COMPETITION AND VERTICAL INTEGRATION IN FINANCIAL EXCHANGES

Craig Pirrong

C.T. Bauer College of Business, University of Houston
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 Wright 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. 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. More recently, the DOJ released a letter arguing that vertical integration between exchanges and clearinghouses was anticompetitive. The 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.

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. 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
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.

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.
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. These trading costs include the bid-ask spread and the price impact of trades. The extant empirical evidence is consistent with these predictions.

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 deflections to the larger venue. Absent strong clientele effects, in equilibrium this process results in the survival of a single exchange.

Empirical evidence is consistent with this tipping hypothesis.

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. 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. 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. Sunk costs in physical trading infrastructure and human capital, switching costs, and coordination costs all impair the contestability of the trade execution venue.
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”),\textsuperscript{14} which required an exchange to direct orders to another venue if the latter offered better prices.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{20}

\textbf{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,\textsuperscript{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.  

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. 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. 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.

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,
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.

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. 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.

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; 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. 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.

\[ \text{This model clearly does not fit the facts in the clearing-execution case.} \]

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, 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.

\[ \text{These models work best to explain ties in highly technologically dynamic industries where monopolies are likely to be short-lived in any event.} \]

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. 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 \( Q_m \). 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. 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. 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.

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 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,\(^4\)\(^2\) 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 clearinghouses, 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 clearinghouses). 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 clearinghouses 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, particularly 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.
8 Cross-border trading restrictions are one potential source of clientele effects. Geographic proximity gave rise to clientele prior to the advent of telegraphic and telephonic communications.
10 Pirrong, Securities Market Macrostructure, supra note 9.
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 JENNIFER TOBIAS, 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.

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.


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.


22 See Craig Pirrong, Burd 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.


24 Peter Norman, The Risk Controllers (Wiley 2011).


27 Bob Hills et al., Central Counterparty Clearing Houses and Financial Stability, 6 FIN. STA.B. 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).

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).


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.