The Informational Role of Upstairs and Downstairs Trading
Author(s): Sanford J. Grossman
Source: *The Journal of Business*, Vol. 65, No. 4 (Oct., 1992), pp. 509-528
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/2353195
Accessed: 04-04-2018 01:28 UTC

Sanford J. Grossman
University of Pennsylvania

The Informational Role of Upstairs and Downstairs Trading*

I assume that customers do not continuously participate in all markets and that intermediaries are repositories of information about the ("unexpressed") demands of currently nonparticipating customers. A model of downstairs (i.e., centralized) versus upstairs (i.e., fragmented) markets is developed based on the assumption that upstairs (and not downstairs) dealers may possess information about unexpressed demand. The equilibrium liquidity of both markets is characterized by the trade-off between the benefits of information about unexpressed demand and the cost to the customer of trading in a fragmented market.

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 article, I will analyze the consequences of the fact that intermediaries play a fundamental role as repositories of information.

The structure of this article is as follows. Section II discusses the reasons for the failure of all individuals to participate continuously in all

* I am grateful to Krishna Ramaswamy and Avraham Beja for helpful discussions and to members of the financial community who would probably prefer to remain anonymous.

(*Journal of Business*, 1992, vol. 65, no. 4)
© 1992 by The University of Chicago. All rights reserved.
0021-9398/92/6504-0002$01.50

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 III 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 that 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 III models equilibrium market making that 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 *expressed* order flows. This information is distinct from an upstairs broker's knowledge of the willingness to trade of customers who have not placed orders, that is, *unexpressed* order flow. The current order flow thus creates an externality that, 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 IV discusses arbitrage between the upstairs and downstairs markets.

Section V 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 VI 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.

The reader may find it useful to contrast my view of the role of upstairs markets with the view that the upstairs market is a place where a trader can certify that he is not informed and hence face a lower effective bid-asked spread (see, e.g., Seppi 1991). It is argued that, since upstairs trading is not anonymous, a person can develop a reputation with a market maker for being an ''informationless'' trader (i.e., someone not selling because he knows that price will fall before the market maker can offset the position acquired from this person). Thus, in this view, upstairs markets screen for informationless traders, and upstairs market makers are repositories of information about honest, uninformed traders. Informed traders and liquidity traders who desire immediacy use the downstairs market. Much of the intuition for this view comes from observing upstairs trading of large blocks of common stock. However, there are many (larger) upstairs markets that are surely not consistent with this view. For example, in the currency markets, the most important informed participants are the central banks and very large speculators, and these entities use the upstairs (interbank) market, rather than downstairs (futures markets). I agree that the absence of anonymity permits reputations to be developed and that this can lower trading costs. However, I think that upstairs brokers serve the additional function of being a repository of information about unexpressed demand.

It should also be noted that a trader can forgo anonymity and trade on an organized exchange. The trader can always use the same floor broker and instruct the floor broker to tell all the other brokers that the order is for the trader. The floor broker can develop a reputation for his customer on the exchange floor. Therefore, though I recognize the importance of reputational effects, I do not believe that it provides the crucial distinction between upstairs and downstairs markets.

II. 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 possibility 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 that give away free options and are thus very costly.
5. As a consequence, most investors' preferences are not continuously represented on organized downstairs markets.

For example, a customer might be willing to sell shares of a particular stock at a price that depends on the price of the Standard and Poor's (S & P) 500, the term structure of interest rates, various macroeconomic announcements, the price of stocks in the same industry, and so on. He could explain all of this to a broker, who would then stand around at the trading post all day waiting for this to be an executable limit order. It is quite costly for the customer to preplan such an order for every contingency, it is costly for the broker to constantly monitor all the factors entering the contingency (in anticipation of an event that may never occur), and it is costly to verify ex post that the order was indeed executed in the correct contingency. The customer could leave a standard limit order, for example, offering to buy 100 shares if the price falls to $5. A standard limit order not only has the problem that it does not express the customer's true demand, but it also gives the market a free option at the customer's expense. The president of the United States could be assassinated, and the price will fall, leaving the customer with an executed limit order in a state of nature for which he would not have purchased the stock at that price. In effect, a resting (uncontingent) limit order to buy stock gives the market a free put option to sell the stock to the customer in the states of nature when it is clearly worth less than his limit price.

Consider the following example of the broker's informational function in the presence of intermarket spreads. Consider an institutional customer who wants to buy 20,000 "in-the-money" calls on a stock that is rumored to be a takeover target. Suppose this is the type of stock that 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. In contrast, 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 four 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."

The above "upstairs" execution of the trade should be compared with a possible scenario for "downstairs" execution. The call buyer could have placed an "at-market" order for the calls on the exchange. The market makers on the floor of the exchange are not aware of the fact that there is an interested seller of the stock, so they raise their offer price to write the desired options (expecting to pay more for the stock that they must purchase to hedge the calls). Indeed, the option market maker will buy the required stock at market, and this will bid up the price of the stock on the stock exchange. The customer who was interested in selling the stock will see the price rise and offer his stock for sale. The overall transaction is effected by prices conveying information to customers, rather than by brokers conveying the information. The two mechanisms give different price paths because all customers did not have resting limit orders, which were known to market makers in all markets. If the seller of stock had a resting limit order on the floor of the stock exchange, and this was known to the market makers on the floor of the options exchange, then the option writers would not have been required to raise their prices to accommodate the option buyer.

Another interesting example of the informational role of brokers arises in considering currency forward and futures markets. The organized (downstairs) futures markets for currencies have remained minuscule relative to the (upstairs) 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 not for forwards.[2] A more important reason concerns the information flow to which interbank market makers are privy. First, note that almost all

[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.

[2] The position limits may exist at their current levels because of the particular current levels of open interest in the currency contracts.

of the hour-to-hour volatility in 1–3-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 businesses engaged in international trade. A U.S. firm that 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. In addition, since central banks carry out currency interventions in the interbank market (often by placing informal limit orders with dealers), interbank dealers will possess information about a very important customer's unexpressed demand.

III. 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:

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

where $\tilde{g}_2$ represents public information about the future payoff stream to the asset that 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 $\tilde{x}_2$, the net order flow aggregated over both period 1 and period 2, and over both upstairs and downstairs markets, affects the date 2 price. Date 2 is the last date of trading, and I am assuming that the equilibrium price is given exogenously by equation (1). The date 1 price is affected by the distribution of orders between upstairs and downstairs markets and by the distribution of demand between "expressed" and "unexpressed" date 1 demand (and will be further explained below). However, date 2 represents the time at which the upstairs and downstairs market prices converge, and the assignment of orders between various markets is irrelevant. In a sense to be described below, the date 1 price

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

is a short-run equilibrium price, while the date 2 price is a long-run equilibrium price.

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 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, I will take the order flow fractions to the upstairs and downstairs markets as exogenous. Specifically, let $\sqrt{f}$ be the fraction of the order flow expressed at date 1, and let $\sqrt{q}$ be the fraction of the date 1 order flow that is expressed in a downstairs market. More precisely, define

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

and

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

where $\tilde{y}_u$, $\tilde{y}_d$, $\tilde{y}_2$ are independent and identically distributed normal random variables with mean zero and variance σ_y^2. Note that $\sqrt{f}$ and $\sqrt{q}$ are designed to reallocate the variance of order flow while maintaining the total variance of order flow constant. In addition, I assume that $\tilde{g}_2$ is normally distributed and independent of $(\tilde{y}_u, \tilde{y}_d, \tilde{y}_2)$.

If the order flow $\tilde{x}_2$ was expressed at date 1, and the upstairs and downstairs markets were unified into one Walrasian auction market, then $\tilde{P}_2$ given in equation (1) would clear the market. The fact that only some of the order flow is expressed at date 1, and that it is distributed between multiple markets, causes the date 1 prices (which will be analyzed below) to temporarily diverge from their "long-run" equilibrium value of $\tilde{P}_2$.

I assume that upstairs market makers observe $\tilde{y}_{\mathrm{u}}$ and downstairs market makers observe $\tilde{y}_d$. In addition, as explained in Section II, 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

In order to focus on the "disequilibrium" between markets, we assume that it is impossible for a market maker to simultaneously trade in an upstairs and downstairs market. Instead, if a market maker is, for example, bidding for yen futures in a downstairs market and someone hits his bid, then it takes some time to sell the yen in the upstairs forward market. Further, the market maker does not know the price he will obtain when he sells the yen. He does not know whether the large "at-market" sell order that he purchased in the futures pit was simultaneously accompanied by such sell orders in the upstairs forward market. We assume that a market maker who acquires a position in one market can expect to offset that position only at the price $\tilde{P}_2$, which is the price at which the distribution of orders between upstairs and downstairs markets is irrelevant.

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

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

where

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

and W_1 is his exogenous initial wealth. Note that, at the instant orders arrive downstairs, the downstairs market makers do not know the extent to which orders are simultaneously arriving on other markets. Thus they condition their beliefs only on observing the realization of $\tilde{y}_d$.

Using the normality assumption, the optimal Z_d is

$$Z_d = \frac{E[\tilde{P}_2|y_d] - P_1}{a \operatorname{var}[\tilde{P}_2|y_d]}. \tag{6}$$

If M_d is the number of downstairs market makers, then date 1, downstairs market clearing, requires that P_1 satisfy

$$M_d\left[\frac{E[\tilde{P}_2|y_d] - P_1}{a \operatorname{var}[\tilde{P}_2|y_d]}\right] = y_d\sqrt{q}, \tag{7}$$

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

$$P_{1d} = E[\tilde{P}_2|y_d] - \frac{a\sqrt{q}}{M_d}\operatorname{var}[\tilde{P}_2|y_d]y_d. \tag{8}$$

As would be anticipated, a large value of customer supply drives down price. Note that an identical equilibrium price would have been computed if I 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' initial wealth is W_0, and it is assumed that entry of market makers occurs to the point where their expected utility of net wealth is unchanged by their decision to enter the market-making business, that is,

$$EU[W_0 - C_d + (\tilde{P} - P_{1d})Z_d] = EU(W_0). \tag{9}$$

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

$$\sqrt{1 + t_d} = e^{acd}, \tag{10}$$

where

$$t_d \equiv a^2 \frac{\text{var}[\tilde{P}_2 | y_d] q}{M_d^2} \sigma_y^2. \tag{11}$$

Equations (10) and (11) 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 both (10) and (11) are 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

$$U_c(W) = -e^{-hW}. \tag{12}$$

Note that

$$EU_c[\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)] = E\Big(E\{U[\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)] | x\}\Big). \tag{13}$$

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

$$\sigma_{\Delta Pd}^2 \equiv \text{var}(\tilde{P}_{1d} - \tilde{P}_2) = \frac{a^2 q}{M_d^2}\{\text{var}[\tilde{P}_2 | y_d]\}^2 \sigma_y^2 + \text{var}[\tilde{P}_2 | y_d]. \tag{14}$$

Thus,

$$E\{U[\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)] | x\} = -\exp\left[h^2 \frac{x^2}{2} \sigma_{\Delta Pd}^2\right]. \tag{15}$$

Using the moment-generating function of the chi-squared distribution,

$$EU[\tilde{x}(\tilde{P}_{1d} - \tilde{P}_2)] = -[1 - h^2\sigma^2_{\Delta P_d}\sigma^2_x]^{-1/2}. \tag{16}$$

I assume that σ^2_x is sufficiently small that (16) is well defined and finite.

It follows from (16) 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 (11) may be used to eliminate the endogenous M_d from (14) and thus to obtain

$$\sigma^2_{\Delta P_d} = \text{var}[\tilde{P}_2 | y_d]e^{2ac_d}. \tag{17}$$

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 that (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 that is a random perturbation from the average price prevailing at a particular point in time.[5] I denote this extra volatility in realized price by σ^2_u. It is straightforward to verify that σ^2_u causes (6) to be replaced by

$$Z_u = \frac{E[\tilde{P}_2 | y_u, y_2] - P_1}{a\{\text{var}[\tilde{P}_2 | y_u, y_2] + \sigma^2_{\text{u}}\}}, \tag{18}$$

[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 modeled by modeling 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 customers is at a disadvantage because his 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.

where Z_u is the market maker's 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

$$M_u Z_u = y_u \sqrt{1 - q}, \tag{19}$$

which generates an equilibrium price P_{1u}:

$$P_{1u} = E[\tilde{P}_2 | y_u, y_2] - \frac{a\sqrt{1 - q}}{M_u} \{\text{var}[\tilde{P}_2 | y_u, y_2] + \sigma_u^2\} y_u. \tag{20}$$

Note that we could have replaced P_1 in (18) by $P_{1u} + \epsilon_i$ to represent the idea that a particular market maker trades at a price that randomly deviates from the average upstairs price P_{1u}. In that case, ϵ_i would sum to zero across market participants and (20) would be obtained for the average price.

The equilibrium number of upstairs market makers can be obtained in a manner analogous to equations (9)–(11):

$$\sqrt{1 + t_u} = e^{ac_u}, \tag{21}$$

where c_u is the cost of upstairs market making and

$$t_u \equiv \frac{a^2\{\text{var}[\tilde{P}_2 | y_u, y_2] + \sigma_u^2\}(1 - q)\sigma_y^2}{M_u^2}. \tag{22}$$

Similarly, the effective variance of the upstairs price change to a customer is, analogous to (14),

$$\begin{aligned} \sigma_{\Delta P u}^2 &\equiv \text{var}(\tilde{P}_{1u} - \tilde{P}_2) + \sigma_u^2 \\ &= \frac{a^2(1 - q)}{M_u^2} \{\text{var}[\tilde{P}_2 | y_u, y_2] + \sigma_u^2\}^2 \sigma_y^2 \\ &\quad + \text{var}[\tilde{P}_2 | y_u, y_2] + \sigma_u^2. \end{aligned} \tag{23}$$

Using (21) and (22) to eliminate M_u from (23), we obtain

$$\sigma_{\Delta P u}^2 = \{\text{var}[\tilde{P}_2 | y_u, y_2] + \sigma_u^2\} e^{2ac_u}. \tag{24}$$

An argument exactly like (16) shows that the quality of the customer execution is a monotone decreasing function of $\sigma_{\Delta P u}^2$. Hence, $\sigma_{\Delta P d}^2$ is an appropriate measure of the quality of upstairs executions.

Upstairs versus Downstairs Equilibrium

Comparing (24) and (17) makes the trade-off 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

$$\begin{aligned} H &\equiv e^{-2ac}\,(\sigma^2_{\Delta P_u} - \sigma^2_{\Delta Pd}) \\ &= \mathrm{var}[\tilde{P}_2 | y_u, y_2] + \sigma^2_u - \mathrm{var}[\tilde{P}_2 | y_d] \end{aligned} \tag{25}$$

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

For a fixed f, equation (25) can be used to find an equilibrium q. I 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, that is, $q = 0$. This is because, if $q = 0$, then $\mathrm{var}[\tilde{P}_2 | y_d] > \mathrm{var}[\tilde{P}_2 | y_u, y_2]$ 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 than 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

$$\left.\begin{aligned} H(q) &= b^2(\mathrm{var}[\tilde{x}_2 | y_u, y_2] - \mathrm{var}[\tilde{x}_2 | y_d]) + \sigma^2_u, \\ H(q) &= b^2\sigma^2_y\{fq - [f(1-q) + (1-f)]\} + \sigma^2_u, \\ &\text{and} \\ H(q) &= b^2\sigma^2_y[2fq - 1] + \sigma^2_u. \end{aligned}\right\} \tag{26}$$

An equilibrium q^* must satisfy either

$$a)\; H(q^*) = 0, \quad 0 < q^* < 1; \tag{27}$$

$$b)\; H(q^*) > 0, \quad q^* = 1; \tag{28}$$

or

$$c)\; H(q^*) < 0, \quad q^* = 0. \tag{29}$$

Note that strict inequality is required in (28) and (29) 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 $H(q^*_I)$ for an interior solution yields

$$q^*_I = \frac{1}{2f}\left[1 - \frac{\sigma^2_u}{b^2\sigma^2_y}\right]. \tag{30}$$

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:

$$1\colon q^* = 0,$$

$$2\text{a}\colon q^* = \min[q^*_I, 1],$$

and

$$2\text{b}\colon q = 1.$$

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

The conclusion that the interior equilibrium is unstable can be reversed if we modify the model to make σ_u^2 dependent on q (representing this dependence by a function $\sigma_u^2(q)$).[6] If upstairs search costs are reduced when a greater proportion of customers use the downstairs market, then σ_u^2 will be a declining function of q.[7] If the derivative of $\sigma_u^2(q)$ with respect to q is very negative, then $H(q)$ will be a declining

[6] I am grateful to Avraham Beja for pointing this out to me.

[7] Of course, it could well be that greater usage of upstairs markets reduces upstairs search costs because of increasing returns to scale in producing electronic display screens and other communications technology used to facilitate search.

function, instead of an increasing function of q. With $H'(q) < 0$, if $\sigma_u^2(0) - b^2\sigma_y^2 < 0$, as in case 1, the only equilibrium will be $q = 0$. If $\sigma_u^2(0) - b^2\sigma_y^2 \geq 0$, and $H'(q) < 0$, then there will be a unique equilibrium described as follows: If $H(1) \geq 0$, then $q^* = 1$; if $H(1) < 0$, then there is an interior equilibrium q^* such that $H(q^*) = 0$.

My model should be contrasted with Pagano (1989), who presents an explicit model of search equilibrium. He uses a search technology to prove that a market will be more liquid and prices less volatile when more traders decide to use it. He reaches the conclusion that the coexistence of multiple markets is unstable, unless there are differential costs associated with trading in the markets. I have focused on the case where the direct costs of trade are the same across upstairs and downstairs markets, but there can be differential benefits associated with upstairs dealers possessing information about unexpressed demands.

IV. Intermarket Arbitrage

It may appear that I have ignored the linkage across markets associated with intermarket 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 exist only 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, I analyze the implications of customers expressing their demands in only one market.

A customer who demands immediacy by sending a large sell (''at-market'') order to the downstairs market will cause the downstairs price to fall relative to the upstairs price. Riskless arbitrage will not prevent this fall. 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 downstairs market maker cannot make his downstairs bid a function of the executable upstairs bid at each instant in time; he cannot be certain that, if his bid is hit, he can turn around and sell at the last upstairs bid. The same event that caused his downstairs bid to be hit may cause the upstairs bids to be hit and replaced by lower bids. The inability to simultaneously trade on both markets is modeled in this article 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 markets. Under this assumption, upstairs-downstairs 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 $\tilde{P}_2$, that is, 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 instantaneous discrepancies between two markets are 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. In the notation of Section III, the downstairs market maker facing a large y_d does not know y_u (i.e., the extent to which the current downstairs shock is idiosyncratic) or y_2 (the extent to which the downstairs shock is temporary and will be offset by unexpressed demand at date 2). This limits the ability of market makers to prevent intermarket price volatility.[8]

[8] The model of this article can also be used to understand how equilibrium is maintained in two closely related markets like the S & P 500 futures market and the New York Stock Exchange (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 that 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 III, 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 & P 500 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 important, 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

V. Additional Considerations in a Customer's Choice of Markets

Customers are both the source of intermarket arbitrage opportunities and also the natural arbitrageurs to provide the solution. Since customers initiate the liquidity event that 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 that serve to restrict customers to one market or the other. One important factor that 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 that increase the relative cost of using these markets to customers. For example, downstairs currency futures markets impose position limits and (through the Commodites Futures Trading Commission) various disclosure and reporting requirements along with the potential legal liability of improperly complying with vague regulations. The "upstairs" currency forward market is totally unregulated.

Another feature of upstairs markets that distiguishes them from downstairs markets is the larger extent to which contracts are tailored to individual customers in the upstairs market. For example, in the (downstairs) currency futures markets, each foreign currency contract traded is for delivery against U.S. dollars, at an exchange-specified delivery date, and for an exchange-specified quantity of foreign currency. The (upstairs) interbank market offers a customer any delivery date for any amount and any denomination of foreign currency. A commercial U.S. customer delivering computer equipment for sterling on a particular date will often prefer to sell the sterling forward for U.S. dollars to settle on the same date as his anticipated receipt of sterling. The use of a futures contract with a different settlement date would have caused the customer to face some basis risk. In the nota-

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.

tion of this article, the reduction of basis risk will have the effect of reducing σ_u^2 for some customers. A corporate treasurer may thus never receive authorization to use futures rather than the interbank market for currency hedging. His superiors may feel that the treasurer will be in the position to ''speculate'' in the futures market, rather than engaging ''hedging'' by an exact offset of exchange rate risk in the interbank market.

An additional distinction between upstairs and downstairs markets involves the fact that the downstairs exchange guarantees a downstairs customer that the opposite side of his trade will be fulfilled, while, in contrast, an upstairs customer accepts the credit risk of a dealer, in that the dealer will fulfill its side of the trade. Thus, the relative credit worthiness of the exchange versus the dealers will affect the customer's relative risk of exchange versus upstairs trading.

VI. Summary and Conclusions

It is too costly for some investors to continuously participate in all markets. These investors have state- and price-contingent demands for trades that are not expressed continuously in the marketplace. Upstairs brokers and dealers are repositories of information about the unexpressed demand of investors. This information will increase the effective liquidity of an upstairs market since an upstairs dealer faces less risk when making a bid to a particular customer if the dealer knows of the unexpressed demand of another customer. This positive aspect of the upstairs market can be offset by the negative aspect that an upstairs market is fragmented, so that a customer (*a*) must search for the best price, (*b*) has no assurance that trades do not take place below his bid, and (*c*) is at an informational disadvantage relative to the dealer since the dealer sees the order flow. In a downstairs market, the downstairs market maker has less information about unexpressed demand, and this may decrease liquidity there. However, on the positive side, a customer (*a*) does not have to search for the best price and (*b*) is assured that trades do not take place below his bid unless he participates in the trade, and (*c*) the order flow to the floor of the exchange is visible to the customer's broker, who, acting as the customer's agent, reveals that information to the customer.

I have defined and characterized the equilibrium fraction of customers that choose upstairs instead of downstairs markets based on the above considerations. Downstairs markets will tend to prevail when knowledge of unexpressed demand is relatively unimportant. However, there is a strong externality that can cause downstairs market failure even when it is the more efficient trading environment. This externality is associated with the informational advantage of observing the order flow. If most of the order flow goes upstairs, then downstairs market makers will be relatively uninformed, not only about the unexpressed order flow, but also about the expressed order flow. As a consequence, the upstairs market may prevail even when there are upstairs search costs and no information about unexpressed demands, simply because more trade is taking place there. This externality also makes unstable an "interior" equilibrium where both upstairs and downstairs markets coexist.

Some exchanges have taken extraordinary steps to make this "interior" equilibrium less tenuous. For example, the NYSE allows its member to prearrange a trade upstairs but requires that the trade take place publicly on the floor of the NYSE and that the public be permitted to participate in the trade.

The futures exchanges do not permit prearranged trades in the futures contracts.[9] 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 III, 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.

It is interesting to note that the currency futures markets have not succeeded in displacing the interbank currency forward market as the major market for trading currencies. I suspect that this is because banks are repositories of information about their corporate clients' foreign currency demands, and banks can tailor the contract specifications to each customer's needs. The currency futures markets were essentially the first financial futures markets in the United States. They were set up with some of the restrictions of commodity futures markets including position limits and daily price limits. Since the interbank market continues to trade when there is a limit move on the futures market, futures customers are at a disadvantage relative to interbank customers. The position limits were set to be so restrictive that many large funds could make only very limited use of the futures markets,

[9] There is one important exception to this statement, relating to the use of "exchange for physicals." Further, some futures exchanges have recently relaxed prohibition against some forms of prearranged trading.

and at times arbitrage between the futures and the interbank markets was inhibited by these limits.[10] This lack of success of the currency futures market should be contrasted with the enormous success of the futures markets in long bonds and Eurodollar deposits.

The coexistence of upstairs and downstairs markets in currency and fixed-income markets may appear to contradict my conclusion that such coexistence is unstable. However, I have already noted that the uniformity of exchange-traded instruments may be inadequate for customers who have very specific needs relating to contract size and maturity. The futures markets have served the "generalized" risk transfer needs of customers rather than the specific investments needs of customers. Customers use the downstairs bond futures markets to transfer general long bond risk, while they use the upstairs bond dealer market to purchase particular coupons and maturities of government bonds. Similarly, the very successful Eurodollar futures market serves as a place where generalized short-term dollar interest rate risk is transferred, while the upstairs bank Eurodollar market serves as a place where customers can trade the credit risk of particular banks and financial institutions. A similar coexistence exists for equities, where the equity futures markets have successfully displaced the stock exchanges as the place where general market risk is traded. However, the upstairs equity market has displaced the downstairs stock exchanges as the place where institutional investors arrange to transfer the risk of particular equities.

The relative success of downstairs markets for trading generalized rather than specific instruments can be understood using the model of this article and the hypothesis that brokers have relatively more information about the unexpressed demand for particular instruments than they have information about the unexpressed demand for generalized instruments. Investors tend to express relatively more of their demand for generalized instruments than for particular instruments. If an investor wants to find a particular coupon and maturity of a government bond, he will have to find a holder of that particular instrument. It is hard to imagine a market where holders of infrequently traded instruments constantly express their limit orders and thus negate the usefulness of an upstairs broker's knowledge about their willingness to part with the instrument. However, investors are more frequently rebalancing their generalized positions. It is almost a tautology that investors will rebalance their generalized positions more frequently than they rebalance their holdings of any specific instrument. Investors, therefore, express relatively more of their generalized than their

[10] The currency futures markets have been gradually eliminating position limits and daily price limits. At present, there is a price limit only at the opening of trade, and position limits have just been eliminated for some currency futures contracts.

specific demands, and, thus, according to the model of this article, downstairs markets will be relatively more successful for generalized instruments.

References

Grossman, S. J., and Miller, M. H. 1988. Liquidity and market structure. *Journal of Finance* 43 (July): 617–37.

Pagano, M. 1989. Trading volume and asset liquidity. *Quarterly Journal of Economics* 104 (May): 255–74.

Seppi, D. J. 1991. Equilibrium block trading and asymmetric information. *Journal of Finance* 45 (March): 73–94.