Market Microstructure and Market Efficiency

Paul G. Mahoney*

I. INTRODUCTION ........................................................................................................ 541
II. MARKET STRUCTURE I: INTERMEDIARIES ............................................................ 543
III. MARKET STRUCTURE II: INFORMATIONAL ASYMMETRY ........................................... 546
IV. CONCLUSION ........................................................................................................... 549

I. INTRODUCTION

One of the most important developments in economic theory over the past twenty years is a focus on institutions as constraints on economic agents. Laws, social norms, and other rules that channel behavior into particular directions have become central objects of study. Prior work in economics tended either to abstract away from institutions or to take them for granted. A justification for these approaches was the view that institutions arise when they are efficient in the sense of economizing on transactions costs.¹ In that framework, institutions are an output of the economic system, like prices or production.

A more recent tradition treats institutions as inputs that help explain some of the cross-sectional variation in prices, economic growth, financial development, or other output variables.² This is not to say that institutions are exogenously imposed, but simply to note that institutions (however they may arise) may survive even if they are inefficient.³ Earlier analyses posited that institutions, like individuals, are subject to the competition of the market and only those that are efficient will survive over the long run.⁴ This ignores, however, the fact that an existing regime may generate rents or settled expectations for some subset of agents, and those agents may resist change.

Although the institutional approach is often associated with economic history, it has more recently taken deep root within financial economics. Economists have devoted considerable attention to the role that corporate law, for example, plays in the operation

---

¹ See RONALD COASE, The Theory of the Firm, 4 ECONOMICA 38b (1937).
² Douglass North is one of the main proponents of this view. See, e.g., DOUGLASS C. NORTH, INSTITUTIONS, INSTITUTIONAL CHANGE, AND ECONOMIC PERFORMANCE (1990).
³ For one example of a model in which social norms survive an evolutionary process despite their inefficiency, see Paul G. Mahoney & Chris Sanchirico, Competing Norms and Social Evolution: Is the Fittest Norm Efficient?, 149 U. PA. L. REV. 2027 (2001).
⁴ See Arman A. Alchian, Uncertainty, Evolution, and Economic Theory, 58 J. POL. ECON. 211 (1950).
of the financial system. This much is well-known to legal academics, who have written extensively on the “law matters” hypothesis and its relevance to various debates within corporate law scholarship.

By contrast, legal academics have paid far less attention to another important strand of finance scholarship that examines how rules and conventions affect financial markets. The literature on market microstructure considers the detailed interactions between traders that culminate in transactions. Actual securities markets are highly-structured environments with detailed rules and procedures that govern the way orders travel from customers to brokers to the place at which execution (the creation of a binding contract of purchase and sale) takes place. Securities markets also involve diverse personnel such as retail investors, institutional investors, brokers, market makers, exchange specialists, and so on. Market microstructure models attempt to capture some of these important features of actual securities markets and study the price formation process at the level of individual interactions.

My objective in this Essay is to sketch out how Gilson and Kraakman’s discussion of the processes that produce market efficiency could be expanded to take account of the market microstructure literature. I will argue that two main features of the Gilson and Kraakman framework require revision in light of intervening developments. One is the central notion that efficiency is, in effect, a measure of how rapidly a new equilibrium price is established after the release of new information. This remains a powerful heuristic device, but, as I will note, it cannot be literally true once we consider time scales measured in minutes rather than days. On such scales, equilibrium is not simply a function of information, but also of the detailed structure of the trading market. Moreover, at that scale we cannot consider the market value of the stock to be identical to the price of the last trade.

The second revision is to one of the most creative ideas in Gilson and Kraakman’s analysis, which is the notion that four different mechanisms—universally informed trading, professionally informed trading, derivatively informed trading, and uninformed trading—operate in different circumstances to produce the new equilibrium, each at a characteristic speed. In fact, once we put the price-discovery process into a game theoretic framework and consider the response of a trader to the possibility that his counterparty is informed, we see that a fifth mechanism is at work which is both more powerful and more rapid. I will call this “order flow informed price setting” and argue that it is more important than professionally informed trading or derivatively informed trading as Gilson and Kraakman describe them. Order flow informed price setting refers to the attempts of the specialist or market maker (who do not play an explicit role in

Gilson and Kraakman's analysis) to adjust prices based on the buy and sell orders he observes. One important lesson from considering the market maker's optimization problem, and the dynamic response of informed and uninformed traders, is that even relatively small trades can move prices. Gilson and Kraakman, by contrast, focus on large trades and price pressure as the primary channels for "information leakage."  

II. MARKET STRUCTURE: INTERMEDIARIES

Like the early financial literature on which it builds, Gilson and Kraakman's article does not take explicit account of the structure and rules of the markets on which equities trade. Prices are the starting point for the analysis, and those prices emerge from a black box known as "the exchange" or "the market." Whether different prices might emerge from a market with a different design is not considered.

This feature of the analysis reflects the empirical origins of the literature on market efficiency. The classic early studies of market efficiency relied on daily closing prices. From the perspective of a market like the New York Stock Exchange ("NYSE"), on which millions of transactions can take place in a single trading day, daily data is low-frequency data. It is not surprising, then, that early work on market efficiency took closing prices themselves as the object of study and abstracted away the detailed mechanics of individual trades.

In the last decade, new sources of intraday price data have become available, such as the New York Stock Exchange's Trade and Quotation (TAQ) database, which provides details of intraday trades and specialist quotes. This has enabled economists to study market efficiency on a time scale that encourages attention to the trading process. It is hard to talk about prices reflecting new information within fifteen seconds, as found in a recent study, without considering how a trader can manage the steps leading to execution of a trade within fifteen seconds of hearing a news report. Communications and trading technology, such as the NYSE's SuperDOT automated order-execution system, makes such rapid execution possible.

Paying explicit attention to the trading system can enrich our understanding of how efficiency comes about, but it also complicates the very concept of market efficiency. To take a simple example of such a complication, imagine a company with 100 million shares outstanding that trade on the NYSE. Let us further imagine that at a particular point in time, the spread quoted by the exchange specialist is $0.10—that is, the specialist stands ready to buy at $20.00 and to sell at $20.10. Consider two consecutive trades. One originates as a customer's sell order and executes at the specialist's bid price of $20.00. The next, a few seconds later, originates as a customer's buy order and executes at the ask price of $20.10.

From the perspective of market efficiency, what do we make of the ten cent change in price from one trade to the next? Do we conclude that the "fundamental" value of the company has somehow increased by $10 million in a matter of seconds? If not, then must

---

9. See Gilson & Kraakman, supra note 8, at 572-75.
we conclude that the market is inefficient because it produced a price change that does not reflect a change in the present value of future dividends? The first seems thoroughly implausible, especially because the sell and buy orders could easily have arrived in the opposite order. The second, though, seems perverse—it would be bizarre to observe the market accommodating a pair of traders within seconds of one another at a transaction cost of approximately one-half of one percent each and call it evidence of inefficiency.

This simple example implicates one of the two main concerns of the market microstructure literature. Intermediaries are a central feature of real securities markets, and the literature seeks to make them a feature of the theory of markets. Clearly intermediaries must earn a return or they will not remain in the market. Assume for the sake of argument that intermediaries contribute something valuable (a point to which I will return momentarily). The fact that intermediaries earn a return sufficient to cover the costs of providing that service, then, cannot be considered a source of inefficiency. Or, put differently, a definition of "efficiency" that rules out such a return will not be of much use in understanding actual securities markets.

We can easily resolve the above paradox by considering the "true" value of the security to be, not the price of the most recent trade, but rather some figure between the market maker's bid and ask price, perhaps the midpoint of $20.05. The five cent difference between that value and the price paid by the purchaser or received by the seller is, in effect, a fee paid to the market maker in return for his services.

What are those services? First, the market maker smoothes random variation in the arrival of buy and sell orders. It may occur by chance that on day one, more sell orders than buy orders arrive at the market, and on day two the reverse is true. By his willingness to hold inventory and trade when there is no other counterparty available, a market maker offers "immediacy." Other traders know that they can purchase or sell at the current market price at any time the market is open.

Rather than rely on (and pay) the market maker for immediacy, a customer might submit a limit order. Unlike a market order, which is an offer to transact at the then-current price, a limit order is an offer to transact at a specified price. If the market maker's current bid and ask prices are $20.00 and $20.10, respectively, a would-be buyer might either purchase at the ask price or submit a limit order at, for example, $20.05, hoping that a seller comes along soon and takes the higher price rather than the market maker's bid. By going the latter route, the trader foregoes the certainty of execution by the market maker in return for the possibility of transacting at a better price. Roughly speaking, it is analogous to selling one's used car directly to another consumer through a newspaper ad rather than to a dealer—the purchaser and seller avoid the dealer's markup, but the seller loses the assurance of a prompt transaction. Indeed, no matter how many traders there are, there is always a probability of less than one that a limit order will execute. Thus, the market maker provides a service by offering certain execution.

In order to offer that service, however, the market maker must hold inventory,

15. See O'Hara, supra note 7, at 46.
thereby subjecting herself to market risk and financing costs, which she must recover in equilibrium. For present purposes, it is important to note that the market maker’s pursuit of her own utility may have effects on market prices that seem, at a sufficiently fine scale, to be inconsistent with efficiency.

Imagine, for example, a market with a monopolist market-maker who posts bid and ask prices for all securities. Assume that trading takes place at discrete intervals, with one trade per interval. Further assume that there is no release of information over some period of interest, so that the true value of a particular stock remains $25.00 at all relevant times. Our market maker has solved her own optimization problem and, given her risk preferences, the volatility of the stock, and other relevant variables, has concluded that her optimal inventory position is 10,000 shares and the optimal spread is $0.20.

At time $t$, the market maker holds her preferred inventory and sets the spread so that the true value of the stock is centered between her bid price ($24.90) and ask price ($25.10). In the first trading period, a customer buys 3000 shares, so that the market maker is a net seller, and at time $t+1$ her inventory is 7000 shares. The true value of the stock is unchanged, but nevertheless the market maker would prefer to buy rather than sell in the next period. Accordingly, she re-centers her bid and ask prices to $24.95 and $25.15. Note that the spread is unchanged—only the location is different. If we assume, reasonably, that the relative probability that the next customer will wish to sell rather than buy is increasing in the price, then it is more likely that the market maker will be a buyer than a seller in the second trading period.

Now consider what happens as this process evolves. A buy (sell) order in any period induces the market maker to re-center her bid and ask prices at the end of the period in order marginally to encourage a sell (buy) order in the next period. Because customer buy orders execute at the bid and sell orders at the ask price, this means that over the short run, prices may seesaw back and forth between the bid and ask. The result is (negative) serial correlation of price changes, which over a larger time scale would be considered evidence against efficiency. On a trade-by-trade time scale, however, it is a predictable consequence of the market maker’s inventory management problem.

We cannot look at price changes over very small time periods and be confident that they reflect information to the extent we can over larger time periods. This, in turn, complicates the effort to define efficiency as the speed with which prices adjust to information. The negative implication of the definition is that prices do not adjust absent information, which at small time scales is not the case. At the level of trade-by-trade data, we have to look not at the prices of completed trades, but at the interval inside the bid-ask spread, as our measure of the value of a security.

So far, we have considered the price effects of the market maker’s efforts to provide immediacy. The market maker offers an additional, and less obvious, service. By offering to trade at announced prices, the market maker in effect offers to lose money to better-informed traders. The market maker’s spread, then, must also provide sufficient compensation for those losses. This implicates the second principal concern of the market microstructure literature: informational asymmetry.

17. See Joel Hasbrouck, Trades, Quotes, Inventories and Information, 22 J. FIN. ECON. 229 (1988).
III. Market Structure II: Informational Asymmetry

Market efficiency implies that prices move in response to information. Because prices are generated by trades, in order to talk meaningfully about the process by which prices change in response to information, we must talk about how trades respond to information and how prices respond to trades. The latter point is a problem if we adopt a simple microeconomic model of perfect competition, because in such a model any individual's purchases or sales do not affect the market clearing price.

Gilson and Kraakman's discussion of professionally informed trading and derivatively informed trading largely assumes that trades move prices through price pressure. That is, they implicitly depart from the perfectly competitive framework by assuming that some trades are "large" in relation to the size of the market. These large trades cause prices to rise or fall, which in turn causes other traders to infer that they are informed. Thus, Gilson and Kraakman note:

In theory, at least, the logic of price decoding is simple. When trading on inside information is of sufficient volume to cause a change in price, this otherwise inexplicable change may itself signal the presence of new information to the uninformed.\(^\text{18}\)

The market microstructure literature, beginning with a foundational article published shortly after Gilson and Kraakman's article, has focused on how traders' attempts to maximize profits in the face of informational asymmetry cause prices to converge to a fully-informed equilibrium.\(^\text{19}\) The beginning insight is that trading with someone who has superior information imposes an opportunity cost. If I sell to a better-informed trader today, I forego the (higher) price I would receive once her information is revealed. Traders—particularly frequent traders like market makers—will try to protect themselves against those losses. The dynamic interaction between informed and uninformed traders causes prices to adjust.

So long as market makers realize that some traders are informed, it does not take a large trade to move prices. Indeed, we can start by looking at a limiting case in which all trading is informed. Imagine that a market maker trades in a stock that has a value either of $20.00 or $30.00, each with probability 0.5. The market maker does not know the true value, but other traders may, and the market maker is aware of the latter fact. Further, assume that all trades are motivated by information—there is no liquidity trading, and this is also known to the market maker. Finally, assume that traders do not behave strategically—that is, they do not attempt to deceive the market maker as to the content of their private information.

In this simple setting, a trader's desire to purchase fully conveys the private information that the value of the stock is $30.00, and the desire to sell shows that the value is $20.00. The dealer, then, should set his bid at $20.00 and ask at $30.00 to avoid regret. Note that the midpoint of the spread, or $25.00, is the expected value of the stock.

---

18. Gilson & Kraakman, supra note 8, at 575 (emphasis added).
prior to any trading. However, the purchase or sale of a single share will move the price to the true value. The informed trader, moreover, earns no profit because she buys at $30.00 or sells at $20.00. This simple market can accordingly be considered strong-form efficient because traders are unable to profit from private information. Yet it does not require large trades to move prices.

Of course, the same fact guarantees that this simple market cannot survive. An early, and intuitive, result in the market microstructure literature is that if the only motivation for trading is information, there will be no trades because an informed trader’s desire to trade (regardless of the size of the order) reveals that the price is incorrect. A more realistic and useful model must incorporate some actors who trade even though they suffer losses to informed traders. These critical inhabitants of microstructure models are called “liquidity” traders, capturing the notion that they trade to convert excess income into savings or savings into consumption. Although their motivations are not typically examined in detail, we might say that these traders implicitly view their losses to the better informed as a cost of using securities as a savings vehicle.

In a market that includes liquidity traders, the market maker cannot be certain that a buy order indicates the stock is undervalued. Instead, the market maker is left with a Bayesian updating problem. Prior to trading, he has a belief as to the value of the stock (or, more realistically, its distribution). To avoid regret, he must determine the conditional expectation of the stock price, given the order flow he observes. Intuitively, a preponderance of buy over sell orders causes the market maker to increase his estimate of the stock’s value, while a preponderance of sell orders causes the reverse. More formally, the market maker wishes to calculate $E[V|B]$, where $V$ is the value of the stock, $B$ is the ratio of buy to sell orders, and $E$ is the expectation operator. If we assume the market maker can estimate the conditional density $f(B|V)$ (that is, the relative probabilities of observing the realized ratio $B$ given possible values of $V$), he can use Bayes’ Theorem to calculate the desired conditional expectation. Having done so, he sets bid and ask prices so as to generate zero expected profits, assuming competition among market makers. The zero expected profit condition means that the profits the market maker earns from uninformed traders just equal the losses to informed traders. In game theory parlance, informed and uninformed traders “pool” in the sense that the market maker cannot distinguish them ex ante and quotes the same prices to each. This feature, however, is not entirely realistic. Uninformed traders will not wish to subsidize the market maker’s losses from informed traders if they can help it. Thus, they may engage in efforts to distinguish themselves—to create a separating rather than a pooling equilibrium.

This dynamic response guarantees that the market maker cannot simply use the size of an order as a proxy for information. Consider a large and frequent trader, such as a pension fund, that happens to be uninformed. The fact that it buys and sells in large quantities means that it is worth investing in demonstrating to the market maker that it is uninformed—perhaps even worth investing in a reputation for being uninformed—if the result will be a smaller recognized spread on its trades.

21. See O’HARA, supra note 7, at 77-88 for a discussion of Bayesian learning models.
This is one function of the block trading mechanism. Unlike the anonymous market on the floor of the exchange, the block trading market operates through negotiation off the floor, between market makers and brokers who may reveal the identity of their customer and the customer's reason for trading. The customer may disclose information to the market maker to assure the market maker it is not informed. For example, a pension fund might disclose that it needs to sell a large quantity of securities to cover a monthly distribution to its beneficiaries. Alternatively, a large trader may buy or sell a basket of securities rather than a single security to reduce the potential benefit of firm-specific information about any one stock. In either case, the customer's objective is to minimize the market maker's adverse movement of the bid price. Ideally, an uninformed trader may even buy or sell "inside the spread," or at a more attractive price than that which the market maker has published.

Different trading mechanisms may also produce a tradeoff between rapid execution and lower transaction costs. An informed trader needs to trade before the information is revealed, but an uninformed trader can afford to be more patient. Thus, as discussed above, an uninformed trader may use a limit order rather than a market order, accepting a delay in execution in return for a better price. Similarly, an uninformed trader may be willing to wait until the market closes and trade through an after-hours crossing network that executes trades at the closing price at reduced transaction costs.

Each of these strategies makes most sense for a large uninformed trader, for whom a small per-share transaction cost saving can amount to a large sum of money. But that means that a large informed trader will tend to tip his hand by not using one of these strategies. Accordingly, such traders face a tradeoff between trading in larger quantities at less attractive prices or smaller quantities at more attractive prices. The result, not surprisingly, is that the price change associated with a particular trade is largest for medium-sized trades (500 to 9999 shares).

These details can be integrated into a more general lesson. Although it is certainly true that the reactions of investors who "watch the ticker" for signs of information are relevant to the process by which a new equilibrium price is established, the more direct and important reaction is that of the counterparty to a particular trade. The desire to minimize opportunity losses to better-informed traders is a constant in the lives of market makers and other frequent traders. They use the order flow as a statistic for the information possessed by the traders arriving at the market. Even before a trade occurs, a specialist, market maker, or other trader will consider the possibility that the buy or sell

25. Note that the market maker's bid and ask prices are automatically available only for a particular quoted size. Any larger order must be individually negotiated, and potentially at a less favorable price than the quoted bid or ask. One possible response of a market maker who believes that the degree of informational asymmetry has temporarily increased is to reduce his quoted size.
order is motivated by information and may adjust the price at which he is willing to transact accordingly. This adjustment process does not require that the order be large enough to generate price pressure independent of its informational content. Indeed, the adjustment process creates the counterintuitive result that large trades are more likely to be uninformed, because they can execute at an attractive price only if the trader is able to offer some credible evidence that it is uninformed.

This mechanism of order flow informed price setting should be reasonably intuitive to anyone who has bought or sold an asset with a counterparty who appears more expert with respect to that type of asset. The very frequency with which I might be offered a “genuine Rolex watch” on the streets of certain cities, and the eagerness of the sellers to conclude a trade rapidly, will lead me to suspect that the watches are not genuine, even though I have no particular skill at distinguishing the real item from a fake. Similarly, a house seller who is inundated with offers immediately after his for-sale sign goes up will conclude that he has set his price too low. These insights, too, are a function of considering the process of efficiency on a trade-by-trade basis.

IV. Conclusion

The more attention we pay to the personal optimization problems of individual actors in the equity market, the more complex and fascinating the process of equilibration looks. The early literature on market efficiency looked at prices in isolation, largely ignoring the trading process that produced those prices. It is therefore entirely natural and understandable that Gilson and Kraakman focused on information itself—that is, the raw data relevant to the value of a business—and how that information makes its way from a small group of agents inside the business to the broad mass of traders. The market microstructure literature, by contrast, focuses on orders, or offers to transact, as statistics from which others might infer the existence of information. The resulting updating of beliefs can cause prices to adjust even when trading volumes are relatively modest.

One reason this process works so rapidly is the fact that market makers engage in it for a living. They presumably learn over time to distinguish informed from uninformed traders and, therefore, to avoid large opportunity losses. The existence of intermediaries, however, also means that prices are not determined solely by information, but also by the cost to the intermediaries of providing their services to the market. Thus, a more complete discussion of efficiency cannot consider only the prices of completed transactions, but must also consider the bid-ask spread and its components.