BID, ASK AND TRANSACTION PRICES IN A SPECIALIST MARKET WITH HETEROGENEOUSLY INFORMED TRADERS*

Lawrence R. GLOSTEN

Northwestern University, and University of Chicago, Chicago, IL 60637, USA

Paul R. MILGROM

Yale University, New Haven, CT 06520, USA

Received August 1983, final version received September 1984

The presence of traders with superior information leads to a positive bid–ask spread even when the specialist is risk-neutral and makes zero expected profits. The resulting transaction prices convey information, and the expectation of the average spread squared times volume is bounded by a number that is independent of insider activity. The serial correlation of transaction price differences is a function of the proportion of the spread due to adverse selection. A bid–ask spread implies a divergence between observed returns and realizable returns. Observed returns are approximately realizable returns plus what the uninformed anticipate losing to the insiders.

1. Introduction

The usual economic view of markets is as a place where buyers and sellers come together and trade at a common price, the price at which supply equals demand. Securities exchanges are often singled out as excellent examples of markets that operate this way. In fact, however, trading on exchanges takes place over time, and some institutional arrangements are necessary to help match buyers and sellers whose orders arrive at different times. On exchanges like the New York Stock Exchange, the economic function of the specialists and the floor traders is that of middlemen: they hold the inventories that facilitate trade when trading occurs over time.

The problem of matching buyers with sellers is most acute in trading shares of small companies, where the volume of trade is relatively low. A common problem in this environment involves the number of insiders who trade in the shares relative to the total trading volume. Classical price theory, which has

*Milgrom's research was supported by National Science Foundation Grant IST-8208600. We wish to thank Rolf Banz, Dan Galai, Jay Ritter, Ravi Jagannathan, the participants in the University of Chicago Finance Workshop, the University of Michigan Finance Workshop and the participants in the University of Southern California conference on transaction prices for helpful comments. We also gratefully acknowledge comments by the editor, Clifford W. Smith, and two referees.

Published in: ***Journal of Financial Economics***

little to say about the dynamics of matching buyers and sellers, offers few hints about the consequences of insider trading. It leaves unanswered such questions as: How completely do prices reflect insider information [Fama (1970)]? How large are insider profits? How does the specialist behave in this environment? How might the existence of insider trading alter the return characteristics of the stock?

A number of researchers have examined the optimal behavior of a specialist and how it leads to a bid–ask spread. The usual approach examines the management of inventory by a monopolist specialist, concentrating on the effect that inventory costs have on the bid–ask spread; e.g. Ohara and Oldfield (1982), Ho and Stoll (1981), Amihud and Mendelson (1980) and Garman (1976).

The approach taken in this paper is based on the idea that a bid–ask spread can be a purely informational phenomenon, occurring even when all the specialist's fixed and variable transactions costs (including his time, inventory costs, etc.) are zero and when competition forces the specialist's profit to zero. The core idea is that the specialist faces an adverse selection problem, since a customer agreeing to trade at the specialist's ask or bid price may be trading because he knows something that the specialist does not. In effect, then, the specialist must recoup the losses suffered in trades with the well informed by gains in trades with liquidity traders. These gains are achieved by setting a spread. This informational source of the spread has also been suggested by Bagehot (1971) and formally analyzed by Copeland and Galai (1983).

In this paper, we use a formal model to show how the spread arises from adverse selection. We analyze the determinants of the magnitude of the spread and investigate the informational properties of the transaction prices. We then show that, depending on how returns are measured, the information-based spread may cause realizable returns to be overestimated relative to the returns that are actually available to a trader without inside information.

Some of our results are restatements and generalizations, within the context our model, of results presented in Copeland and Galai (1983). The major ›al of this paper, however, is examination of some of the dynamic properties ıhe spread and transactions prices, with particular references to the question how specialist markets process privately available information. Whereas opeland and Galai assume that private information is revealed immediately fter each trade, we allow there to be further trading until such time as ıformation is revealed resolving the informational differences between insiders nd the rest of the market. In this way, we can explicitly deal with questions elated to the information contained in prices, the behavior of transaction ›rices and how the spread in turn responds to this market generated and other ›ublic information.

Our model posits a risk-neutral competitive specialist who faces no transac-ions costs (fixed or variable), that is, a specialist whose expected profit from ;ach transaction is zero. We do not specify why the specialist should be

competitive. It may be that he must compete for the right to conduct transactions with the floor traders or with another specialist in the same stock at another exchange. If any explicit model of Bertrand-style price competition among the market makers is to account for the zero profit condition, then one must assume that no single market maker can, by refusing to trade, ever cause a shortage of cash or securities to occur. The analogous assumption here is that the specialist has unlimited inventories of cash and securities with which to transact and the holding cost of these inventories is zero. All of this is subsumed in the assumption that the bid and ask prices at each trade are set to yield zero profits to the specialist.[1]

The model described and presented in section 2 is structured to emphasize the short-term price effects that may occur just after an event that gives an informational advantage to insiders. There is no discounting over this short period, and much of our focus is on how information is assimilated by the market, as reflected by the changing spread. In this model, prices exhibit a semi-strong form of efficiency: indeed, they may reflect slightly more information than was available to the specialist at the time he set the bid and ask prices. The explanation for this seemingly strange conclusion lies in the observation that the specialist does not set a single price. The ask price, for example, specifies what the price will be if the next customer is a buyer. Consequently, the ask price can be (and at equilibrium is) set using both the current information and the information that will be inferred if the next customer turns out to be a buyer.

There are five propositions established in section 2. We review them here in terms of the interpretations we intend, but the reader should be aware that the underlying assumptions vary slightly across the propositions, though the assumptions are mutually consistent.

The first of the five propositions simply asserts that the bid and ask prices straddle the price that would prevail if all traders had the same information as the specialist. This does not call for much comment; it is simply a mathematical affirmation of the logic of adverse selection.

The second proposition establishes that the prices at which transactions actually occur form a martingale. This result contradicts the idea that the negative serial correlation observed in microdata is a necessary consequence of the existence of spreads and the vibration of transactions between the bid and ask prices. Negative serial correlation does arise from spreads that cover the specialist's costs or that generate expected profits. In fact, the serial correlation coefficient of price changes can be used to determine the relative magnitudes of

[1] We shall see that this assumption results in the specialist sometimes accumulating large inventories of stock and sometimes large inventories of cash. It follows that if we were to recognize binding inventory constraints, we could not have a zero profit condition. Nevertheless, if the specialists' and floor traders' inventory carrying capacity were large enough, we anticipate that the expected profit from any given trade in a Bertrand model would be 'near to zero most of the time', so that our results are approximations of the results that would obtain.

these two sources of the spread – adverse selection and specialist costs or profits.

Proposition 3 gives a bound on the size of the spread that can arise from adverse selection. Specifically, the expected value of the squared average spread times volume has a uniform bound (independent of the pattern of trade) that is related to the variance of the underlying uncertainty. The proposition is proved using the observations that the variance of the price at each trade is roughly proportional to the squared spread, and the total variance of prices from trade to trade is bounded by the variance of the underlying value of the security. This proposition is less specific than one might hope, since it gives only a bound, rather than a precise order of magnitude, for the mean spread. As the example in section 3 makes clear, the mean spread depends on how the informed arrivals are distributed through the trading period, so one cannot make any headway on computing the mean spread from adverse selection without making strong assumptions about the unobservable arrival pattern of insiders.

The fourth proposition is that, over time, the value expectations of the specialist and the insiders tend to converge. This is our way of showing that insider information tends to be fully disseminated into the market prices.

The last proposition in section 2 investigates how the spread at a given trading date and with a given trading history responds to variations in the parameters of the model. The results accord well with what one might expect. Generally, ask prices increase and bid prices decrease if the insiders' information becomes better, or the insiders become more numerous relative to liquidity traders, or the elasticity of the expected supply and demand of a liquidity trader increases.

One of the interesting features of our model is that there can be occasions on which the market shuts down. Indeed, if the insiders' are too numerous or their information is too good relative to the elasticity of liquidity traders' supplies and demands, there will be no bid and ask prices at which trading can occur and the specialist can break even. Then, the equilibrium bid price is set so low and the ask price so high as to preclude any trade.[2] A situation like this feeds on itself. Insiders have information that results in a wide spread that precludes trade that prevents insiders from revealing their information through their trading behavior. Since trade itself brings information into the market, this shutting down of the market may worsen subsequent adverse selection problems and cause the next bid price to be lower and the next ask price to be higher than would otherwise be the case. Furthermore, a market, once closed, will stay closed until the insiders go away or their information is at least partly disseminated to market participants from some other information source.

Thus, the failure of a customer to trade may work an externality on future traders which is not accounted for by either the specialist or the current trader:

[2] We argue later that this cannot happen, however, if liquidity supply and demand are inelastic.

it deprives them of potentially valuable information. This opens the possibility that another way of arranging trade may exist which is Pareto superior to the competitive specialist system. The welfare loss we have described is at least partly due to the requirement that the specialist must break even on each trade. Everyone could be made beter off in our example if the specialist were required to make losses on some trades and permitted to recoup the losses on other trades. This might be accomplished in many ways, for example by allowing the specialist to have some monopoly power while requiring him to keep the difference between the bid and ask prices within some range. That similar restrictions are placed on market makers at some exchanges may not be entirely coincidental.

In section 3, we present an example to illustrate the model of section 2. We focus especially on how the proportion of insiders affects trading when enough trading occurs to allow essentially all inside information to be assimilated in the price. We also illustrate how an excess of insider trading can lead to a partial market breakdown.

Finally, in section 4, we examine a variation of the market model in which the uninformed traders expect a fixed positive return. The existence of a spread implies that, over any period, measured returns based on transactions prices tend to overstate the actual returns realizable by a liquidity trader. The reason for this is that the initial price, which may be either a bid or ask price, is on average lower than the ask price which the liquidity trader must pay. Similarly, the final price over the relevant period is on average higher than the bid price which the liquidity trader could expect to receive. We show that the measured return over a 'normal holding period' is approximately a normal return plus a return available only to insiders.

Thus, the bid–ask spread arising from adverse selection provides a possible route to explaining the small firm effect [Banz (1981)] and the ignored firm effect [Arbel and Strebel (1981)]. The evidence suggests that much of the excess returns on small firm investments occur in January. If the annual report for small firms tend to contain considerable new information and if insiders have early access to that information, then our analysis would predict that spreads are especially large in the period before the report is made public and presumably after the end of the firm's fiscal year. This combined with a 'required rate of return' assumption will lead to large measured returns but normal realizable returns. These returns should be observed in January for firms whose fiscal years coincide with the calendar year but in other months for firms with other fiscal years.[3]

[3] Blume and Stambaugh (1983) also show that the spread can lead to an upward bias in the measurement of expected returns due to the mismeasurement of 'true' prices. Our model may not be providing the whole story, since as Keim (1981) has noted, much of the excess January return occurs in the first few days of January – before insiders might be presumed to have accurate information. Also, Blume and Stambaugh do not measure a larger bias for January returns. This is consistent with a stationary spread, but is not necessarily inconsistent with our comments as Proposition 2 suggests.

We offer some concluding remarks and indicate possible directions for additional research in section 5.

2. The basic model

The market that we are modelling is a pure dealership market, i.e., the specialist performs no brokerage services, and in effect all orders are market orders. Trade occurs according to the following sequence of events. The specialist sets a bid and ask price with the interpretation that he is willing to sell one unit of stock at the ask and buy one unit of stock at the bid. An investor arrives at the market and is informed of the bid and ask at which time he is free to buy one unit at the ask or sell one unit at the bid or leave. The specialist is free to (and in general will) change the bid and ask at any time after an arriving investor has made a decision and before the next arrival of an investor. That is, if an arriving order leads to a trade, the trade takes place at the quoted bid or ask. After the trade, the specialist may revise the bid and ask.

The primary differences between the above description and the operation of the stock market are the limitation on the types of orders considered and the restriction to unit trades. In practice, an investor can submit a limit order and, by so doing, in effect compete directly with the specialist. To the extent that limit orders represent competition with the specialist, our assumption that the bid and ask are set competitively implicitly includes the possibility that investors can submit limit orders. However, there is a difference between limit orders and the quotes of the specialist in that limit orders typically have a prespecified lifetime (if unexecuted), whereas the specialist can change his bid and ask relatively freely. This fact implies that inclusion of limit orders may well alter the characteristics of transaction prices. We have not included limit orders because such a model should include investors optimally choosing the type of order to submit, and such detailed description of individual behavior is beyond the scope of this paper.

The assumption that only unit trades take place is restrictive, but it yields a corresponding benefit: It allows us to analyze a model that places no restrictions on the form of the traders' information. Of course, on the NYSE specialist quotes are valid for a specified number of units, and typically, the quotes are for only one unit (typically 100 shares). What we rule out is the specialist revising his quote for some other requested quantity. However, the fact that we put relatively few restrictions on the arrival process of traders makes this assumption somewhat less objectionable.

To examine the informational characteristics of such a market, we assume that there are informed investors and purely 'liquidity' traders. At some time T_0 in the future, some random dollar value V $[V \geq 0, \text{var}(V) < \infty]$ per share will be realized, and the informed have information about this random variable V. Time T_0 may be interpreted as the time at which no trader has an

informational advantage – just after an earnings announcement, for example. At that time, there will be agreement on the value of the firm and the informational differences between insiders and outsiders will be minimal. The informed may receive information about occurrences after time T_0, but this information will be public information at time T_0. This specification implies that there will be no informational asymmetries at time T_0, and hence it is meaningful to specify, exogeneously, the random value, V, which represents the consensus value of the stock given all public information.[4]

The informed receive information and place their orders. We do not rule out the possibility that any one informed investor may decide to submit several orders, each for a unit amount. The informed trader may be speculating based on private information or superior analysis, or he may simply have a 'liquidity' reason for trading, but in any event, his decision to buy, sell or leave is based on his information. We will refer to the informed traders as insiders though other interpretations are possible, for example, they may merely be individuals who are particularly skillful in processing public information.

To motivate the active participation of the uninformed traders in a model where everyone is rational, there must be some disparity of preferences or endownments across individuals. This disparity may arise from predictable life cycle needs or from less predictable events such as job promotions or unemployment, deaths or disabilities, or myriad other causes. We have chosen to suppress the details of the uninformed traders' motives from our formal model. Instead, we simply assign to each a time preference parameter which, together with his expectations, determines how much a trader is willing to pay to buy and to accept to sell a single unit of the stock. Specifically, all participants, informed, uninformed and the specialist, are risk-neutral. Each participant assigns random utility to shares of stock, x, and current consumption, c, as $\rho x V + c$, where ρ is a parameter of the individual investor's utility function representing his personal trade-off between current and future consumption derived from ownership of the asset. For the specialist, we take $\rho = 1$; this is just a normalization. Generally, a high ρ indicates a desire to invest for the future; a low ρ indicates a desire for current consumption. This 'liquidity parameter' could be the result of imperfect access to capital markets or it could represent differential subjective assessments of the distribution of the random variable V. The risk neutrality assumption implies that in order for there to be trade, there must be some variation in ρ across market participants, for otherwise the 'no trade theorem' of Milgrom and Stokey (1982) implies that the spread will be set large enough to preclude all trade. Since ρ is to be unknown to the specialist, and a pure preference parameter, we treat it as a

[4] This formulation is the usual one in the rational expectations models of a single market [for example, Grossman (1976)] and the proofs of our results use this formulation. A general equilibrium model would involve modeling the cash flows from the security and the consumption decisions of investors over an infinite horizon.

random variable independent of V and any information about V and independent across traders. We allow the possibility that ρ might follow a different distribution for the informed and the uninformed.

We assume that investors arrive one by one, randomly and anonymously at the specialist's post. For most of our analysis, the only restriction that we place on the arrival process is that there be only one arrival at any instant. Thus, the arrival process may depend on the history of trade. For example, the level of insider activity, and hence the nature of the arrival process might be a function of how much information has been made public relative to the information known by the insiders. All we require is that the specialist knows the probabilistic structure of the arrival process. For example, in response to private signals, there may be a bunching of insider orders (as seen by an omniscient observer). Our assumption that the specialist knows the probabilistic structure implies that he makes correct statistical inferences from observed data.

Investors, upon arriving at the market and hearing the bid and ask, maximize expected utility given their information to date. For uninformed investors, this information consists of all past transaction prices, the current bid and ask as well as any publicly available information. The informed also have access to the previous transaction price sequence, the current bid and ask, and all public announcements, but in addition they have been able to see some private signal. Formally, let H_t denote the information available publicly up to clock time t. If an uninformed investor arrives at time t, then his information, upon arrival, is H_t joined with the information generated by the quoted bid and ask. If an arrival at time t is informed, then his information includes both his private information, J_t, and the public information, H_t, and the information generated by the quoted bid and ask.[5] By including the specification of who is informed in the sample space, we can generally represent the information of an arrival at time t by F_t, a refinement of H_t which includes the information conveyed by the quoted bid and ask.

Putting the utility functions and information structures together, the optimal decision of an investor arriving at time t, given bid B and ask A is given by

$$\text{buy} \quad \text{if} \quad Z_t > A,$$

$$\text{sell} \quad \text{if} \quad Z_t < B,$$

where Z_t is given by

$$Z_t = \rho_t \mathrm{E}[V|F_t] = \rho_t(1 - U_t)\mathrm{E}[V|H_t, J_t, A, B] + \rho_t U_t \mathrm{E}[V|H_t, A, B], \tag{1}$$

where U_t is one if the individual arriving at t is uninformed and zero otherwise.

[5] Note that since only one individual arrives at any time, we can identify an individual by specifying his time of arrival.

Given the above behavior of the market participants, the specialist chooses bid and ask prices. Let the information available to the specialist at time t be represented by S_t. Assuming anonymity, the specialist cannot know when the bid and ask prices are set whether the next customer will be an insider or an outsider. Given the investors' behavior, the information available to the specialist at time t, S_t, and bid and ask prices B and A, the specialist's expected profit from an arrival at time t is

$$\mathrm{E}\left[(A-V)I_{\{Z_t>A\}}+(V-B)I_{\{Z_t<B\}}|S_t\right], \tag{2}$$

where $I_{\{Z_t>A\}}$ and $I_{\{Z_t<B\}}$ are, respectively, the indicator functions of the events $\{Z_t>A\}$ and $\{Z_t<B\}$, i.e., $I_{\{Z_t>A\}}$ is one if the event $\{Z_t>A\}$ occurs; otherwise it is zero. The expression, (2), may be rewritten as

$$\begin{aligned}&(A-\mathrm{E}[V|S_t, Z_t>A])\mathrm{P}\{Z_t>A|S_t\}\\&-(B-\mathrm{E}[V|S_t, Z_t<B])\mathrm{P}\{Z_t<B|S_t\},\end{aligned} \tag{3}$$

where $\mathrm{E}[\cdot|\cdot]$ is the (conditional) expectation operator derived from the probability measure P.

The above holds as long as there are zero costs associated with all short positions in cash or stock. Our central assumption about the specialist is that he earns zero expected profits on each purchase and each sale, and he faces no transaction costs. To illustrate how competition might lead to such a description, suppose there are two specialists in this one stock. Both have the same information and face the same population. Suppose the first specialist sets an ask price A^1 so that $A^1 > \mathrm{E}[V|S_t, Z_t > A^1]$. The second specialist will rationally undercut the first by choosing an ask $A^2 < A^1$ and $A^2 \geq \mathrm{E}[V|S_t, Z_t > A^2]$. The zero expected profit equilibrium at time t (if it exists) consists of a pair of functions A_t and B_t satisfying

$$\begin{aligned}A_t(\omega)&=\mathrm{E}\left[V|S_t, Z_t>A_t(\omega)\right](\omega),\\B_t(\omega)&=\mathrm{E}\left[V|S_t, Z_t<B_t(\omega)\right](\omega),\end{aligned} \tag{4}$$

where $Z_t = \rho_t \mathrm{E}[V|F_t]$ and A_t and B_t are measurable with respect to F_t (i.e., the customer knows the bid and ask prices).

General existence of such functions would be difficult to show, since it involves a 'rational expectations' type of fixed point condition. The definition is not vacuous, however, as the following examples show. If the specialist's information, S_t, is a finer partition than the information of the informed, then A_t and B_t will both be equal to the conditional mean of V given the information S_t. If, on the other hand, the specialist's information is the same as

the publicly available information H_t, then A_t and B_t are given by

$$A_t = \inf\{a: a \geq \mathrm{E}[V|H_t, Z_t > a]\},$$
$$B_t = \sup\{b: b < \mathrm{E}[V|H_t, Z_t < b]\}. \tag{4'}$$

Our notion of equilibrium requires that the specialist not regret, ex post, any trade that he is obliged to make. For example, suppose that an investor arrives at time t and buys at the ask. After the trade, the information available to the specialist is S_t and the event that Z_t exceeded the ask. The specialist will update his expectation of V given this new bit of information, the probability that the trader was informed (given the past history) and the likelihood of a purchase given that he is informed. As long as the amount the specialist received was greater than or equal to this revised expectation, he does not regret the trade. Thus, the ask defined above is a reservation price. We assume that (unmodelled) competition drives the quoted ask to this reservation level.

This discussion of the equilibrium highlights an important interpretation of the bid and ask. The ask price is what the revised expectation of V will be if the specialist sells, and the bid is what the revised expectation will be if the specialist buys. Thus, once the bid and ask prices are specified, not only do we know that the possible transaction prices are, but we also know what the possible revised expectations of V are.

To insure that the customer's decision rule is formally well defined, and to illustrate the source of the spread, we must prove that at all times the ask exceeds the bid and, if insider trading actually occurs – or more precisely if it could occur – that the expectation of V lies strictly between the bid and ask. This proof and a later one rely on these two related facts from probability theory: (i) $\mathrm{E}[X|X > a] \geq \mathrm{E}[X]$ with a strict inequality when $0 < \mathrm{P}\{X > a\} < 1$, and (ii) $\mathrm{E}[X|X > a]$ is non-decreasing in a, being strictly increasing on any interval in the support of X.

Henceforth, we shall use E_t to denote conditional expectations given the common knowledge at time t, i.e., $\mathrm{E}_t[\,\cdot\,] = \mathrm{E}[\,\cdot\,|S_t \wedge H_t]$ where the 'meet' $S_t \wedge H_t$ denotes the events which are in both S_t and H_t. Notice that $A_t = \mathrm{E}_t[A_t]$ and $B_t = \mathrm{E}_t[B_t]$, since the bid and ask prices are common knowledge at time t. Also, our informal assumptions about ρ_t (that it conveys no information about V nor about an informed trader's opinions) can be adequately formalized by

(i) $\mathrm{E}_t[V|F_t, \rho_t] = \mathrm{E}_t[V|F_t]$,

(ii) $\mathrm{E}_t[\mathrm{E}_t[V|F_t]|\rho_t] = \mathrm{E}_t[V]$.

Proposition 1. Suppose equilibrium bid and ask prices exist satisfying the zero

expected profit conditions:

$$A_t = \mathrm{E}[V|S_t, Z_t > A_t],$$

$$B_t = \mathrm{E}[V|S_t, Z_t < B_t].$$

Then the ask price is greater and the bid price is less than the expectation of V: $A_t \geq \mathrm{E}_t[V] \geq B_t$. *The inequalities are strict if adverse selection is possible, i.e., if*

$$\mathrm{P}\{Z_t > \mathrm{E}_t[V], \mathrm{E}_t[V|F_t] > \mathrm{E}_t[V]\} > 0,$$

$$\mathrm{P}\{Z_t < \mathrm{E}_t[V], \mathrm{E}_t[V|F_t] < \mathrm{E}_t[V]\} > 0.$$

Proof. We prove only the first inequality, since the proof of the second is similar. Also, for brevity, we omit the time subscripts. Let C be the event that the customer makes a purchase, i.e., the event that Z is greater than A,

$$C = \{Z > A\} = \{\mathrm{E}[V|F] > A/\rho\}.$$

Then, by definition, $A = \mathrm{E}[V|S, C]$ so

$$\begin{aligned} A &= \mathrm{E}[A|C] = \mathrm{E}[\mathrm{E}[V|S,C]|C] = \mathrm{E}[V|C] = \mathrm{E}[\mathrm{E}[V|C,\rho]|C] \\ &= \mathrm{E}[\mathrm{E}[\mathrm{E}[V|F,C,\rho]|C,\rho]|C] = \mathrm{E}[\mathrm{E}[\mathrm{E}[V|F,\rho]|C,\rho]|C] \\ &= \mathrm{E}[\mathrm{E}[\mathrm{E}[V|F]|C,\rho]|C] \geq \mathrm{E}[\mathrm{E}[\mathrm{E}[V|F]|\rho]|C] \\ &= \mathrm{E}[\mathrm{E}[V]|C] = \mathrm{E}[V]. \end{aligned}$$

If the additional condition stated in the proposition holds, then the inequality is strict. Q.E.D.

To stress again an interpretation of the bid and ask, it should be noted that in proving the reasonable proposition that the ask exceeds the bid, we have proved that it is necessarily the case that expectations of V are revised upward in response to specialist sales, and revised downward in response to specialist purchases. This is so, because the ask and bid are the revised expectations in the respective cases, and we have assumed that transaction prices are public information.

Define H_t^+ and S_t^+ to be, respectively, the information available to the uninformed and the specialist just after a trade at time t. These information sets include information about whether a trader has arrived at time t, whether he bought or sold, and the price at which trade occurred.

Let T_k be the times at which trades occur. The above discussion shows that the T_k are stopping times relative to $\{S_t^+\}$ and $\{H_t^+\}$, and hence we can define

S_k and H_k by $S_k = S^+_{T_k}$ and $H_k = H^+_{T_k}$. (Also, any process subscripted with a k will be understood to be the value of the process at time T_k.) If the kth trade takes place at the ask at time t, i..e, there is an arrival at time t and Z_t exceeds A_t, then the transaction price will be the ask price, which in this event is equal to the revised expectation of V given this event. Similarly, if there is a trade at the bid, the transaction price is the bid price which in this case is also the revised expectation of V given this (different) event. Mathematically, the transaction price is given by $A_k I_{\{Z_k > A_k\}} + B_k I_{\{Z_k < B_k\}}$. This is, by definition, equal to

$$\mathrm{E}\left[V|S_{T_k}, Z_{T_k} > A_{T_k}\right] I_{\{Z_{T_k} > A_{T_k}\}} + \mathrm{E}[V|S_{Tk}, Z_{Tk} < B_{Tk}] I_{\{Z_{T_k} < B_{T_k}\}}. \tag{5}$$

But (5) is just $\mathrm{E}[V|S^+_{T_k}] = \mathrm{E}[V|S_k]$. This observation allows us to write the kth transaction price as $p_k = \mathrm{E}[V|S_k]$, as long as k trades take place (i.e., as long as T_k is less than T_0). Thus, if N trades actually take place, then $p_1, \ldots, p_N$ are the prices at which trades occur. Notice, however, that p_{N+1} is also well defined, and represents some intermediate value between the bid and ask prices at the end of the trading period. If one thinks in terms of computing returns based on daily data, this use of an imaginary transactions price is not far from the way CRSP calculations are done for days in which there is no trade. The specification of H_t^+ from the preceding paragraph implies that p_k is measurable with respect to H_k which allows us to prove the following proposition:

Proposition 2. The sequence of transaction prices $\{p_k\}$ *forms a martingale relative to the specialist's information,* $\{S_k\}$, *and the public information,* $\{H_k\}$.

Proof. From (5), $p_k = \mathrm{E}[V|S_k]$. Thus,

$$\mathrm{E}[p_{k+1}|S_k] = \mathrm{E}\left[\mathrm{E}[V|S_{k+1}]|S_k\right] = \mathrm{E}[V|S_k] = p_k.$$

Since H_k is contained in S_k, and since p_k is measurable with respect to H_k, the sequence of transaction prices $\{p_k\}$ forms a martingale relative to $\{H_k\}$ as well. Q.E.D.

This result is slightly stronger than the usual statement of the semistrong form of the efficient markets hypothesis – prices form a martingale relative to all public information and the information known to the specialist. The assumed competition among equally informed specialists implies that there are no profit opportunities arising from the information known by the specialist. Furthermore, at the instant that a trade occurs and the price is announced, the

specialist and all outsiders agree on the expected value of V, since $\mathrm{E}[V|p_k] = \mathrm{E}[V|\mathrm{E}[V|S_k]] = \mathrm{E}[V|S_k]$.

Another implication of Proposition 2 is that, in the environment that we have described, first differences of the transaction price process will be serially uncorrelated. This follows from the fact that the increments of a martingale are uncorrelated. Thus, spreads due only to adverse selection are qualitatively different from spreads due to specialist transactions costs, risk aversion or monopoly power. The latter sources of spread lead to negative serial correlation, while spreads due solely to adverse selection do not.

Further intuition into the nature of this result can be gained by considering an environment in which there is both asymmetric information and specialist transaction costs. Specifically, suppose every trade costs the specialist c dollars. The zero profit equilibrium bid and ask then satisfy

$$A = \mathrm{E}[V|S_t, Z_t > A] + c,$$

$$B = \mathrm{E}[V|S_t, Z_t < B] - c.$$

Following the development above [in (5)], transaction prices are then given by $p_k = \mathrm{E}[V|S_k] + Q_k c$ where Q_k is one if the kth transaction involved the specialist selling at the ask, and is minus one if the specialist purchased at the bid. Under the assumption that $\mathrm{E}[Q_{k+1}|H_t] = \mathrm{E}[Q_k|H_t]$ for all $t < T_k$ where T_k is the time of the kth transaction, it is easy to show that transaction price changes will exhibit negative serial correlation.

To indicate how the negative serial correlation depends upon the relative magnitudes of the spread due to adverse selection and the spread due to the costs of transacting, suppose that the probabilities of a buy and a sell are equal so that $\mathrm{E}[Q_k|H_t] = 0$ for $t < T_k$. Let Ψ be the part of the spread due to adverse selection,

$$\Psi = \mathrm{E}[V|S_t, Z_t > A] - \mathrm{E}[V|S_t, Z_t < B];$$

hence the total spread is $\Psi + 2c$. Some straightforward calculations show that the covariance of adjacent price changes is given by: $-\frac{1}{2}c\Psi - c^2$. The calculation is similar to the one in Roll (1984). We obtain a different result because Roll implicitly assumes that no part of the spread is due to adverse selection. If, as here, the spread is due in part to adverse selection, then a specialist sell, for example, leads to an upward revision of expections and hence future prices are not independent of the current transaction as Roll assumes.

The variance of the price change is given by the following: $\theta^2 + (\Psi/2)^2 + c\Psi + 2c^2$, where θ^2 is the variance of public information arriving exogenously between trades.[6] Let β be the proportion of the spread due to transaction costs; i.e., $\beta = 2c/(\Psi + 2c)$. Then the correlation coefficient, R, is given by the following:

$$R = -\beta/(\delta + \beta^2), \tag{6}$$

where $\delta = 1 + (2\theta/(\Psi + 2c))^2$. The correlation coefficient thus moves toward zero if the proportion of the spread due to informational asymmetries increases. We can invert (6) to get the proportional of the spread due to trading costs as a function of the correlation coefficient and δ (a measure of how much public information arrives relative to the spread),

$$\beta = \left(-1 + (1 - 4\delta R^2)^{1/2}\right)/2R. \tag{7}$$

Unfortunately, (7) alone does not define a useful statistic since in general δ will not be known. It does suggest the possibility that transaction data might be used to find a measure of informational assymetry.

In the introduction, we described the theoretical possibility that markets might close down entirely, with the bid price being set so low and the ask price so high as to discourage any trade. This problem is identical to the famous lemons problem of Akerlof (1970), in which adverse selection can destroy the market. For the next proposition to make sense, it must also be possible for markets to function well without suffering such a breakdown. The conditions that determine the functioning of markets depend in part on the supply and demand elasticities of uninformed traders. The expected demand when the ask price is A and the common knowledge value expectation is E is given by $1 - G(A/E)$, where G is the distribution function for an uninformed trader's preference parameter ρ. Similarly, the expected uninformed supply at a bid price of B is $G(B/E)$. If uninformed demand and supply is inelastic, then the closing of markets due to adverse selection can never occur. To prove this, it suffices to note that as A is raised, the expected losses suffered by the specialist to the informed decline and approach zero, while his expected profits from the

[6] The variance calculation is accomplished by dividing the change in expectations between the kth trade and the $(k+1)$st trade into the change in expectations as a result of public information arriving between the two trades and the information contained in the $(k+1)$st trade and noting that these are uncorrelated. The change in expectations squared as a result of the $(k+1)$st trade is one fourth of the spread due to adverse selection squared since a buy and sell are equally likely. Furthermore, Q^2 is identically one, and finally, $\mathrm{E}[Q_k \mathrm{E}[V|H_k]] = 1/2\Psi$. All other expectations are zero.

uninformed are positive and rising. Hence, the specialist's total profits from selling must be positive for some finite A. A symmetric argument applies to the bid price B.

We now turn our attention to the properties of the bid and ask prices when trading does not break down. Imagine performing the following experiment. Every time there is a transaction, record the bid and ask that prevailed just prior to the trade. Do this for a unit of time, and denote the number of trades observed by N. The following proposition shows that if trade is reasonably balanced (i.e., the probability of a purchase given that a trade occurred is bounded away from zero and one) then the expectation of the number of trades times the average spread squared is bounded by a number that is independent of the pattern of trade.[7]

Proposition 3. Define

$$1/\gamma_k = \mathrm{P}\{Z_k > A_k | S_{T_k}\}\mathrm{P}\{Z_k < B_k | S_{T_k}\},$$

and let γ^* *be the mean value of* γ_j *over the* $N+1$ *observations; i.e.,* $\gamma^* = \sum^{N+1}\gamma_k/(N+1)$ *(if* $N+1$ *trades do not occur prior to* T_0*, put* $\gamma_{N+1} = 4$*). Further, define* Ψ_N *to be the average spread over N trades, i.e.,*

$$\Psi_N = \sum^{N}(A_k - B_k)/N.$$

Then,

$$\mathrm{E}\left[(N/(N+1))\Psi_N^2 N/\gamma^*\right] \leq \mathrm{var}(V).$$

In particular, if there is a number γ *such that* $\mathrm{P}\{\gamma^* < \gamma\} = 1$*, then*

$$\mathrm{E}\left[(N/(N+1))N\Psi_N^2\right] \leq \mathrm{var}(V)\gamma,$$

and hence, if N is almost surely positive, the expected value of the volume times average spread squared is bounded by a number that is independent of the pattern of trade,

$$\mathrm{E}\left[N\Psi_N^2\right] \leq 2\,\mathrm{var}(V)\gamma.$$

Proof. Since p_{N+1} is a conditional expectation of V, the variance of V

[7]We remind the reader that we are returning to an environment in which there are no costs associated with trade, and the specialist earns a zero profit on average.

exceeds the variance of p_{N+1}. Taking p_0 to be E[V], we have

$$\begin{aligned}\operatorname{var}(p_{N+1}) &= \operatorname{var}\left(\sum^{N+1}(p_k - p_{k-1})\right) = \mathrm{E}\left[\left(\sum^{N+1}(p_k - p_{k-1})\right)^2\right] \\ &= \mathrm{E}\left[\sum^{N+1}(p_k - p_{k-1})^2\right] \\ &\quad + 2\mathrm{E}\left[\sum_{k}^{N+1}\sum_{i}^{k-1}(p_k - p_{k-1})(p_k - p_{i-1})\right] \\ &= \mathrm{E}\left[\sum^{\infty}(p_k - p_{k-1})^2 I_{\{k-1 \le N\}}\right] \\ &\quad + 2\mathrm{E}\left[\sum^{\infty}\sum^{k-1}(p_k - p_{k-1})(p_i - p_{i-1}) I_{\{k-1 \le N\}}\right].\end{aligned}$$

Since the transaction price sequence forms a square integrable martingale, we can move the expectation inside the summation in each case. In the second term, condition first on H_{k-1}. Since the increments of a martingale are uncorrelated and have mean zero, the second expectation is zero.

Define e_k by

$$e_k = \mathrm{E}\left[(p_k - p_{k-1})^2 | H_{k-1}, A_k, B_k, \gamma_k\right].$$

Thus, we have

$$\operatorname{var}(V) \ge \mathrm{E}\left[\sum^{N+1} e_k\right].$$

Some algebraic manipulation shows that $e_k \ge (A_k - B_k)^2/\gamma_k$ (see footnote 6), and hence for any N,

$$\sum^{N+1}(e_k\gamma_k)^{1/2} \ge \sum^{N+1}(A_k - B_k)^2.$$

By the Cauchy–Schwarz inequality, $\sum^{N+1} e_k \ge (\sum^{N+1} A_k - B_k)^2/\gamma^*(N+1)$, and thus,

$$\operatorname{var}(V) \ge \mathrm{E}\left[\left(\sum^{N} A_k - B_k\right)^2 \Big/ (N+1)\gamma^*\right] = \mathrm{E}\left[(N/(N+1))N\Psi_N^2/\gamma^*\right].$$

If $\gamma^* \leq \gamma$ a.s., then

$$\mathrm{E}\left[\left(N/(N+1)\right)N\Psi_N^2\right] \leq \mathrm{var}(V)\gamma.$$

If N is positive a.s., then $N/(N+1)$ is greater than or equal to $\frac{1}{2}$, so $\mathrm{E}[N\Psi_N^2] \leq 2\,\mathrm{var}(V)\gamma$. Q.E.D.

Since the bound in Proposition 3 is independent of the pattern of trade (given that N is a.s. at least one and $\gamma^* < \gamma$) the proposition suggests a relation between average volume and average spread. Loosely speaking, markets in which there is, on average, large volume will have small average spreads and vice versa. This is consistent with the usual explanation of the relation between spreads and volume which focuses on the specialist's need to recoup fixed costs from the market participants. However, the fixed cost explanation implies that the average spread will be proportional to one over average volume, while proposition 3 suggests that the average spread will be proportional to one over the root of average volume.

It should be clear from the above proof, that the conclusion of the proposition is true when N is some deterministic number of trades. In this case, the proposition states that the average spread squared tends to decline for large N. Indeed, if T_0 is very large relative to the interarrival times of traders, so that N can be very large, the proposition implies that the spread will go to zero a.s.

Within our model, the fact that spreads decline with the number of trades reflects the assimilation by the market of the insiders' information. This leads eventually (if there are sufficiently many trades) to an approximate consensus expectation of value (V) in the market. Proposition 4 offers a formal statement of that tendency toward consensus.

Proposition 4. If trade is reasonably balanced in the sense of Proposition 3, i.e., the probability of a purchase is bounded away from zero and one, then the expectations of the specialist and the traders converge as the number of trades increases, i.e., $\mathrm{E}[V|S_k] - \mathrm{E}[V|F_k]$ *converges to zero in probability (where* F_k *is the information of the* kth *trader to arrive).*

Proof. We use the notation and results of Proposition 1 and two general facts from probability theory cited earlier. The suppressed time subscript is now $t = T_k$.

$$A = \mathrm{E}[V|C] = \mathrm{E}[\mathrm{E}[V|F,\rho]|C] = \mathrm{E}[\mathrm{E}[V|F]|\mathrm{E}[V|F] > A/\rho]$$
$$> \mathrm{E}[\mathrm{E}[V|F]|\mathrm{E}[V|F] > \mathrm{E}[V]/\rho].$$

Define D by

$$D = \mathrm{E}[V|F] - \mathrm{E}[V].$$

Then,

$$A - \mathrm{E}[V] \geq \mathrm{E}[D|D > \mathrm{E}[V](1-\rho)/\rho].$$

By the Chebyshev inequality,

$$\mathrm{E}[D|D > \mathrm{E}[V](1-\rho)/\rho] \geq \varepsilon \mathrm{P}\{D \geq \varepsilon | D > \mathrm{E}[V](1-\rho)/\rho\}$$

$$\geq \mathrm{P}\{D \geq \varepsilon\}\varepsilon.$$

By Propositions 1 and 3, $A_t - \mathrm{E}_t[V]$ converges almost surely to zero, so $\mathrm{P}\{D_t \geq \varepsilon\}$ must also converge to zero for all positive ε. A similar argument using bid prices shows that $\mathrm{P}\{D_t \leq -\varepsilon\}$ goes to zero. Thus, $\mathrm{E}_k[V|F_k] - \mathrm{E}_k[V]$ converges in probability to zero. Also, $\mathrm{E}_k[V|S_k] = \mathrm{E}_k[V]$ (both are equal to A_k if the customer buys and B_k if he sells). Q.E.D.

As the consensus described in Proposition 4 emerges, there comes to be a balance between expected supply and expected demand:

Corollary 1. The specialist's inventory of stocks tends to a driftless stochastic process; i.e.,

$$\lim \mathrm{P}\{Z_k < B_k|H_k\} - \mathrm{P}\{Z_k > A_k|H_k\} = 0.$$

Proof. Using the results of the Proposition 4, let $\hat{p}$ be the limit of $\mathrm{E}[V|S_k]$ and let $\hat{Z} = \rho\hat{p}$. Then,

$$\lim(\mathrm{P}\{Z_k < B_k|H_k\} - \mathrm{P}\{Z_k > A_k|H_k\}) = \mathrm{P}\{\hat{Z} < \hat{p}|\hat{S}\} - \mathrm{P}\{\hat{Z} > \hat{p}|\hat{S}\}$$

$$= \mathrm{P}\{\rho < 1\} - \mathrm{P}\{\rho > 1\} = 0,$$

where $\hat{S}$ is the specialist's limiting information. Q.E.D.

That the specialist's inventory will not drift on average is only true in the limit. To see why inventories might drift, suppose that the insiders may have early access to information about a takeover attempt which would be favorable for stockholders of the firm, but is generally considered unlikely to transpire. There is then an asymmetry in the adverse selection problem faced by the specialist; active buying by insiders might be considered quite informative while active selling would convey less information. In this case, the ask price might include a premium (over the share's expected value) that is larger than the corresponding discount in the bid price. An expected inventory accumulation might result. The fact that there may be an inventory build-up is largely an artifact of our assumption of risk neutrality on the part of the specialist. As

Ho and Stoll (1981) and Ohara and Oldfield (1982) have demonstrated a risk-averse specialist will adjust the level and/or the magnitude of the spread to avoid inventory build-ups. The corollary indicates that the accumulation of inventory will not be too extreme even in the risk-neutral case because eventually, supply and demand will be approximately equal.

We now turn to our analysis of the determinants of the size of the spread. For this analysis, it is useful to distinguish sharply between insiders and liquidity traders, so we assume that insiders have no liquidity motive ($\rho = 1$). We also assume that the specialist has only public information; i.e., $S_t = H_t$. Proposition 5 provides proof of results also reported in Copeland and Galai (1983, p. 1463).

Proposition 5. For any given time t, the ask price A_t increases and the bid price B_t decreases when, other things being equal,

(i) *the insiders's information at time t becomes better (i.e., finer),*
(ii) *the ratio of informed to uninformed arrival rates at t is increased, or*
(iii) *the elasticity of uninformed supply and demand at time t increases.*

Proof. If there is an arrival at time t, we can express the reservation price Z_t by

$$Z_t = (1 - U_t)\mathrm{E}[V|H_t, J_t] + U_t\rho_t\mathrm{E}[V|H_t],$$

where [as in (1) above] U_t is one if an arrival at t is uninformed and zero otherwise, and J_t is the information of an insider if an insider arrives at t (for the succeeding discussion, time subscripts have been dropped to simplify the notation). Define M by $M = \mathrm{E}[V|J]$ where J is the information set of an arriving insider, and let G be the distribution function of the liquidity parameter (and let g be its density). Denote demand and supply elasticities of the uninformed by e_{D} and e_{S}. Then e_{D} and e_{S} are given by

$$e_{\mathrm{D}} = Ag(A/\mathrm{E}[V])/\mathrm{E}[V](1 - G(A/\mathrm{E}[V])),$$

$$e_{\mathrm{S}} = Bg(B/\mathrm{E}[V])/\mathrm{E}[V]G(B/\mathrm{E}[V]).$$

Recalling [from (4a)] that A and B, the ask and bid, are the smallest a and largest b satisfying

$$a \geq \mathrm{E}[V|Z > a], \qquad b \leq \mathrm{E}[V|Z < b],$$

then, for ask price A and bid price B,

$$A\mathrm{P}\{Z > A\} - \mathrm{E}\left[VI_{\{Z > A\}}\right] \geq 0,$$

$$\mathrm{E}\left[VI_{\{Z < B\}}\right] - B\mathrm{P}\{Z < B\} \geq 0.$$

The left sides can be expanded to

$$\mathrm{E}\left[(1-U)(A-M)I_{\{M>A\}} + UAg(A/\mathrm{E}[V])(A/\mathrm{E}[V]-1)/e_{\mathrm{D}}\right]$$

$$= \mathrm{E}\left[\phi(A, U, M, e_{\mathrm{D}})\right],$$

$$\mathrm{E}\left[(1-U)(M-B)I_{\{M<B\}} + UBg(B/\mathrm{E}[V])(1-B/\mathrm{E}[V])/e_{\mathrm{S}}\right]$$

$$= \mathrm{E}\left[\psi(B, U, M, e_{\mathrm{S}})\right].$$

The functions $\phi(\cdot)$ and $\psi(\cdot)$ are concave in M and increasing in U. Also, $\phi(\cdot)$ is decreasing in e_{D} and $\psi(\cdot)$ is decreasing in e_{S}.

For (i), let A' and B' be the ask and bid prices associated with insider information J' and define M' by $M' = \mathrm{E}[V|J']$ where J' is finer than J. For (ii), let A'' and B'' be the ask and bid prices when the arrival of uninformed is governed by U'', where $U''(\omega) \leq U(\omega)$. For (iii), let A''' and B''' be the ask and bid prices when the demand and supply elasticities are e_{D}''' and e_{S}''' respectively with $e_i''' > e_i$ for $i = \mathrm{D}, \mathrm{S}$.

If J' is finer than J, then $M = \mathrm{E}[M'|J]$. This and Jensen's inequality allow us to conclude:

$$\mathrm{E}\left[\phi(A', U, M, e_{\mathrm{D}})\right] \geq \mathrm{E}\left[\phi(A', U, M', e_{\mathrm{D}})\right].$$

Since ϕ is increasing in U, and $U'' \leq U$,

$$\mathrm{E}\left[\phi(A'', U, M, e_{\mathrm{D}})\right] \geq \mathrm{E}\left[\phi(A'', U'', M, e_{\mathrm{D}})\right].$$

Also, since ϕ is decreasing in e_{D},

$$\mathrm{E}\left[\phi(A''', U, M, e_{\mathrm{D}})\right] \geq \mathrm{E}\left[\phi(A''', U, M, e_{\mathrm{D}}''')\right].$$

[Similar inequalities hold for the function $\psi(\cdot)$.] By the definitions of A', A'' and A''' the right-hand sides are all non-negative. But, $A = \inf\{a|\mathrm{E}[\phi(a, U, M, e_{\mathrm{D}})] \geq 0\}$, so A', A'' and A''' all exceed A. The same argument will prove the corresponding bid inequalities. Q.E.D.

Intuitively, the adverse selection problem is worse the greater the fraction of informed traders and the better their information. The specialist is forced to set a higher spread if there are more informed or if they have better information, in order to avoid losses. On the other hand, the greater the desire of the uninformed to trade (measured by the elasticities), the easier is the specialist able to make back his losses to informed traders. The zero-profit condition then results in a smaller spread.

Alternatively, Proposition 5 can be interpreted as an analysis of the determinants of the specialist's updating of his expectations. That the spread is small when the probability that the next arrival is informed is small implies that when this probability is small, updated expectations in response to trade will differ only slightly from the prior expectations. Similarly, when uninformed demand and supply is very inelastic, trade leads to relatively small revisions in expectations.

This proposition also suggests a lagged statistical relation between volume and the spread. Specifically, suppose the level of insider activity is positively related to volume. When the specialist sees unexpectedly high volume, he will revise upward his estimate of the probability of an insider arrival and increase the spread accordingly. Thus, this proposition would suggest a positive correlation between past volume and current spread.

It is easy to make the mistake of interpreting Proposition 5 as a comparative equilibrium result. What the proposition says is that *other things equal*, including the past history of trade, certain changes have determinate effects on the size of the spread. Since the history of trade after the time when insiders gain their information is endogenous and depends on the same parameters as those studied in the proposition, the proposition has a comparative equilibrium interpretation only for the time just after the insiders have become informed. For example, an increase in the frequency of insider arrivals has the immediate effect of increasing the spread. However, as long as trade continues, the increase in insider activity means more information will be conveyed by transaction prices. This in turn may mean that spreads in the future will be smaller because the informational differences between insiders and outsiders will be decreased. This intuition is indicated in the proof of Proposition 3. Recall that there we showed that the expected average spread squared times the volume is bounded by a number that is independent of the level of insider activity. Thus, if spreads increase now, they must be reduced later on as long as there is sufficient trading activity. We pursue this intuition in the following examples.

3. An example

To illustrate the theory of section 2, we present two simple examples. The examples show how bid and ask prices are determined, how the proportion of insiders affects the spread and the informativeness of prices, and how market breakdown can occur.

Suppose that the stock can have either of two values, $V = 1$ or $V = 11$, and that the higher value has prior probability π. The expectation of V is then $E = 11\pi + 1(1 - \pi)$.

Insiders in this example have perfect information about the value of V and have preference parameters $\rho = 1$. Liquidity traders have only public informa-

tion. We shall consider two distributions of the liquidity traders' preference parameters ρ: the case where ρ equals 0 or ∞, each with probability one-half, which is the case of perfectly inelastic liquidity supply and demand, and the case where ρ is uniformly distributed on $(0,2)$. No information arrives exogenously during the trading period.

The inelastic demand case serves as an interesting benchmark. The uniform case is richer in the range of phenomena that can arise. In particular, it allows the logical possibility that markets may shut down on one side (trader's buying at the ask) but not the other, since the liquidity trader's demand is elastic at high prices but their supply has unitary elasticity at all prices.

Let the proportion of insiders in the trading population be designated by the parameter α. For these examples, we assume that insiders arrive independently of their realization of information, and hence the *a priori* probability that any particular trader is an insider is α.

It is clear that with our standing assumptions and in the absence of adverse selection the price would be the expected value E. Suppose that the ask price is set at A (where $A < 11$) and that the next trader is an insider who wants to buy at that price. Then the specialist will lose $11 - A$ to the insider. The event just described will occur with probability $\alpha\pi$, since the event that the buyer is an insider is independent of the value of V. The probability that a liquidity trader will buy at an ask price of A is precisely the probability that the preference parameter ρ exceeds A/E, which in the inelastic case is 0.5 and in the uniform case is $1 - A/(2E)$. So the specialist's break-even condition in the inelastic case is the linear equation (in A),

$$\alpha\pi(11 - A) = 0.5(1 - \alpha)(A - E), \tag{8}$$

and in the uniform case is the quadratic equation,

$$\alpha\pi(11 - A) = (1 - \alpha)[1 - A/(2E)](A - E), \tag{9}$$

provided $A \leq 2E$. The smaller root of the quadratic equation is the relevant one here, since the ask price is the *lowest* price at which the specialist breaks even.

A similar argument shows that the bid price for the inelastic case solves the linear equation

$$\alpha(1 - \pi)(B - 1) = 0.5(1 - \alpha)(E - B), \tag{10}$$

while the bid price for the uniform case is the larger root of the quadratic equation

$$\alpha(1 - \pi)(B - 1) = (1 - \alpha)[B/(2E)](E - B). \tag{11}$$

We have described how the specialist in this model determines a bid and ask price as a function of the parameter α and his beliefs π. To complete the description of his dynamic behavior, it only remains to show how π changes as a result of market behavior. We let π^+ denote the posterior beliefs of the specialist after a trade has just occurred. These posterior beliefs also serve as the prior beliefs for determining the bid and ask prices for the next trader. π^+ is determined from Bayes' Theorem using the formula

$$[\pi^+/(1-\pi^+)] = [\pi/(1-\pi)]\ \textit{Factor},$$

where *Factor* is the likelihood ratio whose numerator is the probability of the trader's action given that $V = 11$ and whose denominator is the probability of the action given $V = 1$. Note that *Factor* may depend on the action taken, the bid price, the ask price, π and α, as well as on the model used – uniform or inelastic.

For any given model and any value of the parameter α, the endogenous variable π determines the expected value E, the bid and ask prices B and A, the spread, the probability that the next trader will buy or sell, etc. Then, from the formula for π^+ above, it is clear that the stochastic process of values of π over time is a Markov process; given the current value of π, its future distribution is independent of its history. Moreover, since π_k is a conditional probability given all information up to time T_k, the Markov process is a martingale. In addition, in the inelastic case, $\log(\pi_k(1-\pi_0)/\pi_0(1-\pi_k))/\log(1+\alpha)/(1-\alpha)$ is, given V, a random walk whose value is the accumulated excess of purchases over sales by arriving customers.

Let us say that "nearly all of the insider information has been assimilated in the prices" when we reach the point where $\pi/(1-\pi) < 1/Odds$ when V is low or $\pi/(1-\pi) > Odds$ when V is high. For the case of perfectly inelastic demand, one can show[8] that the expected number of trades that must take place before nearly all information is revealed is approximately[9]

$$\frac{\log(Odds) + (1-\pi_0)\log(\pi_0/(1-\pi_0))}{\alpha\log((1+\alpha)/(1-\alpha))}. \tag{12}$$

For small values of α, (12) is approximately proportioned to $1/\alpha^2$. Also for small values of α and for any π, the size of the spread determined in (8) and (10) is approximately proportional to α. Thus, the effect of doubling α from, say, 0.1 to 0.2 is roughtly to double the spread at each level of π and to divide by four the time taken until nearly all insider information has been assimilated in prices.

[8] The proof uses Wald's lemma separately for the cases $V = 1$ and $V = 11$.

[9] The expression is approximate because $(\log(Odds) - \log(\pi_0/(1-(\pi_0))))/\log((1+\alpha)/(1-\alpha))$ and $(\log(Odds) + \log(\pi_0/(1-\pi_0)))/\log((1+\alpha)/(1-\alpha))$ may not both be integral.

Of course, the spread after nearly all insider information has been assimilated is nearly zero. So, if enough trading occurs in the period of interest to assimilate information for either value of α, then the expected average spread over the period for the case $\alpha = 0.1$ would be about twice that for the case $\alpha = 0.2$, but the expected squared total spread would be about the same for both cases. Thus, given plenty of trading volume, one can make strong statements about the expected squared total spread without knowing anything about the fraction of insiders α, but statements about the expected mean spread require that extra information. This formal analysis is example specific, but it does accord nicely with Proposition 3, which derives a bound on the expectation of spread squared times volume which is independent of the proportion or arrival pattern of insiders.

When α is small, simulation results for the uniform case are not dissimilar from those for the inelastic case. This is as might be expected, since for small α the ratios A/E and B/E which enter into the break-even equations are approximately equal to 1. Substituting 1 for these ratios yields precisely the equations of the inelastic case.

As α grows larger, however, the spread can become so large as to deter most, and eventually all, potential liquidity buyers. In the numerical example at hand, if $\alpha = 0.3$, then there is no ask price at which the specialist can break even [as evidenced by the fact that the discriminant of the quadratic equation for A, (9), is negative].

In this example, there is still a bid price at which both insiders and liquidity traders will trade and the specialist will break even. It is a feature of our example that the demand function of liquidity traders is quite elastic for high ask prices but the supply function has unitary elasticity. For such cases, as we noted in the previous section, market breakdown (on the supply side) can never occur.

Of course, by choosing ρ, the liquidity parameter, to be uniformly distributed on $(0.5, 1.5)$, one can create an example in which both sides of the market break down for large values of α and intermediate values of the prior expectation. When a market breakdown does occur in this model, since we have assumed stationarity and no exogenous flow of information, the problem will persist. The market will remain closed indefinitely. As we observed in the introduction, this possibility strongly suggests that the trading institution we have been describing is not socially optimal.

4. A model with discounting

The model discussed in the previous section is based on a particular normalization of reservation prices that was mathematically convenient. This normalization took the form of the specialist having a ρ of one, while the median of the ρ's of the traders was one. Another normalization that is of

economic interest is the following: the reservation price of an individual arriving at time t is Z_t^* given by $Z_t^* = \exp(r_t(T_0 - t))Z_t$ where Z_t is as defined in the previous section, and T_0 is the time of the informational event. The parameter r_t may arise from other unmodeled market opportunities and depends only on time, not on any personal characteristics. The zero profit condition for the specialist now becomes a zero excess return condition and may be stated as (if solutions exists):

$$A_t^* = \exp(-r_t(T_0 - t))\mathrm{E}\left[V|S_t^*, Z_t^* > A_t^*\right],$$
$$B_t^* = \exp(-r_t(T_0 - t))\mathrm{E}\left[V|S_t^*, Z_t^* < B_t^*\right]. \tag{13}$$

Since the market now being described is merely a renormalization of the one described in section 2, it is straightforward to show that A_t^* and B_t^* are given by $A_t^* = \exp(-r_t(T_0 - t))A_t$; $B_t^* = \exp(-r_t(T_0 - t))B_t$, where A_t and B_t satisfy (as above)

$$A_t = \mathrm{E}[V|S_t, Z_t > A_t], \qquad B_t = \mathrm{E}[V|S_t, Z_t < B_t].$$

To insure that outsiders have an incentive to be involved in the market, the following hypothesis is offered. Let τ be a holding period. The expected gross holding period return of someone buying at time t and holding for τ periods of time is $\mathrm{E}_t[B_{t+\tau}^*]/A_t^*$. It is assumed that at any time t, $\mathrm{E}_t[B_{t+\tau}^*]/A_t^* = \mathrm{e}^{i\tau}$, where i is an exogenously given rate of return. Although this is implicitly a hypothesis about the exogenous variables, it is stated in terms of market parameters and appears to be testable. The variable i might be taken to be a required return consistent with the risk of the stock. The important limitation such a condition imposes on the data is that i be unrelated to the magnitude of the spread and constant through time. In effect, this assumption defines r_t.[10] Since $B_{t+\tau}^*$ is a function of $r_{t+\tau}$ and A_t^* is a function of r_t, a terminal condition and the above expected holding period return condition will define r_t. The proof of the following proposition is tedious, and is relegated to an appendix.

Proposition 6. Let the expected realizable return of an uninformed trader over the normal holding period be i, i.e., $\mathrm{E}_t[B_{t+\tau}^*/A_t^*] = \mathrm{e}^{i\tau}$ for all t. Assume that after the informational event at T_0, V becomes known so that for $t \in [T_0 - \tau, T_0]$,

[10] Stoll and Whaley (1983) have shown that transactions costs including the bid–ask spread may explain part of the small firm effect. Their analysis makes use of a 'holding period' such as specified here.

$B^*_{t+\tau} = Ve^{-i(T_0-t-\tau)}$ *(i.e.,* $r_{t+\tau} = i$ *for* $t \in [T_0 - \tau, T_0]$*). Then* r_t*, the discount rate at* t*, is the normal return* i *plus a premium,*

$$r_t = i + (n+1)/(T_0 - t)\log(k_t),$$

where

$$\frac{1}{k_t} = \left\{ E_t\left[\frac{B_{t+\tau}}{A_t} \frac{B_{t+2\tau}}{A_{t+\tau}} \cdots \frac{B_{t+(n+1)\tau}}{A_{t+n\tau}} \right] \right\}^{1/(n+1)} \leq 1,$$

and

$$t + n\tau \in [T_0 - \tau, T_0).$$

The discount rate applied at time t, r_t, has a particularly interesting interpretation. Notice that $(1/k_t)$ is the expected geometric mean gross return per τ units of time earned by an investor that follows a strategy of buying and selling every τ periods of time in a market with no discounting. The log of this is thus the continuously compounded expected return from such a strategy. Obviously, such a return is negative. Recall from the definition of A_t and B_t that the specialists sets the bid and ask so that on average what he loses to the informed is made up by what he gains from the uninformed liquidity traders. Thus, $(n+1)/(T_0 - t)\log(k_t)$ (a positive number) is, in return (per unit time) terms, what the uninformed on average lose to the informed. Thus, r_t represents the expected holding period return, i, plus the return that the uninformed anticipate losing to the informed. Note that r_t depends upon the holding period τ. In particular, $n+1$ in Proposition 6 is approximately $(T_0 - t)/\tau$, and hence r_t is approximately $i + (1/\tau)\log(k_t)$.

The above proposition, with i specified exogenously, closes the model in the sense that Z_t^*, A_t^*, B_t^* are now specified. The resulting price process will be $\{p_t^*\}$ with $p_t^* = e^{-r_t(T_0-t)}p_t$ where p_t is as specified in section 2. The observed holding period return will be $p^*_{t+\tau}/p^*_t$. If $T_0 - t$ is large relative to τ, then r_t and $r_{t+\tau}$ will be approximately equal, in which case $p^*_{t+\tau}/p^*_t$ will be on average approximately equal to $e^{r_t\tau}$. The observed returns will be larger than i, the hypothesized holding period return, since $p^*_{t+\tau} \geq B^*_{t+\tau}$ and $p^*_t \leq A^*_t$, and hence $p^*_{t+\tau}/p^*_t \geq B^*_{t+\tau}/A^*_{t+\tau}$, which is equal to $e^{i\tau}$ in expectation. That is, returns calculated by observing transaction prices will always be at least as large as the returns that one could realize by buying at time t and selling at time $t+\tau$.

On the other hand, it is easy to see that the existence of a bid–ask spread is less important the longer is the investment horizon. Intuitively, this spread can be amortized over a larger number of periods. To see this, the expected value

of the return that can be realized long-term is

$$\mathrm{E}_t\left[\left(\frac{V}{A_t^*}\right)\right] = \mathrm{E}_t\left[\left(\frac{p_t^*}{A_t^*}\frac{p_{t+\tau}^*}{p_t^*}\cdots\frac{p_{t+n\tau}^*}{p_{t+(n-1)\tau}^*}\frac{V}{p_{t+n\tau}^*}\right)\right]$$

$$= \mathrm{E}_t\left[\left(\frac{p_t^*}{A_t^*}\right)\left(\frac{p_{t+\tau}^*}{p_t^*}\cdots\frac{p_{t+n\tau}^*}{p_{t+(n-1)\tau}^*}\frac{V}{p_{t+n\tau}^*}\right)\right]$$

$$= \mathrm{E}_t\left[\left(\frac{p_t}{A_t}\right)\left(\frac{p_{t+\tau}^*}{p_t^*}\cdots\frac{p_{t+n\tau}^*}{p_{t+(n-1)\tau}^*}\frac{V}{p_{t+n\tau}^*}\right)\right]. \tag{14}$$

Since $p_t \leq A_t$, the above expected return is less than the observed return. If $T_0 - t$ is large, however, then $(p_t/A_t)^{1/T_0 - t}$ will be close to one, and the long-term per period mean return will be close to the observed (from the transaction price sequence) per period return.

These observations may provide some insight into such 'anomalies' as the 'small firm effect' and the 'ignored firm effect'. In both cases it may be reasonable to conjecture that informational differences between market participants may be significant. In the case of the small firm effect, it may be the case that insiders hold a larger proportion of the stock. As the results in section 2 show, this will indicate (other things equal) a larger spread earlier in the period of time when there are informational asymmetries, and hence a larger divergence between r_t and i. In the latter case, the lack of public reporting on a firm may imply that there is a larger informational difference between insiders and outsiders. This will also mean a larger spread and hence a greater difference between r_t and i. The above results suggest that the measured 'excess returns' are not realizable in a short-run basis. Rather, the spread, which represents the expected loss of the uninformed to the informed, leaves an outsider with a 'normal' rate of return. In the long run, returns will indeed be larger on average, but these higher returns can only be realized by buying and holding.

5. Conclusion

We have analyzed a model of a securities market in which the arrival of traders over time is accommodated by a specialist. Adverse selection, by itself, can account for the existence of a spread between the ask and bid prices, and the average magnitude of the spread depends on many parameters, including the exogenous arrival patterns of insiders and liquidity traders, the elasticity of supply and demand among liquidity traders, and the quality of the information

held by insiders. Furthermore, the transaction prices are informative, and hence spreads tend to decline with trade.

We do not claim that adverse selection is the sole source of the bid–ask spread. Even if there were free entry into the specialist and floor trading business, the expected profit of a specialist need not be zero – or even constant – from trade to trade. Free entry and risk neutrality can only imply that the expected profit of a new entrant, net of inventory holding costs, the opportunity cost of the entrant's time, etc. must be zero [Phillips and Smith (1980)].[11] However, the spread from such sources has a qualitatively different effect on the serial correlation of price changes, and the correlation coefficient can be used to determine the relative magnitudes of the sources of the spread. Moreover, the average spread from sources other than informational asymmetries declines as one over the average volume of trade, whereas the average spread from adverse selection need only decline as one over the square root of the average volume of trade.

The spread can be important both because of its welfare implications, which we have hinted at but not fully analyzed in this paper, and because it offers a potential explanation of the measured excess returns on small firms just after their fiscal year ends. To the extent that these fiscal year ends differ from the tax years of investors in small firms, this explanation is distinguishable from explanations based on the tax consequences of investing.

Appendix: Proof of Proposition 6

First consider $t \in [T_0 - \tau, T_0)$. Then

$$\mathrm{E}_t\left[B^*_{t+\tau}/A^*_t\right] = \mathrm{e}^{-i(T_0-t-\tau)}\left(\mathrm{E}_t[B_{t+\tau}]/\mathrm{e}^{-r_t(T_0-t)}A_t\right) = \mathrm{e}^{i\tau}.$$

That is,

$$r_t = i + \log\left[\left(\frac{1}{\mathrm{E}_t[B_{t+\tau}]/A_t}\right)^{T_0-t}\right].$$

For $t + \tau < T_0$, and $t + n\tau \in [T_0 - \tau, T_0)$, suppose

$$r_{t+\tau} = i + \left(n/(T_0 - t - \tau)\right)\log(k_{t+\tau}).$$

Then,

$$\mathrm{e}^{i\tau} = \frac{\mathrm{E}_t\left[\exp\left(-\left(i + \left(n/(T_0 - t - \tau)\right)\log(k_{t+\tau})\right)(T_0 - t - \tau)\right)B_{t+\tau}\right]}{A_t \exp\left(-r_t(T_0 - t)\right)},$$

[11] It is hard to reconcile free entry and risk-averse specialists without also including transactions costs. Risk aversion of the specialists would also contribute to the spread [Ho and Stoll (1981)].

and

$$r_t = i + \frac{1}{T_0 - t}\log\left(\frac{A_t}{\mathrm{E}_t\left[B_{t+\tau}/k_{t+\tau}^n\right]}\right)$$

$$= i + \log\left[\left(\frac{1}{\mathrm{E}_t\left[B_{t+\tau}/A_t k_{t+\tau}^n\right]}\right)^{1/T_0 - t}\right].$$

Now,

$$\mathrm{E}_t\left[\frac{B_{t+\tau}}{A_t}\left(\frac{1}{k_{t+\tau}}\right)^n\right] = \mathrm{E}_t\left[\frac{B_{t+\tau}}{A_t}\mathrm{E}_{t+\tau}\left[\frac{B_{t+2\tau}}{A_{t+\tau}} \cdots \frac{B_{t+(n+1)\tau}}{A_{t+n\tau}}\right]\right]$$

$$= \mathrm{E}_t\left[\frac{B_{t+\tau}}{A_t}\frac{B_{t+2\tau}}{A_{t+\tau}} \cdots \frac{B_{t+(n+1)\tau}}{A_{t+n\tau}}\right]$$

$$= (1/k_t)^{n+1} \leq 1,$$

since

$$\mathrm{E}_{t+(k+1)\tau}\left[\frac{B_{t+(k+1)\tau}}{A_{t+k\tau}}\right] \leq \mathrm{E}_{t+k\tau}\left[\frac{P_{t+(k+1)\tau}}{A_{t+k\tau}}\right] = \frac{P_{t+k\tau}}{A_{t+k\tau}} \leq 1.$$

Thus, $r_t = i + ((n+1)/(T_0 - t))\log(k_t)$. The (backwards) induction argument shows that r_t is as claimed. Q.E.D.

References

Akerlof, G.A., 1970, The market for 'lemons', qualitative uncertainty and the market mechanism, Quarterly Journal of Economics 84, 488–500.

Arbel, Avner and Paul Strebel, 1981, Neglected firm effect and the inadequacy of the capital asset pricing model, Unpublished working paper no. 81-08 (State University of New York at Binghamton, Binghamton, NY).

Amihud, Yakov and Haim Mendelson, 1980, Dealership market, Journal of Financial Economics 8, 31–53.

Bagehot, Walter (pseud.), 1971, The only game in town, Financial Analysts Journal 22, 12–14.

Banz, Rolf, 1981, On the relationship between return and market value of common stocks, Journal of Financial Economics 9, 3–18.

Blume, Marshall E. and Robert F. Stambaugh, 1983, Biases in computed returns: An application to the size effect, Journal of Financial Economics 12, 387–404.

Copeland, Thomas and Dan Galai, 1983, Information effects on the bid ask spread, Journal of Finance 38, 1457–1469.

Fama, Eugene F., 1970, Efficient capital markets: A review of theory and empirical work, Journal of Finance 25, 383–417.

Garman, Mark B., 1976, Market microstructure, Journal of Financial Economics 3, 257–275.

Grossman, Sanford, 1976, On the efficiency of competitive stock markets when traders have diverse information, Journal of Finance 31, 573–585.

Ho, Thomas and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, Journal of Financial Economics 9, 47–73.

Keim, Donald B., 1983, Size related anomalies and stock return seasonality: Further empirical evidence, Journal of Financial Economics 12, 13–32.

Milgrom, Paul R. and Nancy Stokey, 1982, Information, trade and common knowledge, Journal of Economic Theory 26, 17–27.

Ohara, Maureen and George Oldfield, 1982, Microeconomics of market making, Working paper (Graduate School of Business, Cornell University, Ithaca, NY).

Phillips, Susan M. and Clifford W. Smith, Jr., 1980, Trading costs for listed options: The implications for market efficiency, Journal of Financial Economics 12, 179–201.

Roll, Richard, 1984, A simple measure of the effective bid/ask spread in an efficient market, Journal of Finance 39, 1127–1139.

Stoll, Hans R. and Robert E. Whaley, 1983, Transactions costs and the small firm effect, Journal of Financial Economics 12, 57–79.