Competition and Collusion in Dealer Markets
Author(s): Prajit K. Dutta and Ananth Madhavan
Source: *The Journal of Finance*, Vol. 52, No. 1 (Mar., 1997), pp. 245-276
Published by: Blackwell Publishing for the American Finance Association
Stable URL: http://www.jstor.org/stable/2329563
Accessed: 25/11/2009 10:04

THE JOURNAL OF FINANCE • VOL. LII, NO. 1 • MARCH 1997

Competition and Collusion in Dealer Markets

PRAJIT K. DUTTA and ANANTH MADHAVAN*

ABSTRACT

This article develops a game-theoretic model to analyze market makers' intertemporal pricing strategies. We show that dealers who adopt *noncooperative* pricing strategies may set bid–ask spreads above competitive levels. This form of "implicit collusion" differs from explicit collusion, where dealers cooperate to fix prices. Price discreteness or asymmetric information are not required for collusion to occur. Rather, institutional arrangements that restrict access to the order flow are important determinants of the ability to collude because they reduce dealers' incentives to compete on price. Public policy efforts to increase interdealer competition should focus on such restrictions.

> Among the plays which men perform in taking different parts in this magnificent world theater, the greatest comedy is played at the Exchange. There, . . . hiding places, concealment of facts, quarrels, provocations, mockery, idle talk, violent desires, collusion, artful deceptions, betrayals, cheatings, and even tragic end are to be found.—Joseph de la Vega, *Confusion de Confusiones* (1688), describing the Amsterdam Stock Exchange.

IN TWO WIDELY PUBLICIZED articles, Christie and Schultz (1994) and Christie, Harris, and Schultz (1994) conclude that Nasdaq market makers implicitly collude to set bid–ask spreads for their stocks above competitive levels. Christie and Schultz (1994) document an absence of odd-eighth quotes for Nasdaq stocks, and offer implicit collusion among dealers as the most likely explanation for their results. Christie, Harris, and Schultz (1994) find a significant drop in spreads following the public disclosure of the findings of Christie and Schultz (1994). These assertions have triggered numerous class-action law suits against Nasdaq dealers and have led to investigations of Nasdaq dealer pricing by the U.S. Department of Justice and the Securities and Exchange Commission.

* Dutta is from the Department of Economics, Columbia University. Madhavan is from the School of Business Administration of the University of Southern California. We thank William Christie, David Easley, Michael Fishman, Larry Harris, Joel Hasbrouck, James Shapiro, and an anonymous referee for their helpful comments. We have also benefitted from the comments of seminar participants at The Ohio State University, the University of Pittsburgh, the JFI Symposium at Northwestern University, the 1995 European Finance Association Meetings, and the 1996 American Finance Association Meetings. Support from National Science Foundation Grant SBR-9410485 (Dutta) is gratefully acknowledged. The comments and opinions expressed in this article are those of the authors alone.

The extent to which the findings of Christie and Schultz (1994) and Christie, Harris, and Schultz (1994) actually reflect collusive behavior remains highly controversial. Some studies conclude that collusion is infeasible in the Nasdaq market, and that the Christie-Schultz results can be explained by other factors. For example, Grossman, Miller, Fischel, Cone, and Ross (1995) assert that the avoidance of odd-eighth quotes arises from the stock price clustering first analyzed by Harris (1991), while Kleidon and Willig (1995) and Furbush (1995) argue that there are minimal entry barriers in the Nasdaq market.

Alternatively, some studies acknowledge that spreads in Nasdaq stocks may be above competitive levels and provide explanations for this result. Godek (1995) focuses on structural features of the Nasdaq market as an explanation for the Christie-Schultz results; Bhasin and Brown (1995) show that the discounted value of expected future profits acts as a bond that induces dealers to fulfill their agreements to trade, thereby providing a justification for "excess" spreads; and Kandel and Marx (1995) argue that spreads in excess of competitive levels may arise from discrete prices.

Finally, there is disagreement regarding the extent to which the empirical evidence supports the hypothesis of implicit collusion. For example, Barclay (1995) finds that effective and quoted spreads for Nasdaq stocks fall after they are listed on an exchange. The drop is greatest for those stocks previously quoted on even-eighths, which appears consistent with the implicit collusion hypothesis. By contrast, Doran, Lehn, and Shastri (1995) and Laux (1995) suggest factors other than implicit collusion may explain the absence of odd-eighths.

This article develops a game-theoretic model to analyze the dynamic pricing strategies of competing market makers in a dealer market. Our objectives are (i) to analyze whether market makers who compete on price can sustain spreads above competitive levels, even if they do not explicitly cooperate to fix prices; (ii) to examine the relation between institutional arrangements specific to the Nasdaq market and dealer pricing behavior, and (iii) to discuss the role of public policy in facilitating price competition among dealers.

At this stage, it is worth considering why a formal model of dealer pricing behavior is needed. Indeed, the theoretical literature on infinite-horizon repeated games shows that there are typically equilibria where rival oligopolistic firms in a product market act as a cartel. However, there are several difficulties in generalizing these results to conclude that dealers in a security market may set prices equal to those of a profit maximizing cartel.

First, for basic economic reasons it is not appropriate to model trading as a repeated game. In particular, a repeated game is literally the repetition of the same one-shot game over and over again, and the market environment (and hence payoffs) are identical from period to period. Put differently, if a dealer market were modeled as a repeated game, then dealers' profits would be constant from period to period if they were to maintain constant bid and ask prices. This is not economically plausible. Intuitively, demand and supply depend on not just price alone, but on factors such as the population of investors, which may change either deterministically or stochastically over time. A dynamic game, by contrast, allows for such changes in the market environment. For example, dealers' profits may fluctuate with stochastic changes in volume even if bid and ask prices were constant. Thus, trading is best modeled as a *dynamic* game, not a repeated game. Since there is no general result like the Folk Theorem for dynamic games, collusive equilibria may not be attainable even in an infinite horizon game.

Second, collusion appears at first glance to be far more difficult among security dealers than among firms in a typical product market. Dealers in a security market compete on the basis of their quotes, and the best intermarket quotes can be easily found. In contrast, firms in product markets may compete on dimensions other than price, and customers may find it difficult or costly to locate the firms offering the lowest prices. Further, unlike many product markets, there are not, at first glance, significant barriers to entry in market making. For example, Kleidon and Willig (1995), citing the ease of entry and exit for Nasdaq dealers who are already providing market making services, conclude that "the Nasdaq market is characterized by the virtual absence of barriers to entry." They note that active dealers provide a large pool of potential entrants (425 in 1991) into any stock where spreads exceed competitive levels because of collusion.

Finally, institutional arrangements unique to security markets play an important role in dealer pricing. These include agreements between brokers and dealers to direct (or preference) customer order flow to preferred dealers, payments for order flow, and price discreteness. The impact on security prices of these institutional features is not well understood.

We demonstrate conditions under which market makers who act *noncooperatively* in setting prices may still set bid–ask spreads above competitive levels. We term such behavior "implicit collusion" because dealers, while acting in their own self-interest, earn abnormal profits. Implicit collusion differs from explicit collusion in that dealers do not cooperate to fix prices, i.e., they follow Nash strategies. As a result, the observed relation between prices and volume can differ between the two forms of collusive behavior. This is important for public policy because antitrust law distinguishes between implicit and explicit collusion.[2]

There are cases when dealers' pricing strategies under implicit collusion coincide with those under explicit collusion. This situation is especially interesting from a public policy viewpoint. On the positive side, it may be easier for regulators to distinguish between collusive and competitive pricing using data, as discussed below. Conversely, however, even if regulators conclude that

[1] See, e.g., Dutta (1995) for further discussion of this issue. Indeed, the literature on dynamic games suggests that firms generally cannot sustain the same prices as a cartel in all economic environments.

[2] The Sherman Act (1890) prohibits "contracts, combinations, or conspiracies" in restraint of trade such as price fixing agreements or explicit collusion. By contrast, the behavior that we refer to as implicit collusion need not be per se illegal since it may reflect "parallelism" where dealers independently make the same pricing decisions. Evidence of collusive behavior is therefore likely to require proof of an explicit pricing agreement among dealers.

dealer behavior is collusive, they may not find any evidence of explicit collusion (such as evidence of interdealer communications regarding pricing) since it is not possible to discriminate between implicit and explicit collusion.

Implicit collusion, as defined here, is not feasible in all markets. Indeed, if dealers are sufficiently impatient or there are no entry barriers, competitive pricing is the only equilibrium. Contrary to common belief, collusive behavior may be easier to sustain in more actively traded stocks than infrequently traded stocks. Further, implicit collusion is more difficult to sustain in a market where some dealers dominate the order flow as opposed to a market where dealers are roughly equal in size. Thus, simple measures of concentration (such as the four-firm concentration ratio or Herfindahl index) are unlikely to accurately measure the potential for collusion.

We also examine the effect of institutional features of the Nasdaq market on bid–ask spreads. We show that price discreteness is not necessary to support excess spreads. Even so, a reduction in the minimum price variation can narrow spreads by making implicit collusion more difficult to sustain. Of greater importance are institutional arrangements that allow dealers to receive directed (or preferenced) order flow provided they match the inside spread, even if they do not offer the best quotes themselves. This reduces dealers' incentives to compete on price and increases the range under which collusive prices can be sustained. In the absence of price competition, dealers may compete in other dimensions, paying to secure access to the order flow. Our results suggest that policy makers seeking to increase price competition among dealers should focus primarily on easing restrictions on access to the order flow.

The article proceeds as follows: Section I describes the model; Section II analyzes the dynamic pricing strategies of dealers; Section III extends the basic model to allow for endogenous entry and heterogenous dealers; Section IV models the impact of various institutional features; Section V discusses our results and their implications for public policy; and Section VI concludes. All proofs are contained in the Appendix.

I. The Theoretical Framework

A. Trading Protocols in a Dealer Market

Consider a risky stock that trades in a dealer market. To focus attention on the effects of dealers' intertemporal strategies, we assume that there are no transaction costs or adverse selection costs. At the start of each period, M dealers quote firm bid and ask prices at which they will purchase or sell a risky stock. For now, we will assume that M is exogenously determined.

Individual dealers simultaneously determine their bid and ask prices at the start of each trading round (indexed by $t = 1, 2, \ldots$) and these prices are irrevocable. Dealers maximize the expected present value of profits. Let a_t^j denote the ask price quoted by dealer j in round t and similarly define the bid price b_t^j. Let $a_t = \min[a_t^1, \ldots, a_t^M]$ denote the best (lowest) ask price in time t,

$b_t = \max[b_t^1, \ldots, b_t^M]$ the best (highest) bid price, and define the *inside bid–ask spread* as $\lambda_t = a_t - b_t$. In each trading round, traders submit their orders to dealers for execution at the best inside quotes in accordance with strict price priority. Initially, we will assume that the dealers are homogeneous in the sense that if N ($N \leq M$) dealers quote the inside spread, they share the order flow equally.

In round t, let $d_t \geq 0$ and $s_t \geq 0$ denote the number of shares bought and sold, respectively. The actual quantities traded depend on three factors: (i) The best ask and bid prices, (ii) A volume shock, z_t, which represents a transitory market-wide shift in the placement of demand and supply schedules, and (iii) Unpredicted demand and supply shocks, denoted by ϵ_t and γ_t respectively. Thus, the actual quantities demanded and supplied are given by:

$$d_t \equiv d(a_t, z_t) + \epsilon_t \tag{1}$$

$$s_t \equiv s(b_t, z_t) + \gamma_t, \tag{2}$$

where $s(b_t, z_t)$ and $d(a_t, z_t)$ are the expected supply and demand in period t, respectively. We assume that the volume shock z_t is an independent and identically distributed random variable which takes any nonnegative value. We discuss below the case where z_t exhibits time-dependence. The unpredictable demand and supply shocks, ϵ_t and γ_t, are also assumed to be independent and identically distributed and have mean zero.[3]

The state variable, z_t, represents the current market environment. The fact that z_t is stochastic makes this a dynamic game, rather than the traditional repeated game. Dealers observe the volume shock z_t (but not the realizations of the shocks ϵ_t and γ_t), just before quotes for period t are determined. There need not be any asymmetry of information between market participants, i.e., traders may also observe the state at the time they decide to trade.

Let v denote the market-clearing price, i.e., $s(v, z) = d(v, z)$, for all z. In what follows, we implicitly assume that the market clearing price v is independent of the shock z, but this assumption can replaced, at the only cost of additional notation, by a shock-dependent market clearing price, $v(z)$. We also assume that the expected value of the security is the market clearing price v. This is a natural assumption since in a rational competitive market the price will in fact be v, but is not required for what follows.

[3] Some care should be taken in the interpretation of equations (1) and (2) to ensure that d_t and s_t are nonnegative. Formally, suppose that the actual *nonnegative* demand shock is ϵ' and the associated quantity demanded is $d'(a, z) + \epsilon'$. Let μ denote the expectation of the shock ϵ'. Then, writing $d(a, z) \equiv d'(a, z) + \mu$ and $\epsilon \equiv \epsilon' - \mu$, we obtain equation (1). Note that it is the expected (not realized) demand and supply functions that are relevant for our analysis. Hence, the unpredictable demand and supply shocks could be multiplicative (rather than additive) without altering our conclusions.

B. Supply and Demand

Given the generality of the expected demand and supply functions, we need to impose some restrictions on their functional form:

(R1) *Symmetry:* Demand and supply functions exhibit symmetry around the market clearing price, i.e., for all α and z, $s(v - \alpha, z) = d(v + \alpha, z)$.

The assumption of symmetry is made purely to simplify the exposition. Under symmetry, every restriction placed on the demand function implies an analogous restriction on the supply function. This assumption is easily relaxed, but at the cost of additional notation.

(R2) *Monotonicity:* For every fixed volume shock z, demand decreases with price; $a' \leq a$ implies $d(a', z) \geq d(a, z)$. On the other hand, for every fixed ask a demand increases with the volume shock; $z' \leq z$ implies $d(a, z') \leq d(a, z)$.

The monotonicity assumption simply states that the demand (supply) curve slopes downward (upward), and that volume shocks shift both curves outward.

(R3) *Monotone Demand Ratio:* Consider two prices a' and a, where $a' \leq a$. The ratio of expected demands, $d(a', z)/d(a, z)$, decreases in the shock z; i.e., if $z' \leq z$, then $d(a', z')/d(a, z') \geq d(a', z)/d(a, z)$.

The monotone demand ratio is a standard assumption and includes several natural cases of interest. In particular, it holds under multiplicative separability, i.e., $d(a, z) \equiv d_1(a)d_2(z)$, as well as additive separability, i.e., $d(a, z) \equiv d_1(a) + d_2(z)$.

(R4) *Concavity:* For every value of the volume shock, z, the expected profit function, $(a - v)d(a, z)$, is a strictly concave function of the ask price.

This is also a standard assumption, and is satisfied in a variety of cases including that of linear demand and supply. Conditions (R1)–(R4) will be the maintained assumptions of this article.

To understand the economic nature of these restrictions, it is helpful to consider the following primitive model as the one underlying our formulation. Suppose that there are z investors, each of whom has an idiosyncratic reservation price, denoted by w, drawn from a density function $g(w)$. Heterogeneity in the reservation price may arise, for example, because of differences in beliefs, information, risk-aversion or liquidity factors. An investor buys (or sells) one unit of the security if the ask (or bid) price is below (or above) his or her reservation price. Define the representative investor's expected demand function by $d(a)$. Formally, $d(a) = \Pr[w > a]$, i.e., the probability that the investors' reservation price exceeds the ask. The expected demand function is just $d(a, z) = d(a)z$.

Then, (R1) simply implies that the density function of the reservation price, $g(w)$, is symmetric around the mean $v = \int wg(w)dw$. Since $\Pr[w > a]$ is nonnegative and decreasing in a, (R2) follows immediately. Restriction (R3) holds because the state variable z is just the investor population and enters the expected demand function multiplicatively. Finally, (R4) places conditions on the shape of the distribution of reservation prices (i.e., on demand elasticity) to ensure that there is an interior ask price that maximizes expected profits.

II. Security Prices in Equilibrium

A. Benchmark Equilibria

We begin by examining the two polar cases of perfect competition and overt collusion.

Perfect Price Competition. With perfect competition in every trading round, the unique ask and bid prices are $a_t = b_t = v$, and the inside bid–ask spread in period t is $\lambda_t = 0$.[4] In a single-period model, the competitive solution is the only Nash equilibrium where dealers price noncooperatively. Clearly, at these prices, no dealer will quote a lower ask price or a higher bid price. If all dealers quote an ask price strictly above v, a dealer can earn higher expected profits by undercutting the prevailing spread and thus capturing all the order flow. Indeed, even if dealers interact over many trading rounds, the strategy of "pricing at value" in all periods remains a noncooperative Nash equilibrium.

Overt or Explicit Collusion. Overt collusion is the polar opposite of perfect competition. Under overt collusion, dealers cooperate to maximize the expected present value of their joint trading profits. When dealers act as a cartel, the solution to dynamic maximization problem is equivalent to each period's individual maximization problem.

Let $\pi_a(a_t; z_t)$ denote the expected profits to dealers as a group from selling at the ask price a_t when the state is z_t and similarly define $\pi_b(b_t; z_t)$ as the expected profits from trading at the bid price b_t, so

$$\pi_a(a_t; z_t) = (a_t - v)d(a_t; z_t) \tag{3}$$

$$\pi_b(b_t; z_t) = (v - b_t)s(b_t; z_t). \tag{4}$$

The cartel's maximization problem in period t reduces to:

$$\max_{a_t \geq v \geq b_t} [\pi_a(a_t, z_t) + \pi_b(b_t, z_t)]. \tag{5}$$

In contrast to the competitive case, therefore, the collusive bid and ask prices are functions of the volume shock z_t and hence change from period to period. The following proposition characterizes the behavior of the bid and ask prices, as a function of the volume shock:

PROPOSITION 1: *Under overt or explicit collusion:*

- *For every realization of the volume shock z_t, there is a unique price pair $\{a^c(z_t), b^c(z_t)\}$ that maximizes cartel profits. The market clears on average at these prices.*
- *The ask (bid) price weakly increases (decreases) in the volume shock. In particular, if $z_t' \geq z_t$, then $a^c(z_t') \geq a^c(z_t)$ and $b^c(z_t') \leq b^c(z_t)$.*
- *The expected profits of dealers increase with the volume shock.*

[4] Of course, if dealers incurred real costs to providing liquidity, the competitive spread would be nonzero.

The proposition shows that the market clears on average at the optimal ask and bid prices. Further, the bid–ask spread is a weakly increasing function of the volume shock. Intuitively, in a dealer market, a higher volume shock implies a greater demand for market making services. Since the spread is the price dealers charge for providing liquidity, a cartel raises the spread when faced with a higher volume. It is worth emphasizing that this result obtains under quite general conditions regarding the nature of the demand and supply functions and the distribution of the stochastic volume shocks.

B. Dynamic Noncooperative Strategies

Overt collusion requires a high level of cooperation among dealers in setting their prices, and may lead to antitrust violations. In addition, individual dealers have strong incentives to undercut the prevailing spread, especially during high volume periods. Thus, overt collusion appears unlikely unless dealers could severely penalize defectors.[5]

In this section, we investigate Nash equilibria where dealers act *noncooperatively*. Let $\{a^i(z), b^i(z)\}$ denote dealer i's quotation strategy as a function of the current economic state, where we omit hereafter the time subscripts for notational simplicity. Clearly, the strategy $\{a^i(z), b^i(z)\} = \{v, v\}$ is a Nash equilibrium, since no dealer can benefit from deviating from this strategy if others also price at value. However, there may well be other, more complex, Nash equilibria for the dynamic game.

Specifically, we focus on those noncooperative equilibrium strategies that yield the highest possible profits. The resulting equilibrium is the so-called "second-best" equilibrium, since the "first-best" equilibrium (from dealers' viewpoint) corresponds to overt collusion. The second-best strategies are typically both nonlinear and history-dependent, but are of the most interest since they yield the most "natural" Nash equilibria in a dynamic context. These strategies exploit the intuition that dealers acting in self-interest may quote uncompetitive prices if they believe this will prompt others to behave similarly in future trading rounds. These intertemporal price dependencies may result in noncooperative strategies that resemble a form of implicit collusion in the sense that they yield above-normal expected profits.

But what is the optimal strategy? Ideally, dealers would set the same prices as under overt collusion. However, these prices cannot always be sustained because the economic environment changes from period to period. To see this, observe that the spread under overt collusion is an increasing function of the volume shock, z_t. When trading volume is abnormally large, the expected profits from undercutting the collusive bid and ask prices may exceed the loss of future profits from the collapse of cooperative pricing. Consequently, the collusive spread is not sustainable for large volumes. However, if the spread narrows with volume, this may deter individual dealers from unilaterally

[5] Such penalties might include a refusal to trade with a dealer who defects, thereby imposing a large cost disadvantage because the defector would be unable to easily offset large inventory imbalances.

raising their bid price or lowering their ask price, allowing dealers to earn excess profits. The question then is whether such a strategy can be supported in equilibrium, and if so, under what conditions. We now turn to a formal analysis of this problem.

C. The Optimal Pricing Strategy

Denote the total expected present value of future profits to all dealers under the optimal (second-best) strategy by $J(z)$. Suppose the current realization of the volume shock is z'. Using standard dynamic programming arguments, we can then express the value function $J(z')$ as

$$J(z') = \max_{\{a',b'\}}\{\pi_a(a'; z') + \pi_b(b'; z') + \rho E[J(z)]\} \tag{6}$$

subject to

$$(1/M)\{\pi_a(a'; z') + \pi_b(b'; z') + \rho E[J(z)]\} \geq \pi_a(a; z') + \pi_b(b; z') \tag{7}$$

for all $a \leq a'$ and $b \geq b'$, where ρ denotes the one-period discount factor.

To understand the statement of the maximization problem, first consider the incentive compatibility constraint (7). A dealer can set prices equal to the "suggested" or "proposed" prices (a', b'). If the dealer conforms to the proposed prices, and expects the tacit pricing arrangement to continue into the future, he or she receives an immediate expected profit of $[\pi_a(a'; z') + \pi_b(b'; z')]/M$ and the expected discounted continuation payoff of $\rho E[J(z)]/M$, where the expectation is taken over all possible realizations of z. Alternatively, if the dealer were to undercut the proposed prices, he or she would choose a price pair (a, b) to maximize $\pi_a(a; z') + \pi_b(b; z')$, expecting other dealers thereafter to price at value, the worst possible threat. Then, equation (7) simply states that the expected profits from continuing to price under the normal strategy (i.e., from implicit collusion) dominate the one-time profits from undercutting the other dealers and causing the tacit pricing agreement to collapse.

The optimization problem (6) is to maximize the aggregate intertemporal profits of all dealers subject to the incentive constraint (7). Intertemporal profits consist of: (i) Total expected current profits, and (ii) The total expected present value of future continuation payoffs to all dealers, i.e., $\rho E[J(z)]$. The optimization problem in equation (6) is simply the definition of the second-best equilibrium, i.e., to select bid and ask profits that maximize intertemporal profits among those prices that will not be undercut by individual dealers.

The nature of the optimal strategy can be understood by considering the underlying economics of the decision problem. It is clear that π_a is equal to zero if $a = v$ and achieves its maximum at $a^c(z)$, where $a^c(z) > v$. Given a volume shock z', suppose that the proposed prices are the prices under overt collusion, i.e., $(a^c(z'), b^c(z'))$. Given assumption (R4), the most profitable way to undercut these prices is to post an ask (bid) price marginally below $a^c(z')$ (above $b^c(z')$). In other words, $(a^c(z'), b^c(z'))$ is *sustainable* (in the sense that no dealer is willing to undercut this spread and set prices equal to v thereafter) if and only if

$$\frac{K}{M} \geq \left(1 - \frac{1}{M}\right)[\pi_a(a^c(z'); z') + \pi_b(b^c(z'); z')], \tag{8}$$

where $K \equiv \rho E[J(z)]$ denotes the discounted expected continuation value to dealers as a group. Thus, K/M represents the expected payoff to an individual dealer given that no market maker chooses to deviate from the "proposed" ask and bid prices. Intuitively, K/M is the deterrence that prevents individual market makers from raising the bid or lowering the ask price, thus capturing all the order flow.

Since the collusive expected profit is an increasing function of the volume shock z, it follows that for sufficiently large z, the incentive constraint will be violated if prices are set at the collusive levels of $(a^c(z), b^c(z))$. Thus, spreads under the second-best strategy must be narrower to prevent defection. For z below this level, however, collusive prices may be sustainable. Intuition suggests that the extent to which collusive prices are sustainable when dealers act noncooperatively will be a function of the discount rate. If dealers are patient, collusive pricing may be sustainable in a wide range of economic environments, because dealers are unwilling to sacrifice future profits by undercutting the current spread. In contrast, if dealers are sufficiently impatient, they will act more competitively, and collusive pricing may only be sustainable over a very small range, if at all.

The following proposition characterizes the second-best equilibrium.

PROPOSITION 2: *There exists a constant $\rho_0 < 1$ such that if the discount factor $\rho > \rho_0$ (i.e., dealers are sufficiently patient), the optimal (second-best) strategy takes the following form:*

- *If the volume shock is below a unique critical level, i.e., if $z_t < z_c(\rho)$, where $0 < z_c(\rho) \leq \infty$ is a constant, the bid and ask prices for dealer i are the collusive prices, i.e., $\{a_t^i(z_t), b_t^i(z_t)\} = \{a^c(z_t), b^c(z_t)\}$. If the volume shock is above the critical level (i.e., $z_t \geq z_c(\rho)$), the bid and ask prices for dealer i are $\{a_t^i(z_t), b_t^i(z_t)\} = \{a^*(z_t), b^*(z_t)\}$, where either $a^*(z_t) < a^c(z_t)$ or $b^*(z_t) > b^c(z_t)$, or both. In the event of a deviation from this strategy, dealers set prices $a_t = b_t = v$.*
- *The inside spread in period t, $\lambda_t^*(z_t)$, is an increasing function of the volume shock z_t for $z_t < z_c(\rho)$ and a decreasing function of the volume shock thereafter.*

When $\rho \leq \rho_0$, however, the unique equilibrium is always to price at value, i.e., $a_t = b_t = v$, for all t, regardless of the previous trading history.

Proposition 2 provides a complete characterization of the optimal (second-best) Nash equilibrium. The proposition shows that if the discount factor is sufficiently low (i.e., dealers are impatient), the only equilibrium is perfect competition. However, if dealers are sufficiently patient, dealers earn strictly positive expected profits. The equilibrium characterized in Proposition 2 provides the highest expected present value of profits when dealers act noncooperatively. However, dealers do not collude in the strict sense of the word by fixing prices. Rather, they act in their own self-interest while recognizing their dynamic interdependence. In this sense, this noncooperative pricing behavior can be described as "tacit" or "implicit" collusion. Other dealers need not know the identity of the "defector," i.e., the dealer who undercuts the inside spread.

Implicit collusion differs from overt collusion or price fixing. Under overt collusion, spreads are constant or increase with volume as demonstrated by Proposition 1. In contrast, spreads under implicit collusion can exhibit a different pattern. When expected volume is less than the critical threshold level, $z_c(\rho)$, dealers' noncooperative behavior results in prices that are equal to those set under overt collusion. In this case, spreads weakly rise with volume in this range. However, if the volume shock is above the critical level, the prices corresponding to overt collusion cannot be sustained. Consequently, the bid–ask spread must *narrow* as volume rises to prevent individual dealers from defecting from the proposed equilibrium prices.[6] Even so, the bid–ask spread is still positive, although it is smaller than the collusive spread.

Interestingly, there are conditions under which implicit collusion yields the same prices as explicit collusion in all economic environments, i.e., the bound z_c can be infinite even if $\rho < 1$. Further, as shown below, certain institutional features of securities markets may make it easier to sustain purely collusive prices without any explicit pricing agreements. The case where implicit and explicit collusion coincide poses somewhat of a dilemma for policy makers. On one hand, it may be easier for regulators to distinguish between collusive and competitive pricing using data, as discussed below. On the other hand, without evidence of price fixing (in the form of, say, interdealer communications regarding prices), regulators may never be able to prove that dealer behavior is collusive (and hence a violation of antitrust statutes) even if they correctly conclude that spreads exceed competitive levels.[7]

Finally, it is worth considering the robustness of the model to changes in the underlying assumptions. In particular, although collusion holds under relatively general specifications of demand and supply, the assumption that the state variable z_t is independent and identically distributed warrants further consideration. If z_t is interpreted as the investor population (as above) or a volume variable, it seems reasonable to consider alternative stochastic processes that exhibit persistence, i.e., where the expected value of z_t is a function of past values of the state variable. Specifically, consider a stationary Markovian process for z_t, and assume that $E(z_{t+k}|z_t)$ is nondecreasing and concave. This process covers many important cases, including the one where z_t follows an AR(1) process. In this case, it can be shown that collusive prices are more difficult to sustain for higher rather than for lower values of z_t, as in the *iid*

[6] Rotemberg and Saloner (1986) model oligopolistic pricing over the business cycle and derive a similar result.

[7] We thank the referee for making this point to us.

case considered above. Thus, the same qualitative conclusions hold in the more general case where the state variable follows a persistent process.

III. Market Conditions and Dealer Behavior

In this section we analyze the effect of changes in market conditions on dealer pricing. Specifically, we examine the impact of (i) differences in market size and trading frequency, (ii) heterogeneity among dealers arising from reputation or size, and (iii) free entry into market making.

A. Market Size and Trading Frequency

It is of interest to ask how dealer pricing behavior varies across stocks as a function of the size of the market and the frequency of trading. The interesting case of continuous trading occurs in the limit as frequency increases while scale simultaneously decreases to keep total expected volume constant.

To analyze the effect of changes in the size of the market, let k index the scale of the market, and consider expected demand and supply functions of the form $d^k(a; z) = kd(a; z)$ and $s^k(b; z) = ks(b; z)$. In other words, in a market of size k, the volume of buy and sell orders for any given price pair (a, b) and state z is simply k-times that of the base case. Alternatively, we can ask how dealer behavior varies with the frequency of trading. Let n denote the number of trading rounds in a calendar period or "day" of unit length, where $d(a, z)$ and $s(b, z)$ are the expected demand and supply functions in each period. Then, since the time between trades is $1/n$, more frequent trading is associated one-to-one with higher n.[8]

The following proposition characterizes the nature of dealer pricing behavior across markets that differ in scale and the frequency of trading.

PROPOSITION 3: *Under the optimal (second-best) strategy:*

- *The optimal bid and ask prices are unaffected by the scale of the market, i.e., the price pair $(a^*(z), b^*(z))$ and the bound z_c are independent of k.*
- *As trading frequency, n, increases, the range under which collusive pricing can be supported, i.e., z_c, increases. In the limit, as $n \to \infty$, collusive prices can be sustained in all environments:*

$$\lim_{\rho_n \to 1} z_c(\rho_n) = \infty.$$

Proposition 3 shows that the optimal bid and ask prices, and the range under which collusive prices are maintained (i.e., z_c) are unaffected by scale. From an economic viewpoint, this result shows that it is not market size that is the primary determinant of dealers' ability to collude but rather the underlying demand and supply elasticities.

[8] When there are n trading rounds within every calendar period, we will normalize the discount factor appropriately; the discount factor ρ_n will satisfy $\rho_n \equiv \rho^{1/n}$.

Proposition 3 also shows that as trading frequency rises, the range under which collusive pricing is sustainable increases. Although the result appears paradoxical, the economic intuition underlying this result is straightforward. The expected future profit from maintaining collusive pricing relative to the payoff from deviation is larger in an active stock than an inactive stock. Since the optimal prices are independent of the market's scale k, it follows immediately that in the limit with continuous trading (i.e., as $n \to \infty$ and $k \to 0$ keeping nk constant), collusive prices are sustainable for all z. Thus, the claim by Christie and Schultz (1994) that there is collusion even in actively traded Nasdaq stocks such as Apple Computer cannot be simply dismissed.

B. Heterogeneous Dealers

So far, we have assumed that dealers share the order flow equally when quoting the inside spread. In practice, however, the order flow may not be shared equally among dealers. This may reflect differences in dealers' size, reputation, or relationships with brokers. It is also possible that some dealers may purchase order flow, a question we discuss below. These differences may affect the ability of market makers to implicitly collude. In this section, we investigate the impact of heterogeneity among market makers on bid–ask spreads, trading volume, and equilibrium profits. In particular, we are interested in the question of whether collusion is more likely to be sustained in a market dominated by a few large dealers than in a market where dealers are of roughly equal size.

Let $\phi_i \geq 0$ denote the fraction of order flow handled by market maker i when all dealers quote identical prices, where $\Sigma_{i=1}^{M} \phi_i = 1$. In the analysis above, we assumed that $\phi_i = 1/M$ for all i, but now we allow for differences among dealers. If dealers do not quote identical prices, assume that those dealers posting the inside spread share the aggregate order flow according to the proportions given by $(\phi_1, \ldots, \phi_M)$. For example, if dealers 1 to N $(N < M)$ quote the lowest ask prices, dealer i $(i = 1, \ldots, N)$ receives a fraction $\phi_i/\Sigma_{j=1}^{N} \phi_j$.

Let $\Phi = (\phi_1, \ldots, \phi_M)$ denote the total allocation among market makers, and let $J_i(z; \Phi)$ denote the lifetime profits of dealer i in the highest sustainable noncooperative (second-best) equilibrium. We restrict attention to equilibria in which all dealers quote the same bid and ask prices and no dealer has an incentive to deviate from the common price. Since each dealer receives a constant fraction of the order flow and hence the total profits, it follows immediately that $J_i(z; \Phi) = \phi_i E[\Sigma_{k=1}^{M} J_k(z; \Phi)]$. Let $K(\Phi)$ denote the aggregate expected present value of future trading profits to all market makers, where

$$K(\Phi) = \rho E\left[\sum_{i=1}^{M} J_i(z;\, \Phi)\right], \tag{9}$$

where the expectation is taken over future realizations of z. Thus, the expected lifetime future profits for dealer i is just $\phi_i K(\Phi)$.

We are now in a position to analyze the equilibrium when dealers face an exogenously determined allocation of order flow. The following proposition characterizes the effect of heterogeneity among market makers.

PROPOSITION 4: *If the dealer with the smallest share of order flow chooses not to improve the prevailing prices in the market, then no other dealer will offer a lower ask price or a higher bid price.*

Proposition 4 demonstrates that the "smallest" dealer (in the sense that this dealer receives the lowest share of order flow when all dealers quote identical prices) is the one most likely to undercut the prevailing spread. Intuitively, the smallest dealer has the most to gain from narrowing the prevailing spread and capturing the order flow. This dealer also has the least to lose, in terms of future profits, from a breakdown of tacit collusion. Proposition 4 thus implies that it is the smallest dealer who determines the equilibrium bid–ask spread.

Without loss of generality, index market makers so that dealer M is the smallest dealer, i.e., $\phi_M = \min[\phi_1, \ldots, \phi_M]$. Then, we have:

PROPOSITION 5: *Consider two different allocation vectors Φ and Φ':*

- *If the minimum shares are identical under two allocations Φ and Φ' (i.e., $\phi_M = \phi'_M$) then the aggregate expected present value of profits are also identical, i.e., $K(\Phi) = K(\Phi')$. Further, in the unique second-best strategy, bid and ask prices are identical, i.e., $a^*(z; \Phi) = a^*(z; \Phi')$ and $b^*(z; \Phi) = b^*(z; \Phi')$* for all $z \geq 0$.
- *If $\phi_M > \phi'_M$ (i.e., the smallest dealer under the allocation Φ receives a larger fraction of the order flow than the smallest dealer under the allocation Φ'), then the aggregate expected lifetime profits under allocation Φ are larger than under Φ', i.e., $K(\Phi) > K(\Phi')$. Further, the ask (bid) price under Φ is larger (smaller) than under Φ', i.e., $a^*(z; \Phi) \geq a^*(z; \Phi')$ and $b^*(z; \Phi) \leq b^*(z; \Phi')$ for all $z \geq 0$.*

Proposition 5 implies that the share of order flow of the smallest dealer determines profits and the spread. The higher the share of the smallest dealer, the higher the spread and the profits of all dealers.[9] This is especially important given that low measures of dealer concentration (in the form of a low Herfindahl-Hirschmann Index for dealers' share of volume) has been cited (see, e.g., Grossman *et al.*, 1995) as evidence that implicit collusion is unlikely to have occurred in the stocks studied by Christie and Schultz (1994). Our results, however, suggest that implicit collusion is easier to sustain in a market where dealers are of roughly equal size and the Herfindahl-Hirschmann concentration index is low.

[9] Similarly, if dealers differed in their discount rates (perhaps because of differences in risk aversion or in the cost of capital), the highest cost dealer will determine the inside spread.

C. Endogenous Entry

The conclusions of Christie and Schultz (1994) have been criticized on the grounds that entry into the Nasdaq market is relatively easy and that collusion appears unlikely in active stocks with large numbers of dealers.[10] If there are no barriers to entry, it is clear that dealers will enter as long as expected profits are positive. From equations (6) and (7), it is immediately clear that as M grows large, z_c tends to zero, and the bid and ask prices converge to v. Under free entry, no form of collusion is possible.

The explicit costs of entering the Nasdaq market, as noted by Kleidon and Willig (1995), are relatively low. However, the implicit costs associated with securing access to the order flow may be more substantial. In particular, dealers bear entry costs in the form of establishing reputation and in forming long-term relationships with brokers, who act as the customer's agent by directing order flow to dealers. Indeed, as described by NASD (1991), dealers in Nasdaq markets provide a variety of cash and noncash inducements (e.g., research services, reciprocal trades (swaps), and assistance in clearing and settlement) to brokers to preferentially direct order flow to them. These implicit costs, although typically ignored in discussions of industry structure, may represent a significant barrier to entry. To investigate this further, we assume that dealers must incur a one-time cost $c > 0$ to enter the market. The cost c captures not only the explicit costs associated with market making, such as capital costs, but also the costs of establishing a reputation and securing access to the order flow.

Before the start of trading, the expected present value of future trading profits is $K(\Phi) \equiv \rho E[J(z)]$. From the analysis in the previous section, the second-best profit level is a function of the minimum share alone, so that when dealers share the order flow equally, we can write $K(\Phi) = K(1/M)$. From the previous section, it is clear that $K(1/M)$ decreases in M. Thus, ignoring integer constraints, the equilibrium number of dealers solves the following equation:

$$\frac{1}{M} K\left(\frac{1}{M}\right) = c. \tag{10}$$

It is clear from this representation M is finite when $c > 0$, so that tacit collusion is sustainable with endogenous entry. Indeed, tacit collusion can be viewed as a mechanism that allows dealers to overcome the problem introduced by the fixed costs of entry. In the absence of tacit collusion, dealers may rationally conjecture that they could not recover their entry costs, leading to either a monopoly equilibrium or a no trade equilibrium. In this situation, tacit collu-

[10] Kleidon and Willig (1995) find that there are, on average, 22 dealers in each of the stocks examined by Christie and Schultz (1994), well above the average of 10 dealers per stock for all Nasdaq stocks. Wahal (1995) documents decreases (increases) in spreads in Nasdaq markets following the entry (exit) of market makers.

sion may increase the welfare of outsiders by permitting trades or risk sharing that would not otherwise occur.[11]

In the symmetric case discussed above, all dealers earn zero expected profits net of entry costs. We show below that preferencing arrangements that direct order flow to certain dealers allow some dealers to make strictly positive profits. A similar result would obtain if dealers differed in their costs of entry. In that case, the marginal entrant would make zero expected lifetime profits (net of the costs of entry), but established dealers with lower entry costs would make strictly positive net profits.

IV. Institutional Features and Price Competition

In this section we generalize the model to incorporate some key institutional features of the Nasdaq stock market: order preferencing arrangements, price matching agreements (i.e., Nasdaq's Preference Trade Rule), and price discreteness.

A. Order Preferencing Arrangements

In the Nasdaq market, a portion of the order flow is directed (or preferenced) by brokers (acting in their capacity as agents of the customer and as members of the NASD) to preferred dealers. Preferencing is similar to internalization, where a broker-dealer acts as a dealer for its retail customers. Internalization and preferencing arrangements are important because they limit the ability of other dealers to attract order flow through price competition.[12] We begin by analyzing the general form of such arrangements, and in the next section examine the specific form of these arrangements in the Nasdaq National Market.

Formally, suppose that a fraction θ (where $0 \leq \theta \leq 1$) of the order flow is directed by brokers to a group of N dealers, where $N \leq M$. Each dealer gets a fraction $\theta_i \geq 0$, where $\Sigma_{i=1}^{N} \theta_i = \theta$. The remaining portion of the order flow, $(1 - \theta)$, is shared equally among all M dealers. Thus, a dealer i ($i = 1, \ldots, N$) who receives preferenced orders has a share of $\phi_i = \theta_i + (1 - \theta)/M$ of the total order flow. The share of a dealer j ($j = N + 1, \ldots, M$) who does not receive any directed order flow is simply $\phi_j = (1 - \theta)/M$. Observe that equal routing is a special case of preferencing where $\theta = 0$.

We are interested in how preferencing affects the ability of market makers to maintain bid–ask spreads above competitive levels. Proposition 6 shows that order preferencing can restrict entry and hence increase bid–ask spreads and trading profits.

PROPOSITION 6: *The equilibrium number of market makers is smaller and bid–ask spreads are larger when some dealers receive preferenced order flow than*

[11] We thank the anonymous referee for suggesting this point to us.

[12] See also Christie and Schultz (1994), Godek (1995), Kandel and Marx (1995), and Huang and Stoll (1996).

under equal routing. Further, the expected profits, net of entry costs, of those dealers who receive preferenced order flow is strictly positive.

The proposition shows that institutional arrangements to direct order flow to preferred dealers can restrict entry into market making. Even if all dealers face identical fixed costs, the dealers to whom order flow is routed make profits net of these entry costs. This is not the case if dealers share the order flow equally. Further, preferencing arrangements may deter new dealers from entering the market if their residual share is so low as to not cover their entry costs.

As an aside, larger broker-dealers may choose to direct orders to smaller dealers, provided this does not trigger additional entry. To see this, suppose we are in an equilibrium with M dealers, where the marginal entrant (i.e., dealer $M + 1$) cannot recover her cost of entry given the residual order flow. Recall from Proposition 5 that (before the start of trading) the total expected lifetime profits of dealers 1 through $M - 1$ is given by $(1 - \phi_M)K(\Phi)$. The total expected lifetime profits of all dealers, i.e., $K(\Phi)$, depends only on the share of the smallest dealer, denoted by ϕ_M. Thus, we can write $K(\Phi)$ as an increasing function of ϕ_M, denoted by $K(\phi_M)$. Let $\phi^* = \text{argmax}(1 - \phi_M)K(\phi_M)$ subject to $\phi_M \leq 1/M$. Then, it follows immediately that if $\phi_M < \phi^*$, the larger dealers would be better off if an incremental fraction $\phi^* - \phi_M$ of the order flow were preferenced to dealer M. Intuitively, directing order flow to the smallest dealer may make sense for larger dealers because this decreases the likelihood the dealer will undercut their prices, raising their own profits. However, if reallocation induces further entry, this may result in spreads falling to competitive levels. This would not occur if dealers have heterogenous costs of entry, or if the nonpreferenced portion of the order flow were so small as to deter entry for the marginal dealer.

B. Price Matching Arrangements

In some cases, the preferencing arrangements described above take the form of price matching (or best price) agreements. For example, under Nasdaq's Preference Trade Rule, a dealer who does not post the inside spread can nevertheless receive preferenced orders provided he or she matches the best intermarket prices. By contrast, under the preferencing arrangements considered above, dealers must quote the inside quotes before any orders can be directed to them.

Although a price matching agreement closely resembles the general preferencing rule considered above, our analysis shows that the seemingly small difference in terms of quoted prices can have a large impact on dealer pricing. Indeed, under a price matching or best price preferencing arrangement, collusive prices can be maintained even in a one-shot market. To see this, observe that each dealer earns $(1/M)[\pi(a^c(z); z) + \pi(b^c(z); z)]$ under collusive pricing. If dealer i undercuts to, say, prices of (a', b') his profits are only $(1/M)[\pi(a'; z) + \pi(b'; z)]$ since all other dealers can match his prices. Evidently it does not pay to undercut the collusive prices. As a result, the collusive prices are sustainable in a one-shot market. It follows then that collusive pricing can also be maintained in a dynamic model.

The analysis above shows that preferencing arrangements reduce dealers' incentives to compete on the basis of price. In turn, this suggests the possibility that competition among dealers takes other forms. Specifically, since customer order flow in Nasdaq stocks is directed to dealers by the broker, dealers may have incentives to pay brokers for order flow or to enter into vertical agreements with brokers to secure access to customer orders. To see this, observe that every share directed to a dealer yields an expected profit of $E[\lambda(z)/2]$, i.e., half the inside spread. Thus, dealers can increase their expected profits without competing on price. If order flow payments are unobserved, competition will reduce or possibly even eliminate dealers' expected profits from tacit collusion.

If dealers do collude to elevate bid–ask spreads above competitive levels, the excess spread must be at least equal to the minimum tick of one-eighth. Competition among dealers to secure access to the order flow would then imply order flow payments of half the excess spread or roughly 6.25 cents. Yet, Kleidon and Willig (1995) note that order flow payments are typically only 1 to 2 cents per share, arguing against the collusive hypothesis. However, it is important to keep in mind that dealers use a variety of noncash inducements to attract order flow. Specifically, as described by NASD (1991), dealers may provide research services to brokers, arrange reciprocal trades (swaps), provide clearing and settlement, as well as other services. Such noncash inducements to brokers reduce the per share price of order flow so that one cannot make simple inferences about the competitiveness of prices from the magnitude of order flow payments. Such actions are also consistent with the hypothesis that dealers engage in nonprice competition to secure access to the order flow rather than directly undercutting one another's quotes.

C. Price Discreteness

Security prices in financial markets are typically quoted in discrete units determined by the minimum price variation or tick. Discreteness is interesting because it has been suggested that reducing the size of the minimum price variation or even decimalization would increase competition among dealers.[13] Further, the original study by Christie and Schultz (1994) emphasizes the issue of price discreteness by focusing on the avoidance of odd-eighths quotes by Nasdaq dealers. In this section, we examine the effect of price discreteness on the dynamic pricing behavior of dealers.

Price discreteness affects dealer pricing since it effectively amounts to a restriction on the set of actions available to market makers. On an a priori basis, it is difficult to assess whether this limitation facilitates collusion or not. For example, suppose initially that the tick is one-eighth and that under the

[13] Harris (1994) discusses of the effects of changes in the minimum price variation for NYSE stocks.

(restricted) second-best strategy the bid–ask spread is \$0.25. If the tick is lowered to one-sixteenth (\$0.0625), it appears possible that dealers may be able to support a spread of \$0.3125 (although they cannot support \$0.375, the next highest spread under an eighth tick), widening the equilibrium bid–ask spread. Conversely, the ability to marginally undercut the quoted spread (i.e., to \$0.1875) may lead to narrower spreads. For this reason, a more formal analysis of discreteness is warranted.

Formally, let $P(\Delta) = \{0, \Delta, 2\Delta, \ldots\}$ denote the set of potential prices when the minimum variation is $\Delta > 0$. Suppose for simplicity that the fundamental value v lies on the price grid, i.e., $v \in P(\Delta)$.[14] Now consider the impact of coarsening the price grid by increasing the minimum variation to $\Delta' > \Delta$. The following proposition provides conditions under which a coarsening of the price grid increases the set of sustainable prices and hence dealer's expected profits.

PROPOSITION 7: *If a price strategy lies on both $P(\Delta)$ and $P(\Delta')$, and is sustainable under $P(\Delta)$, then it is also sustainable under $P(\Delta')$, where $\Delta' > \Delta$. In particular, if the second-best pricing strategy when the minimum variation is zero is feasible with price discreteness, then it remains an equilibrium price strategy. Further, dealers' highest equilibrium profit under price discreteness is larger than in the second-best equilibrium when prices are continuous.*

Proposition 7 shows that a coarsening of the price grid by a widening of the minimum tick will increase the size of sustainable bid–ask spreads and hence will increase dealers' expected profits. The proposition also holds for more general grid structures. For example, consider two sets of possible prices P and P', and say that P' is coarser than P if $P' \subset P$. Then, Proposition 7 holds for P and P'.

The intuition underlying Proposition 7 is straightforward. When the minimum price variation is increased, the amount by which a dealer must undercut rivals' prices to capture the order flow is increased. Hence, the size of the immediate profits that a dealer obtains by undercutting the prevailing prices is smaller than before. Conversely, the expected future profits that a dealer anticipates if he or she were not to undercut the prevailing spread are higher under a larger minimum variation. As a result, any spreads that were sustainable before the grid was coarsened continue to be sustainable, making implicit collusion easier. This widening of spreads is quite distinct from dealers "rounding up" when selling and "rounding down" when buying. Such actions will also tend to widen spreads on average when the tick size increases.

From a policy viewpoint, Proposition 7 suggests that bid–ask spreads may be narrowed by a reduction in the minimum price variation. However, implicit collusion can occur even with perfectly divisible prices. In particular, Proposition 2 establishes that spreads are strictly above competitive levels even though the set of available price choices are not constrained in any way. Hence,

[14] This assumption simplifies the proofs of the results that follow but is not necessary for them to go through.

reductions in the tick size are not likely to completely eliminate excess spreads if dealers are colluding.

Our results also suggest that regulators exercise caution in interpreting the absence of odd-eighth quotes as evidence for collusive behavior. Proposition 7 shows that if dealers can agree to restrict the set of available quotes to a coarser grid (avoiding odd-eighths, say), they would increase their expected profits. However, it is difficult to see how such an agreement can be maintained, or, for that matter, why dealers choose to avoid odd-eighths instead of even-eighths. After all, if dealers can cooperate to avoid odd-eighths, then they also should be able to explicitly fix spreads without restricting their available actions. It could be argued that dealers avoid odd-eighths because this facilitates the monitoring of defections. However, since information on prices in NASD markets is readily available, this hypothesis appears unlikely.

Alternatively, the avoidance of odd-eighths may reflect the clustering that arises as a way to reduce negotiation costs, as suggested by Harris (1991). Clustering on even-eighths, combined with spreads that typically exceed the minimum tick of one-eighth, may explain the empirical results of Christie and Schultz (1994). Indeed, Huang and Stoll (1996) find no association between the frequency of even eighths and the size of spreads after correcting for the fundamental factors determining these two variables. Thus, the avoidance of odd-eighths per se does not imply collusion, but is consistent with this hypothesis since it implies spreads of at least a quarter in even the most active stocks.

V. Discussion

A. Policy Issues

Our results have several implications for policy makers and regulators. First, our results show that despite the differences between financial and product markets, competing dealers may engage in pricing behavior that result in excess spreads. It is important to note that the pricing strategies described here arise from noncooperative behavior, and are quite distinct from overt collusion where dealers cooperate to fix prices. This distinction is important because the two forms of behavior have different antitrust implications and require different burdens of proof. Regulators investigating dealer pricing behavior should recognize that excess spreads can arise without any explicit (verbal or written) price fixing agreements among dealers.

Second, our analysis shows that implicit collusion can be sustained under relatively weak conditions. Indeed, Proposition 2 was derived under quite general assumptions regarding the stochastic environment and the nature of the underlying demand and supply functions. However, it is important to keep in mind that competitive pricing is always an alternative Nash equilibrium, and in some cases may be the only equilibrium.

Third, contrary to popular belief, collusion appears feasible in frequently traded securities as shown by Proposition 3 and in markets with roughly evenly matched dealers, as demonstrated by Proposition 5. Further, our characterization of implicit collusion does not depend on other dealers knowing the identity of the defector. This is relevant since it is argued that dealers can trade anonymously using nonpublic quotation dissemination systems such as Instinet or SelectNet at prices that differ from their own posted Nasdaq quotes, thus making collusion impractical.

Fourth, institutional arrangements such as dealer preferencing and price matching agreements are crucial to sustaining "excess" spreads. These practices restrict access to the order flow and are inconsistent with time priority. In particular, price matching preferencing arrangements eliminate the incentive to compete on the basis of price, increasing spreads. In the absence of price competition, dealers may compete in other dimensions, perhaps paying to secure access to the order flow. However, the absence of odd-eighth quotes documented by Christie and Schultz (1994) need not imply collusion; this may simply arise because of wide spreads and clustering. Similarly, there may be other factors (see, e.g., Bhasin and Brown (1995)) that cause spreads to be wider than competitive levels, but are not the result of collusion among dealers.

Finally, our results suggest a variety of measures that policy makers and regulators can use to facilitate greater competition among dealers. Eliminating structural impediments such as dealer preferencing and price matching arrangements will increase dealers' incentives to compete for order flow on the basis of price. Similarly, discouraging internalization and the use of cash and noncash inducements for order flow would increase the number of orders exposed to the entire market and increase dealers' incentives to improve the prevailing bid and ask prices. Competition also could be enhanced by facilitating the use of public limit orders, allowing investors to directly compete with market makers to narrow the inside spread.[15] Such changes raise difficult issues concerning price and time priority, but may ultimately reduce the costs of trading for retail investors by fostering greater competition.

It is worth questioning the robustness of our conclusions for public policy. Clearly, our model does not incorporate some factors that reduce the likelihood of collusive behavior. These include penalties imposed by regulators if collusion were detected, competition from exchanges for new listings, competition from limit orders, and bilateral bargaining between dealers and large institutional traders. These factors would limit the ability of dealers to widen the spread but would not eliminate the motive for implicit collusion, especially for smaller, retail orders. On the other hand, our analysis omits other considerations that may make collusion easier. For example, if dealers periodically trade with one another to offset their excess inventories to reduce risk, the threat of refusing to trade in the future with any dealer who undercuts spreads will facilitate collusion. Similarly, the threat of retaliation is larger if dealers make markets in multiple stocks, again making collusion easier to sustain.

[15] Public limit orders were not given priority in the Nasdaq market in the period studied by Christie and Schultz (1994), but Nasdaq has recently proposed changing this rule.

Overall, we view these considerations as second-order effects, and incorporating them formally into the analysis would not alter our general conclusions.

B. Empirical Issues

We turn now to a discussion of the empirical implications of our analysis. Under implicit collusion, spreads coincide with those set by a cartel for all volumes below a certain critical bound. In this case, the spread is either constant or is higher in periods where volume is higher. In periods where volume exceeds the critical bound (which can theoretically be very large) spreads under implicit collusion narrow with volume to prevent individual dealers from undercutting the prevailing quotes. This nonmonotonic time-series price-volume relation is distinct from the cross-sectional inverse relation between bid–ask spreads and volume documented in the previous empirical literature. It is also worth noting that in the special case where implicit and explicit collusion coincide, spreads are either constant or increasing with volume in contrast to the competitive case.

However, there are several difficulties in testing this prediction using publicly available databases. First, the measurement of volume is not as straightforward as it may appear. It makes sense to test the model by aggregating transaction volume over finite time intervals. Measuring volume at the transaction level is problematic because individual transactions are often bilateral in nature, and as a result, the relations we seek to identify may be obscured. In particular, some individual trades should be excluded from the analysis. These include interdealer trades to offset inventory imbalances and large-block trades negotiated by institutional traders that need not transact at the best bid or offer prices.[16] However, the appropriate time interval over which volume is aggregated is an open question. For active issues, relatively short time horizons of, say, an hour or less may be appropriate, while for inactive issues the interval may be as long as a day. This is an empirical issue.

It may also be difficult to distinguish the effects of collusive behavior from other factors affecting spreads such as asymmetric information and dealer inventory.[17] Further, some institutional trades in Nasdaq stocks do not involve commissions payments that are implicitly factored into the transaction price, so that the quoted spread overstates the true cost of trading.[18]

Of the issues discussed above, the most important problem in testing the model's predictions is to control for other factors that may confound the relation between spreads and volume. This problem may be partly overcome by

[16] It is straightforward to extend the theoretical framework to incorporate institutional traders with bargaining power; the model would then be interpreted as applying to the "retail" order flow from investors who do not directly negotiate with dealers.

[17] See Stoll (1976), Reinganum (1990), Blume and Goldstein (1992), Neal (1992), Shapiro (1993), Christie and Huang (1994), and Huang and Stoll (1996) for a discussion of the determinants of spreads in Nasdaq stocks.

[18] Keim and Madhavan (1995) use data from institutional equity traders and find evidence of differences in total execution costs, including commission costs, by market after controlling for other factors.

forming a control sample of stocks with similar characteristics (i.e., market capitalization, trading frequency, volume, risk, etc.) to the sample being investigated. Again, considerable care must be taken in forming the control sample. Specifically, the use of stocks from a different exchange as a control may be misleading because differences in spreads may reflect institutional factors as well as differences in behavior.[19] A better approach may be to use Nasdaq stocks where no allegations of collusion have been raised as a control, taking suitable care to obtain a representative sample. This approach, however, necessitates outside information (such as allegations of collusion reported to regulatory authorities) to partition stocks into the control group. Alternatively, it is possible to use stocks that trade on odd-eighths as a control, using an econometric model (see, e.g., Laux, 1995) to handle possible endogeneity issues.

Given a suitable control sample, the time-series relation between the inside spread and volume *relative to that of the control sample* can be estimated using nonlinear regression analysis with market-specific interaction terms on the volume coefficients. By controlling for confounding effects, this would isolate the effect of dealer behavior on spreads. A stock-specific trade size filter may eliminate many institutional or inter-dealer trades present in publicly available data. It would also be interesting to see if the sharp decreases in bid–ask spreads discussed by Christie, Harris, and Schultz (1994) occurred in periods with abnormally large volume, as suggested by the model.

It is much easier to test the finer predictions of the model with data that identify the trade's participants because we could then measure dealer profits on a transaction level. This would enable us to directly test the hypothesized relation between dealer profits and volume, as well as the model's specific predictions regarding the level of "excess" spreads, dealer concentration and trading frequency. Finally, information on order preferencing and the number of dealers would allow us to assess (through a cross-sectional regression) the impact of these institutional arrangements on entry and hence on dealer profits. Given the limitations of publicly available data, the best hope for testing the model lies in data that identify the participants to a transaction.[20]

VI. Conclusions

This article develops a game-theoretic model of dynamic dealer pricing. Our primary objective is to examine whether implicit collusion is possible in a dealer market, as suggested by Christie and Schultz (1994), and if so, to characterize this behavior. In addition, we analyze the impact of institutional

[19] For example, Huang and Stoll (1996) form a control sample of exchange-listed stocks and find that Nasdaq stocks have wider spreads. However, they attribute these differences to institutional features of the Nasdaq market rather than to collusive behavior among dealers.

[20] Experimental asset markets offer an interesting alternative approach to testing the model's predictions. See, e.g., Cason (1996), who finds that screen-based trading systems and interdealer communications facilitate collusion.

arrangements of the Nasdaq market so as to better understand how public policy can enhance competition among dealers.

We demonstrate conditions under which market makers who act *noncooperatively* may set spreads above competitive levels. We term this behavior "implicit collusion." While implicit collusion is not always possible, it can occur under relatively general conditions. Further, frictions such as price discreteness or asymmetric information are not required for collusive behavior.

Our characterization of implicit collusion may prove useful for policy makers. Some authors have argued against implicit collusion on the grounds that this is implausible in heavily traded stocks with large numbers of market makers. However, our analysis suggests that collusion may be easier in actively traded stocks than in inactive stocks. This is because the gains from undercutting rival dealers' prices relative to the loss of future expected profits decrease as the stock is more frequently traded. Similarly, we show that collusion is more difficult to maintain when volume is heavily concentrated among a few dealers. This occurs because the smallest dealer has the most to gain and least to lose from undercutting the prevailing quoted spreads.

Interestingly, our analysis shows that institutional features such as order preferencing rules and internalization play a critical role in sustaining supranormal spreads. These arrangements are inconsistent with time priority and reduce dealers' incentives to compete for order flow on price. In the absence of price competition, however, dealers may compete in other dimensions by offering brokers cash and noncash inducements to secure access to the order flow. From a policy viewpoint, the analysis suggests that efforts to enhance competition among dealers should be focused on restrictions on access to the order flow.

Appendix

Proof of Proposition 1: The symmetry assumption (R1) implies there is an identical maximand for both problems. By (R4), there is a unique maximand for the problem; i.e., there exists $\alpha^c(z_t) \geq 0$ such that the unique profit maximizing ask in round t is $v + \alpha^c(z_t)$ while the profit maximizing bid is $v - \alpha^c(z_t)$ at which the market clears on average.

Now suppose that $\alpha^c(z_t') < \alpha^c(z_t)$, i.e., the ask decreases with z. Profit maximization requires that at z_t', $\alpha^c(z_t)$ is a feasible but suboptimal action—and vice-versa. In other words,

$$\alpha(z_t')d(v + \alpha(z_t'), z_t') > \alpha(z_t)d(v + \alpha(z_t), z_t')$$

$$\alpha(z_t)d(v + \alpha(z_t), z_t) > \alpha(z_t')d(v + \alpha(z_t'), z_t) \tag{A.1}$$

The inequalities in equation (A.1) in turn imply that

$$\frac{d(v + \alpha(z_t'), z_t')}{d(v + \alpha(z_t), z_t')} > \frac{d(v + \alpha(z_t'), z_t)}{d(v + \alpha(z_t), z_t)} \tag{A.2}$$

By hypothesis, $z_t' \geq z_t$ and $v + \alpha^c(z_t') < v + \alpha^c(z_t)$. Equation (A.2) is then seen to contradict the monotone demand ratio condition (R3).

The conclusion that expected profits increase in the volume shock follows from the monotonicity assumption (R2). This is because one possible price at a higher volume shock is to post exactly the same price that is posted at a lower volume shock; clearly, quantities and therefore profits are nevertheless higher than at the lower volume shock. Q.E.D.

Proof of Proposition 2: Step 1. It is well known (see, e.g., Abreu (1988)) that all equilibrium pricing strategies can be decomposed into two parts: (i) A *normal* pricing rule (followed in the event that there have been no departures from it in the past), and (ii) A *fallback* rule (which is adopted in the event of at least one departure from the normal pricing rule in a previous period.) Further, the fallback rule can be taken to be the pricing rule that generates the "worst" (i.e., lowest expected profit) equilibrium. A normal pricing rule is sustainable as an equilibrium strategy if and only if no single-period deviation from it yields higher expected lifetime profits. To establish the proposition, we must derive the normal and fallback rules.

Observe that the worst equilibrium (regardless of the discount factor) occurs when dealers price at value regardless of the current state and previous trading history. Thus, we can take the fallback rule to be $a_t = b_t = v$ in each trading round t. Now consider the normal pricing rule. Let $K = \rho E[J(z)]$ be the present value of the continuation payoffs to dealers as a group under the normal strategy, where $J(z)$ is the value function under the second-best strategy. Suppose that $K > 0$, and the current volume shock is z. Recall from the incentive compatibility condition (8) that the collusive prices are sustainable if and only if

$$\frac{K}{M} \geq \left(1 - \frac{1}{M}\right)[\pi_a(a^c(z); z) + \pi_b(b^c(z); z)]. \tag{A.3}$$

From Proposition 1, we know the right-hand side of equation (A.3) is increasing in z. Define a constant $z_c \geq 0$ as the solution to

$$\frac{K}{M} = \left(1 - \frac{1}{M}\right)[\pi_a(a^c(z_c); z_c) + \pi_b(b^c(z_c); z_c)], \tag{A.4}$$

provided a solution exists. If $K/M < (1 - 1/M)[\pi_a(a^c(0), 0) + \pi_b(b^c(0), 0)]$, then define $z_c = 0$. On the other hand, if K/M is greater than the right-hand side of equation (A.3) for all values of z, define $z_c = \infty$.

Define a normal pricing rule as follows: (i) For $z < z_c$, the normal pricing rule is to price at the collusive prices $(a^c(z), b^c(z))$; (ii) For $z \geq z_c$, the prices are $(a^*(z), b^*(z))$, which solve

$$\frac{K}{M} = \left(1 - \frac{1}{M}\right)[\pi_a(a^*; z) + \pi_b(b^*; z)]. \tag{A.5}$$

The second-best strategy is exactly this pricing rule. To see this, observe that the pricing strategy is clearly deviation proof if play continues with the second-best pricing strategy from the next period onwards. Further, it achieves the highest profits among the set of prices that are sustainable.

From the discussion above, the expected current profit in a period when $z < z_c$ is $\pi_a(a^c(z); z) + \pi_b(b^c(z); z)$, whereas when $z \geq z_c$, the expected current profit is (from equation (A.5)) $K/(M - 1)$. In both cases, the discounted continuation value is K. It follows that

$$J(z) = \begin{cases} \pi_a(a^c(z);\, z) + \pi_b(b^c(z);\, z) + K & \text{for} \quad z < z_c \\ KM/(M - 1) & \text{for} \quad z \geq z_c \end{cases} \tag{A.6}$$

So far, we have analyzed the optimal strategy under the assumption that the continuation payoffs (i.e., K) are positive. In Steps 2 and 3, we examine the conditions under which K is positive.

Step 2: Here we show that if dealers are sufficiently impatient (in particular, if $\rho < 1 - 1/M$), the unique equilibrium is the competitive equilibrium where $a_t = b_t = v$. To see this, observe that the claim of uniqueness for the competitive equilibrium is equivalent to a claim that the expected profits in the second-best equilibrium are exactly zero, i.e., $K = 0$. We shall establish this latter claim.

Suppose that $K > 0$. From above, the second-best value function is given by equation (A.6). For all z,

$$J(z) \leq \frac{MK}{M - 1}. \tag{A.7}$$

This inequality holds because for $z < z_c$, $J(z) = [\pi_a(a^c(z); z) + \pi_b(b^c(z); z)] + K$, and by the incentive constraint the current expected profits (the term in brackets) is strictly less than $K/(M - 1)$. For $z \geq z_c$, $J(z) = MK/(M - 1)$. Taking expectations in equation (A.7), and recalling that $K = \rho E[J(z)]$, we obtain the inequality $\rho \geq 1 - 1/M$. Thus, if $\rho < 1 - 1/M$, $K = 0$, and the only equilibrium is to price at value.

Observe that when $\rho > 1 - 1/M$, the critical value $z_c > 0$ when $K > 0$. To see this, note that if $z_c = 0$, then $J(z) = (M - 1)K/M$ which implies that $\rho = M/(M - 1)$. We now demonstrate that there exist conditions (provided dealers are sufficiently patient) such that $K > 0$, i.e., that the competitive equilibrium is not the only equilibrium.

Step 3: We begin by showing that the second-best profits are increasing in the discount factor. In particular, suppose that the second-best expected payoffs, denoted K_1, are positive for $\rho_1 < 1$. Then, for any discount factor $\rho_2 > \rho_1$, the associated second-best payoffs, say K_2 are larger, i.e., $K_2 > K_1$.

To establish this result, note that the second-best pricing rule for ρ_1 is to price according to $(a^c(z), b^c(z))$ for $z \in [0, z_c(\rho_1)]$ and according to $(a^*(z), b^*(z))$ for $z > z_c(\rho_1)$. This rule is incentive compatible at a higher discount factor ρ_2.

The payoffs to following this strategy are therefore higher at ρ_2, because each period's profits are positive and hence sum to a higher number when the discount factor increases. The second-best payoffs under ρ_2 must be, by definition, even higher than this value, establishing the claim.

Step 4: Lastly, we show that the second-best profits must eventually be positive for sufficiently high discount factors. In particular, there exists $\rho_0 < 1$ such that $K > 0$ for $\rho \geq \rho_0$. To see this, consider the following modification of the second-best equilibrium: Pick any $z_0 \geq 0$ and for $z < z_0$, price according to the first-best or collusive pricing strategy while for $z \geq z_0$, price at value. We show that this strategy is an equilibrium for an appropriate discount factor. Evidently, the prices for $z \geq z_0$ are equilibrium prices. What happens when $z < z_0$? Let K_0 be the expected lifetime returns to this strategy, i.e.,

$$K_0 = \frac{\rho}{1-\rho}\left\{\int_0^{z_0} [\pi_a(a^c(z);\, z) + \pi_b(b^c(z);\, z)]\; \mathrm{d}F(z)\right\}. \tag{A.8}$$

The strategy is incentive compatible for $z < z_0$ if and only if

$$\frac{1}{M} K_0 \geq \left(1 - \frac{1}{M}\right)[\pi_a(a^c(z);\, z) + \pi_b(b^c(z);\, z)], \tag{A.9}$$

for $z < z_0$. This equation holds for an appropriate discount factor $\rho_0 < 1$. Since the profits to this strategy are strictly positive, it follows that the second-best profits are strictly positive as well.

Collecting steps 1–4, we have established that there exists a discount factor $\rho_0 \in [1 - 1/M, 1)$ such that: (i) For $\rho \leq \rho_0$, the unique equilibrium is to price at value, so that $K = 0$; and (ii) For $\rho > \rho_0$, the second-best pricing rule is as described above and $K > 0$.

Finally, observe that for $z < z_c$, the spread is the collusive spread and hence increases in z. For $z \geq z_c$, we know that $K/(M - 1) = [a^*(z) - b^*(z)]Q(a^*, b^*; z)$, where $Q(a^*, b^*; z)$ is the expected trading volume. Since Q is an increasing function of z, the spread must decline with z.

It is possible that implicit and explicit collusion coincide. Specifically, when $K > 0$ (i.e., for $\rho > \rho_0$), the incentive constraint (A.3) may be satisfied for all z (i.e., z_c is infinite), and collusive pricing is always sustainable. In this case,

$$K = \frac{\rho}{1-\rho} E_z[\pi_a(a^c(z);\, z) + \pi_b(b^c(z);\, z)] < \infty. \tag{A.10}$$

The critical bound $z_c(\rho)$ is infinite for sufficiently high ρ if and only if

$$\lim_{z\to\infty}[\pi_a(a^c(z);\, z) + \pi_b(b^c(z);\, z)] < \infty.$$

To see this, note that $z_c(\rho) = \infty$ must imply from the incentive constraint (A.3) that $\lim_{z\to\infty}[\pi_a(a^c(z); z) + \pi_b(b^c(z); z)]$ is finite. Conversely, if the limit is finite,

$z_c(\rho) = \infty$ for a sufficiently high ρ. This can be seen from equation (A.10), since the lifetime payoffs $K \to \infty$ as $\rho \to 1$. Therefore, in the limit, as $\rho \to 1$, the normal pricing rule and the overt collusive pricing rule coincide, and collusion is sustainable in all cases. Q.E.D.

Proof of Proposition 3: Size. Let $J^k(z)$ denote the value function when the market scale is given by k. Now consider equations (6) and (7). It is clear that the profit functions are homogeneous of degree one in k. Thus, if a price pair (a', b') maximizes the original value function, $J_1(z)$, this also maximizes the value function $J^k(z) = kJ_1(z)$. Thus, k does not affect the choice of $\{a, b\}$ that maximizes $J(z)$.

Frequency. Denote by $\bar{r}$ the discount rate over a calendar period of unit length with n trading rounds. Let r_n denote the discount rate applied to each trading period, where $r_n = (1 + \bar{r})^{1/n} - 1$. For an actively traded stock, there are more trading rounds in a given *calendar* interval of time. As trading frequency, n, increases, the discount factor applied to future trading rounds, $\rho_n = 1/(1 + r_n)$, increases monotonically. From Step 3 in the proof of Proposition 2, we know that as ρ increases, both K and $z_c(\rho)$ increase. Hence, as n increases, the range under which collusion is sustainable and the second-best profits increase. In the limit, as $\rho_n \to 1$, overt collusion is sustainable for all z. Q.E.D.

Proof of Proposition 4: Let $p(z) = (a(z), b(z))$ denote any equilibrium pricing rule, and let dealer i's associated profits be denoted by $W_i(z; \Phi)$. Denote by K_i the expected discounted profits, i.e., $K_i = \rho E[W_i(z; \Phi)]$. We consider equilibria in which all dealers quote identical prices on the equilibrium path. The price quote (a', b') is sustainable if

$$\phi_i[\pi_a(a'; z) + \pi_b(b'; z) + K(\Phi)] \geq \pi_a(a'; z) + \pi_b(b'; z) \quad \text{(A.11)}$$

for all $i = 1, \ldots, M$. The highest payoffs to undercutting is given by the right-hand side of equation (A.11) since dealer i can attract the entire order flow by marginally undercutting the prevailing ask or bid price. The incentive constraint (A.11) is equivalent to

$$K(\Phi) \geq \left(\frac{1 - \phi_i}{\phi_i}\right)[\pi_a(a'; z) + \pi_b(b'; z)]. \quad \text{(A.12)}$$

It is clear that the right-hand side of equation (A.12) is maximized when $\phi_i = \min[\phi_1, \ldots, \phi_M]$. It follows immediately that if market maker M does not undercut the quoted prices, neither do any of the others. This establishes the proposition. Q.E.D.

Proof of Proposition 5: We proceed in several steps, first developing an algorithm to compute the second-best payoffs for any fixed share structure Φ, and then using this algorithm to show how a change in the share structure affects payoffs, and hence spreads.

Step 1: Consider the lifetime payoffs when market makers act as a cartel. Denote these payoffs $J_0(z;\,\Phi)$, where:

$$J_0(z;\,\Phi) = \pi_a(a^c(z);\,z) + \pi_b(b^c(z);\,z) + \rho E[J_0(z;\,\Phi)]. \qquad \text{(A.13)}$$

Thus, J_0 represents the first-best value function, which is independent of the share structure Φ; we have written it as such to maintain consistency with the notation that follows.

The iterative algorithm used to construct the second-best value function generalizes the approach used by Abreu, Pearce, and Stachetti (1990). Consider the following "one-shot" game: If dealers set prices $(a',\, b')$, dealer i gets a payoff $\phi_i\{\pi_a(a';z) + \pi_b(b';z) + \rho E[J_0(z;\Phi)]\}$, where $J_0(\,\cdot\,)$ (the lifetime profits under overt collusion) can be interpreted as the payoff dealers expect for not undercutting the price pair $(a',\, b')$. If a dealer marginally undercuts the prevailing prices, he or she receives expected profits equal to $\pi_a(a';\,z) + \pi_b(b';\,z)$. The highest equilibrium payoffs to this one-shot game is defined by the mapping T:

$$TJ_0(z;\,\Phi) = \max_{\{a',b'\}}[\pi_a(a';\,z) + \pi_b(b';\,z) + \rho E[J_0(z;\,\Phi)]] \qquad \text{(A.14)}$$

subject to

$$\rho E[J_0(z;\,\Phi)] \geq \frac{1-\phi_M}{\phi_M}[\pi_a(a';\,z) + \pi_b(b';\,z)]. \qquad \text{(A.15)}$$

Let $J_1(z;\,\Phi)$ denote the value function defined by the mapping TJ_0.

Now continue the iteration one more step by considering another one-shot game: If dealers set prices $(a',\, b')$, dealer i receives $\phi_i\{\pi_a(a';\,z) + \pi_b(b';\,z) + \rho E[J_1(z;\,\Phi)]\}$, whereas if the dealer deviates, he or she receives an expected profits of $\pi_a(a';\,z) + \pi_b(b';\,z)$. Let $J_2 \equiv TJ_1$ denote the highest sustainable profits to this second game. Equivalently, J_2 can be interpreted as the highest sustainable profits in a market with a three period horizon where trading proceeds to higher rounds if and only if no dealer has undercut prices in the previous round and where the terminal (round three) payoff is the lifetime collusive profit J_0.

More generally, define $J_{n+1}(z;\,\Phi) = TJ_n(z;\,\Phi)$. Suppose, for every z, that there is a function $J(z;\,\Phi)$ such that $J(z;\,\Phi) = \lim_{n\to\infty} J_n(z;\,\Phi)$. (We establish that such a limit exists below.) Then, by continuity, $J(z;\,\Phi) = TJ(z;\,\Phi)$. The mapping TJ yields the highest payoffs from incentive-compatible prices with J as the continuation rewards. Hence, by the logic explained earlier, the fixed point for this mapping, J, must be the second-best value function.

Step 2: Note that $J_1(z;\,\Phi) \leq J_0(z;\,\Phi)$. This must be the case since J_0 solves the maximization problem (A.14) without the additional constraint imposed by (A.15). Then, if $J_2(z;\,\Phi) = TJ_1(z;\,\Phi)$, it follows that $J_2 \leq J_1$. To see this, observe that $J_0 \geq J_1$ implies that the set of $(a',\, b')$ which satisfy the incentive constraint under J_1 must also satisfy this under J_0. This fact, combined with the fact that the maximand in equation (A.14) is smaller with J_1 rather than J_0, implies that $J_2 \leq J_1$. More generally, recall that $J_{n+1}(z;\, \Phi) = TJ_n(z;\, \Phi)$. By the arguments above, it follows that $J_{n+1} \leq J_n$.

Step 3: Define, for every z, the limit $J(z;\, \Phi) = \lim_{n\to\infty} J_n(z;\, \Phi)$. The limit is well defined because of the monotonicity property described above. Further, it is immediate that the limit satisfies the following property:

$$J(z;\, \Phi) = \max_{\{a',b'\}}[\pi_a(a';\, z) + \pi_b(b';\, z) + \rho E[J]] \tag{A.16}$$

subject to

$$\rho E[J] \geq \frac{1 - \phi_M}{\phi_M}[\pi_a(a';\, z) + \pi_b(b';\, z)]. \tag{A.17}$$

Thus, as argued above, J represents dealers' second-best expected lifetime profits.

Step 4: Now consider an alternative share structure Φ'. From Proposition 4, we need only focus on the smallest dealer. Suppose that $\phi_M \geq \phi_M'$, and $J_0(z;\, \Phi') = J_0(z;\, \Phi)$. It follows immediately from equation (A.14) that the set of bid–ask price pairs that are sustainable in equilibrium is smaller under Φ'. Hence, in turn, $J_1(z;\, \Phi') \leq J_1(z;\, \Phi)$. Repeating this argument yields $J_n(z;\, \Phi') \leq J_n(z;\, \Phi)$ for all n and z. Hence, the second-best payoffs with share structure Φ' are smaller than those under Φ, i.e., $J(z;\, \Phi') \leq J(z;\, \Phi)$, for all z, and $\rho E_z[J(z;\, \Phi')] \equiv K(\Phi') \leq K(\Phi) \equiv \rho E[J(z;\, \Phi)]$. From the incentive constraint it follows that any price pair that is incentive compatible at Φ' will also be incentive compatible at Φ.

There are three cases to consider: (i) For small values of z, the collusive prices $a^c(z)$ and $b^c(z)$ are incentive compatible for both Φ and Φ'; (ii) For larger values of the volume shock, the collusive prices are incentive compatible for Φ but not for Φ'; and (iii) For large values of z, the collusive prices cannot be sustained under either Φ or Φ'. In cases 2 and 3, the sustainable ask (bid) must be lower (higher) than the collusive ask (bid). In particular, the ask under Φ is higher and the bid lower than under Φ'. This completes the proof. Q.E.D.

Proof of Proposition 6: Proposition 4 implies that bid–ask spreads, and hence total expected lifetime profits, depend only on the share in order flow of the smallest market maker. Suppose initially that $M > N$. Then, the share of the smallest dealer is $(1 - \theta)/M$, so that the equilibrium number of dealers in this case solves (ignoring integer constraints):

$$\frac{(1 - \theta)}{M} K\left(\frac{(1 - \theta)}{M}\right) = c \tag{A.18}$$

However,

$$\frac{1}{M} K\left(\frac{1}{M}\right) > \frac{(1-\theta)}{M} K\left(\frac{(1-\theta)}{M}\right). \tag{A.19}$$

It follows therefore that the equilibrium number of dealers in the case where some order flow is directed by brokers is less than in the case where dealers share order flow equally.

Now consider the case where $M = N$, i.e., no dealers choose to enter other than those who already receive directed order flow. Without loss of generality, suppose that dealer N receives the smallest share of the directed order flow, i.e., $\theta_N = \min[\theta_1, \ldots, \theta_N]$. If $M = N$, then it must be the case that $(\theta_N + (1 - \theta)/N)K(\theta_N + (1 - \theta)/N) = c$. Again, since the order flow share of dealer N is at most $1/N$, it follows that the equilibrium number of dealers in the case where no order flow is directed to dealers is larger than N. Q.E.D.

Proof of Proposition 7: Let $U(z)$ denote the aggregate expected present value of trading profits associated with a pricing strategy (not necessarily the second-best strategy) $\hat{p}(z) = (\hat{a}(z), \hat{b}(z))$. Thus,

$$U(z) = \pi_a(\hat{a}(z); z) + \pi_b(\hat{b}(z); z) + \rho E[U(z)]. \tag{A.20}$$

Since π_a and π_b are single-peaked functions (with maxima attained at the collusive (first-best) price levels a^c and b^c), the incentive constraint is

$$(1/M)U(z) \geq \pi_a(\hat{a}(z) - \Delta; z) + \pi_b(\hat{b}(z) + \Delta; z), \tag{A.21}$$

where $\Delta > 0$ is the minimum price variation. In other words, a dealer's expected share of the total expected lifetime profits must exceed the one-time profits from undercutting the cartel prices. Equation (A.21) implies that

$$(1/M)U(z) > \pi_a(\hat{a}(z) - \Delta'; z) + \pi_b(\hat{b}(z) + \Delta'; z), \tag{A.22}$$

for $\Delta' > \Delta$. This follows from the fact that π_a and π_b are strictly concave functions with maxima attained at $a^c(z)$ and $b^c(z)$. Equation (A.22) is sufficient to establish that $(\hat{a}, \hat{b})$ is also sustainable under $P(\Delta')$. Now suppose that $\hat{p}(z)$ is the second-best under $P(\Delta)$ and is feasible under $P(\Delta')$. The proposition then implies that this pricing strategy is also sustainable under Δ'. Hence, by definition, $J(z; \Delta') \geq J(z; \Delta)$. Q.E.D.

REFERENCES

Abreu, Dilip, 1988, On the theory of infinitely repeated games with discounting, *Econometrica* 56, 383–396.

Abreu, Dilip, David Pearce, and Ennio Stachetti, 1990, Towards a theory of discounted repeated games with imperfect monitoring, *Econometrica* 58, 1041–1063.

Barclay, Michael, 1995, Bid–ask spreads and the avoidance of odd-eighth quotes on Nasdaq: An examination of exchange listings, Working paper, University of Rochester.

Bhasin, Vijay, and David Brown, 1995, The organization of competitive financial exchanges, Working paper, Indiana University.

Blume, Marshall, and Michael Goldstein, 1992, Displayed and effective spreads by market, Working paper, University of Pennsylvania.

Cason, Timothy, 1996, The opportunity for conspiracy in asset markets organized with dealer intermediaries, Working paper, University of Southern California.

Christie, William, and Roger Huang, 1994, Market structure and liquidity: A transactions data study of exchange listing, *Journal of Financial Intermediation* 3, 300–326.

Christie, William, and Paul Schultz, 1994, Why do Nasdaq market makers avoid odd-eighth quotes?, *Journal of Finance* 49, 1813–1840.

Christie, William, Jeffrey Harris, and Paul Schultz, 1994, Why did Nasdaq market makers stop avoiding odd-eighth quotes?, *Journal of Finance* 49, 1841–1860.

Doran, Lynn, Kenneth Lehn, and Kuldeep Shastri, 1995, Do Nasdaq market makers collude? Evidence from 19c-3 stocks, Working paper, University of Pittsburgh.

Dutta, Prajit, 1995, A folk theorem for stochastic games, *Journal of Economic Theory* 60, 1–32.

Furbush, Dean, 1995, The Nasdaq market: Reconsidering the collusion hypothesis, Mimeo, Economists Incorporated.

Godek, Paul, 1995, Why Nasdaq market makers avoid odd-eighth quotes, Mimeo, Economists Incorporated.

Grossman, Sanford, Merton Miller, Daniel Fischel, Kenneth Cone, and David Ross, 1995, Clustering and competition in asset markets, Mimeo, Lexicon Incorporated.

Harris, Lawrence, 1991, Stock price clustering and discreteness, *Review of Financial Studies* 4, 389–416.

Harris, Lawrence, 1994, Minimum price variations, discrete bid–ask spreads, and quotation sizes, *Review of Financial Studies* 7, 149–178.

Huang, Roger, and Hans Stoll, 1996, Dealer versus auction markets: A paired comparison of execution costs on Nasdaq and the NYSE, *Journal of Financial Economics* 41, 313–357.

Kandel, Eugene, and Leslie Marx, 1995, Nasdaq market structure and spread patterns, Working paper, University of Rochester.

Keim, Donald, and Ananth Madhavan, 1995, Transactions costs and investment style: An inter-exchange analysis of institutional equity trades, Working paper, University of Southern California.

Kleidon, Allan, and Robert Willig, 1995, Why do Christie and Schultz infer collusion from their data?, Mimeo, Cornerstone Research.

Laux, Paul, 1995, The bid–ask spreads of Nasdaq stocks that quote on even eighths, Working paper, Case Western Reserve University.

National Association of Security Dealers, Inc., 1991, *Inducements for Order Flow*, Washington, D.C.

Neal, Robert, 1992, A comparison of transaction cost between competitive market maker and specialist market structures, *Journal of Business* 65, 317–334.

Reinganum, Marc, 1990, Market microstructure and asset pricing: An empirical investigation of NYSE and Nasdaq Securities, *Journal of Financial Economics* 28, 127–147.

Rotemberg, Julio, and Garth Saloner, 1986, A supergame-theoretic model of business cycles and price wars during booms, *American Economic Review* 76, 390–407.

Shapiro, James, 1993, Recent competitive developements in U.S. equity markets, Working paper, New York Stock Exchange, Inc., New York, NY.

Stoll, Hans, 1976, Dealer inventory behavior: An empirical investigation of Nasdaq stocks, *Journal of Financial and Quantitative Analysis* 11, 359–380.

Wahal, Sunil, 1995, Entry, exit, market makers, and the bid–ask spread, Working paper, University of North Carolina.