Securities Market Macrostructure: Property Rights and the Efficiency of Securities Trading

Craig Pirrong
Oklahoma State University
Stillwater, OK 74078
405-744-1243
pirrong@okstate.edu

December 10, 2001

Abstract. This article derives securities market macrostructure from microstructural foundations under a variety of assumptions regarding property rights. Because liquidity effectively makes securities trading a network industry, intermediaries can exercise market power by restricting access to the trading mechanism. Fragmentation, cream skimming and free riding reduce the inefficiency that results from this market power, but welfare would be improved further by requiring open access to all trading venues. Implementing open access in practice must confront a trade-off between reducing market power and potentially impairing the incentives of the operators of trading systems to reduce cost and improve quality. Other network industries, notably telecoms and electricity transmission, have faced similar dilemmas, and the path to the creation of a more efficient property rights structure in financial markets could benefit from the experiences of other network markets.

JEL Classification: L11, L12, L31, G10, G20. Key Words: Securities market structure, financial exchanges.

1 Introduction

Controversy over the structure of securities markets is a hardy perennial. Technology changes, faces change, but the market structure debate has remained surprisingly static since the debates over the Securities and Exchange Act in 1933-1934. Through the early years of the SEC, the Special Study on Securities Markets in 1963, the formation of the National Market System in the 1970s, and the recent controversy over the structure of future electronic securities markets, two themes have defined the debate: fragmentation and competition. On the one hand, it has been argued vociferously that fragmentation of trading in securities is inefficient, especially when off-exchange trading venues "cream skim" uninformed order flow; critics of fragmentation typically advocate measures to centralize securities trading. On the other hand, it has been argued equally vociferously that fragmentation creates competition absent which exchanges would exercise market power to the benefit of their members and to the detriment of the trading public; advocates of this position view regulatory measures designed to centralize trade (such as the creation of a mandatory central limit order book or "CLOB") as unwarranted checks on competition.

This article attempts to reconcile these seemingly irreconcilable views by constructing a model of the "macrostructure" of a securities market– the number of trading venues, their size, their market shares, and the policies they adopt–from fundamental microstructural considerations. I derive the macrostructure of a securities market under alternative property rights regimes from basic microstructural factors including information asymmetry

and risk sharing.

Three fundamental conclusions flow from this analysis. First, if exchanges can restrict their size, the creation of cream skimming off-exchange trading venues actually improves welfare, but does not achieve a first-best outcome. Second, if exchanges cannot restrict size, but instead must provide open access to liquidity suppliers, cream skimming off-exchange trading venues may exist even though this fragmentation is inefficient. Third, if exchanges must provide open access, banning off-exchange trading produces a first-best outcome. That is, rules that preclude fragmentation are efficient if and only if access to the central trading venue is truly unrestricted. Thus, the efficiency implications of fragmentation depend crucially on whether or not access to the primary exchange is restricted. This means that property rights exert a decisive influence on the efficiency of securities trading.

These results derive from the nature of liquidity. Liquidity creates network effects that induce centralization of trading. Traders operating on an exchange can exploit this centripetal tendency and increase their profits by limiting access to the exchange. Cream skimming off-exchange trading venues arise in response to the restrictive policies of the exchange, and provide competition and additional risk bearing capacity that off-sets in part the deadweight costs associated with exchange restrictions. When access to the exchange is not restricted, cream skimming third markets sometimes can survive by offering better terms of trade to some uninformed market participants even though it would be first best to centralize all trading. Under these conditions, forced centralization improves welfare as long as access to the exchange is unrestricted.

The issues of open access and cream skimming are not unique to securities trading. Indeed, they are the primary sources of controversy in virtually all network industries, including telecommunications and electricity transmission. The similarities between securities markets and telecom or transmission markets should not be surprising because liquidity effectively makes security trading a network industry. Thus, regulation of security market structure is a piece with the regulation of other network industries and must confront the same basic issues. Most important, as in telecoms or electricity transmission, mandated open access to securities markets is a desirable public policy in the abstract. In reality, however, mandated open access raises serious practical issues that do not admit easy solution. In particular, treating the market as a public good can lead to underproduction and overconsumption of key attributes of the trading system. Thus, analysis of security market structure needs to come to better grips than it has heretofore with property rights issues that have absorbed students of "public utility" regulation for decades. This research represents a first step in that process.

The remainder of this article is organized as follows. Section 2 lays out the formal framework of the analysis. Section 3 analyzes security market structure under varying assumptions about exchange access, centralization, and free riding. Section 4 discusses some of the difficulties of implementing an open access trading mechanism. Section 5 summarizes the article.

2 Micro Foundations

A primary objective of this article is to build a model of the macrostructure of a securities market on a microstructural foundation. To do so, I employ a

variant of the canonical Kyle microstructure model.

Specifically, consider trading in a risky security or financial contract. The true value of the traded instrument (which is not public knowledge) is v. The unconditional distribution of v is normal with mean of 0 and variance σ^2.

Two types of agents desire to trade the instrument. First, there are K risk neutral informed traders who know v. Second, there is a large (but finite) number of uninformed traders–"noise traders"–who trade for portfolio balancing or hedging purposes. Net noise trader demand (noise trader buys minus noise trader sells) for the asset is perfectly inelastic, and is a normal random variable with mean 0 and variance S. Individual noise trader demands are uncorrelated, so the variance of the sum of several noise trader's demands is equal to the sum of the variances of their individual demands. Noise trader demand and the value of the asset are orthogonal.[1]

Noise traders, in turn, come in two varieties. The first variety–the "U1" type–are verifiably uninformed; by implementing a screening technology, liquidity suppliers (described more fully below) can determine whether a trader is of the U1 type and therefore uninformed. In contrast, the other variety–the "U2" type–are not verifiably uninformed; the screening technology cannot distinguish the U2s from the informed.[2] Fraction $q^* < .5$ of the noise traders are U1s, and fraction $1 - q^* > .5$ are U2s.[3]

As an example of a screening technology, small noise traders may be able to represent credibly that they are uninformed, whereas large noise traders may not. Reputation and trading constraints (such as the "no bagging" constraint analyzed in Seppi, 1990) are other means by which some (but not all) *large* uninformed traders can identify themselves as such. Other

mechanisms, such as using periodic auctions rather than continuous trading, may also serve to segment some uninformed traders.

This assumption of an imperfect screening mechanism is crucial to understanding market macrostructure. It is well known that such mechanisms exist in practice.[4] So-called "third market" dealers (such as Madoff Securities) explicitly attempt to limit their dealings to the uninformed and use algorithms to analyze the profitability of trading with certain counterparties to screen out those who are more likely to be informed. Moreover, some trading mechanisms (such as crossing networks) are designed to limit participation by the informed by imposing high costs on them; the informed incur higher costs to wait to trade than some of the uninformed because their information may depreciate rapidly. Similarly block trading (Seppi, 1990) and "sunshine" trading (Admati-Pfleiderer, 1991) allow identification of some, but not all uninformed traders. Thus, imperfect screening mechanisms are ubiquitous in financial markets.[5]

Moreover, this assumption has empirical content; as will be seen, it generates predictions that are consistent with salient features of security market macrostructure, whereas alternative assumptions lead to counter-factual predictions. For instance, if no screening is possible, the model implies that only one trading venue survives in equilibrium; such a model could not explain the existence of Madoff or crossing networks. Alternatively, if market makers can identify *all* uninformed traders, they would restrict their dealings to the uninformed; securities prices and trading would be uninformative in this case. This is inconsistent with overwhelming evidence. In contrast, the analysis will demonstrate that the partial screening assumption permits

the existence of multiple trading venues (such as third market dealers and crossing networks). Moreover, with partial screening, the model implies that off-exchange prices are less informative than exchange prices (because off-exchange venues limit informed trading whereas exchanges do not). This is consistent with the empirical evidence. Easley et al. (1996) show that orders executed on one third market (Cincinnati) are substantially less informative than orders submitted to the NYSE. Hasbrouck (1997) estimates that NYSE trades account for 93 percent of the information revealed by trading. Huang and Stoll (1994) and Bessembinder and Kaufman (1997) demonstrate that off-exchange trades are substantially less informative than exchange trades for listed NYSE stocks. Smith et al. (2001) show that "upstairs" trades in listed Toronto Stock Exchange equities have virtually no information content, whereas trades conducted on the exchange trading mechanism proper do. Thus, the partial noise trader screening assumption generates empirically supported predictions, whereas alternative assumptions do not.

In addition to the noise traders and the informed traders, there is a set of potential liquidity suppliers (also referred to as market makers) $\mathbf{L} = \{1, 2, \ldots, N\}$. Each liquidity supplier $j \leq N$ is risk averse, with a constant absolute risk aversion coefficient α_j. Equivalently, the risk tolerance of intermediary j is $t_j = 1/\alpha_j$. Moreover, wlog $t_j > t_k$ for $j < k$. That is, intermediaries are ordered by decreasing risk tolerance. The total supply of risk bearing capacity (i.e., aggregate risk tolerance) is $T^A = \sum_{i=1}^{N} t_i$.

The assumption of risk averse market makers is realistic and important.[6] Limits on the capital of market makers constrain their ability to bear inventory risk and induce them to act as if they are risk averse. It is well

documented that market makers in securities are compensated for bearing risk, which would not occur if they were risk neutral. Moreover, the existence of limits to market makers' risk bearing capacity implies that the size of exchange size has efficiency implications; risk is borne inefficiently if exchanges restrict membership to a suboptimally small number.

The next section analyzes market structure under varying assumptions about (1) the ability of liquidity suppliers to form coalitions with restricted membership and (2) the ability of some liquidity suppliers to restrict their dealings to those who can prove they are uninformed. Variations in these assumptions generates four distinct regimes.

In the first regime, liquidity suppliers can form coalitions that restrict membership. That is, some market makers may be excluded from a coalition. I refer to a group of market makers as an exchange. In the first regime, all coalitions of intermediaries are obligated to trade in a non-discriminatory fashion. That is, they must accept market orders from all traders and cannot refuse to deal with those that they believe to be informed; equivalently they cannot use the screening technology to restrict their dealings to those they know to be noise traders.

In the second regime, liquidity suppliers can form coalitions that restrict membership. Unlike the first regime, however, in the second regime intermediaries can refuse to deal with customers. In particular, in the second regime, market makers can implement the screening technology and restrict their dealings to the demonstrably uninformed U1s. I refer to trading only with the verifiably uninformed as "cream skimming." Moreover, cream skimming market makers can condition their trades on the prices determined in trading

on exchanges where the market makers deal with all on a non-discriminatory basis. Since trading by the informed reveals some of their information, a price determined in a non-discriminatory auction reduces uncertainty about the value of the asset. I refer to the ability to condition trades on prices from markets where the informed trade as "free riding" on price discovery.

In the third regime, groups of liquidity suppliers cannot form exclusive coalitions that restrict membership. That is, every every exchange must be open to all market makers. In this regime, cream-skimming and free riding are permitted.

In the fourth regime, exclusive coalitions of liquidity suppliers are precluded; exchanges must admit as members all market makers who care to join. In this regime, cream skimming and free riding are precluded.

In all the regimes, noise traders choose where to trade non-cooperatively. Noise traders choose to trade where their expected execution costs are minimized. The informed trader can trade with any and all coalitions that do not restrict their dealings to the demonstrably uninformed. Once noise traders and the informed trader have submitted market orders to the liquidity supplier coalition(s) of their choice, all markets clear in a batch auction, with the auctions of coalitions that do not restrict dealings to the demonstrably uninformed clearing immediately before those who do so restrict their trading. Due to this difference in the timing of trading, cream skimming markets can observe–and free ride on–prices determined on exchange.[7] In the auctions, participating liquidity suppliers condition their trades on observed net order flow (noise trader net order flow plus the informed orders).

In regimes three and four, where exchanges cannot restrict entry, liq-

uidity suppliers choose the exchange they trade on non-cooperatively and simultaneously.

The next section analyzes equilibrium market structure in these four regimes. The analysis derives the number of exchanges and cream skimming coalitions in each regime. It also derives total surplus. Given the assumption of inelastic noise trader demand, total surplus is determined by the total cost of operating the market, where total cost equals noise trader execution costs minus informed trader profit minus the risk adjusted profit of market makers. A first best market macrostructure minimizes total cost.

3 Macrostructure Under the Four Regimes

3.1 Regime One

Consider the trading process when two coalitions–exchanges–form; the analysis can be extended readily to incorporate an arbitrary number of exchanges. The total risk tolerance (the sum of the risk tolerances) of the members of exchange 1 is T_1, and the total risk tolerance of exchange 2 is $T_2 < T_1$. Assume initially that fraction q_1 of the noise traders have chosen to trade on exchange 1, and $q_2 = 1 - q_1$. Due to the independence of noise trader demands, the variance of noise trader order flow on exchange 1 is $S_1 = q_1 S$, and the variance of noise trader order flow on exchange 2 is $S_2 = q_2 S$.

Analysis of equilibrium proceeds in the standard way. Upon learning v the informed traders conjecture that the price on exchange i, $i = 1, 2$ is a linear function of order flow:

$$P_i = \lambda_i(\sum_{k=1}^{K} w_{ik} + z_i) \tag{1}$$

where w_{ik} is the order that the informed trader k submits to exchange i, z_i is net noise trader demand on exchange i, and λ_i is a constant. λ_i measures the sensitivity of the security's price to variations in order flow. Its reciprocal is referred to as market "depth;" greater depth (smaller λ_i) desirable because it implies lower transactions costs for noise traders.

Given this conjecture of a linear price function, the informed trader l chooses w_{il}, $i = 1, 2$ to maximize:

$$V_i = w_{il}E[v - \lambda_i(w_{il} + z_i + \sum_{k \neq l} w_{ik})] \tag{2}$$

where the expectation is taken over z_i. Since v and z_i are orthogonal, the symmetric solution of the informed traders' maximization problems implies:

$$w_{il} = \beta_i v = \frac{v}{(K+1)\lambda_i} \quad \forall \quad l \leq K \tag{3}$$

That is, $\beta_i = 1/[(K+1)\lambda_i]$. β_i measures the intensity of informed trading.

Conditional on order flow liquidity supplier j on exchange i chooses his trade y_j to maximize his certainty-equivalent profit. Given the strategies of the informed and the market makers, the analysis in the appendix shows that in equilibrium:

$$\lambda_i = \frac{\hat{\sigma}^2}{T_i} + \frac{K\beta_i\hat{\sigma}^2}{S_i} \tag{4}$$

where $\hat{\sigma}^2$ is the variance of the asset value conditional on order flow (which is also derived in the appendix). Since order flow communicates information about v (because the informed buy (sell) more when v is high (low)), $\hat{\sigma}^2 < \sigma^2$. Expression (4) shows that the sensitivity of price to order flow in exchange i consists of two parts. The first part is the cost that intermediaries incur to absorb the risk of order flow imbalances. The second term is the adverse selection cost incurred when trading with the informed.[8]

Taking the derivative of expression (4) after substituting $\beta_i = 1/(K+1)\lambda_i$ implies that $d\lambda_i/dS_i < 0$. Moreover, since $dS_i/dq_i = S > 0$, $d\lambda_i/dq_i < 0$. This means that the sensitivity of price on exchange i to order flow is smaller, the larger the fraction of noise traders who select to trade on exchange i. Equivalently, market depth on exchange i is greater, the larger the fraction of noise traders who trade there. It is also straightforward to show that $d\lambda_i/dT_i < 0$. That is, price on exchange i is less sensitive to order flow, the larger the total risk tolerance of its members. Finally, $d\hat{\sigma}^2/dS_i > 0$; conditional variance is increasing in the variance of noise trader order flow.

These results determine where noise traders choose to transact. Each noise trader takes the expected cost of execution on each exchange as a given and chooses to trade where the per-noise trader cost of execution is smallest. The per-noise trader expected execution cost on exchange i is given by:

$$x_i(q_i, T_i) \equiv \frac{E(P_i - v)z_i}{q_i} = \frac{\lambda_i S_i}{q_i} = \frac{\lambda_i q_i S}{q_i} = \lambda_i S \tag{5}$$

Since λ_i is decreasing in q_i, exchanges are subject to increasing returns to scale; per uninformed trader expected execution costs are smaller, the larger the number of noise traders that choose to trade on that exchange. Liquidity effects create economies of scale. In the presence of informed traders, a noise trader prefers to trade where the largest number of other noise traders congregate in order to minimize losses from adverse selection.[9]

This analysis implies that there are three possible equilibria in this market when noise traders choose where to trade simultaneously. Figure 1 illustrates these equilibria. The horizontal axis in the figure is q_1, the fraction of noise traders that choose to trade on exchange 1. The downward sloping curve is

$\lambda_1 S$, the average noise trader execution cost on exchange 1; the downward slope indicates the economies of scale. The upward sloping curve is $\lambda_2 S$, the average noise trader execution cost on exchange 2. The upward slope also indicates economies to scale, as an increase in q_1 implies a decrease in q_2, and thus a rise in execution costs on that exchange.

Place Figure 1 Here

The first equilibrium, which is unstable, occurs at the intersection of the two curves. The second equilibrium occurs at $q_1 = 1$, i.e., all noise traders congregate at exchange 1. The third equilibrium is $q_1 = 0$, i.e., all noise traders choose to trade on exchange 2.

This analysis indicates that exchange markets with informed trading are "tippy." That is, all traders choose one exchange or the other. The intermediate equilibrium with $1 > q_1 > 0$ is not stable; any perturbation of q_1 away from this point tends to "tip" the noise traders towards one exchange or the other. Thus, stable equilibria in this market are monopoly equilibria.

Although this particular model is novel, the prediction that trading will concentrate in a single market in the absence of cream skimming is not. Admati and Pfleiderer (1988) and Chowdhry and Nanda (1991) present models in which trading activity "clumps" for reasons similar to those driving the foregoing result. Pagano (1989) derives a model in which risk-sharing considerations lead to concentration of trade on a single market.

Although these models predict "clumping" they are incomplete because they make no predictions about exchange size and the efficiency implica-

tions thereof. Exchange size is irrelevant (or indeterminate) in Admati-Pfleiderer and Chowdhry-Nanda because they assume risk neutral market makers. Pagano's model is "intermediary free" in the sense that there are no market makers; instead, all agents trade directly without intermediation. Thus, Pagano's theory cannot generate predictions about the size of exchanges in economies (such as ours) where intermediaries play an important role in securities trading. Moreover, it cannot analyze whether restrictions on the number of intermediaries have welfare implications. Furthermore, since there are no privately informed traders in the Pagano model, it is of limited utility in understanding the economics of free riding and cream skimming.

In contrast, by utilizing an equilibrium selection criterion, the present model makes predictions about exchange size and welfare. The selection criterion determines which exchange all noise traders choose. Knowing how noise traders behave, a subset of liquidity suppliers can form an exchange that maximizes their profits.

I utilize the standard criterion that the noise traders coordinate their choice to minimize their costs.[10] If $T_1 > T_2$, the fact that execution costs are decreasing in an exchange's total risk tolerance implies that $x_1(1, T_1) < x_2(1, T_2)$. Therefore, in this case, the lowest cost equilibrium involves all noise traders choosing to trade on the exchange with the greatest risk bearing capacity–exchange 1.

This fact influences the equilibrium allocation of intermediaries among exchanges. This allocation must satisfy several equilibrium conditions.[11] First, in equilibrium no additional exchanges must be able to enter profitably. That is, no coalition of intermediaries outside the equilibrium exchange(s) can earn

a profit for each of its members by forming an exchange. Second, the members of an equilibrium exchange cannot increase their profits by altering the size of their exchange's membership. Third, if a total of $\hat{L}$ intermediaries belong to exchanges, then the equilibrium allocation requires intemediaries $\{1, \ldots, \hat{L}\}$ to belong to exchanges. This condition reflects the fact that exchange memberships are transferrable. If intermediary j is a member of an exchange, and intermediary $i < j$ is not, there is a price at which i could buy the membership from j that makes both parties better off.[12]

The only coalition of intermediaries that satisfies these conditions is $\mathbf{L}^* = \{1, 2, \ldots, L^*\}$, where $\sum_{j=1}^{L^*} t_j > .5T^A$, and $\sum_{j=1}^{L^*-1} t_j < .5T^A$. The intermediaries in $\mathbf{L}^*$ account for just over half of the total risk tolerance; if intermediary L^* were excluded from the coalition, the exchange would offer less than one-half of total risk tolerance. This exchange can attract all noise traders because execution costs are higher on every other possible coalition (since every other exchange has lower total risk tolerance). Moreover, an exchange consisting of some strict subset of the intermediaries in $\mathbf{L}^*$ would attract no business because another exchange with greater total risk tolerance would enter, capture all of the order flow, and earn a profit; such a subset cannot be an equilibrium exchange.

Furthermore, the members of $\mathbf{L}^*$ are harmed by the addition of more members. The appendix shows that exchange member j's profit is:

$$E(\Pi_j) = \frac{.5t_j\sigma^2 S}{T_1^2} \tag{6}$$

Since (6) implies that $dE(\Pi_j)/dT_1 < 0$, the profitability of an exchange member $j \in \mathbf{L}^*$ declines if additional members are added; increasing membership

beyond $\mathbf{L}^*$ increases the competition faced by those in $\mathbf{L}^*$, and thereby reduces their profits.[13]

Together, these results imply that in equilibrium, the exchange consists of the intermediaries $j \in \mathbf{L}^*$. Consequently, total equilibrium risk tolerance is $T_1^* = \sum_{j \in \mathbf{L}^*} t_j \approx .5T^A$. Given the formation of such a coalition, no other exchange can enter profitably. Moreover, both increases and decreases in the membership of this coalition reduce the profits of its members.

Thus, in the absence of cream skimming, the equilibrium exchange is a monopoly that limits the number of intermediaries it admits to increase the profits of its members. Limits on the number of members are a near universal feature of financial exchanges. This article derives these limits endogenously from fundamental microstructural considerations.[14]

Note that optimal risk bearing requires the exchange to admit all intermediaries $\{1, 2, \ldots, N\}$. The appendix shows that total cost with the monopoly exchange is $.5\sigma^2 S/T_1$. Total costs equal execution costs minus certainty-equivalent member profits minus informed trading profits. The cost of operating the market is minimized, and welfare is maximized, when $T_1 = T_A$. The exchange has no incentive to grow this large, however. By limiting membership to $\mathbf{L}^*$, it is immune from competitive entry by another exchange and does not dissipate profits as would be the case if more intermediaries were admitted. Therefore, limits on exchange size cause deadweight losses. They also generate profits for exchange members. The model implies that exchange members should earn economic rents. Pirrong (1999) provides evidence of substantial economic rents accruing to members of US equity and derivatives exchanges.

In this regime, therefore, all trading is centralized even though centralization is not compelled; it occurs naturally due to the centripetal force of liquidity. The resulting equilibrium is *not* first-best, however, because a suboptimally small coalition of liquidity suppliers can create a monopoly exchange that supplies too little risk bearing capacity. Thus, centralization is *not* equivalent to efficiency when exchanges can restrict membership. In essence, the model shows that the nature of liquidity makes securities trading a network industry. A suboptimally small "network" survives as a natural monopoly in equilibrium because no other network can compete on equal terms. Restricting the size of the network raises the profits of those intermediaries who can trade on it because they face less competition than would prevail in a first-best world. Thus, centralization of trading is not necessarily a good thing. This raises the possibility that fragmentation can improve efficiency. The next subsection considers this possibility.

3.2 Regime Two

The preceeding analysis shows that only one exchange that trades in a non-discriminatory fashion can survive in equilibrium. I next show that a cream skimming trading venue can survive in competition with a non-discriminatory exchange.

In this regime, liquidity suppliers who are excluded from the exchange can trade off-exchange in what I will refer to as a third market. Note that intermediaries who trade on the third market *must* restrict their dealings to the fraction q^* of the noise traders who are U1s who can be identified using the screening technology. That is, only cream skimming intermediaries

can survive in competition with a non-discriminatory exchange. This is true because the analysis of Regime One implies that if (1) the exchange membership offers total risk tolerance $T_1 > .5T_A$, and (2) the third market dealers do *not* restrict their dealings to the demonstrably uninformed, all noise traders choose to trade on the exchange because it offers greater risk tolerance.

In what follows, I assume that all intermediaries who are excluded from the exchange trade in the third market. That is, I assume that entry to the third market is open and unrestricted.[15] I also assume that the exchange continues to restrict membership to $\mathbf{L}^*$.[16]

Recall that cream skimming dealers can free ride on the exchange's price discovery. That is, third market dealers' estimate of the variance of the price of the traded asset is $\hat{\sigma}^2$, not σ^2. Since there is no informed trading in the third market, an analysis like that used to derive (4) implies that the λ of the third market is $\lambda_3 = \hat{\sigma}^2/T_3$, where T_3 is total risk tolerance on the third market. Therefore, in the first regime, the expected execution cost of each trader who chooses to trade in the third market is:[17]

$$x_3(T_3) = \frac{\hat{\sigma}^2 S}{T_3} \tag{7}$$

Assuming that exchange membership is given by the coalition $\mathbf{L}^*$, where as before this coalition offers just more than half of the total risk tolerance, and there is free entry onto the third market, total risk tolerance thereon is $T_3 = T_A - T_1^* \approx T_1^*$. A comparison of (7) to (4)-(5) shows immediately that average execution costs on the exchange assuming all noise traders trade there is higher than average execution cost on the third market. That is, $x_1(1, T_1^*) > x_3(T_3)$. Moreover, since $x_1(q_1, T_1^*)$ is decreasing in q_1, $x_1(1 -$

$q^*, T_1^*) > x_1(1, T_1^*) > x_3(T_3)$. Average execution costs are lower on the third market than on exchange because those who trade in the third market bear no adverse selection costs.

This analysis implies that all noise traders who can use the third market–the demonstrably uninformed U1s–will do so if the membership of the exchange remains unchanged. When exchange membership is $\mathbf{L}^*$ and third market dealers can observe the outcome of exchange trading, switching to the third market reduces noise trader execution costs. Thus:

Result 1 *Fraction q^* of noise trading takes place on the free-riding third market.*

The foregoing implies that the third market attracts all the demonstrably uninformed, whereas all others trade on exchange. This analysis implies that prices on the third market should be less informative than trading on the exchange. As noted earlier, there is substantial empirical evidence consistent with this prediction. Result 1 and the fact that λ_1 is increasing in q^* together imply that the creation of a third market reduces execution costs for the noise traders who can switch to the third market, and raises the execution costs of those who cannot. The effect of the entry of a third market on total noise trader execution costs depends on which effect dominates.

Total noise trader execution costs on exchange and third market are:

$$x^*(T_1^*) = S[(1-q^*)\lambda_1(T_1^*, 1-q^*) + q^*\lambda_3(T_3)] \tag{8}$$

where λ_1 is given by (4) with $S_1 = (1-q^*)S$, and λ_3 is given above; the notation is expanded to recognize the dependence of the λ's on q^* and T_1

and T_3. After some substitutions, this expression becomes:

$$x^*(T_1^*) = S\frac{\hat{\sigma}^2(1-q^*)}{T_1^*} + \beta_1(1-q^*)\hat{\sigma}^2(1-q^*) \tag{9}$$

where $\hat{\sigma}^2$ and β_1 are now written as functions to recognize explicitly their dependence on q^*. Recall that $\hat{\sigma}^2(1-q^*) < \hat{\sigma}^2(1)$ and $\beta_1(1-q^*) < \beta_1(1)$. Therefore, $x^*(T_1^*) < x_1(1, T_1^*)$. This proves:

Result 2 *Introduction of a free-riding open entry third market unambiguously reduces total noise trader execution costs.*

Thus, although the third market harms some noise traders, in aggregate noise traders are better off when a free riding third market is introduced.

Indeed, the third market increases total surplus if the third market free rides. The appendix shows that with free riding total cost equals:

$$TC_3 = \frac{.5\sigma^2(1-q^*)S}{T_1^*} + \frac{.5\hat{\sigma}^2 q^* S}{T_3}$$

Since $\hat{\sigma}^2 < \sigma^2$ and $T_3 \approx T_1^*$, TC_3 is smaller than the total cost incurred when there is no third market, $.5\sigma^2 S/T_1^*$. Thus:

Result 3 *The free-riding open entry third market unambiguously improves welfare.*

This improvement is attributable to the fact that the third market improves the efficiency of risk bearing. The third market dealers supply additional risk bearing capacity to the market. Although this reduces the profits of the exchange members, their loss is more than offset by the gains realized by noise traders and third market dealers.[18]

Although equilibrium surplus in the second regime is larger than in the first, the second regime equilibrium is not first best. Note that

$$TC_3 > \frac{.5t\sigma^2 S}{T_A}$$

if

$$.5 > q^*(1 - \frac{\hat{\sigma}^2}{\sigma^2})$$

Since $\hat{\sigma}^2 < \sigma^2$, this expression holds for $q^* < .5$. Since $.5\sigma^2 S/T_A$ is the cost of operating the market when all liquidity suppliers trade on the exchange, total costs are not minimized in the second regime even though they are lower than in the first regime.

These results imply that an open entry third market that free rides on exchange prices improves market performance. This may seem counterintuitive as it implies that an externality–the free acquisition of costly trade information by the third market–improves welfare.[19] This result obtains because we are in the world of the second best. The "tippiness" of the exchange market leads to a natural monopoly that restricts the supply of risk bearing to enhance its members' profits. This is inefficient. The externality reduces the costs of enhancing the supply of risk bearing and mitigates the inefficiency.

3.3 Regime Three

When access to any trading venue must be open, liquidity suppliers must choose which one to trade on. The analysis of section 3.1 implies that only a single non-cream skimming venue can exist. Therefore, liquidity suppliers must choose between trading on the exchange (which does not cream skim) and the third market (which does).

There may be several equilibria in this regime. Note that as exchange risk tolerance T_1 increases, execution costs fall on the exchange and rise on the cream skimming market (because a rise in T_1 implies a decline in T_3). Thus, there is some critical value of T_1, $\hat{T}_1$, such that if $T_1 > \hat{T}_1$ (and hence $T_3 < T_A - \hat{T}_1$) the third market cannot survive.

The fact that the third market must achieve some critical mass to survive implies that under most conditions one equilibrium is for all liquidity suppliers to join the exchange. Specifically, this is an equilibrium if $T_A - t_1 > \hat{T}_1$. To see why, assume initially that all market makers join the exchange. If any single market maker leaves the exchange, third market risk tolerance $T_3 \leq t_1 < T_A - \hat{T}_1$. Therefore, the sole third market dealer gets no business, and earns a profit of zero. This is smaller than his profit on the exchange. Thus, there is no incentive to defect and $T_1 = T_A$ is an equilibrium.

Equilibria that exhibit fragmentation may exist as well. This is most easily depicted graphically, as in Figure 2. The figure depicts two curves. The downward sloping curve depicts $\Pi_j^1(.)$, the profit of market maker j if he joins the exchange when its total risk tolerance is T_1'. This curve is downward sloping because the profitability of belonging to the exchange declines as the quantity of risk bearing capacity its members can supply increases. The upward sloping curve is $\Pi_j^3(.)$, market maker j's profit of trading on the third market when exchange risk tolerance is T_1'. This curve is upward sloping because an increase in T_1' lowers the supply of the risk bearing capacity of third market dealers, which increases the profit of those who remain. There is a discontinuity in each curve at $\hat{T}_1$. There is a downward discontinuity in Π_j^3 because the profit of trading on the third market goes to zero if the third

market does not achieve critical mass. There is an upward discontinuity in Π_1^j because the U1's shift their trading to the exchange if the third market fails to achieve critical mass.

Place Figure 2 Here

In Figure 2 $q^* = .1$, $S = 10$, $\sigma^2 = .75$, $T_A = 8$, and $K = 5$. Given these parameters, the exchange and third market profit functions cross. The intersection of these curves is an equilibrium. For a value of T_1' to the right (left) of the intersection, a third market dealer (exchange member) could increase his profit by joining the exchange (third market). In this case, fragmentation is an equilibrium outcome, but recall it is not the *only* equilibrium. It is possible to show that the fragmented equilibrium is not first-best. Thus, fragementation with an open entry exchange is inefficient.

Figure 3 depicts a situation in which fragmentation cannot occur because it is more profitable to trade on exchange than in the third market for all values of T_1'. In this figure, $q^* = .1$, $S = 10$, $\sigma^2 = 5$, $T_A = 8$, and $K = 5$. In this case joining the exchange is a dominant strategy for all liquidity suppliers. This outcome is efficient.

Place Figure 3 Here

These results provide an interesting contrast to those in Glosten (1994). Glosten assumes the existence of an open access central market. In his model, only this market survives; no cream skimming market can survive. In con-

trast, in the present model a cream skimming market may survive. These different results are attributable to the fact that Glosten assumes that all trading is anonymous, and market makers can attempt to identify uninformed traders using trade size alone. In contrast, if some traders can be identified as uninformed by means other than trade size, a cream skimming market may survive in competition with an open entry exchange.

3.4 Regime Four

Equilibrium in Regime Four is quite simple. Since cream skimming is not permitted, the analysis in section 3.1 implies that all transactions occur on an exchange that conducts a non-discriminatory auction. Moreover, since the exchange is open to all, all liquidity suppliers join it. Thus, the first best is achieved in this regime.

Note that it is unnecessary to *mandate* centralization. In this model, it occurs in equilibrium as a result of the centripetal forces of liquidity. In this regime, there is no competition between *exchanges* in equilibrium. There is, however, the greatest possible competition between liquidity suppliers, all of whom trade on the open access exchange. The exchange provides the infrastructure on which competing intermediaries operate. In an electronic market, the open access exchange could consist of an order execution mechanism and order book facility with an open interface to which liquidity suppliers connect (perhaps through portals provided by ECNs.)[20]

Combined with the analysis of Regime Three, the analysis of Regime Four provides a rationale for restrictions on the operation of cream skimming markets. As noted in section 3.3, if entry to the exchange is open, cream

skimming can lead to inefficient fragmentation. Measures such as requiring all orders be submitted to a non-discriminatory auction market can therefore improve welfare *if* there are no restrictions on liquidity suppliers' access to the central market. That is, to paraphrase Glosten (1994), a central limit order book may *not* be inevitable, and some regulatory constraints may be required to ensure its operation. By the same token, however, mandating that all trade occur on a non-discriminating market can actually reduce welfare if liquidity suppliers' access to that market is restricted.[21]

3.5 Summary

Security market macrostructure depends on the interaction of two key variables: (1) the ability of exchanges to restrict liquidity supplier access, and (2) the ability of off-exchange liquidity suppliers to "skim" some uninformed order flow. If cream skimming cannot occur (due to the inability to screen the uninformed or some regulatory restriction), exchanges that have the right to limit membership can exploit the nature of liquidity to restrict the supply of risk bearing capacity and increase exchange member profits. With limited exchange membership, the entry of a cream skimming third market increases welfare, but does not result in a first best outcome. Conversely, if the exchange cannot restrict its membership, cream skimming can lead to inefficient fragmentation. In the model, a requirement that exchanges admit any liquidity supplier combined with a ban on cream skimming markets produces a first best outcome.

The nature of liquidity drives these results. Liquidity exerts a centripetal force that attracts trading to a central market. Liquidity suppliers can exploit

this force to raise profits by restricting entry if allowed to do so. Fragmentation is one market response to such strategic behavior, but a more efficient result obtains if exchanges cannot limit the number of members.

4 Flies in the Ointment

The results of the foregoing section imply that elimination of off-exchange trading *can* improve welfare, just as the critics of cream skimming argue. However, this desirable outcome occurs only if access to the primary market is unrestricted: eliminating cream skimming without imposing a corrollary duty on exchanges to open admission to all actually reduces welfare. Thus, the analysis suggests that the optimal security market involves the creation of an open access central limit order facility and the simultaneous elimination of any cream skimming markets.

These results obtain because the nature of liquidity creates network effects. Efficiency requires maximization of the size of the network. Self-interested agents may not have an incentive to achieve this outcome because restricting network size can increase their profits; due to network effects, a restricted-size liquidity network need not fear direct competition if it exceeds some size threshold. Thus, mandated open access is required to achieve an efficient outcome.

Viewed in this light, regulation of a securities market bears strong similarities to regulation of other network industries. These include telecommunications, electricity transmission, natural gas transportation, and (perhaps) computer operating systems. In the first three industries concerns about market power due to network effects first led to rate regulation. Dereg-

ulation of these industries in the United States (and elsewhere) has been accompanied by open access requirements; these requirements are intended to prevent network operators from exercising market power by limiting access to the network.

The experience in deregulating network industries sounds a warning to securities market regulators. This experience shows clearly that although open access is easily stated as a goal, it is difficult indeed to implement in practice. The details of open access in network industries are devilish.

One key difficulty is that if the trading network is privately owned by a firm or group of firms that supply liquidity on it, the owner(s) may be able to restrict access to the network through manipulation of interfaces or other technical means. The owner may rationalize these policies on technical grounds, which may include network security or stability. In a securities trading context, the private for-profit operator of the centralized market may use solvency and performance concerns to constrain access by imposing unduly burdensome financial requirements on would-be participants; restriction of access in this way would allow the operator to increase profits if it also provides liquidity on the system.

Disintegration–rules that preclude the owner-operator of the trading network from trading itself–would diminish the incentive for the owner-operator to limit access in this fashion. Notably, disintegration has been a feature of deregulation in several network industries in the US. Disintegration does not eliminate another difficulty, however; the owner-operator may exploit market power derived from the nature of liquidity by charging supracompetitive prices to liquidity suppliers for access. Moreover, disintegration can increase

transactions costs (Joskow, 2000).

The foregoing problems–the difficulty of enforcing open access and the potential for supracompetitive pricing–have been constant themes in discussions of the regulation of network industries. Attempts to resolve these difficulties raise their own problems. In particular, elaborate rules designed to ensure that network operators do not restrict access increase the potential for inefficient gaming behavior. Moreover, rules intended to make networks more accessible may turn the network into a quasi-public good. This tends to reduce the ability of the network owner-operator to internalize benefits from improving the quality of the network or reducing operating costs. Finally, mitigating market power through rate regulation or other means leads to well-known incentive and information problems.

Although the formal analysis does not consider the costs of building, operating, or pricing access to a central market, they are likely to be important practical concerns in securities markets. This is particularly true given the ongoing technological revolution in securities trading. A state-of-the-art securities trading system is capital intensive. Moreover, due to technological change, it is likely that there is considerable scope for innovation and future system enhancement. Under these circumstances, regulations intended to ensure open access may reduce the incentive of the system owner-operator to improve and innovate. Furthermore, the necessity of incurring large fixed costs to create a trading system requires implementation of Ramsey-Boiteux pricing mechanisms to achieve efficiency. Regulators have faced difficulties in implementing such mechanisms in other network industries.[22]

In brief, the experience of other network industries suggests that im-

plementing open access in securities markets raises substantive regulatory questions. Who should own the trading system/network? How should it be governed? How should it be priced? Who should have control rights? What is the right organizational structure for the owner of the system? How must the owner(s)' property rights (notably the right to exclude) be attenuated to achieve open access? How will these attenuations of rights influence incentives to improve system performance and to develop system enhancements? These questions are not exhaustive. Moreover, the answers may be technology dependent.[23]

These are questions that securities regulators have never really addressed because heretofore all securities regulation (at least in the United States) has been undertaken in an environment in which intermediaries own exchanges and can limit entry thereto; in the terminology of this article, Regime Two is the default environment. In this environment, fragmentation and free riding are the contentious issues because the issue of access is not even raised. If such an environment is considered immutable, the fact that such exchanges can profit from network effects by restricting access implies that restrictions on cream skimming and fragmentation are unwise.

However, if regulators and legislators attempt to improve market efficiency by forcing open access (perhaps by creating an open access central limit order book), they must address the serious difficulties that have plagued in other network industries. Although the formal analysis suggests that actions which cause more trading to occur in an open access central market (e.g., eliminating cream skimming) cannot reduce welfare, and may increase it, practical considerations temper this conclusion. In particular, true open

access may be impossible to achieve and exchanges may exercise market power as a result. Under these circumstances, off-exchange markets–third markets–may provide a valuable competitive check on exchanges. Thus, even if an ostensibly open access trading mechanism is created, it may prove wise to permit off-exchange trading venues to operate. This raises the danger of inefficient fragmentation, but serves to mitigate the threat of market power exercised by circumvention of the open access goal or monopoly pricing of access to the trading system.

5 Summary and Conclusions

The macrostructure and efficiency of a securities market depend on the interaction between access and information externalities. When access to all trading venues is unrestricted, free riding on price information generated on exchanges by cream skimming satellite markets may be inefficient. Conversely, if some markets (exchanges) limit access, fragmentation of trading through the creation of cream skimming off-exchange markets can improve welfare. These results derive from the nature of liquidity. Liquidity creates network effects that can be exploited strategically.

Although an open entry central trading facility and the elimination of cream skimming leads to an efficient securities market macrostructure in theory, achieving this outcome is not a trivial task. Securities market regulators who attempt to create an open access system will face the same difficulties that regulators of other network industries have struggled with for years. Open access is difficult to achieve in practice, as the operators of networks may have the incentive and ability to offer access to a suboptimally small

number of participants either explicitly or through supracompetitive pricing. Moreover, rules and regulations designed to combat such strategic behavior may make crucial parts of the central trading facility public goods. If these are not priced properly, there will be overconsumption and underproduction of key attributes of the trading system.

Put differently, securities markets are made, not born.[24] Making them efficient requires the specification of the appropriate property rights. The current property rights regime gives securities exchanges the right to exclude intermediaries from membership and allows considerable free riding on exchange-generated price information; although the second attribute of this regime has received considerable attention, the first has not. The analysis of this article implies that the exclusionary practices of exchanges leads to inefficient risk bearing, but that free riding and cream skimming mitigate these inefficiencies. This article also implies that at a theoretical level, this property rights structure is exactly backwards; an efficient structure would deny exchanges the right to exclude but would prevent off-exchange dealers from free riding on exchange price information and skimming uninformed order flow. In essence, restrictions on property rights similar to those imposed on common carriers under common law can improve securities market efficiency.

Many property rights issues need to be addressed if an open access trading mechanism is adopted, but the securities trade is not the first industry to grapple with them. They have been central to debates in other network markets, including telecommunications and electricity transmission. These industries also show the diversity of institutions and regulations that have developed to address property rights issues in network markets. Only time

will tell what institutions will develop in securities markets. Although the specifics are not yet clear, it is evident that getting the property rights right is essential to the creation of efficient institutions for trading securities.

A Appendix

Derivation of λ_i. Conditional on order flow, liquidity supplier j chooses his trade y_j to maximize his certainty-equivalent profit. Formally:

$$E\Pi_j = \max_{y_j}\{y_j E[v - P|K\beta_i v + z_i] - \frac{.5\hat{\sigma}^2 y_j^2}{t_j}\} \tag{10}$$

where $\hat{\sigma}^2$ is the variance of v conditional on total order flow $K\beta_i v + z_i$, and where P is given by (1). The first term inside the brackets is the market maker's expected profit from a trade of y_j units. The second term adjusts for the risk of holding y_j units; $\hat{\sigma}^2 y_j^2$ is the variance of j's wealth, and $-.5/t_j$ is the cost per unit of variance.

Note that due to the normality of v and z_i, $E[v|K\beta_i v + z_i]$ is given by the regression of v on $K\beta_i v + z_i$. Thus,

$$E[v|K\beta_i v + z_i] = \frac{K\beta_i \sigma^2}{K^2\beta_i^2\sigma^2 + S_i}(K\beta_i v + z_i) \tag{11}$$

Moreover, by (1), $E[P|K\beta_i v + z_i] = \lambda_i(K\beta_i v + z_i)$, and

$$\hat{\sigma}^2 = \frac{S_i \sigma^2}{K^2\beta_i^2\sigma^2 + S_i} \tag{12}$$

Therefore,

$$y_j = \frac{t_j[\frac{K\beta_i\sigma^2}{K^2\beta_i^2\sigma^2+S_i} - \lambda_i](K\beta_i v + z_i)}{\hat{\sigma}^2} \tag{13}$$

Call $\mathbf{L_i}$ the set of intermediaries on exchange i. Market clearing implies:

$$z_i + \sum_{j\in\mathbf{L_i}} y_j + K\beta_i v = 0. \tag{14}$$

Thus,

$$\frac{T_i[\frac{K\beta_i\sigma^2}{K^2\beta_i^2\sigma^2+S_i} - \lambda_i](K\beta_i v + z_i)}{\hat{\sigma}^2} + K\beta_i v + z_i = 0 \tag{15}$$

where $T_i = \sum_{j\in \mathbf{L_i}} t_j$. This, in turn, implies:

$$\lambda_i = \frac{\hat{\sigma}^2}{T_i} + \frac{K\beta_i\hat{\sigma}^2}{S_i} \tag{16}$$

Proof that $\hat{\sigma}^2$ is increasing in S_i. To see that conditional price variance is increasing in S, recall that

$$\hat{\sigma}^2 = \frac{S\sigma^2}{K^2\beta^2\sigma^2 + S} \tag{17}$$

Thus, the sign of $d\hat{\sigma}^2/dS$ is given by the sign of:

$$S + \sigma^2K^2\beta^2 - S(1 + 2\sigma^2K^2\beta\frac{d\beta}{dS}) = K^2[\sigma^2\beta^2 - 2S\sigma^2\beta\frac{d\beta}{dS}] \tag{18}$$

The quadratic that defines β is:

$$K\beta^2\sigma^2 + \frac{(K+1)S\sigma^2\beta}{T_1} - S = 0 \tag{19}$$

Therefore:

$$\frac{d\beta}{dS} = \frac{1 - \frac{(K+1)\sigma^2\beta}{T_1}}{2K\beta\sigma^2 + \frac{(K+1)S\sigma^2}{T_1}} \tag{20}$$

Making further substitutions from the quadratic implies:

$$\frac{d\beta}{dS} = \frac{\beta[1 - \frac{(K+1)\sigma^2\beta}{T_1}]}{S + K\beta^2\sigma^2} > 0 \tag{21}$$

Thus,

$$2S\beta\frac{d\beta}{dS} = \frac{2S\beta^2(1 - \frac{(K+1)\sigma^2\beta}{T_1})}{S + K\beta^2\sigma^2} \tag{22}$$

This implies:

$$\sigma^2\beta^2 - 2S\sigma^2\beta\frac{d\beta}{dS} = \frac{\beta^2\sigma^2}{S + K\beta^2\sigma^2}[K\beta^2\sigma^2 + \frac{2(K+1)S\sigma^2\beta}{T_1} - S] \tag{23}$$

Since

$$K\beta^2\sigma^2 + \frac{(K+1)S\sigma^2\beta}{T_1} - S = 0 \tag{24}$$

$$\sigma^2\beta^2 - 2S\sigma^2\beta\frac{d\beta}{dS} = \frac{\sigma^2\beta^2}{S + K\beta^2\sigma^2}\frac{(K+1)S\sigma^2\beta}{T_1} > 0 \tag{25}$$

The inequality holds because $\beta > 0$.

Derivation of Exchange Member Profit. First note that by (13) and (15), the position of trader $j \in \mathbf{L}^*$ is equal to

$$y_j = -\frac{t_j}{T_1}(K\beta v + z) \tag{26}$$

where subscripts are suppressed because there is only a single exchange.

The expected certainty-equivalent profit of any member $j \in \mathbf{L}^*$ is given by:

$$E(\Pi_j) = E[y_j(v - \lambda_1(K\beta v + z)) - \frac{.5\hat{\sigma}^2 y_j^2}{t_j}] \tag{27}$$

where this expectation is taken over the unconditional joint distribution of v and z. Therefore,

$$E(\Pi_j) = -\frac{t_j K\beta\sigma^2}{T_1} + \frac{t_j}{T_1}[\lambda_1 - \frac{.5\hat{\sigma}^2}{T_1}](S + K^2\beta^2\sigma^2) \tag{28}$$

After some additional substitution, this reduces to

$$E(\Pi_j) = \frac{.5t_j\sigma^2 S}{T_1^2} \tag{29}$$

Derivation of Total Cost. The total cost of operating the market equals noise trader's execution costs minus informed trader profits minus certainty-equivalent market maker profits. Given v and z, exchange execution costs are $z\lambda_1(\beta v + z)$, informed traders' profits are $-K\beta v\lambda_1(K\beta v + z) + Kv^2/(K+1)\lambda_1$ and certainty-equivalent market maker profits are:

$$\sum_{j=1}^{L^*}\{-\frac{t_j}{T_1^*}(K\beta v + z)[v - \lambda_1(K\beta v + z)] - \frac{.5t_j\hat{\sigma}^2(K\beta v + z)^2}{T_1^{*2}}\}. \tag{30}$$

Since $\sum_{j=1}^{L^*} t_j = T_1^*$, this simplifies to:

$$-(K\beta v + z)[v - \lambda_1(K\beta v + z)] - \frac{.5\hat{\sigma}^2(K\beta v + z)^2}{T_1^*}. \tag{31}$$

Substituting for $\hat{\sigma}^2$ and simplifying implies that the total cost of trading on the exchange is:

$$vz + (K^2\beta^2 v^2 + 2K\beta vz + z^2)\frac{.5(1-q)S\sigma^2}{T_1^*[K^2\beta^2\sigma^2 + (1-q)S]} \tag{32}$$

Taking expectations over v and z implies that expected total cost equals:

$$\frac{.5\sigma^2 S(1-q)}{T_1^*} \tag{33}$$

Similar analysis implies that with free riding, the expected total cost of operating the third market is:

$$\frac{.5\hat{\sigma}^2 qS}{T_3} \tag{34}$$

Endnotes

[1] This is a simplified version of the Chowdhry-Nanda (1991) framework. In some versions of their model they preclude some noise traders from choosing where to trade. In contrast, all noise traders in the present model are "discretionary" in their terminology. Chowdhry-Nanda also include a large noise trader who can split orders between exchanges. When all noise traders in their model can choose where to trade, the large noise trader ends up trading on a single market. Since this result would obtain in the present model, I simplify the approach by considering only small discretionary noise traders. Unlike Chowdhry-Nanda, I assume that market makers are risk averse.

[2] Admati-Pfleiderer (1991) also assume the existence of an exogenous number of noise traders that can credibly disclose that they are uninformed.

[3] The analysis can be extended to consider $q^* < 1$, but the model's predictions for $q^* > .5$ are counterfactual so I restrict attention to the more realistic case of $q^* < .5$. Moreover, at the expense of considerable additional formalism, the analysis can be extended to the more realistic case where there are multiple noise trader types. Specifically, all major results derived below hold if the noise trader types are indexed by $q = 1, 2, \ldots N_T$, and where the cost of verifying whether a trader of type q is uninformed is given by the increasing, convex function $c(q)$. Since the results are robust to changes in the screening assumption, the text focuses on the simpler, more transparent case.

[4] See O'Hara (1997) for a discussion of institutions that facilitate the identification of uninformed traders.

[5] Critics of third markets, including Easley et al. (1996) and Mulherin

et al. (1991) assert that third market dealers can identify some of the uninformed, and that this cream skimming is detrimental to market performance. To evaluate these claims, it is necessary to derive the logical implications of cream skimming. The partial screening assumption permits such an evaluation and is hence responsive to the existing literature on security market structure. Moreover, as the analysis in the text demonstrates, this assumption is descriptively accurate and has implications that are consistent with extensive empirical research.

[6] See Diamond-Verrecchia (1991), Admati-Pfleiderer (1991), Subrahmanyam (1991), and Brown-Zhang (1997) for examples of models involving market maker risk aversion.

[7] The assumption of batch auctions is for convenience only. The fundamental factors drive the results of this article–the relation between execution costs and the risk bearing capacity of market makers, and adverse selection costs–have the same effects in continuous markets as in batch markets. Using the standard batch auction model greatly simplifies the analysis and makes the key insights more transparent.

[8] This is similar to the result in Brown and Zhang (1997).

[9] Traditional scale economies due to fixed costs in the creation or operation of a trading system can also lead to centralization. See Pirrong (1999) for a formal model that derives such a result. To generate fragmentation, however, it is necessary to include adverse selection costs and cream skimming.

[10] Fudenberg and Tirole (1999) claim that this is the "standard equilibrium selection in static network models." Shy (2000) similarly notes that this

"no coordination failure" assumption is standard in network models. Moreover, the analysis can be made dynamic. Given the assumptions made here, if noise traders choose where to trade sequentially, Farrell and Saloner (1985) Propositions 1 and 3 imply that the unique perfect equilibrium is for noise traders to choose the exchange that minimizes total execution cost.

[11] See Pirrong (1999) for a more formal statement of these conditions.

[12] Expression (6) below shows that a member's profit is increasing in t_j, which implies the stated result.

[13] Competition between members does not drive their profits to zero because (a) each member's supply curve of risk bearing services is upward sloping due to risk aversion, and (b) the number of market makers is finite. Thus, the exchange supply curve of risk bearing services is upward sloping, and members earn a scarcity rent in equilibrium. Restrictions on entry increase the scarcity rent.

[14] Exchange members may enhance their profits by other means, such as mandating a supracompetitive "tick" size or collusion. Network effects give them the market power required for these arrangements to survive.

[15] This assumption is motivated by the observation that most historical third markets, including the OTC market in listed stocks, bucket shops, and so on, have not restricted entry. See Pirrong (2001) for an analysis of restricted-entry third markets. That analysis shows that restrictions on the size of the third market reduce surplus. This reinforces the basic claim of this article that entry restrictions in financial markets are a source of inefficiency.

[16] The exchange will limit membership to $\mathbf{L}^*$ for some values of q^*, but if q^* is big enough it may respond to the competitive threat of the third market

by expanding membership. See Pirrong (2001) for a detailed analysis of this case. That analysis shows that the exchange will always offer risk tolerance that is less than T_A, so the equilibrium is not first best even if the exchange expands. However, total cost is lower if the exchange expands than if it does not. Therefore, the potential for free riding and cream skimming improves welfare relative to a regime where exchanges can limit entry and need fear no competition from cream skimming markets.

[17] The execution price on the third market has mean $E(v|P_1)$, where P_1 is the exchange price. The third market price varies randomly around this mean with random variations in noise trader order flow because third market dealers require compensation for bearing the risk taken on when they absorb noise trader order imbalances.

[18] If information is costly to obtain, and therefore the number of informed traders is endogenous, there is another welfare gain from the third market. The third market reduces the returns from information, and consequently leads to reduced expenditures on information. This is beneficial because informed trading is a form of rent seeking in this model (and other microstructure models as well). The benefits from the third market are reduced to the extent that screening is costly.

[19] Pirrong (2001) also shows that prices are more informative when a third market exists.

[20] Of course an exchange may perform other functions, such as clearing and market oversight.

[21] The formal analysis assumes that all noise traders are homogeneous, and care only about execution costs. In fact, liquidity demanders may be het-

erogeneous and may have different preferences for execution speed and other transaction attributes. In this case, multiple trading systems with differing attributes may coexist to accommodate customer heterogeneity even though this fragments liquidity, thereby increasing average execution costs. If all trading venues adhere to the principles of open access and non-discrimination (i.e., no cream skimming) the market is likely to offer near optimal variety. It should be noted that some existing trading mechanisms that offer different attributes (e.g., crossing networks that cater to patient traders who do not demand immediacy) may serve as mechanisms for screening out the informed. For instance, if patient traders have less information (as is plausible–the informed may want to trade quickly fearing that others will acquire the relevant information), periodic batch auctions or crossing systems may facilitate cream skimming. Thus, in Regime Two it is difficult to determine whether some trading systems succeed because they accommodate diverse customer needs, or because such accommodation facilitates cream skimming.

[22] See Laffont-Tirole (2000) for a discussion of obstacles facing implementation of Ramsey-Boiteux pricing in telecommunications.

[23] Pirrong (2000) shows that there is likely to be a linkage between trading technology and the efficient form of organization and governance of exchanges.

[24] Mulherin et al. (1991) emphasize this point.

References

[1] Admati, Anat, and Paul Pfleiderer. 1988. "A Theory of Intraday Patterns: Volume and Price Variability," *Review of Financial Studies* 3-40.

[2] Admati, Anat, and Paul Pfleiderer. 1991. "Sunshine Trading and Financial Market Equilibrium," 4 *Review of Financial Studies* 443-481.

[3] Bessembinder, Hendrik, and Harold Kaufman. 1997. "A Cross-Exchange Examination of Trading Costs and Information Flow for NYSE-listed Stocks," 46 *Journal of Financial Economics* 293-319.

[4] Brown, David, and ZhiMing Zhang. 1997. "Market Orders and Market Efficiency," 52 *Journal of Finance* 277-307.

[5] Chowdhry, Bhagan, and Vikram Nanda. 1991. "Multi-Market Trading and Market Liquidity," 3 *Review of Financial Studies* 483-511.

[6] Diamond, Douglas, and Robert Verrecchia. 1991. "Disclosure, Liquidity, and the Cost of Capital," 46 *Journal of Finance* 1325-1359.

[7] Easley, David, Nicholas Kiefer, and Maureen O'Hara. 1996. "Cream skimming or Profit-Sharing? The Curious Role of Purchased Order Flow., 51 *Journal of Finance* 811-833.

[8] Farrell, Joseph, and Garth Saloner. 1985. "Standardization, Compatibility, and Innovation," 16 *Rand Journal of Economics* 70-83.

[9] Fudenberg, Drew, and Jean Tirole. 1999. "Pricing Under the Threat of Entry by a Sole Supplier of a Network Good." Unpublished Manuscript, Harvard University.

[10] Glosten, Lawrence. 1994. "Is the Electronic Limit Order Book Inevitable?" 49 *Journal of Finance* 1127-1161.

[11] Hasbrouck, Joel. 1995. "One Security, Many Markets: Determining the Contributions to Price Discovery," 50 *Journal of Finance* 1175-1199.

[12] Huang, Roger, and Hans Stoll. 1994. "Competitive Trading of NYSE Listed Stock: Measurement and Interpretation of Trading Costs." Vanderbilt University Financial Markets Research Center Working Paper 94-13.

[13] Joskow, Paul. 2000. "Transaction Cost Economics and Competition Policy." Unpublished Manuscript, Massachussetts Institute of Technology.

[14] Kyle, Albert. 1985. "Continuous Auctions and Insider Trading," 53 *Econometrica* 317-355.

[15] Laffont, Jean Jaques, and Jean Tirole. 2000. *Competition in Telecommunications.* Cambridge: MIT Press.

[16] Mulherin, Harold, Jeffrey Netter, and James Overdahl. 1991. "Prices are Property: The Organization of Financial Exchanges from a Transaction Cost Perspective," 34 *Journal of Law and Economics* 591-644.

[17] O'Hara, Maureen. 1997. *Market Microstructure Theory.* Malden, MA: Blackwell Business Publishers.

[18] Pagano, Marco. 1989. "Trading Volume and Asset Liquidity," 104 *Quarterly Review of Economics* 255-274.

[19] Pirrong, Craig. 1999. "The Industrial Organization of Financial Markets" Theory and Evidence," 2 *Journal of Financial Markets* 329-358.

[20] Pirrong, Craig. 2000. "A Theory of Financial Exchange Organization," 43 *Journal of Law and Economics* 437-472.

[21] Pirrong, Craig. 2001. "Third Markets and the Second Best." Unpublished manuscript, Oklahoma State University.

[22] Shy, O. 2000. *The Economics of Network Industries.* Cambridge: Cambridge University Press.

[23] Smith, Bruce, Alasdair Turnbull, and Robert White. 2001. "Upstairs Market for Principal and Agency Trades: Analysis of Adverse Information and Price Effects," 56 *Journal of Finance* 1723-1746.

[24] Subrahmanyam, Avanidhar. 1991. "Risk Aversion, Market Liquidity, and Price Efficiency," 4 *Review of Financial Studies* 417-441.

Figure 1

Craig Pirrong

Figure 2

Craig Pirrong

Figure 3

Craig Pirrong

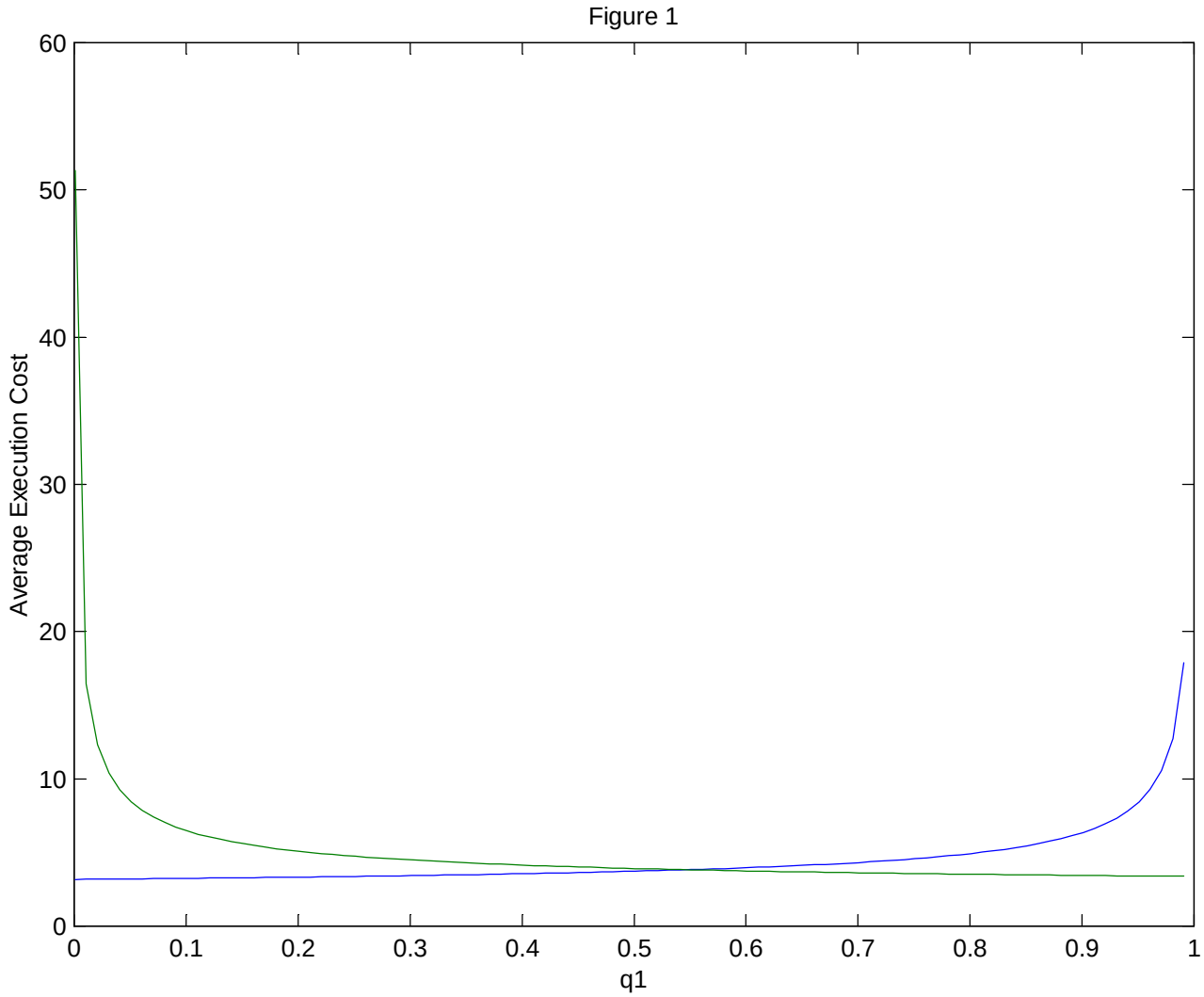
Figure 1
Average Execution Cost
q1
0
10
20
30
40
50
60
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

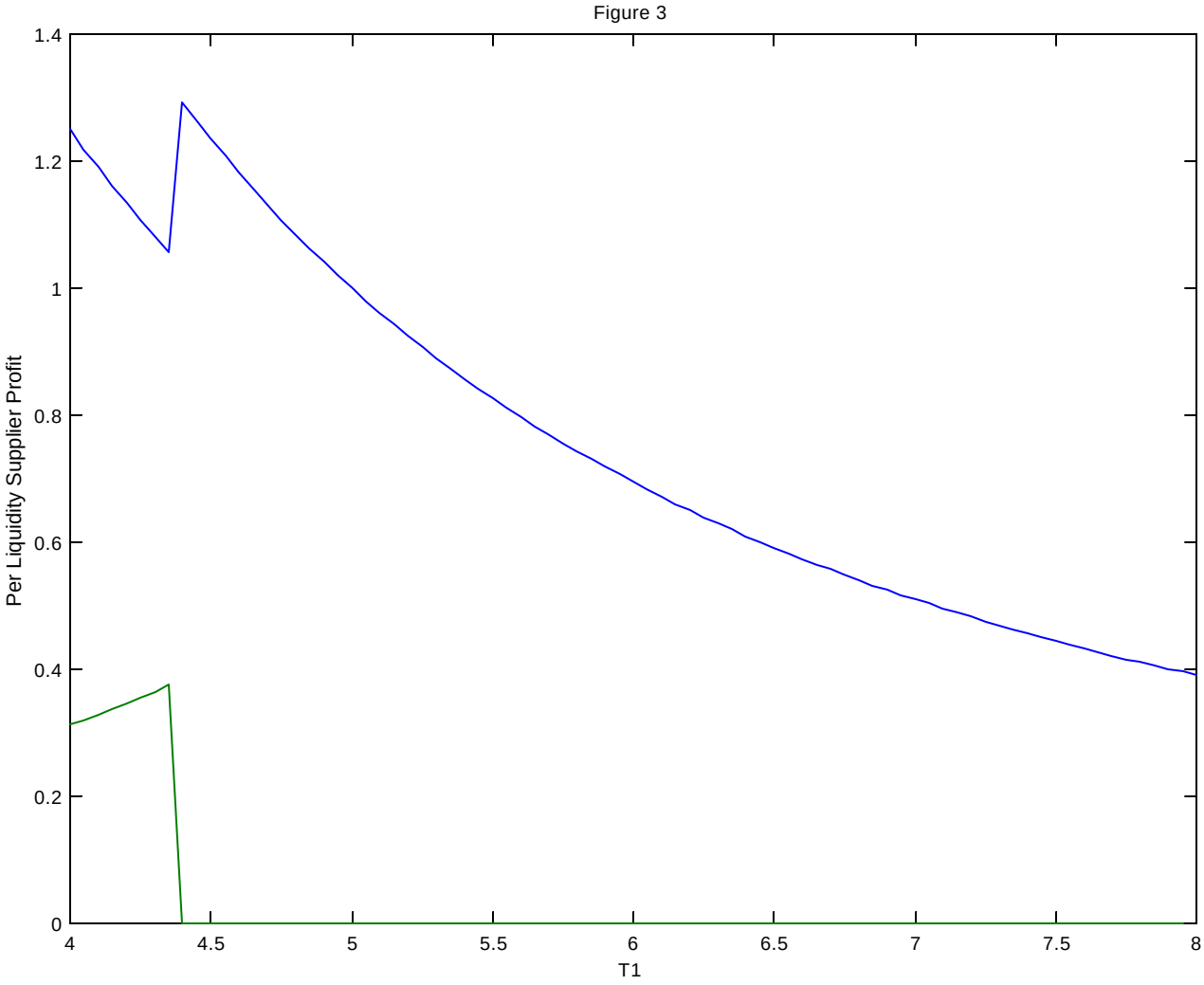
Figure 3
Per Liquidity Supplier Profit
T1
1.4
1.2
1
0.8
0.6
0.4
0.2
0
4
4.5
5
5.5
6
6.5
7
7.5
8

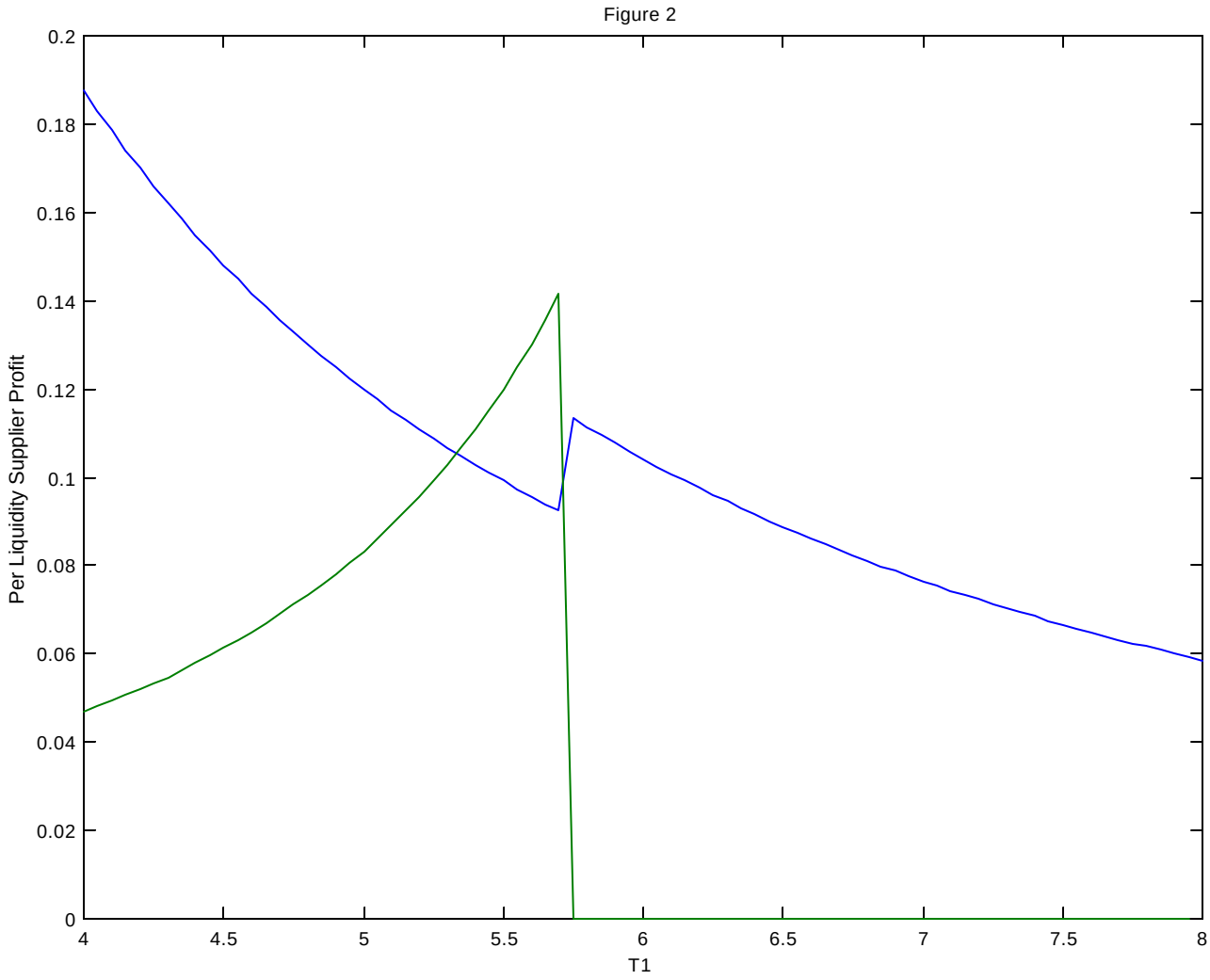
Figure 2
Per Liquidity Supplier Profit
T1
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
4
4.5
5
5.5
6
6.5
7
7.5
8