattern of dual membership predictably created very strong pressures for standardization in equipment, procedures, and format. Intersystem rivalry has not completely disappeared; but the opportunity and incentive for such rivalry, particularly in technological innovation, has greatly diminished. This regrettable loss of competitive structure was avoidable but is now probably irreversible, for political reasons if for no others.

Contributing to this irreversibility is the fact that technological changes in the intervening years have facilitated a great degree of interbank competition within a particular system than appeared possible in 1975. Improvements in communications technology have made it possible for a subgroup of banks within a system, subject to only minimal standardization, to differentiate the financial service they offer or even to deploy a differentiated set of terminals and yet continue to operate within the system network.

Of course the more obvious but nevertheless important forms of interbank competition—for card-holder accounts and for servicing merchants—continue. Although the loss of intersystem rivalry is unfortunate, and although such rivalry should be carefully preserved if a new opportunity, in the form of a new card system, arrives on the scene, the industry appears to be functioning competitively.

1 Like "transactional paper," for the purpose of this article "bank" is an abstraction for financial intermediaries. It includes savings and loan associations that process "NOW account" paper and credit unions that process "draft account" paper.

2 I say at least four parties because often additional banks or clearing houses participate in the process, facilitating the flow of the transactional paper from the merchant's bank to the purchaser's bank. For the most part, whether additional parties participate is irrelevant to the basic points.

3 Note that although *P* and *M* have a consumer-supplier relationship with respect to one another, they are both **consumers** with respect to transactional services, which in my nomenclature are supplied by banks.

4 Another way of viewing the problem is to consider the transactional services provided to *P* and those provided to *M* as separate products that are jointly consumed, analogously to joint consumption of public goods. It is now widely recognized that the analytical apparatus long used in dealing with joint-cost problems also has application to peak-load pricing problems and to public good problems. The critical common feature is that the demand schedules of consumers must be summed vertically rather than horizontally in order to derive aggregate demand. This technique can be traced in the literature at least as far back as Howard R. Bowen, *The Interpretation of Voting in the Allocation of Economic Resources*, 58 Q. J. Econ. 27 (1943).

5 Indeed, in any real-world setting there may be no such intersection, although in my diagrams I have drawn the separate curves so as to produce one. It is not unlikely that in the real world the demand curve of merchants lies everywhere above, or perhaps everywhere below, the demand curve of purchasers, in which case there is no possible equilibrium that entails an equal division of transaction costs.

6 The assumption that there are precisely two banks is adopted to facilitate discussion. In actuality there will be some number of purchaser-merchant transactions in which both parties to the transaction happen to have their banking relationships with the same financial institution. Some of the problems discussed in this paper arise in that context. There will be other transactions in which more, perhaps many more, than two banks will be involved—for example, when transactional paper is forwarded through a series of correspondent relationships for ultimate clearance. While these cases present additional problems, substantially all of the analytically difficult problems that arise on the supply side are present in the two-bank situation. Accordingly, I ignore the possibility of multibank clearance chains.

7 The analysis would be significantly affected if *C* exhibited negative slope over a very wide range. That would be the result if both c_M and c_P had negative slope over that range or if either c_M or c_P had negative slope over that range to a degree that exceeded the positive slope of the other. If *c* had negative slope through the range of equilibrium output, the existence of natural monopoly conditions would be strongly suggested.

8 By construction, $q^*e = q^*a + q^*d = q^*b + q^*c$; hence, rearranging, $q^*d - q^*b = q^*c - q^*a$. But $q^*d - q^*b = bd$, the revenue deficiency of *P* bank; and $q^*c - q^*a = ac$, the revenue excess of *M* bank. It should be clear that nothing turns on the fact that I have drawn the diagram in such a way that *CP* lies above cm in the range q^* or that d_M lies above d_P in that range. No matter what combination of these relationships exists, as long as the sum of the revenues equals the sum of the costs, then notwithstanding that *P* bank's revenues from its purchasers do not equal its costs, there is some transfer payment between the two banks that will bring revenues into equality with costs for each.

9 The phenomenon discussed in the text occurs in any four-party transaction in which each of two transacting principals is represented by an independent agent or broker, each of whom also incurs costs. The costs of the two brokers must be paid out of the theoretically possible gains from trade between the two principals. Tradition and transaction-cost considerations may require that the selling principal compensate the selling broker and the buying principal compensate the buying broker; yet there may be no equivalence between the height of each principal's demand curve for brokerage services and the costs incurred by his broker. Often a side payment between principals in the form of an adjustment to the underlying sale price will be used to achieve equilibrium. In such a situation the form of the side payment obscures its very existence and also obscures the complexity of the equilibrium that is being attained. Many brokered real estate transactions answer this description. In four-party payment mechanisms, too, a side payment between *P* and *M*, coupled with payment by each *P* and *M* to *P* bank and *M* bank, respectively, in amounts equal to respective bank costs but not to respective marginal utilities of *P* and *M*, is theoretically sufficient to attain equilibrium. That in practice side payments between banks occur instead is strong evidence that higher transaction costs characterize side payments that take the form of price adjustments between the principals.

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

11 *See* Sec. III *infra*.

12 Assume that credit cards are issued to card holders only by a single bank, *P* bank, which is effectively sheltered from competition by law; and assume that merchants are serviced by a competitive set of merchant banks. Then *P* bank can maximize profits by restricting output to a level *q"* below *q**, at which the total marginal cost curve, *c* in Figure 4, equals the marginal revenue curve (not shown in Figure 4) pertaining to the aggregated demand curve *d*. But since there must be some particular rate *q"* at which transactions are conducted, the output restriction implies a higher price in equilibrium to card holders as well as to merchant banks and merchants. An increase in the interchange fee without an increase in card holder fees would result in a decrease in the number of card transactions that merchants were willing to enter without reducing the number that card holders were desirous of entering. This would reduce the aggregate utility of the card system to card holders simultaneously with increasing the utility to card holders of the marginal transaction each was able to enter. Thus *P* bank would be forgoing the opportunity to exploit, through card holder fees, that higher marginal utility. This pattern would create incentives for card holders to make side payments to merchants to induce additional transactions. Because those side payments must be presumed to involve higher transaction costs, *P* bank would be squandering its monopolistic potential. Assume, more realistically, that credit cards are issued by a group of banks that own the card system as a cooperative venture and share in the profits of the system proportionately to the dollar volume of charge transactions executed by each member's card holders. Now any attempt to exploit merchant banks (and merchants) by increasing the interchange fee is doomed to failure, quite apart from competition from rival payment mechanisms, unless the member banks also act collectively to exploit card holders. If member banks compete actively for card holders, as they would have strong incentives to do, to increase their share of interchange monopoly profits, they will simultaneously dissipate the monopoly profits and create incentives, even stronger than those previously described, for card holders to make side payments to merchants. Equilibrium is attained at zero monopoly profits, needlessly high transaction costs, and a smaller industry than under competition. Cartelization with respect to the merchant's demand function without simultaneous cartelization with respect to the card holder's demand function would not appear to be feasible; and cartelization with respect to both demand functions is difficult by unusually high information requirements about the relative positions of the two demand functions, in addition to the usual difficulties of policing cheating by cartel members through rivalry for card holders.

13 The use of checks in America had its origins in the operation of "the fund at Boston" in 1681. A person could direct the manager of the fund, in writing, to transfer part of his deposit to the credit of another. However, the use of deposit currency, or checks proper, did not become common until a century later. W. E. Spahr stated, in his excellent history of checks, that deposit currency did not develop until after the Revolutionary War, for the following reasons: (1) The colonists had very little specie to deposit. (2) The country was sparsely settled, and deposit banking implies that the inhabitants be in close touch with their banks in order to test the validity of their checks. (3) There was not the requisite security of personal and property rights and confidence in government and banking institutions. Walter E. Spahr, The Clearing and Collection of Checks 38-43 (1926).

14 The use of checks for local payments accelerated after the Revolution. There is substantial evidence of the use of checks in the nation's commercial centers before the creation of the first United States Bank in 1791. Id. at 43. Spahr estimated the amount of check use in America by examining the relation between deposits and currency in circulation. Deposits passed bank note currency in 1855. *Id*. at 60. In 1867 the public held $1.20 in deposits for every dollar of currency and, by 1872, held $2.00 for every dollar of currency. After 1880 the ratio began a long-term climb; it was twelve to one in 1929. Milton Friedman & Anna Schwartz, A Monetary History of the United States 16 (1963).

15 *Federal Reserve Bank of Richmond, Letter No. 4*, Mar. 1922, reprinted in Readings In Money, Credit and Banking Principles 377, 379 (Ivan Wright ed. 1926)

16 Broy Hammond, Banks and Politics in America: From the Revolution to the Civil War 549 (1957).

17 However, the banks in Boston, under the leadership of the Suffolk Bank, were able to institute a system that discouraged the discounting of New England Bank notes. *Id*. at 549- 56; V. Longstreet, Currency Systems of the United States in Banking Studies 65, 69 (Federal Reserve ed. 1941). *See* note 45 *infra* and accompanying text.

18 *See* note 14 *supra*. Bank notes were far more important to country banks, especially those in the southern and western states, than for the city banks. In 1841, "Gallatin pointed out that deposits constituted the principal currency in the larger cities but that country banks could not exist unless they had the right to issue bank notes." Spahr, *supra* note 13, at 63.

19 Although there is a consensus that the draft was the principal means by which a buyer in the country paid a long-distance debt during the early part of the nineteenth century, there is disagreement about the duration of the practice. Thatcher C. Jones, Clearing and Collections 172-74 (1931); *Testimony on Par Collection of Checks: Hearings on H.R. 12379 Before the House Comm. on Banking and Currency*, 66th Cong., 2d Sess. (1920), indicates the importance of the use of drafts up until the 1890s. But Claudius B. Patten, writing on the mid-1880s, stated that although the use of drafts was common thirty to forty years previously, "Nowadays no country trader, no matter whether he is located in Deadwood or St. Augustine, thinks he is in fashion unless he 'pays' his New York or Boston bills by sending there his individual checks on his local bank, which gets all the advantage of his deposit until the checks come around for collection from the city banks, which have given their dealers immediate credit for them, and made no charge for their collection." Claudius B. Patten, The Methods and Machinery of Practical Banking 1100-01 (11th ed. 1902).

20 The fee charged by *P* bank was referred to as the "charge for exchange" or, often, "exchange." The amount of this exchange varied greatly with the circumstances of the case, but generally speaking it was large enough to cover the cost to *P* bank of sending currency to *M* bank, including the transportation charges, insurance, and interest on the money in transit. Federal Reserve Bank of Richmond, *supra* note 15, at 380.

21 The average price of southern and western exchange on New York markets in 1859 was estimated to vary from 1 to 1.5 percent. After 1890 the charges varied from one-tenth to one-fourth of 1 percent. Spahr, *supra* note 13, at 102.

22 In 1863 Congress passed "An Act to provide a national Currency, secured by a Pledge of United States Stocks, and to provide for the circulation and Redemption thereof." Act of Feb. 25, 1863, ch. 58, 12 Stat. 665. The 1863 law was replaced by the Act of June 3, 1864, ch. 106, 13 Stat. 99. This Act established the National Banking System and is commonly known as the National Bank Act.

23 A tax of "ten per centum on the amount of notes of any state bank, or state banking association" was levied by Congress. Act of Mar. 3, 1865, ch. 78, § 6, 13 Stat. 484. One year later the tax was reenacted by Congress with a more extended application. Act of July 13, 1866, ch. 184, § 9, 14 Stat. 146. The Supreme Court upheld the constitutionality of the tax in *Veazie Bank v. Fenno*, 75 U.S. (8 Wall.) 533 (1869). Because of widespread evasion of the law by banks, corporations, and municipalities, Congress repealed the Act and substituted a more comprehensive prohibition. Act of Feb. 8, 1875, ch. 36, §§ 19-21, 18 Stat. 311. The tax, which was intended not only to eliminate state bank notes but also to force the state banks to become national banks, did not achieve the second purpose. State banks managed to survive by increased reliance on deposit currency. See Hammond, *supra* note 16, at 753. Although the tax initially caused many banks to become national banks, the decline (as measured by the decreasing size of state and private bank deposits) ceased in 1867. By 1871 the deposits in nonnational banks had expanded to the point where they equaled the deposits of the national banks. *See* Friedman & Schwartz, *supra* note 14, at 19. *See also* Kenneth W. Dam, *The Legal Tender Cases*, 1981 Sup. Ct. Rev. 367, for a treatment of the causes and consequences of the legislation in this period.

24 Country banks used their reserves as a means of clearing their checks without paying remittance charges. After the banks in New York City started charging for the collection of these out-of-town checks, the reserve balances were transferred to other cities. Spahr, *supra* note 13, at 110-11; Charles F. Dunbar, The Theory and History of Banking 50 (4th rev. ed. 1922).

25 "By taxing State bank notes out of existence in 1865, a vacuum was created which gave an added impetus to the use of deposit currency. Other factors which were responsible for the increasing use of deposit currency, and consequently checks, were the inelastic note currency, better means of communication, the cheap and uniform postage rates, and the denser population." Spahr, *supra* note 13, at 84. Spahr explains the greater use of out-of- town checks in the following manner. "As the banks grew in numbers and the use of checks in payment of foreign (out of town) bills became more general, the banker found he could charge the collecting bank a maximum rate with less compunction than he could charge his depositor a minimum rate on drafts, and so he encouraged the use of the check." *Id*. at 103. These comments leave unexplained why *P* was expected to pay for exchange but *M* bank was expected to pay when checks were used.

26 Competition soon forced banks into the practice of crediting immediately the uncollected checks to the depositor's account and paying interest on those uncollected funds. Spahr, *supra* note 13, at 110.

27 One check traveled 1,500 miles and passed through eleven banks in an attempt to avoid remittance charges. James C. Cannon, *Clearing House Methods and Practice* 74-78 (1900), reprinted in U.S. National Monetary Commission, Clearing Houses and Credit Instruments 70-74 (Publications of the Nat'l Monetary Comm'n No. 6, 1910). *See also* Spahr, *supra* note 13, at 105.

28 Spahr, *supra* note 13, at 18. Current explanations also use conflict-of-interest explanations, for example, Hal Scott, *The Risk Fixers*, 91 Harv. L. Rev. 737 (1978).

29 *See generally* Cannon, *supra* note 27. The first clearinghouse was established in New York City in 1853. During the following five years clearinghouses were established in Boston, Philadelphia, Baltimore, and Cleveland. By the mid-1870s clearinghouses were established in most of the leading cities in the United States. In 1899, there were 31 clearinghouses in the United States. Dale H. Hoffman & Melvin Miller, Origin and Development of Charges for Banking Services 10-14 (1942).

30 *Compare* Wright, *supra* note 15, at 380-81.

31 Albert Gallatin first proposed establishing a clearing system in 1841 as a means of reducing the costs of exchanging checks and notes. *See* Hammond, *supra* note 16, at 705- 07; Spahr, *supra* note 13, at 79-82.

32 In 1899 the banks of Boston organized a system for the collection of country checks. The Boston Plan was intended to force all banks in New England to clear checks at par. The plan resulted in 97 percent of the checks in New England being collected at par. Under the Boston Plan the cost of collection was reduced from a rate which varied from $1.00 to $1.50 per thousand dollars to a charge of six or seven cents per thousand. Spahr, *supra* note 13, at 128. *See* Federal Reserve Bank of Richmond, *supra* note 15, at 382-83; note 25 *supra* and accompanying text.

33 *See* Thomas C. Schelling, The Strategy of Conflict 67-80 (1960).

34 *See* Spahr, *supra* note 13, at 103-08. *See also* note 27 *supra* and accompanying text. In the political arena, arguments of doubtful substance were built on the existence of these circuitous routings. Because such routings tended to add to the number of items (and dollar volume of items) outstanding at any point in time, they increased the float—the number of dollars shown as additions to the deposits of *M* bank but not yet deducted from the deposits shown on the books of *P* bank. This phenomenon results in an overstatement, in the aggregate, of deposits in the banking system. Since the aggregate of loans that the banking system is able to make is a percentage of deposits, anything that increases the float increases the money supply and tends to have inflationary effects. The increase in the mean money aggregates would represent a one-time event and would be of doubtful significance, but to the extent that the float is less stable than genuine deposits, a large float might also tend to destabilize the money supply. Banks that did not clear at par were criticized for causing these undesirable macroeconomic effects. Slow and circuitous clearance of checks is also undesirable from the standpoint of banking policy because it facilitates the practice of "kiting"—the deliberate manipulation by an individual of deposits and checks outstanding against nonpar banks—and practices were criticized on this basis too. Although this attack may have had more substance than the money supply attack, both confuse the desirability of standardization with that of par clearance. Spahr, *supra* note 13, at 105-08; Federal Reserve Bank of Richmond, *supra* note 15, at 384-89. *See* note 23 *supra* and accompanying text.

35 Paul F. Jessup, The Theory and Practice of Nonpar Banking 48 (1967).

36 "In many instances throughout the South the exchange revenue of the small or country bank constituted considerably more than half of the bank's income." Federal Reserve Bank of Richmond, *supra* note 15, at 391.

37 Act of May 30, 1908, ch. 229, Pub. L. No. 169, §§ 17-20, 35 Stat. 546, 552.

38 Federal Reserve Act, ch. 6, Pub. L. No. 43, §§ 1-30, 38 Stat. 251 (1913).

39 The National Monetary Commission did not make any specific recommendations about exchange charges. Section 16 of the Federal Reserve Act only prohibited member banks from charging other members remittance charges. Member banks were allowed to charge their customers the actual cost of collection.

40 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 13, 38 Stat. 263 (current version at 12 U.S.C. § 342 (1976)).

41 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 16, 38 Stat. 265, 268 (1913). The only amendment made to the quoted portion of the section is the name of the Federal Reserve Board. The second sentence quoted now reads, "The Board of Governors of the Federal Reserve System. . . ." Act of Aug. 23, 1935, ch. 614, § 302(a), 12 U.S.C. § 360 (1976).

42 Section 19 of the Federal Reserve Act specified the reserve requirements of member banks. The requirements were substantially lowered by the Act of June 21, 1917, ch. 32, Pub. L. No. 25, § 10, 40 Stat. 239. Member banks in central reserve cities were required to maintain reserves of 18 percent against demand deposits (decreased to 13 percent) and 5 percent against time deposits (decreased to 3 percent). Member banks in reserve cities were required to carry reserves of 15 percent against demand deposits (decreased to 13 percent). The reserves of country banks were fixed at 12 percent for demand deposits (decreased to 7 percent) and 5 percent for time deposits (decreased to 3 percent). The reserve requirements were lowered to stimulate membership in the Federal Reserve System. *See* Federal Reserve Bank of Richmond, Letter No. 5, Apr., 1922, reprinted in Wright, *supra* note 15, at 391-404.

43 Federal Reserve Act, ch. 6, Pub. L. No. 43, § 13, 38 Stat. 263 (current version at 12 U.S.C. § 342 (1976)).

44 Act of Sept. 7, 1916, ch. 461, Pub. L. No. 270, 39 Stat. 752 (current version at 12 U.S.C. § 342 (1976)) (emphasis added).

45 Federal Reserve Bank of Richmond, *supra* note 42, at 402.

46 In 1916 the number of member banks actually underwent a slight decline from 7,631 to 7,614. Spahr, *supra* note 13, at 218.

47 Federal Reserve Bank of Richmond, *supra* note 42, at 400.

48 Act of June 21, 1917, Pub. L. No. 25, 40 Stat. 234 (current version at 12 U.S.C. § 342 (1976)).

49 Excerpt in Federal Reserve Bank of Richmond, *supra* note 42, at 406.

50 *Id*.

51 At the end of 1918 there were 8,692 member banks of the Federal Reserve System and 10,305 nonmember banks remitting at par, and 10,247 nonmember banks not on the par list. Federal Reserve Bank of Richmond, *supra* note 42, at 407.

52 *Id*. at 408; Spahr, *supra* note 13, at 234-35.

53 Federal Reserve Bank of Richmond, Letter No. 6, May 1922, reprinted in Wright, *supra* note 15, at 410-12. This tactic of going to the window of the noncomplying bank and demanding full payment had been used before as a means of achieving a system of par clearance. The Suffolk Bank System in the 1820s (see Justice Story's decision in *Suffolk Bank v. Lincoln Bank*, 22 Mass. 106 (1827)) and the Country Checks Department of the Boston Clearing House in the 1890s (**see** note 32 *supra*) both used the same tactic to force par clearance. The Suffolk Bank System was primarily designed to prevent the discounting of bank notes. *See* Spahr, *supra* note 13, at 73-78, 126-29; Federal Reserve Bank of Richmond, *supra* note 15, at 379.

54 *See* Spahr, *supra* note 13, at 103-04.

55 In 1919 the number of par banks increased from 18,905 to 25,486 and the number of nonpar banks decreased from 10,191 to 4,015. Federal Reserve Bank of Richmond, Letter No. 5, *supra* note 42, at 410.

56 Federal Reserve Bank of Richmond, *supra* note 53, at 4125-16.

57 For an excellent discussion of the specific statutes see Spahr, *supra* note 13, at 251-54.

58 *Id*. at 256-90.

59 *American Bank & Trust Co. v. Federal Reserve Bank of Atlanta*, 262 U.S. 643 (1923).

60 *Farmers & Merchants Bank v. Federal Reserve Bank of Richmond*, 262 U.S. 649 (1923).

61 Jessup, *supra* note 35, at 23.

62 Federal Reserve System, Memorandum on Exchange Charges (September 1, 1980).

63 *Id*.

64 Pub. L. No. 96-221, § 1, 94 Stat. 132 (codified at 12 U.S.C. 226 (1980)).

65 *See Broadcast Music, Inc. v. Columbia Broadcasting Sys., Inc.*, 441 U.S. 1 (1979); *Continental T.V., Inc. v. GTE Sylvania Inc.*, 433 U.S. 36 (1976). However, the Supreme Court has on occasion failed to recognize the significance of maximum price fixing where the product has joint-demand characteristics. *See Albrecht v. Herald Co.*, 390 U.S. 145 (1968). *See also* Frank H. Easterbrook, *Maximum Price Fixing*, 48 U. Chi. L. Rev. 886 (1981).

66 *See generally* Note, New Directions in Bankcard Competition, 30 Cath. U. L. Rev. 65 (1980).