THE INFORMATIONAL ROLE OF UPSTAIRS AND DOWNSTAIRS TRADING

by

Sanford J. Grossman

22-90

RODNEY L. WHITE CENTER FOR FINANCIAL RESEARCH
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6367

The contents of this paper are the sole responsibility of the author(s).

Copyright @ 1990 by S. J. Grossman

THE INFORMATIONAL ROLE OF UPSTAIRS AND DOWNSTAIRS TRADING

Sanford J. Grossman[1]
Steinberg Trustee Professor of Finance

The Wharton School
University of Pennsylvania
Philadelphia, PA 19104

April 1990

Revised August 1990

[1] Preliminary and incomplete draft; please do not quote or circulate without the author's permission. I am grateful to Krishna Ramaswamy for helpful discussion, and to members of the financial community who probably prefer to remain anonymous.

Abstract

Much of economic theory is concerned with understanding price determination in competitive markets. Such theories assume that all individuals continuously participate in one giant market where they can express their demands for all assets simultaneously as a function of a giant price vector. This assumption of simultaneous and continuous participation in all markets is inconsistent with two important facts: First, it is costly for an individual or an institution to continuously express demands in any single market, and second it is simply impossible to trade in all markets simultaneously. These two facts create a need for intermediaries. Much is known about the role of intermediaries as principals who add liquidity to markets by trading on their own account. However, far less is known about the informational role of intermediaries. In this paper, we will analyze the consequences of the fact that intermediaries play a fundamental role as repositories of information.

The Informational Role of Upstairs and Downstairs Trading

I. Introduction

Much of economic theory is concerned with understanding price determination in competitive markets. Such theories assume that all individuals continuously participate in one giant market where they can express their demands for all assets simultaneously as a function of a giant price vector. This assumption of simultaneous and continuous participation in all markets is inconsistent with two important facts: First, it is costly for an individual or an institution to continuously express demands in any single market, and second it is simply impossible to trade in all markets simultaneously. These two facts create a need for intermediaries. Much is known about the role of intermediaries as principals who add liquidity to markets by trading on their own account. However, far less is known about the informational role of intermediaries. In this paper, we will analyze the consequences of the fact that intermediaries play a fundamental role as repositories of information.

The structure of this paper is as follows. Section 2 discusses the reasons for the failure of all individuals to participate continuously in all markets simultaneously, and explains how this creates an informational role for intermediaries. It also explains how intermediaries with information about customer preferences use this information to facilitate upstairs market making.

Section 3 presents a model of upstairs and downstairs market making. A downstairs market refers to an organized Exchange where all members agree that trades take place publicly in a central place. An upstairs market refers to a market where trades take place privately; two parties can agree on a price at the same time as two other parties are trading at a different price. The upstairs dealer is potentially a repository of customer information; he may know what states of nature are likely to induce a customer to trade. The upstairs dealer can give one selling customer a higher bid than prevails downstairs when he knows that another customer is an interested buyer. The informational advantage which may be possessed by the upstairs dealer can be offset by the fact that a customer must search for the best price, and indeed has no way of knowing that at the instant he makes a deal, he is indeed receiving the best price available. Section 3 models equilibrium market making which simultaneously occurs on upstairs and downstairs markets. Customers choose whether to be upstairs or downstairs customers. The model computes the equilibrium fraction of customers who choose to trade on upstairs and downstairs markets. Both upstairs and downstairs market makers get information from their respective current <u>expressed</u> order flows. This information is distinct from an upstairs broker's knowledge of the willingness to trade of customers who have <u>not</u> placed orders, i.e. <u>unexpressed</u> order flow. The current order flow thus creates an externality which for example implies that a downstairs market may fail even though upstairs market makers

possess no private information about their customers' unexpressed orders. For example, if most customers decide to use upstairs markets, then any single customer thinking of switching to the downstairs market will face an illiquid market. The downstairs market is illiquid because downstairs market makers must make wide markets if they observe only a small portion of the order flow.

Section 4 discusses arbitrage between the upstairs and downstairs markets.

Section 5 discusses the implications of the model for the issues surrounding the regulation of Off-Exchange trading, prearranged trading, "crossing" of orders, and related issues.

Section 6 contains conclusions and explains how the model can be used to understand the linkages between any two related markets, such as the stock market and the index futures market.

2. The Failure of Continuous Participation and its Implications

Economic theory assumes that each investor presents a demand function for every asset as a function of all prices, and that this demand is continuously updated as new information arrives about relevant states of nature. In reality, no investor can produce these continuously updated functions. There are a number of reasons for the difference between reality and economic theory:

1. It is expensive to preplan behavior.

2. It is expensive for an investor to remain on the Exchange all day because of the possiblity that he will want to trade.

3. As a consequence, an investor will delegate his order to

a broker who is able to achieve economies of scale in participation. However, it is more difficult to explain state contingent plans to a broker (in a manner so that his failure to carry out the instruction is verifiable) than it is to execute and develop the plans for oneself.

4. As a consequence, the contingent orders left with brokers are very simple, and therefore very risky to the investor. Simple limit orders are used which give away free options, and are thus very costly.

5. As a consequence, most investors' preferences are not continuously represented on organized downstairs markets.

The above points are best understood with some examples. Consider the market for OEX stock index options on the Chicago Board Options Exchange (CBOE). A customer can place a standard limit order which specifies his desire to buy as a function of the option price. This type of limit order is extremely risky to the customer, since the value of the option to him depends on the underlying value of the OEX stock index. The customer would prefer to place an order which states that he is willing to pay, say $1 for an OEX call if the OEX index is above 300. Though the CBOE's limit order book cannot handle the above type of order, a customer can ask his broker to take the above order. Brokers will even take orders for options based upon the implied volatility of the option (e.g., "buy the call if its Black-Scholes volatility falls to 15%"). A more sophisticated customer will recognize that the OEX

index can lag the actual value of the component stocks due to the fact that not all stocks in the index trade simultaneously. Such a customer may decide that the S&P500 futures price will, in some circumstances, lead the OEX cash index price. He may then decide that his options order should be conditioned not only on the option price but also the S&P500 futures price. Here, even a very sophisticated trader will have trouble giving an order, even to a sophisticated broker, regarding the contingencies which should cause the option to be bought. Sometimes the OEX index moves in a manner not perfectly correlated with the S&P500 index, even when all stocks in both indexes trade at the same time. Some customers will be able to leave a resting contingent limit order with a broker which expresses their demands, but others will find it impossible to specify all the contingencies under which their order will get executed. The more contingencies which are specified, the more difficult it is for the broker to carry out the customers desires, and the more difficult it will be for the customer to verify whether indeed the broker has carried out his instructions.

A broker who has a long term relationship with a customer will come to understand the types of contingencies which can bring forth the customer's orders. For example, some customers would like to buy OEX volatility when it falls in price relative to S&P500 volatility. But there may be other contingencies which affect the customer's desires so that a resting limit order does not exist. With the above in mind, consider a customer who wants to buy a large amount of out-of-the-money calls. Suppose he needs the order

executed very quickly because of some speculative or hedging requirement. If he sends a large market order down to the floor of the Exchange, then he will bid up the price (and implied volatility) of the infrequently traded option. However, his broker may ask the customer to wait a moment while the broker solicits the other side of the trade from his other customers. The broker may know that some of his customers would be willing to sell volatility at a price slightly above the last price. These latter customers do not leave resting limit orders with their brokers for the reasons given above. The broker is a repository of information about their preferences. For this reason the broker can:

(a) take the other side of his customer's buy order at a better price than the customer could immediately achieve on the Exchange floor, with the knowledge that the broker can turn around and buy the options from the customers who are likely sellers of volatility; or

(b) delay the buying customer while he finds selling customers, and then bring both orders down to the Exchange floor simultaneously, and "crossing" them at the price agreed upon upstairs.

Note that in the above example, the broker gets two commissions, rather than one, and this extra commission is a return for his information. Note also that if the customer had merely bought the options "at market," then he would have caused a temporary price rise. The upstairs broker would have alerted his selling customers after the price rise, and their selling will tend to return the price to its previous level. In the event that the

buyers bid up the price to attract customers, then the informational function of the market price is being substituted for the informational function of the broker.

An even more extreme example of the broker's informational function appears in the presence of intermarket spreads. For example, consider an institutional customer who wants to buy 20,000 "in the money" calls on a stock which is rumored to be a takeover target. Suppose this is the type of stock which normally trades 10,000 shares a day and 200 calls a day. The broker can sell the calls directly to the customer, and then hedge the position by buying stock. If the broker has to buy 20,000 shares of stock at market, then it will bear a large market impact cost. On the other hand, brokers have lists of all institutional holders of individual stocks. The broker or some other salesman may recognize that one of the institutional holders of the stock has been a willing seller in similar circumstances in the past. The broker can buy the stock from that customer, and sell the call simultaneously to the first customer. Of course, the broker will try to find a call writer to cross with the first customer's order, thereby getting two commissions and avoiding the risk of hedging the options. Some brokers have even found a method of getting <u>four</u> commissions out of the first customer's order, as follows. The broker may know customers who follow "buy-write" strategies. A "buy-write" customer is interested in buying stock and selling "in the money" calls (of course, he is merely a writer of "out of the money" puts). An insurance company might do a buy-write for extra

"income."[1] If the broker knows of these three customers, then he can effect a cross without taking any positions on his own account. He will receive commissions from (I) the calls purchased by the customer who initiated the trade to buy the calls; (II) the shares sold by the customer owning the stock; (III) the shares purchased by the customer doing the "buy-write;" and (IV) the options sold by the customer doing the "buy-write."

It is very difficult to see how the above three-way trade would have been instantaneously executed without a brokers intervention. Note that the initiator of the trade could have bid up the price of calls on the Exchange floor, and this would cause the market makers on the options floor to buy stock to hedge the calls which they write. This in turn will bid up the price of the stock. Finally, the customer interested in selling stock will buy the stock. The transaction, in that case, is effected via prices transmitting information to customers.

Another interesting example of the informational role of brokers arises in considering currency forward and futures markets. The organized futures markets for currencies have remained miniscule relative to the interbank forward market. This has persisted over many years. One reason for this is that there are position limits and inflexible margin requirements for futures but

---

[1] Buy-writes are presented by brokers to their customers with a "rate of return" computed by assuming that the calls will expire "in the money," and treating the transactions as if the customer currently pays the stock price less the option premium, and will receive the strike price "when" the call is exercised at expiration.

not for forwards.[2] A more important reason concerns the information flow to which forward market makers are privy. First, note that almost all of the hour to hour volatility in one to three month forward rates are due to the volatility of spot rates. This is because a forward transaction involves a spot transaction and then an interest rate spread, and the short term interest rate differential is far less volatile than the level of spot rates (at least at very high frequencies). The market makers in the interbank forward market are privy to an enormous amount of information about their customers desired spot and forward transactions. Most of their customers are business engaged in international trade. A U.S. firm which closes a deal with a foreign firm to supply heavy equipment may agree to a price in the foreign currency to be paid at various times in the future. The U.S. firm's bank may provide short term financing to the U.S. firm for the project and at the same time provide the currency hedging if the firm desires. The bank will also be privy to the currency flows before they occur.

As in the examples given earlier, if one customer needs to purchase say 100 million Canadian dollars forward, then he can place a market order on the futures market for 1000 Canadian dollar contracts (which represent about 45% of the typical day's volume) or he can transact in the forward market. To the extent that banks in the forward market use their information about customer currency

---

[2] However, the position limits are Exchange-imposed and exist at their current levels because of the particular current levels of open interest in the currency contracts.

flows, they will be able to make better bids and offers than downstairs market makers who are privy to less information.

3. Formal Model

For the sake of tractability, I assume that there are two dates at which trade takes place. Further, without any loss of generality, I assume that all trading is for assets in zero net supply (such as forwards or futures contracts).[3]

At date 2 the settlement value $\tilde{P}_2$ of the contract (which was traded at date 1) is represented as follows:

(1) $\tilde{P}_2 = \tilde{g}_2 - b\tilde{x}_2$,

where $\tilde{g}_2$ represents public information about the future payoff stream to the asset which underlies the contract, and $-b\tilde{x}_2$ represents the impact on the date 2 price of liquidity demanders. The coefficient b is taken as exogenous, and represents the extent to which order flow observed by market makers at date 1 impacts the date 2 price. Date 2 is the last date of trading, and I am simply assuming that the equilibrium price is given exogenously by equation (1).

Most of the analysis will focus on the equilibrium at date 1. At date 1, a liquidity event occurs. This is the event that some customers desire to trade. The magnitude of the event is represented by the realization of $\tilde{x}_2$. No one directly observes the

---

[3] See Grossman and Miller [1988] for a simple method to transform futures equilibria into "cash" market equilibria.

realization of $\tilde{x}_2$ at date 1. It is composed of the current ("at market") order flow to upstairs and downstairs brokers. In this Section we will take the order flow fractions to the upstairs and downstairs markets as exogenous. Specifically, let f be the fraction of the order flow expressed at date 1, and let q be the fraction of the date 1 order flow which is expressed in a downstairs market. More precisely, define

(2) $\tilde{x}_1 = \tilde{y}_d\sqrt{q} + \tilde{y}_u\sqrt{1-q}$

(3) $\tilde{x}_2 = \tilde{x}_1\sqrt{f} + \tilde{y}_2\sqrt{1-f}$,

where $\tilde{y}_u$, $\tilde{y}_d$, $\tilde{y}_2$ are independent and identically distributed Normally random variables with mean zero and variance σ_y^2. In addition, I assume that $\tilde{g}_2$ is Normally distributed and independent of $(\tilde{y}_u, \tilde{y}_d, \tilde{y}_2)$.

I assume that upstairs market makers observe $\tilde{y}_u$, and downstairs market makers observe $\tilde{y}_d$. In addition, as explained in Section 2, a primary business of upstairs market makers is to stay in contact with customers and thus have a good idea about what states of nature would induce them to trade. Hence, I assume that upstairs market makers observe the unexpressed customer orders $\tilde{y}_2$ at date 1, and that is the source of their advantage over downstairs market makers.

Downstairs Equilibrium

I assume that downstairs market makers maximize the expected value of their exponential utility of final (i.e., date 2) wealth. Let Z_d represent the demand of such a market maker, and let P_1 be the date 1 price. They choose Z_d to maximize

(4) $E[U(\tilde{W}_2)\,|\,y_d] = -\exp[-a\tilde{W}_2]$,

where

(5) $\tilde{W}_2 = W_1 + (\tilde{P}_2 - P_1)Z_d$,

and W_1 is his exogenous initial wealth.

Using the Normality assumption, the optimal Z_d is

(6) $Z_d = \dfrac{E[\tilde{P}_2|\,y_d] - P_1}{a\ Var[\tilde{P}_2|\,y_d]}$.

At this point we ignore arbitrage between the upstairs market and the downstairs market. If M_d is the number of downstairs market makers, then date 1, downstairs market clearing, requires that P_1 satisfy:

(7) $M_d\left[\dfrac{E[\tilde{P}_2|y_d] - P_1}{a\ Var[\tilde{P}_2|y_d]}\right] = y_d\sqrt{q}$

where $y_d\sqrt{q}$ represents the customer supply which is equated to the market maker downstairs demand. Equation (7) can be solved for the downstairs equilibrium price P_{1d} :

(8) $P_{1d} = E[\tilde{P}_2|\,y_d] - \dfrac{a\sqrt{q}}{M_d}Var[\tilde{P}_2|\,y_d]\,y_d$.

As would be anticipated, a large value of customer supply drives down price. Note that an identical equilibrium price would have been computed if we did not assume that market makers directly observe y_d, but instead computed a Rational Expectations Equilibrium where they conditioned their "demands" on price.

As in Grossman and Miller [1988], I assume that market makers face an entry cost of c_d. The market makers intial wealth is W_o, and it is assumed that entry of market makers occur to the point where their expected utility of net wealth is unchanged by their decision to enter the market making business, i.e.,

$$(9) \quad EU(W_o - c_d + (\tilde{P} - P_{1d}) Z_d) = EU(W_o)$$

The calculations in Grossman and Miller [1988, p.626], and the assumption that $E\tilde{y}_d = 0$, can be used directly to show that (9) is equivalent to

$$(10a) \qquad \sqrt{1 + t_d} = e^{ac_d},$$

where

$$(10b) \qquad t_d = a^2 \frac{Var[\tilde{P}_2 | y_d] q}{M_d^2} \sigma_y^2 .$$

Equations (10a) and (10b) can be used to solve for M_d as a function of the cost of market making and the other parameters. In what follows, I will always assume that M_d has adjusted so that (10) is true.

Thus far, customer participation has been exogenous. The benefits to a customer from selling x immediately in the downstairs

market is $x(P_1 - P_2)$. For example, a bond dealer may have an inventory of x bonds. If he sells x bond futures contracts in the downstairs market at P_1, and subsequently the price of bonds is P_2, then his gain is $(P_1 - P_2)x$. Assume that a particular customer must choose whether it will have upstairs or downstairs trading facilities before it knows the realization of P_1, P_2, and x. Then its expected utility from using the downstairs market is $EU_c(\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2))$. Assume that this particular trader will have liquidity needs independent of $\tilde{P}_{1d} - \tilde{P}_2$, in particular assume that $\tilde{x}$ is Normal and independent of $(\tilde{g}_2, \tilde{y}_d, \tilde{y}_u, \tilde{y}_2)$, with mean zero and variance σ_x^2. Assume that

(11) $U_c(W) = -e^{-hW}$.

Note that:

(12) $EU_c(\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)) = E\{E[U(\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)) \mid x]\}$.

Using (8), and the fact that $E\tilde{y}_d = 0$, it is clear that $\tilde{P}_{1d} - \tilde{P}_2$ is Normally distributed with mean zero and variance

(13) $\sigma^2_{\Delta P_d} = Var(\tilde{P}_{1d} - \tilde{P}_2) = \frac{a^2 q}{M_d^2}\{VAR[\tilde{P}_2 \mid y_d]\}^2\sigma_y^2 + VAR[\tilde{P}_2 \mid y_d]$.

Thus,

(14) $E[U(\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)) \mid x] = -\exp\left[h^2 \frac{x^2}{2}\sigma^2_{\Delta P_d}\right]$.

Using the moment generating function of the Chi-squared

distribution

(15) $EU(\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)) = -[1 - h^2\sigma^2_{\Delta P_d}\sigma^2_x]^{-\frac{1}{2}}$.

I assume that σ^2_x is sufficiently small that (15) is well defined and finite.

It follows from (15), that the quality of the downstairs market is a monotone decreasing function of $\sigma^2_{\Delta P_d}$. We will thus refer to $\sigma^2_{\Delta P_d}$ as the customers trading downstairs trading cost.

Note that (10b) may be used to eliminate the endogenous M_d from (13) and thus to obtain

(16) $\sigma^2_{\Delta P_d} = Var[\tilde{P}_2 | y_d]\, e^{2ac_d}$.

Thus, customers will receive high quality executions when the order flow y_d is very informative about the future price. This arises because market maker services are more effective when the order flow is more informative.

Upstairs Market Equilibrium

Two important distinguishing characteristics of upstairs markets are: (1) customers and market makers must spend some time searching for contra parties to a trade, and (2) trades are negotiated in private and not publicly displayed immediately to all potential participants. This leads to situations where two trades can take place at the same time but at different prices. The fact

that downstairs markets focus all orders in a single place implies that it is relatively rare for two trades to take place at the same time but at different prices in a downstairs market.[4]

A consequence of the fact that trades take place at different prices at the same time is that a particular market maker or customer will realize a price which is a random pertubation from the average price prevailing at a particular point in time.[5] I denote this extra volatility in realized price by σ_u^2. It is straightforward to verify that σ_u^2 causes (6) to be replaced by

$$(17)\quad Z_u = \frac{E[\tilde{P}_2 | y_u, y_2] - P_1}{a[Var[\tilde{P}_2 | y_u, y_2] + \sigma_u^2]},$$

where Z_u is the market makers demand function, and I am imposing the assumption that the market maker faces price risk associated with incomplete information about current prices.

Upstairs market clearing implies

$$(18)\quad M_u Z_u = y_u\sqrt{1-q},$$

which generates an equilibrium price P_{1u}:

---

[4] One exception is the opening and closing minutes on futures markets, and other periods of very hectic trading.

[5] I reject the often repeated assertion that customers sell at "the bid," and hence a dealer market should be modelled by modelling bid-ask spreads. In my view, a trade takes place when someone's bid crosses someone else's offer. A customer can always offer to sell at a price above a particular dealer's bid. However, in a dealer market, the customer's offer is not displayed to as many potential trading partners as is the dealer's offer. In a downstairs market, a customer's offer and a market maker's offer are displayed to the same set of people.

$$(19)\quad P_{1u} = E[\tilde{P}_2 \mid y_u, y_2] - \frac{a\sqrt{1-q}}{M_u}\left[Var[\tilde{P}_2 \mid y_u, y_2] + \sigma_u^2\right] y_u$$

Note that we could have replaced P_1 in (17) by $P_{1u} + \epsilon_i$ to represent the idea that a particular market maker trades at a price which randomly deviates from the average upstairs price P_{1u}. In that case, ϵ_i would sum to zero across market participants and (19) would be obtained for the average price.

The equilibrium number of upstairs market makers can be obtained in a manner analogous to (9) - (10):

$$(20a)\qquad \sqrt{1+t_u} = e^{ac_u},$$

where c_u is the cost of upstairs market making and

$$(20b)\qquad t_u = \frac{a^2\left[Var[\tilde{P}_2 \mid y_u, y_2] + \sigma_u^2\right](1-q)\,\sigma_y^2}{M_u^2}$$

Similarly, the effective variance of the upstairs price change to a customer is, analogous to (13),

$$(21)\quad \sigma^2_{\Delta P_u} \equiv Var(\tilde{P}_{1u} - \tilde{P}_2) + \sigma_u^2 = \frac{a^2(1-q)}{M_u^2}\left[Var[\tilde{P}_2 \mid y_u, y_2] + \sigma_u^2\right]^2\sigma_y^2 + Var[\tilde{P}_2 \mid y_u, y_2] + \sigma_u^2 .$$

Using (20) to eliminate M_u from (21), we obtain

$$(22)\quad \sigma^2_{\Delta P_u} = \left[Var[\tilde{P}_2 \mid y_u, y_2] + \sigma_u^2\right] e^{2ac_u}.$$

An argument exactly like (15) shows that the quality of the customer execution is a monotone decreasing function of $\sigma^2_{\Delta P_u}$. Hence, $\sigma^2_{\Delta P_u}$ is an appropriate measure of the quality of upstairs executions.

Upstairs vs. Downstairs Equilibrium

Comparing (22) and (16) makes the tradeoff between upstairs and downstairs execution evident. Obviously, if $c_d < c_u$, then this creates a benefit to customers from downstairs execution. To ease the comparison, however, assume that $c_d = c_u = c$. In that case, downstairs relative quality will be determined by whether

(23) $$H = e^{-2ac}\left(\sigma^2_{\Delta P_u} - \sigma^2_{\Delta P_d}\right) - Var[\tilde{P}_2 | y_u, y_2] + \sigma^2_u - Var[\tilde{P}_2 | y_d]$$

is positive (i.e., H > 0 implies that downstairs markets are better).

For a fixed f, equation (23) can be used to find an equilibrium q. We define q*(f) as a q with the property that no customer will want to change the market to which it brings all of its business.

It may be of some interest to note that there can be multiple equilibria. For example, if f = 1 and $\sigma^2_u > 0$, so there is no benefit to an upstairs market, there can be an equilibrium where the downstairs market is shut down, i.e., q = 0. This is because, if q = 0 then $Var[\tilde{P}_2 | y_d] > Var[\tilde{P}_2 | y_u, y_2]$, since y_d is totally uninformative and y_u becomes very informative when q = 0. This effect can outweigh the fact that $\sigma^2_u > 0$. Clearly this is a less

satisfactory equilibrium for customers that the one where q = 1, and the upstairs market is closed.

A precise characterization of equilibrium can be obtained by using (1) - (3). In particular, note that

$$H(q) = b^2(Var[\tilde{x}_2|y_u,y_2] - Var[\tilde{x}_2|y_d]) + \sigma_u^2 ;$$

$$H(q) = b^2\sigma_y^2(fq - (f(1-q) + (1-f)) + \sigma_u^2 ; \tag{24}$$

$$H(q) = b^2\sigma_y^2[2fq - 1] + \sigma_u^2$$

An equilibrium q* must satisfy either

$$\text{(a)} \quad H(q^*) = 0 \qquad 0 < q^* < 1 ; \tag{25a}$$

$$\text{(b)} \quad H(q^*) > 0 \qquad q^* = 1 ; \tag{25b}$$

$$\text{(c)} \quad H(q^*) < 0 \qquad q^* = 0. \tag{25c}$$

Note that strict inequality is required in (25b) and (25c) because H(q) is strictly increasing in q. For example, if H(0) = 0, then q* = 0 is not a sensible equilibrium since a small shift of customer business from upstairs to downstairs will make H(q) > 0.

Solving $H(q_I^*) = 0$ for an interior solution yields

$$q_I^* = \frac{1}{2f}\left[1 - \frac{\sigma_u^2}{b^2\sigma_y^2}\right] \tag{26}$$

Note that the case f = 0 is irrelevant since this is the case where no customers present demands in the date 1 market.

More precisely, we can divide equilibrium outcomes into two cases:

Case 1 $\sigma_u^2 - b^2\sigma_y^2 < 0$, which implies that H(0) < 0.

There are two types of equilibria:

1: $q^* = 0$;

2a: $q^* = Min[q_I^*, 1]$;

2b: $q^* = 1$.

Case 2 $\sigma_u^2 - b^2\sigma_y^2 \geq 0$, which implies that H(0) $\geq$ 0.

There is a unique equilibrium:

$q^* = 1$.

The interior equilibrium in Case 1 is not "stable." A small shift from the downstairs to the upstairs market will make H(q) < 0, and drive all customers further to the upstairs market. A small shift toward the downstairs market will drive all customers to the downstairs market.

Note that b^2 is a parameter which captures the fact that order flow affects the date 2 price. If b = 0, then knowledge of order flow is irrelevant in equilibrium, and the upstairs market loses its advantage.

The following points are an immediate consequence of the above characterization of equilibrium..

A. If f = 1 and $\sigma_u^2 > 0$, then a downstairs market is superior, nevertheless, there is an equilibrium where the downstairs market is shut down. The upstairs market makers have no informational

advantage because all possible order flows are expressed at date 1; there is no unexpressed order flow. Nevertheless, the upstairs market survives because all orders are sent there, and the upstairs market makers observe the expressed order flow. All customers would be better off if downstairs market makers received all the order flow.

B. The fundamental tradeoff between upstairs and downstairs markets is that the higher search costs upstairs cause $\sigma_u^2 > 0$, and this must be offset either by superior customer knowledge about unexpressed order flow or about expressed order flow.

4. Intermarket Arbitrage

It may appear that I have ignored the linkage across markets associated with inter-market arbitrage. However, for reasons to be explained next, quite the reverse is true: the model of the previous Sections can be used to understand the extent to which arbitrage can work. First, it is useful to understand the source of arbitrage opportunities.

The classic case of arbitrage is where there are two markets and an arbitrageur can buy for $5 in market 1 and sell at $6 in market 2. It is crucial to realize that a customer in market 1 sold to the arbitrageur at $5, instead of selling in market 2 at $6. Similarly, the customer in market 2 who bought from the arbitrageur at $6, could have bought for $5 in market 1. Thus, arbitrage opportunities can only exist in situations where customers cannot freely choose the markets in which they trade.

Further, arbitrageurs are utterly unnecessary in a world in which customers can costlessly choose where and with whom to trade. Before explaining why some customers cannot freely so choose, we analyze the implications of customers expressing their demands in only one market.

A customer who demands immediacy by sending a large sell order to the downstairs market will cause the downstairs price to fall relative to the upstairs price. Riskless arbitrage will not prevent this fall.[6] Riskless arbitrage is impossible because the downstairs market maker does not know whether the downstairs order flow is (a) purely idiosyncratic to his market, or (b) represents a fall in demand in the upstairs market as well. In case (a) he would buy instantly at the smallest price fall in the downstairs market, and sell upstairs at the best available price. In case (b) this strategy would not be profitable.

The inability to simultaneously trade on both markets is modelled in this paper by the assumption that the demand function instantaneously expressed in each market is a function only of that market's price. Here "demand function" refers to the actual executable bids and offers made by a market maker, rather than the quantity he would buy if he could simultaneously trade on multiple

---

[6] Riskless arbitrage will prevent the fall if (a) the same customer is simultaneously buying the asset in the upstairs market, or (b) the customer's broker can simultaneously find another upstairs buyer. In either case, the broker can <u>simultaneously</u> execute a downstairs buy-order and an upstairs sell-order for its own account, and the textbook case of a riskless arbitrage will take place. I ignore both cases since there is no genuine demand for immediacy on the downstairs market.

markets. Under this assumption, upstairs-downstairs arbitrage adds nothing. The arbitrage is already built into the model by the assumption that there is a common date 2 price across both the upstairs and downstairs markets. In the model of the last Section, a "downstairs" market maker is buying when P_{1d} is low relative to $\tilde{P}_2$, i.e., when his market's price is out of line with his perception of what is the true price to which both markets will converge. The model of the last Section assumes that arbitrage is perfect over the long run represented by the time from date 1 to date 2. The focus of the model is an analysis of the instantaneous consequences of a liquidity event. If a liquidity event expresses itself as the arrival of "sell at market" orders on the floor of an Exchange, then those orders will get executed before the downstairs agent of the upstairs market maker can discover whether this liquidity event is idiosyncratic to the downstairs market or represents information about where both market's prices are heading (i.e., $\tilde{P}_2$).

The instantaneous discrepancies between two markets is fundamentally due to the "at market" orders of customers. Customers who have the ability to trade in only one market will move the price in that market if they have a large demand for immediacy. A market maker facing that order flow does not know whether the flow is idiosyncratic and temporary, or permanent. This limits the ability of market makers to prevent excessive price volatility.

Clearly, customers are both the problem and also the natural arbitrageurs to provide the solution. Since customers initiate the liquidity event which distorts the two markets, they have the information as to whether their demands are temporary or permanent. In principal, at any instant, they can split their demands across both markets so that P_{1d} and P_{1u} are equal. In practice, this is sometimes impossible.

There are a variety of reasons which serve to restrict customers to one market or the other. One important factor which restricts some customers to a downstairs market is the agency problem that the final customer has with his broker. The downstairs market has public prices. The fact that prices of all trades are public makes it easier for the customer to monitor his broker. On the other hand, downstairs markets are burdened with regulations which will increase the relative cost of using these markets to customers. For example, downstairs currency futures markets impose position limits, margin requirements, and (through the Commodities Futures Trading Commission) various disclosure requirements. The "upstairs" currency forward market is totally unregulated. A money manager who obtains a bank credit line is free of many thousands of dollars of legal fees which would have been required to register as a futures trader.

5. The Regulation of Off-Exchange Trading

In the model described in Section 3, there is a strong tendency for extreme outcomes; either the upstairs market survives

or the downstairs. Both markets can survive together only in a very tenuous equilibrium. Some Exchanges have taken extraordinary steps to make this "interior" equilibrium less tenuous. For example, the New York Stock Exchange (NYSE) allows its member to prearrange a trade upstairs, but requires that the trade take place publically on the floor of the NYSE, and that the public be permitted to participate in the trade.

The futures markets do not permit prearranged trades in the futures contracts.[7] Their desire appears to be toward boundary equilibrium (and they hope that $q^* = 1$ is the boundary reached). It is easy to understand their point of view. Suppose that upstairs brokers have no useful information about their customers' unexpressed demands. This may be the case because most large futures customers are professional hedgers or speculators who are constantly participating in the market. This is in contrast to the stock market where only a tiny fraction of the stock being held actually trades on a given day, so that brokers are important repositories of information about unexpressed customer demand. If upstairs brokers have no useful information, then $f = 1$ in the model of Section 3, and there are still cases where the equilibrium involves no downstairs market, (i.e., $q^* = 0$ is an equilibrium when $\sigma_u^2 < b^2\sigma_y^2$). In such a case, the upstairs market is strictly less efficient than the downstairs market, yet it drives the downstairs market out of business.

---

[7] There is one important exception to this statement, relating to the use of "Exchange for Physicals."

6. Conclusions and Extentions

The model of this paper can also be used to understand how equilibrium is maintained in two closely related markets like the S&P500 futures market and the NYSE. These two markets cannot be perfectly arbitraged at each instant in time because a trader on the floor of one market does not know whether the unusual order flow which he faces is common to both markets or special to his own. This lack of information is similar to the lack of information faced by the two sets of market makers in Section 3, where market makers in the "d" market observe only their own order flow, and market makers in the "u" market observe only their own order flow.

It is feasible for someone to stand in the S&P pit and bid for futures whenever futures trade below their theoretical value computed off of the last observed S&P500 cash price. However, the trader in the pit does not know whether the order flow he is observing at that instant will also hit the NYSE thereby changing the appropriate theoretical value, (and more importantly changing the price at which he can sell stock to hedge his position). A symmetrical statement is true about the trader on the NYSE. If he observes a sell order flow but an unchanged "last" futures price, then he can assume that this order flow is idiosyncratic to the NYSE and buy stock planning to sell futures immediately to hedge the position. If this assumption were correct, then stocks would never trade at a discount to futures. The fact that intermarket trades are not executed simultaneously is identical to the fact

that at the instant in which an order flow occurs in one market, it is not observed in the other market. The prices in the two markets get out of equilibrium because of this fact (among other reasons), and the subsequent observation of these disequilibrium prices leads market makers to trade in such a way that equilibrium returns.

Note that upstairs brokers observing order flow simultaneously on both markets are obviously best situated for intermarket arbitrage activity. This is because they may have a presence on both trading floors as well as information about order flows in both markets. They are thus more knowledgeable than the floor traders in the futures market, and the NYSE floor traders (including specialists) about whether an order flow to one market creates an intermarket arbitrage opportunity or is merely the beginning of the order flow to both markets. Hence, regulations designed to restrict NYSE member firms from index arbitrage activity have the potential of eliminating the most effective providers of liquidity to both markets.

REFERENCES

Grossman, S.J. and M.H. Miller, "Liquidity and Market Structure," The Journal of Finance, 43, No. 3 (July 1988), pp. 617-637.