CHAPTER 2

How Markets Slowly Digest Changes in Supply and Demand

Jean-Philippe Bouchaud
Science & Finance, Capital Fund Management, Paris

J. Doyne Farmer
Santa Fe Institute and LUISS Guido Carli, Rome

Fabrizio Lillo
University of Palermo and Santa Fe Institute

Note: We want to thank Klaus Reiner Schenk-Hoppé and Thorsten Hens for giving us the opportunity to write this chapter and for their patience. We also want to thank our collaborators who helped us to shape our understanding of the subtle world of market microstructure and impact: R. Almgren, S. Ciliberti, C. Deremble, Z. Eisler, A. Gerig, L. Gillemot, E. Henry, A. Joulin, J. Kockelkoren, E. Moro, G. La Spada, Y. Lemperiere, R.N. Mantegna, S. Mike, A. Ponzi, M. Potters, G. Vaglica, H. Waelbroeck, J. Wieslinger, and M. Wyart. JDF gratefully acknowledges support from Bill Miller, Barclays Bank, and NSF grant HSD-0624351.

HANDBOOK OF FINANCIAL MARKETS: DYNAMICS AND EVOLUTION

Abstract

In this chapter we revisit the classic problem of *tâtonnement* in price formation from a microstructure point of view, reviewing a recent body of theoretical and empirical work explaining how fluctuations in supply and demand are slowly incorporated into prices. Because revealed market liquidity is extremely low, large orders to buy or sell can only be traded incrementally, over periods of time as long as months. As a result order flow is a highly persistent long-memory process. Maintaining compatibility with market efficiency has profound consequences on price formation, on the dynamics of liquidity, and on the nature of impact. We review a body of theory that makes detailed quantitative predictions about the volume and time dependence of market impact, the bid–ask spread, order book dynamics, and volatility. Comparisons to data yield some encouraging successes. This framework suggests a novel interpretation of financial information, in which agents are at best only weakly informed and all have a similar and extremely noisy impact on prices. Most of the processed information appears to come from supply and demand itself, rather than from external news. The ideas reviewed here are relevant to market microstructure regulation, agent-based models, cost-optimal execution strategies, and understanding market ecologies.

Keywords: financial markets, market microstructure, price impact, market ecology

2.1. INTRODUCTION

In this chapter we discuss the slow process by which markets “digest” fluctuations in supply and demand, reviewing a body of work that suggests a new approach to the classic problem of *tâtonnement*—the dynamic process through which markets seek to reach equilibrium.

2.1.1. Overview

The foundation of this approach is based on several empirical observations about financial markets, the most important of which is long memory in the fluctuations of supply and demand. This is exhibited in the placement of trading orders, and corresponds to long-term, slowly decaying positive correlations in the initiation of buying vs. selling. It is observed in all the stock markets studied so far at very high levels of statistical significance. It appears that the primary cause of this long memory is the incremental execution of large hidden trading orders. The fact that the long memory of order flow must coexist with market efficiency (at least in a statistical sense) has a profound influence on price formation, causing dynamic adjustments of liquidity that are strongly asymmetric between buyers and sellers.

This has important consequences for market impact. (By market impact we mean the average response of prices to trades; liquidity refers to the scale of the market impact.[1]) We discuss theoretical work predicting the average market impact as a function of both volume and time. The asymmetric liquidity adjustments needed to maintain compatibility between the long memory of order flow and market efficiency can equivalently be interpreted in terms of the temporal response of market impact, leading to a slow decay of market impact with time.

This work also has important consequences about the interpretation and effect of information in financial markets. In particular, the explanation for market impact that we develop here differs from the standard view in the finance literature, which holds that the shape of the impact function is determined by differences in the information content of trades. The body of work reviewed here instead assumes that the impact of trades depends only on their predictability; for example, that highly predictable trades have little impact, as originally postulated by Hasbrouck (1988). We argue that this is a much simpler explanation that produces stronger predictions, is more plausible from a theoretical point of view, and is more in line with what is observed in the data.

The implications of this work range upward from microstructure—that is, at the level of individual price changes—to patterns of price formation on time scales that can be measured in months. At the microstructure level this work makes several predictions, such as the relationship between market impact, the bid–ask spread, and volatility. It also make predictions about the impact of large trades executed over long periods of time as well as the effect such trades may have in causing clustered volatility.

2.1.2. Organization

In the remainder of the introduction we discuss the motivation and scope of the work described here and discuss our approach to creating a theory for market microstructure, which is somewhat unusual within economics. In Section 2.2 we discuss the institutional aspects of the markets that form the basis of our empirical studies and define some of the terms that will be used throughout the paper. In Section 2.3 we lay out some of the main conceptual issues, discussing the concept of information in finance and its relationship to market efficiency and the important role that liquidity (or more accurately, the lack of liquidity) plays in forming markets. We critique so-called “noise trader” models and present an alternative point of view. In Section 2.4 we present the empirical evidence for long memory in order flow, develop a theory for its explanation based on strategic order splitting, and present evidence that this theory is correct. In Section 2.5 we describe the various types of impact and review the empirical evidence. In Section 2.6 we develop a theory for market impact for each type of impact. In Section 2.7 we discuss the problem of explaining the behavior of the bid–ask spread and compare theory and empirical observations. Section 2.8 discusses the close relationship between liquidity and volatility. In Section 2.9 we discuss models for the order book,

[1] Market impact is closely related to the demand elasticity of price and is typically measured as the return associated with a transaction as a function of volume. Liquidity (as we will use it here) measures the size of the price response to a trade of a fixed size and is inversely proportional to the scale of the impact. If trading a given quantity produces only a small price change, the market is liquid, and if it produces a large price change, it is illiquid.

which can be regarded as models for liquidity. Section 2.10 discusses the problem of trading in an optimal manner to minimize execution costs. Section 2.11 describes recent attempts to characterize trading ecologies of market behavior in short time scales, and Section 2.12 presents our conclusions.

2.1.3. Motivation and Scope

Markets are places where buyers meet sellers and the prices of exchanged goods are fixed. As originally observed by Adam Smith, during the course of this apparently simple process, remarkable things happen. The information of diverse buyers and sellers, which may be too complex and textured for any of them to fully articulate, is somehow incorporated into a single number—the price. One of the powerful achievements of economics has been the formulation of simple and elegant equilibrium models that attempt to explain the end results of this process without going into the details of the mechanisms through which prices are actually set.

There has always been a nagging worry, however, that there are many situations in which broad-brush equilibrium models that do not delve sufficiently deeply into the process of trading and the strategic nature of its dynamics may not be good enough to tell us what we need to know; to do better we will ultimately have to roll up our sleeves and properly understand how prices change from a more microscopic point of view. Walras himself worried about the process of *tâtonnement*, the way in which prices settle into equilibrium. While there are many proofs for the existence of equilibria, it is quite another matter to determine whether or not a particular equilibrium is stable under perturbations—that is, whether prices initially out of equilibrium will be attracted to an equilibrium. This necessarily requires a more detailed model of the way prices are actually formed. There is a long history of work in economics seeking to create models of this type (see e.g., Fisher, 1983), but many would argue that this line of work was ultimately not very productive, and in any case it has had little influence on modern mainstream economics.

A renewed interest in dynamical models that incorporate market microstructure is driven by many factors. In finance, one important factor is growing evidence suggesting that there are many situations where equilibrium models, at least in their current state, do not explain the data very well. Under the standard model prices should change only when there is news, but there is growing evidence that news is only one of several determinants of prices and that prices can stray far from fundamental values (Campbell and Shiller, 1989; Roll, 1984; Cutler et al., 1989; Joulin et al., 2008).[2] Doubts are further fueled by a host of studies in behavioral economics demonstrating the strong boundaries of rationality. Taken together this body of work calls into question the view that prices always remain in equilibrium and respond instantly and correctly to new information.

The work reviewed here argues that trading is inherently an incremental process and that for this reason, prices often respond slowly to new information. The reviewed body of theory springs from the recent empirical discovery that changes in supply and

[2] See Engle and Rangel (2005) for a dissenting view.

demand constitute a long-memory process; that is, that its autocorrelation function is a slowly decaying power law (Bouchaud et al., 2004; Lillo and Farmer, 2004). This means that supply and demand flow in and out of the market only very gradually, with a persistence that is observed on time scales of weeks or even months. We argue that this is primarily caused by the practice of order splitting, in which large institutional funds split their trading orders into many small pieces. Because of the heavy tails in trading size, there are long periods where buying pressure dominates and long periods where selling pressure dominates. The market only slowly and with some difficulty "digests" these swings in supply and demand. To keep prices efficient in the sense that they are unpredictable and there are not easy profit-making opportunities, the market has to make significant adjustments in liquidity. Understanding how this happens leads to a deeper understanding of many properties of market microstructure, such as volatility, the bid–ask spread, and the market impact of individual incremental trades. It also leads to an understanding of important economic issues that go beyond market microstructure, such as how large institutional orders impact the price and in particular how this depends on both the quantity traded and on time. It implies that the liquidity of markets is a dynamic process with a strong history dependence.

The work reviewed here by no means denies that information plays a role in forming prices, but it suggests that for many purposes this role is secondary. In the last half of the twentieth century, finance has increasingly emphasized information and deemphasized supply and demand. The work we review here brings forward the role of fluctuations and correlations in supply and demand, which may or may not be exogenous. As we view it, it is useful to begin the story with a quantitative description of the properties of fluctuations in supply and demand. Where such fluctuations come from doesn't really matter; they could be driven by rational responses to information or they could simply be driven by a demand for liquidity. In either case, they imply that there are situations in which order arrival can be very predictable. Orders contain a variable amount of information about the hidden background of supply and demand. This affects how much prices move and therefore modulates the way in which information is incorporated into prices. This notion of information is internal to the market. In contrast to the prevailing view in market microstructure theory, there is no need to distinguish between "informed" and "uninformed" trading to explain important properties of markets, such as the shape of market impact functions or the bid–ask spread.[3]

We believe that the work here should have repercussions on a wide gamut of questions:

- At a very fundamental level, how do we understand why prices move, how information is reflected in prices, and what fixes the value of the volatility?
- At the level of price statistics, what are the mechanisms leading to price jumps and volatility clustering?

[3] Since modern continuous double auction markets are typically anonymous, it is hard to see how the identity of traders could play an important role in the size of the price response to trades. See the discussion in Section 2.3.

- At the level of market organization, what are the optimal trading rules to ensure immediate liquidity and orderly flow to investors?
- At the level of agent-based models, what are the microstructural ingredients necessary to build a realistic agent-based model of price changes?
- At the level of trading strategies and execution costs, what are the consequence of empirical microstructure regularities on transaction costs and implementation shortfall?

We do not wish to imply that these questions will be answered here—only that the work described here bears on all of them. We will return to discussing the implications in our conclusions.

2.1.4. Approach to Model Building

Because this work reflects an approach to model building that many economists will find unfamiliar, we first make a few remarks to help the reader understand the philosophy behind this approach. Put succinctly, our view is that the enormous quantities of data that are now available fundamentally change the approach one should take to building economic theories about financial markets.

In recent years the computer has made it possible to automate markets, has enabled an explosion in the amount of recorded data, and has made it possible to analyze unprecedented quantities of information. Financial instruments are now typically standardized, stable entities that are traded day after day by many thousands of market participants. Modern electronic markets offer an open and transparent environment that allows traders across the world to get real-time access to prices, and most important for science, makes it possible to save detailed records of human decision making. The past decades have seen an explosion in the volume of stored data. For example, the total volume of data related to U.S. large caps on, say, October 2, 2007, was 57 million lines, approximately a gigabyte of stored data. The complete record of world financial activity is more than a terabyte per day. Each market has slightly different rules of operation, making it possible to compare market structures and the way they affect price formation and, most important of all, to look for patterns of behavior that are common across all market structures. The system of world financial markets can be viewed as a huge social science experiment in which profit seekers spend large quantities of their own money to collect enormous quantities of data for the pleasure of scientists.

With so much data it becomes possible to change the style in which economics is done. When one has only a small amount of noisy data, statistical testing must be done with great care, and it is difficult to test and reject competing models unless the differences in their predictions are very large. Data snooping is a constant worry. In contrast, with billions of data points, if an effect is not strong enough to leap out of the noise, it is unlikely to be of any economic importance. Even more important is the effect this has on developing and testing theories. With a small data set inference requires strong priors. This fosters an approach in which one begins with pure theory and tests the resulting models only after they are fully formulated; there is less opportunity to let the data speak for itself. Without great quantities of data it is difficult to test a theory in a fully quantitative manner, and so predictions of theories are typically qualitative.

The work reviewed in this chapter takes advantage of the size of financial data sets by strongly coupling the processes of model formation and data analysis. This begins with a search for empirical regularities, that is, behaviors that under certain circumstances follow consistent quantitative laws. Even though such effects do not have the consistency of the laws of physics, one can nonetheless be somewhat more ambitious than simply trying to establish a set of "stylized facts." An attempt is made to describe regularities in terms that are sufficiently quantitative so that theories have a clear target and can thus sensibly make strongly falsifiable predictions. A key goal of such theories is, of course, to understand the necessary and sufficient conditions for regularities.

The approach for building theory described here is phenomenological. That is, it does not attempt to derive everything from a set of first principles but rather simply tries to connect diverse phenomena to each other to simplify our description of the world. Many economists will be uncomfortable with this approach because it often lacks "economic content,"—that is, the theories developed do not invoke utility maximization. In this sense this body of work lies somewhere between pure econometrics and what is usually called a theory in microeconomics. Even though the models infer properties of agent behavior and connect them to market properties such as prices, there is no attempt to derive the results from theories that maximize preferences. Instead we content ourselves with weaker assumptions, such as market efficiency. Given all the empirical problems surrounding the concept of utility, we view this as a strength rather than a weakness.

The work described here is still in an early stage and is very much in flux; many of these results are quite new, and indeed our own view is still changing as new results appear.

2.2. MARKET STRUCTURE

All of the work described here is based on results from studying stocks from the London, Paris, New York (NYSE and NASDAQ), and Spanish stock markets. These markets differ in their details, but they all do at least half of their trading (and in some cases all their trading) through a continuous double auction. "Auction" indicates that participants may place quotes (also called *orders*) stating the quantities and prices at which they are willing to trade; "continuous" indicates that they can update, cancel, or place new quotes at any time, and "double" indicates that the market is symmetric between buyers and sellers.[4]

[4] There are some small exceptions to symmetry between buying and selling, such as the uptick rule in the NYSE, but these are relatively small effects.

There are some important differences in the way these markets are organized. The NYSE was unusual (until the end of 2007) in that each stock has a designated specialist who maintains and clears the limit order book. The specialist can see the identity of all the quotes and can selectively show them to others. The specialist can also trade for his own account but has regulatory obligations to "maintain an orderly market." The London Stock Exchange, in contrast, has no specialists. It is completely transparent in the sense that all orders are visible to everyone, but it is completely anonymous in the sense that there is no information about the identity of the participants, and such information is not disclosed even to the counterparties of transactions. The Spanish Stock Market is unusual in that membership codes for quotes are publicly displayed. Thus these exchanges are generically similar but have their own peculiar characteristics.

Markets also differ in the details of the types of orders that can be placed. For example, the types of orders in the London Stock Exchange are called "limit orders," "market orders with limiting price," "fill-or-kill," and "execute and eliminate." To treat these different types simply and in a unified manner, we simply classify them based on whether an order results in an immediate transaction, in which case we call it an *effective market order*, or whether it leaves a limit order sitting in the book, in which case we call it an *effective limit order*. Marketable limit orders (also called *crossing limit orders*) are limit orders that cross the opposing best price and so result in at least a partial transaction. The portion of the order that results in an immediate transaction is counted as an effective market order, whereas the nontransacted part (if any) is counted as an effective limit order; thus in this case a single action by the participant gets counted as two separate orders. Note that we typically drop the term *effective* so that, for example, *market order* means *effective market order*. Similarly, a limit order can be removed from the book for many reasons; for example, because the agent changes her mind, because a time specified when the order was placed has been reached, or because of the institutionally mandated 30-day limit on order duration. We will lump all these together and simply refer to them as *cancellations*.

In addition to continuous double auctions, the London Stock Exchange has what is called the *off-book market* and the New York Stock Exchange has what is called the *upstairs market*. These are both bilateral exchanges in which members can interact in person or via telephone to arrange transactions. Such transactions are then reported publicly at a later time. With exceptions noted in the text, all the results obtained are from the continuous markets.

2.3. INFORMATION, LIQUIDITY, AND EFFICIENCY

The aim of this section is to motivate the empirical study of microstructure in a broader economic context—that of the information content of prices and the mechanisms that can lead to market efficiency. We discuss several fundamental questions concerning how markets operate. The discussion here sets the stage for the detailed quantitative investigations that we report in the following sections. Since one of our main subjects here is market impact, we review and critique the standard model for market impact, which is based on informed vs. uninformed trading.

2.3.1. Information and Fundamental Values

It is often argued that there is a fundamental value for stocks, correctly known to at least some informed traders who buy underpriced stocks and sell overpriced stocks. By doing so they make a profit and, through the very impact of their trades, drive back the price toward its fundamental value. This mechanism is the cornerstone of the theory of efficient markets and is often used to justify unpredictable prices. In such a framework, the fundamental value of a stock can only change with unanticipated news. The scenario is then the following: A piece of news becomes available, and market participants work out how this changes the fundamental price of the stock and trade accordingly. After a (supposedly fast) phase of *tâtonnement*, the price converges to its new equilibrium value, and the process repeats itself. To explain deviations from this picture, one can add a suitable fraction of uninformed trades to add some high-frequency noise.

Is this picture fundamentally correct to explain the reason that prices move and to account for the observed value of the volatility? Judging from the literature, it looks as if a majority of academics still believe that this story is at least a good starting point (but see, for example, Lyons, 2001). Recent empirical microstructure studies open the way to testing in detail the basic tenets and the overall plausibility of the standard equilibrium picture. We hope to convince the reader that the story is in fact significantly different. That is, we argue that an alternative way of looking at events provides superior explanatory power based on a simpler set of hypotheses. Before discussing at length the microstructural evidence for a change of paradigm, we would like at this stage to make several general comments that will be relevant—first on the very notion of fundamental value and information and second on various orders of magnitude and time scales involved in the problem.

Is the fundamental value of a stock or a currency a valid concept in the sense that it can be computed, at least as a matter of principle, with arbitrary accuracy with all information known at time t? The number of factors influencing the fundamental value of a company or of a currency is so large that there should be, at the very least, an irreducible intrinsic error. All predictive tools used by traders, based on economic ratios, earning forecasts, or the like, are based on statistical models detecting trends or mean reversion and are obviously noisy and sometimes even biased. For example, financial experts are known to be on the whole rather bad at forecasting the next earning of a company (see, e.g., Guedj and Bouchaud, 2005). News is often ambiguous and not easy to interpret. But if we accept the idea of an intrinsically noisy fundamental value with some band of width Δ within which the price can almost freely wander, the immediate question is, how large is the uncertainty Δ? Is it very small, say, 10^{-3} in relative terms, or quite a bit larger, say, 100%, as suggested by Black (1986)? If Black is right (which we tend to believe) and the uncertainty in the fundamental value is large, then the information contained in a trade is noisy and the amount of information contained in any given trade is necessarily small. Analysis of price impact makes it clear that the standard deviation of impacts is very large compared to their mean, suggesting that this is indeed the case.

2.3.2. Market Efficiency

Market efficiency is one of the central ideas in finance and appears in many guises. A standard definition of market efficiency (in the informational sense) is that the current price should be the best predictor of future prices; that is, that prices should be a martingale. Another closely related notion is arbitrage efficiency, which in its weakest form states that it should not be possible to make a profit without taking risks; in a stronger form it says that two strategies with the same risk should make the same profits, at least once their usefulness for inclusion in a portfolio is taken into account. Steve Ross, among others, has advocated that market efficiency (rather than equilibrium) should be the core postulate for financial theory (Ross, 2004).

We agree with this point of view, at least in so far as it does not imply believing in allocative efficiency; that is, that prices correctly reflect the underying value of the assets. Strictly speaking a market is allocatively efficient if it is Pareto optimal, in the sense that there is no alternative allocation of prices and holdings that makes someone better off without making someone worse off. This is related to whether or not prices are set at their "proper" values. It is entirely possible to imagine a market in which prices are unpredictable and yet in which there is no sense that prices are set correctly. That is, once we depart from neoclassical equilibrium, a market might be informationally efficient yet allocatively inefficient.

A closely related point is that there are two very different possible explanations for market efficiency:

1. The standard view in economics is that perfect efficiency reflects perfect information processing. Traders process each new bit of information as it arrives, and prices immediately go to their new equilibrium values.
2. An obvious alternative is the standard one that explains randomness in many other fields, such as fluid turbulence. Markets are too complicated to be predictable. Under this explanation prices move randomly because investor behavior is complicated, based on many hidden factors, so to an external observer it is "as if" individual investors are just flipping coins.

The correct explanation is likely to be a mixture of both effects. On one hand markets are inherently complicated, but on the other hand, whatever predictability is left over is substantially removed by arbitrageurs. Under this synthetic view, which we take here, one can simply associate an impact with trades, treat all investors as more or less the same, and adjust the expected impact as needed to preserve efficiency based on factors that derive from the predictability of trades.

Finally we want to emphasize that though we believe that market efficiency is a very useful concept and provides an excellent starting point for developing theories, it is inherently contradictory and is at best an approximation. Markets can only be informationally efficient at first order but must necessarily be inefficient at second order. This was originally pointed out by Milton Friedman, who noted that without informed traders to push prices in the right direction, there is no reason that markets should ever be efficient. If markets were truly efficient, informed traders should make the same profits as anyone else, and there would be no motivation for them to remain in the market. Thus markets cannot be fully efficient.

Even if for many purposes it can be a good approximation to assume that markets are efficient, there are other situations in which deviations from efficiency can be quite important. Understanding how markets evolve from inefficient to efficient states, predicting the necessary level of deviations from efficiency that must persist in steady state, and understanding their role in the way markets function remain areas of investigation that are still largely not understood. This is relevant for our discussions on incorporating information into prices because when we speak about information we must have traders to process that information and trade based on it. It is precisely the market impact of these traders that moves prices. Thus while on one hand market impact is a friction, it can also be viewed as the factor that maintains efficiency, and so it is essential that we properly understand it.

2.3.3. Trading and Information

Informational efficiency means that information must be properly incorporated into prices. Under assumptions of rationality, when all traders have the same information, prices should move more or less automatically, with very little trading (Milgrom and Stokey, 1982; Sebenius and Geanakoplos, 1983). But of course that's not true—people don't have the same information, and even if they did, real people are likely to take different views on what the information means. The empirical fact that there is so much trading supports this idea (Shiller, 1981). Grossman and Stiglitz (1980) developed an equilibrium model in which traders have different information that shows that in this situation, trading and price movements are informative (see also Grossman, 1989). If I know that you are rational, and I know that you have different information than I have, when I see you trade and the price rises I can infer the importance of your information and thus I should change my own valuation.

Intuitively the problem with this view is that even small deviations from rationality and perfect information can lead to incorrect prices and instabilities in the price process. Suppose, for example, that you and I both overestimate how much information the other has. Then when I see you trade I change my valuation too much. When I see you buy, I also buy, but I buy more than I should. To make this slightly more quantitative, let the initial price be p_0 and suppose that after Agent A observes new fundamental information the price rises by f, which might or might not be the correct fundamental level. After Agent A trades, the new price becomes $p_1 = p_0 + f$. Agent B sees the price rise by f, and assuming that Agent A has more information than he really does, he buys and causes the price to rise to $p_2 = p_0 + af$. Then B sees the price rise more than f, so he buys, driving it to $p_3 = p_0 + a^2 f$, and so on. This process is clearly unstable if $a > 1$. The agents either need to know the value of a exactly or they need to be able to adapt a based on information that is not contained in the price. It is difficult to understand how they can do this since by definition if they are not rational, not only do they not have full information, they do not know how much information they have, and they thus cannot know *a priori* the proper value of a. Under deviations from rationality, deviations from fundamentals are inevitable. For a beautiful model where copycats lead to such instabilities, see P. Curty and M. Marsili (2006).

In its extreme version, this is just the kind of scenario that occurs during a bubble (see Bouchaud and Cont, 1998, for an explicit model). Any reasonable investor who lived through the millennium technology bubble experienced this problem. Even though high prices seemed difficult to rationalize based on values, prices kept going up. This led many sanguine investors to lose confidence in their own valuations and to hang onto their shares much longer than they thought was reasonable. If they didn't do this they experienced losses as measured relative to their peers. Under this view, bubbles stem from the problem of not knowing how much information price movements really contain and the feedback effects that occur when most people think they contain more information than they really do. This point of view differs from that in the standard literature on rational bubbles. As we argue here, though not entirely different, there are important contrasts between this view and the standard rational expectations/noise trader models.

2.3.4. Different Explanations for Market Impact

Why is there market impact? We will distinguish three possibilities:

1. *Trades convey a signal about private information.* This idea, discussed in the previous section, was developed by Grossman and Stiglitz (1980). The arrival of new private information causes trades, which cause other agents to update their valuations, which changes prices. In this case it is fair to say that trades cause price changes, since even if there happens to be no information, unless this is common knowledge the observation of a trade is still interpreted as information, which causes the price to change.
2. *Agents successfully forecast short-term price movements and trade accordingly.* This can result in measurable market impact even if these agents have absolutely no effect on prices at all. If an agent correctly forecasts price movements and trades based on this forecast, when this agent buys there will be a tendency for the price to subsequently rise. In this case causality runs backward, that is, because the price is about to rise, agents are more likely to trade in anticipation of it, but a trade based on no information will have no effect.
3. *Random fluctuations in supply and demand.* Even in the standard market-clearing framework, if a given agent increases her demand while other agents keep theirs constant, when the market clears that agent buys and the price rises. Fluctuations in supply and demand can be completely random, unrelated to information, and the net effect regarding market impact is the same. In this sense impact is a completely mechanical—or better, statistical—phenomenon. As we will see in Appendix 2.1, the meaning of this can be subtle and may depend on the market framework.

All three of these possibilities result in identical short-term market impact—that is, a positive correlation of trading volume and price movement—but they are conceptually very different. If some traders really know the "true" price at some time in the future (say, the end of the day, after the market closes), the observation of an excess of buy trades allows the market to guess that the price will move up and to change the quotes accordingly (see Section 2.7.2 on the Glosten-Milgrom model). In this sense, information has progressively included in prices as a function of the observed order flow. In this picture, as emphasized in Hasbrouck (2007), "orders do not *impact* prices. It is more accurate to say that orders *forecast* prices." But if the mechanical interpretation is correct, correlation between price changes and order flow is a tautology. If prices move only because of trades, "information revelation" may merely be a self-fulfilling prophecy that would occur even if the fraction of informed traders is zero. The only possible differences between these pictures come about in the temporal behavior of impact, which we discuss in Section 2.6.

2.3.5. Noise Trader Models and Informed vs. Uninformed Trading

In behavioral finance, the problem of irrational investors is typically coped with by introducing "noise trader" models, in which some agents (the noise traders) are stupid while others are completely rational (Kyle, 1985; DeLong et al., 1990; Shleifer, 2000). Noise trading could be driven by the need for liquidity (here meaning the need to raise capital for other reasons), it could be driven by the desire to reduce risk, or it could be "irrational behavior," such as trend following. The assumption is made that such investors lack the skill or information-processing ability to collect and/or make full use of information. The rational investors, in contrast, are assumed to correspond to skilled professionals. Their trading is perfect in the sense that they know everything. Examples of what they must know include the strategies of all the noise traders and the fraction of capital traded by noise traders as opposed to rational investors. In such models, prices can deviate from fundamental values due to the action of the noise traders and the desire of the rational agents to exploit them as much as possible, but the rational agents always keep them from deviating too much. The rational traders make "informed" trades while the noise traders make "uninformed" trades.

There are several conceptual problems with noise trader models that are clear *a priori*. No one can seriously dispute that traders must have different levels of skill, but is the noise trader approach the right way to model this? Though it might be fine to model a continuum of skill levels as "low" and "high," the idea of identifying the "high" level with perfect rationality postulates a level of skill at the top end that is difficult to imagine. The panoply of strategies used by real traders is large, and financial professionals (and even private investors) are sufficiently secretive about what they do that it is difficult to imagine that even the most skilled traders could fully understand everyone else's behavior.

Another problematic issue is the operational problem of measuring information. For example, under the theory that urgency is a proxy for informativeness, empirical work on the subject has often defined an informed trade as one that is executed by a market order and defined an uninformed trade as one that is represented by a limit order. This goes against the fact that many of the most successful hedge funds make extensive use of limit orders.[5] The only alternative is to use data that contains information about the identity of the agents making the trades. Such data does indeed confirm that professionals perform better than amateurs (Barber et al., 2004), but as mentioned, there is no demonstration that this means they are rational, and other than stating that professionals make larger profits it is impossible to determine whether or not professionals are good enough to be considered rational. (On this point, see also Odean, 1999).

2.3.6. A Critique of the Noise Trader Explanation of Market Impact

One of the most important questions to ask about any theory is what it explains that is not explained by a simpler alternative. Noise trader models have been proposed to explain why market impact is a concave function of trading volume. The empirical evidence for this concept is discussed in detail in Sections 2.5 and 2.6; in any case, it is a well-established empirical fact that the market impact as a function of trade size has a decreasing derivative. This can be alternatively stated as saying that the price impact per share decreases with the total size of the trade. The standard explanation for this is that it is due to a mixture of informed and uninformed trading. If more informed traders use small trade sizes and less informed traders use large trade sizes, small trades will cause larger price movement per share than large trades.

There are several problems with this theory:

- A concave market impact function is observed in all markets that have been studied, including many such as the London and Paris markets, where the identities of orders are kept completely anonymous. This rules out any explanation that depends on trades made by some agents communicating more information about prices than others and leaves only the possibility that some traders are able to anticipate short-term price movements better than others; see the discussion in Section 2.3.4.
- The model is unparsimonious in the sense that it requires the specification of a function that states the information that traders have as a function of the size of the trades that they use.
- The model is difficult to test because it requires finding a way to specify the information that various groups of traders have *a priori*. One proposal is to do this based of the average profits of different groups of traders. This proposal suffers from the problem that the time horizon for market impact is typically very different than the time horizon on which traders attempt to make profits. A fund manager who intends to a buy a stock and hold it for three years may make the trade to take up that position in a single day. Though this manager might have great skill in predicting stock price movements on a three-year time horizon, she may have no skill at

[5] We are basing this on personal conversations with market practitioners and so can only place a lower bound: We know many people working in many sophisticated trading operations and all of them at least partially use limit orders. We suspect the correct statement is that "most" or even "nearly all" successful hedge funds use limit orders for at least a substantial part of their trading.

all on a daily horizon. Thus in a large fraction of cases, even under large variations in trader skill, impact may have little correlation with profits.

- If it is indeed advantageous to use small trades, then since this is a trivial strategy, one would think that everyone would quickly adopt it and the effect would disappear. In fact, in the past five years or so there has been a huge increase in algorithmic trading, in which brokers automatically execute large trades for clients by cutting up the trades into small pieces. One would therefore think that in modern times the concavity should have diminished or even eliminated entirely. There is little evidence for this; the impact continues to be highly concave.

Thus we have argued in the preceding that the theory is implausible, but even more important, that it makes weak and untestable predictions. The prediction of concavity requires a set of assumptions that are complicated to specify and impossible to measure. The predictions are purely qualitative, and it is not obvious how they might be extended to other properties of impact, such as temporal behavior.

2.3.7. The Liquidity Paradox: Prices Are Not in Equilibrium

We will argue here that liquidity is an important intermediary that modulates the effect of information. We are defining liquidity in terms of the size of the price response to a trade of a given size. High liquidity implies a small price response. Since trades carry information, if the size of trades in response to a given level of information remains constant, as the liquidity varies the price response to information varies with it.

Under the assumption that trading is an intermediate step in the response of prices to information, one can conceptually decompose it into two terms:

$$\Delta p = \mathcal{T}(I)/\lambda \tag{2.1}$$

where λ is the liquidity and $\mathcal{T}(I)$ is the response of trades to information I. Variations in the liquidity do not tell the full story about the response of prices to information; to do that one would also need to understand $\mathcal{T}(I)$. Nonetheless, as we argue here, the effects of varying liquidity are substantial, and they have the huge advantage of being easily measurable.[6] In contrast, since information is difficult to measure, $\mathcal{T}(I)$ is difficult to measure. Furthermore, the preceding equation should be interpreted rather loosely; as we shall see, impact is in fact neither linear nor permanent.

A very important empirical fact that is crucial to understanding how markets operate is that even “highly liquid” markets are in fact not that liquid. Take, for example, a U.S. large-cap stock. Trading is extremely frequent: a few thousand trades per day, adding up to a daily volume of roughly 0.1 to 1% of total market capitalization. Trading is even more frantic on futures and Forex markets. However, the volume of buy or sell limit orders typically available in the order book at a given instant in time is quite small: only the order of 1% of the traded daily volume—that is, 10^{-4} – 10^{-5} of the market cap for

[6] Of course, liquidity may also depend on information, and indeed in Section 2.6 we will develop this connection.

stocks. Of course, this number has an intraday pattern and fluctuates in time, and it can reach much smaller values in liquidity crises.

The fact that the outstanding liquidity is so small has an immediate consequence: Trades must be fragmented. The theoretical motivations for this were originally discussed by Kyle (1985). It is not uncommon that investment funds want to buy large fractions of a company, often exceeding several percent. One possibility is to arrange upstairs block trades, but this lacks transparency and can be costly. If trading occurs through the continuous double auction market, these numbers suggest that to buy 1% of a company requires at least the order of 100–1000 individual trades. This is under the unrealistic assumption that each individual trade completely empties the order book; more realistically, each trade consumes only a fraction of the order book, and the number of trades is even larger. But because 1000 trades corresponds to roughly the whole daily liquidity, it is clear that these trades have to be diluted over several days, since otherwise the market would be completely destabilized. Thus an informed trader cannot use her information immediately and has to trade into the market little by little.

But why is liquidity, as measured by the number of standing limit orders, so low? Both for similar and for opposite reasons. Too large a buy limit order from an "informed" trader would give her away and raise the price of the sellers. Too large a limit order from a liquidity provider would put him at risk of being "picked-off" by an informed trader. There is a kind of hide-and-seek liquidity game taking place in organized markets, where buyers and sellers face a paradoxical situation: Both want to have their trading done as quickly as possible, but both try not to show their hands and reveal their intentions. As a result, markets operate in a regime of vanishing *revealed liquidity* but large *latent liquidity*; this leads to a series of empirical regularities that we will present here.

From a conceptual point of view, however, the most important conclusion of this qualitative discussion is that prices are typically not in equilibrium in the traditional Marshall sense. That is, the true price is very different than it would be if it were set so that supply and demand were equal as measured by the honest intent of the participants, as opposed to what they actually expose. As emphasized previously, the volume of individual trades is much smaller than the total demand or supply at the origin of the trades. This means that there is no reason to believe that instantaneous prices are equilibrium, efficient prices that reflect all known information. Much of the information is necessarily latent, withheld due to the small liquidity of the market, and only slowly revealing itself (see Lyons, 2001, for similar ideas). At best, the notion of equilibrium prices can only make sense over a long time scale; high-frequency prices are necessarily soiled by a significant amount of noise.

2.3.8. Time Scales and Market Ecology

Consider again the case of a typical U.S. large-cap stock—say, Apple, which (as of November 2007) had a daily turnover of around $8 billion. There are on average six transactions per second and on the order of 100 events per second affecting the order book. These are extremely small time scales compared to the typical time for public news events, in which a hot stock like Apple might be mentioned by name every few hours during a period of fast information arrival. Perhaps surprisingly, the number of large jumps in price is much higher. For example, if we define a jump as a one-minute return exceeding three standard deviations, there are on the order of ten such jumps per day, reflecting the very heavy-tailed distribution of high-frequency returns (Joulin et al., 2008). More often than not such jumps occur in the absence of any identified news. It is obviously a particularly important question to understand the origin and the mechanisms leading to these jumps. The difference between the frequency of news and the frequency of jumps already suggests that something else must be at work, such as fluctuations in liquidity, that may have little or nothing to do with external news entering the market.

What is the typical time scale of the round-trip trades of investors? This depends very much on the style of trading; traditional long-only funds have investment horizons on the scale of years, whereas more aggressive long-short statarbs have time scales of weeks or days, sometimes even shorter. Some empirical results support the existence of a broad spectrum of investment horizons (see Sections 2.4.3 and 2.11.1). The optimal frequency of a trading strategy is a trade-off between the expected profit and the friction and transaction costs. Since the fraction of costs grows with the trading volume, large investment funds cannot trade too quickly. This, again, is directly related to the small prevailing liquidity. So it is reasonable to think that information-based trading decisions have intrinsic frequencies ranging from a few days to years. As we have already emphasized, for large investors a single decision may generate many more trades: A decision to buy or sell may persist for days to months, generating a series of small trades. Again, the important message is that low-frequency, large-volume investment decisions imply high-frequency, small-volume trades and that high-frequency prices cannot be equilibrium prices.

There is, however, a potentially viable high-frequency strategy called *market making* that consists of providing instantaneous liquidity to buyers and sellers and trying to eke out a profit from the bid–ask spread. As originally shown by Glosten and Milgrom (1985), the difficulty is to avoid losses due to adverse price moves. Since market makers are offering either to buy or to sell, they are giving a free option to others who might have better information. The profitability of market-making strategies depends both on the spread, which is beneficial, and on the long-term impact of trades, which is detrimental. This intuition will be made more precise and discussed in detail in Section 2.7. On some exchanges market making is institutionalized, with certain obligations and advantages bestowed to those who take the burden of providing liquidity. However, markets have become more and more electronic, with an open order book allowing each investor to behave either as a liquidity provider by posting limit orders or as a liquidity taker by issuing market orders. Depending on market conditions (for example, the instantaneous value of the spread), investors can choose either type of order. There is both empirical and anecdotal evidence that some players implement high-frequency, market-making strategies. This contribution to order flow is often described as “uninformed.” Although this flow differs from longer horizon trades, which are supposed to be economically informed, these market-making strategies routinely use sophisticated short-term prediction tools and exploit any profitable high-frequency signals. The preceding simplified separation of market participants into two broad classes—speculators/liquidity hunters that trade at medium to low frequencies and market makers/liquidity providers at high frequencies—is both realistic and useful to understand the *ecology* of financial markets (Handa and Schwartz, 1996; Farmer, 2002; Wyart et al., 2008; Lillo et al., 2008b). The competition between these two categories of traders allows one to make sense of a number of empirical facts, we believe much more usefully than noise trader models. In Section 2.11 we present some recent empirical results on the characterization of a market ecology.

2.3.9. The Volatility Puzzle

Given that markets are ecological systems in which participants have a broad distribution of time horizons from seconds to years, it is perhaps not surprising to see long-memory effects in financial markets—for example, in trading volume, volatility, and order flow. What is *a priori* surprising, however, is that despite the fact that high-frequency prices cannot possibly be in equilibrium because of lack of liquidity, and despite the fact that it should take time for the market to interpret a piece of news and agree on a new price, the average volatility is remarkably constant on a wide range of different time scales. As measured by autocorrelation, prices are remarkably efficient down to the fastest time scales. We have argued that news arrival happens on much longer time scales. Given that this is true, how can prices remain so efficient, at least with respect to linear models, even on very fast time scales?

One possible explanation might be that public information as evidenced on news feeds is only a small part of the available information. Instead, suppose there are many sources of private information, which agents are continually processing. As they make their decisions, they trade. Given that heavily traded stocks average many trades per second, this would suggest that a truly staggering amount of information is being processed. We find this explanation implausible.

The alternative is that there is an information-processing cascade from fundamental information on slow time scales to technical information on fast time scales. As we have argued, fundamental information enters at a relatively slow rate and then is processed and incorporated into prices. Under this view, high-frequency strategies play an important role. Such strategies do not directly process external information but rather serve the role of digesting that information and keeping the price stream unpredictable. Such strategies are not processing fundamental information but rather are acting as technical trading strategies, processing information contained in the time history of prices, trading volume, and other information that is completely internal to the market. The ability to substitute information in a time history for state information is well supported in dynamical systems theory (Packard et al., 1980; Takens, 1981; Casdagli et al., 1991). Thus we argue that in the ecology of financial markets, high-frequency strategies are fed by lower-frequency strategies through an information cascade from longer to shorter time scales and from fundamental to technical information, finally resulting in white noise on all scales.

This also suggests that microstructural effects may influence the value of the volatility, as suggested by Lyons (2001); "microstructure implications may be long-lived" and "are relevant to macroeconomics." We will comment on the relation between microstructure and volatility in Section 2.8. This relation is also relevant for the regulator who might attempt to alter the microstructural organisation of markets to reduce the volatility.

2.3.10. The Kyle Model

A classic noise trader model for market impact, which is a natural point of comparison, is due to Kyle (1985). This model assumes that there are three types of traders: noise traders who make random trades, market makers who set prices to guarantee efficiency, and an insider who has access to superior information. Under the most general version of the model the noise traders and insider trade continuously from a starting time until a final liquidation time, at which point everyone is paid the liquidation price for their holdings. The insider has superior information about the final liquidation price p_∞ and an infinite bank, which she uses to maximize profits at the expense of the noise traders.

The optimal amount that the investor should trade is easily found to be proportional to the difference $p_\infty - p_t$, where p_t is the current price. With the assumption of a linear and permanent impact, in Kyle's notation the price evolution is given by:

$$p_{t+1} - p_t = \lambda\,[\Phi_t + \xi_t] + \eta_t \quad \Phi_t = \beta[p_\infty - p_t] \tag{2.2}$$

where Φ_t is the signed demand of the investor, λ, β are coefficients, ξ_t is the noise trader demand coming from all other market participants, and η_t is a noise term accounting for possible changes of prices not induced by trading (news, etc.). This equation can easily be solved and leads to an exponential relaxation of the initial price toward p_∞ plus a bounded noise term.

The impact in this model can be regarded as essentially mechanical. There is an apparently permanent change in price that is linearly proportional to the total amount that the noise traders and insider trade. We say "apparently permanent" because, since there is a final liquidation time, what happens past this point is undefined. Note that in this model the price will move toward p_∞ regardless of whether it is the correct price; all that is necessary is that insiders *believe* it is the correct price. A random assignment of beliefs about p_∞ will result in a corresponding random set of impacts. Thus, referring to our discussion of the various explanations for market impact in Section 2.3.4, while the Kyle model is built in the spirit of explanation (1), that trades convey a signal about private information, it is equally consistent with (3), random fluctuations in supply and demand.

The assumption of a final liquidation price can naïvely lead to erroneous conclusions. For example, this model suggests that one can easily manipulate the price. However, in the absence of a liquidation price where a transaction with a counterparty can be realized without impact, things are not so trivial: As soon as the investor wants to close his position, he will again mechanically revert the price back to its initial value and take losses. (To see this, note that in a single round trip the investor will buy at a high price and sell at the original price.) The preceding impact model, Eq. (2.2), although used very often in agent-based models of price fluctuations (two of us have also developed similar ideas, i.e., Bouchaud and Cont, 1998; Farmer, 2002), is far too naïve to represent the way real markets operate, at least at the tick-by-tick level.

Thus we see that while the Kyle model provides a good starting point for understanding why there should be market impact and why it is useful to trade into a position incrementally, it falls short of making realistic predictions about impact. We feel that the key elements that need to be extended are (1) removing the final liquidation price, (2) eliminating the infinite bank of the insider and replacing it with the more realistic assumption of a finite, predetermined trading size, and (3) eliminating the distinction between the insider and the noise trader. The aim of the following sections is to explain in detail how to construct a model generalizing Eq. (2.2), using an approach based on robust facts observed in empirical data and consistency arguments. We will find that impact is in general *nonlinear* and *transient*—or equivalently, as explained in Section 2.6.5, *history dependent*. It is only after a properly defined "coarse-graining" procedure that such an impact model can possibly make sense.

2.4. LARGE FLUCTUATIONS AND LONG MEMORY OF ORDER FLOW

From a mechanical point of view, price formation process is the outcome of (i) the flow of orders arriving in the market and (ii) the response of prices to individual orders. Since price dynamics are reasonably well described by a Brownian motion, one might naïvely assume that this would be true for order flow as well. In fact, this is far from the truth. As we explain in detail in this section, order flow is a highly autocorrelated long-memory process. As a consequence, to maintain market efficiency the price response to orders must strongly depend on the past history of order flow. This has profound conseqences on the way in which markets incorporate information.

2.4.1. Empirical Evidence for Long Memory of Order Flow

We discuss here the statistical properties of order flow by considering the time series of signs of orders. Specifically, consider the symbolic time series obtained in event time by replacing buy orders with $+1$ and sell orders with -1, irrespective of the volume of the order. We reduce these series to ± 1 rather than directly analyzing the signed series of order sizes to avoid problems created by the large fluctuations in order size.[7] This reduction can be done for market orders, limit orders, or cancellations, all of which show very

[7] Fluctuations in order size are heavy tailed and have long memory themselves, so statistical averages based on them converge only slowly. The essential behavior is captured by the series of signs.

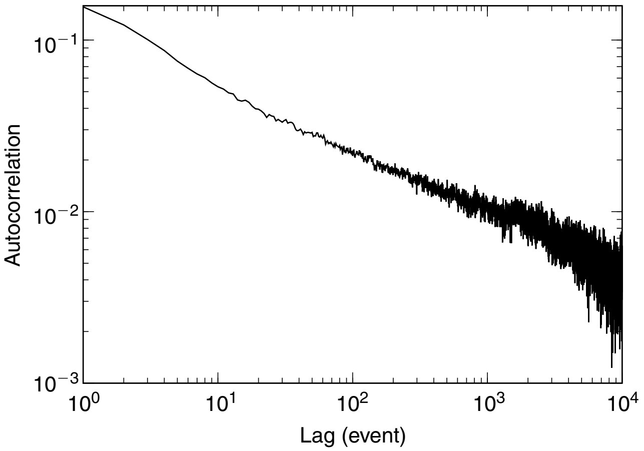


FIGURE 2.1 Autocorrelation function of the time series of signs of orders that resulted in immediate trades (effective market orders) for the stock Vodafone traded on the London Stock Exchange from May 2000 to December 2002, a total of 5.8×10^5 events.

similar behavior.[8] We denote with ϵ_i the sign of the i^{th} market order. Figure 2.1 shows the sample autocorrelation function of the market order sign time series for Vodafone (VOD) in the period 1999–2002 in double logarithmic scale. The figure shows that the autocorrelation function for market order signs decays very slowly. The autocorrelation function is still above the statistical noise level even after 10^4 transactions, which for this stock corresponds to roughly 10 days. This result indicates that if one observes a buy market order now, based on this information alone there is some nonvanishing predictability of the market order signs two weeks from now.

We also note that the autocorrelation function shown in Figure 2.1 is roughly linear in a double logarithmic scale over more than four decades.[9] This suggests that a power-law relation $C_\tau \sim \tau^{-\gamma}$ might be a reasonable description for the sample autocorrelation function.[10]

Stochastic processes for which the autocorrelation function decays asymptotically as a power law with an exponent smaller than one are called *long-memory processes* (Beran, 1994). A precise definition of long memory processes can be given in terms of the autocovariance function Γ_τ. We define a process as long memory if in the limit $\tau \to \infty$

$$\Gamma(\tau) \sim \tau^{-\gamma} L(\tau), \tag{2.3}$$

[8] Long memory is also observed if the side of all activity (bid or ask), including both limit and market orders, is taken together. In contrast, if one assigns to limit orders to sell and cancellations of buy orders a negative sign, corresponding to the fact that the only nonzero price movements it can produce are downward, the combined sequence of signs for market orders, limit orders, and cancellations does not show long memory.

[9] The noisy behavior for large τ comes from the fact that for large lags the statistical errors are remaining roughly constant while the signal decreases, so the relative size of the fluctuations becomes larger.

[10] $f(y) \sim g(y)$ means that there exists a constant $K \neq 0$ such that $\lim_{y\to\infty} f(y)/g(y) = K$.

where $0 < \gamma < 1$ and $L(\tau)$ is a slowly varying function[11] at infinity. The degree of long memory is given by the exponent γ; the smaller γ, the longer the memory. The integral of the autocovariance (or autocorrelation) function of a long-memory process diverges. Long memory can also be discussed in terms of the Hurst exponent H, which is simply related to γ. For a long-memory process, $H = 1 - \gamma/2$ or $\gamma = 2 - 2H$. Short-memory processes have $H = 1/2$, and the autocorrelation function decays faster than $1/\tau$. A positively correlated long-memory process is characterized by a Hurst exponent in the interval $(0.5, 1)$. The use of the Hurst exponent is motivated by the relationship to diffusion properties of the integrated process. For normal diffusion, where by definition the increments do not display long memory, the standard deviation asymptotically increases as $t^{1/2}$, whereas for diffusion processes with long-memory increments, the standard deviation asymptotically increases as $t^H L(t)$, with $1/2 < H < 1$, and $L(t)$ a slow-varying function. In econometrics of financial time series, many variables have the long-memory property. For example, it is widely accepted that the volatility of prices (Ding et al., 1993) and stock market trading volume (Lobato and Velasco, 2000) are long-memory processes. Models of long-memory processes include fractional Brownian noise (Mandelbrot and van Ness, 1968) and the ARFIMA process introduced by Granger and Joyeux (1980) and Hosking (1981).

As Figure 2.1 suggests and as discovered by Bouchaud et al. (2004) and Lillo and Farmer (2004), order flow is also described by a long-memory process. The long-memory of order flow is very robust and is consistently observed for every stock that has so far been examined. Lillo and Farmer tested for long memory in a panel of 20 highly capitalized stocks traded at the London Stock Exchange using Lo's modified R/S test (Lo, 1991), which is known to be a strict test for long memory. They found that even on short samples, in most cases the hypothesis of long memory could not be rejected. The value of H observed in the London Stock Exchange was generally about $H \approx 0.7$, which corresponds to $\gamma = 0.6$. Bouchaud et al. (2004) measured a larger interval of γ values in the Paris Stock Exchange, ranging from 0.2 to 0.7. Long memory has also measured an assortment of stocks in the NYSE; these results are mentioned in Lillo and Farmer (2004) but have not been published in detail.

2.4.2. On the Origin of Long Memory of Order Flow

What causes long memory in order flow? The presence of persistent time correlations in the order flow suggests two possible classes of explanations. The first type of explanation is that this is a property of the order flow of each investor, independent of the behavior of other investors, as proposed by Lillo et al. (2005). The second type of explanation is that investors herd in their trading though an imitation process that involves

[11] $L(x)$ is a slowly varying function (see Embrechts et al., 1997) if $\lim_{x\to\infty} L(tx)/L(x) = 1\ \forall t$. In the preceding definition, and for the purposes of this chapter, we are considering only positively correlated long-memory processes. Negatively correlated long-memory processes also exist, but the long-memory processes we consider in the rest of the chapter are all positively correlated.

an interaction between them, as proposed by LeBaron and Yamamoto (2007). It is of course possible that both effects operate at once, but in any case one would like to know their relative magnitude.

We believe that the evidence gathered so far strongly favors the first explanation. More explicitly, we believe that the dominant cause is the strategic behavior of large investors who split their orders into many small pieces and execute them incrementally. The evidence for this stance comes from two sources. One is the agreement of the properties of the order flow with theory, and the other is additional evidence based on data that gives information about the identity of participants. We summarize both of these here.

2.4.3. Theory for Long Memory in Order Flow Based on Strategic Order Splitting

Lillo et al. (2005) have hypothesized that the cause of the long memory of order flow is a delay in market clearing. To make this clearer, imagine that a large investor decides to buy ten million shares of a company. It is unrealistic for her to simply state her demand to the world and let the market do its job. It is unlikely that enough sellers will be present, and even if there were, revealing the intention to buy a large quantity of shares will very likely push the price up substantially. Instead the large investor keeps her intentions as secret as possible and trades incrementally over an extended period of time, possibly through intermediaries. As already discussed, the strategic reasons for doing this were made clear by Kyle (1985), who investigated a model in which an insider with information about a final liquidation price tries to maximize profits. In simple terms, the motivation is that by splitting the hidden order into small pieces, the investor is able to execute much of the hidden order at prices that do not reflect the full price movement that it will eventually cause.

Our perspective differs from Kyle's in that we assume that the size of the order, which we call the *hidden order*, is given at the outset when the initial trading decision is made. We believe that the size of such orders is largely determined by the fund manager *a priori* and is influenced by a combination of the funds under management and the time scale of the strategy, which is typically much longer than the time scale for completing the trade. The other notable differences are that we do not assume a final liquidation price and we do not make a distinction between informed traders and noise traders. When taken together these differing assumptions create key differences in the predictions of the model in comparison with Kyle.

In several studies based on data giving the identity of hidden orders, about a third of the dollar value of such institutional trades took more than a week to complete (Chan and Lakonishok, 1993, 1995; Vaglica et al., 2008). This conflicts with the standard model of market clearing presented in textbooks, which assumes that agents fully state their supply and demand and that prices are set so that supply and demand are evenly matched. The fact that large orders are kept secret and executed incrementally implies that at any given time there may be a substantial imbalance of buyers and sellers. Effective market clearing is delayed by variable amounts that depend on fluctuations in the size and signs of the unrevealed hidden orders.

We now describe a recently proposed simple model of order flow that postulates the independence of trading activity of investors and which is able to reproduce the long-memory properties of order flow (Lillo et al., 2005). In the simplest version of the model, assume that at any time there are K hidden orders present in the market. Initially the size V of these hidden orders is drawn from a distribution $P(V)$ and the sign ϵ_i is randomly chosen. For simplicity we assume that V is an integer number. We indicate with $V_i(t)$ the volume of hidden order i that has not yet been traded at time t. At each time step t an existing hidden order i is chosen at random with uniform probability, and a unit volume of that order is traded, so that $V_i(t+1) = V_i(t) - 1$. This generates a revealed order of unit volume and sign ϵ_i. A hidden order i is removed if $V_i(t+1) = 0$; that is, when the hidden order is completely traded. When this happens a new hidden order is created with a random sign and a new size.

It is possible to find a closed expression for the autocorrelation function of the trade sign C_τ as a function of the hidden order size distribution $P(V)$. The asymptotic behavior of C_τ can be obtained through a saddle point approximation. If the hidden order size is asymptotically Pareto distributed, that is,

$$P(V) \sim \frac{\alpha}{V^{\alpha+1}} \tag{2.4}$$

the autocorrelation function of order sign behaves asymptotically as (Lillo et al., 2005)

$$C_\tau \sim \frac{K^{\alpha-2}}{\alpha} \frac{1}{\tau^{\alpha-1}} \tag{2.5}$$

Thus the model makes the falsifiable prediction that the exponent γ of the power-law asymptotic behavior of the autocorrelation of order sign is determined by the exponent α of the power-law asymptotic behavior of the hidden order size distribution through

$$\alpha = \gamma + 1 \tag{2.6}$$

Since we observe that $\gamma \simeq 0.5$, this model predicts that $\alpha = 1.5$.

It is worth noting that Lillo et al. (2005) also introduced a more general model in which the number of hidden orders is not constant in time. Specifically, at each time t a new hidden order is generated with probability $0 < \lambda < 1$ if $K(t) > 0$, or probability 1 if $K(t) = 0$. Although this model is not solved analytically, numerical simulation shows that the relation between the exponent of the autocorrelation of order sign and the exponent of order size distribution is the same as in the simpler model where the number of hidden orders is fixed.

This model for the origin of correlation in order flow is in principle empirically testable. The main difficulty arises from the lack of large and comprehensive databases of the hidden orders of investors. There are two ways to check the consistency of the theory. The first one is to compare the distribution of trade sizes in block markets to the autocorrelation function of order signs in order book markets. In block markets, trades are made bilaterally and the identity of counterparties is known. Brokers do not like order splitting and strongly discourage it. Thus block markets can be considered a crude proxy for observing the distributional properties of hidden orders.[12] In the next section we discuss evidence that suggests that block trade volume is indeed asymptotically power-law distributed with an exponent $\alpha \simeq 1.5$. For comparison[13] the average measured values of γ for LSE stocks is $\gamma = 0.57$, close to $\hat{\gamma} = 0.59$ as predicted by $\hat{\gamma} = \alpha - 1$. The second supporting evidence comes from a study of the Spanish Stock Exchange by Vaglica et al. (2008), who have inferred hidden orders using data with membership codes. This study is discussed in Section 2.11.1.

2.4.4. Evidence Based on Exchange Membership Codes

Empirical testing is difficult due to the fact that it is not easy to collect data on the behavior of individual investors. Nonetheless, partial information about the identity of participants can be obtained using data that identifies the broker or the member of the exchange who executes the trade, which we will simply call the *membership code*. In many stock markets, such as the LSE, the Spanish Stock Exchange, the Australian Stock Exchange, and the NYSE, it is possible to obtain data containing this information. It is important to stress that knowing the membership code is not the same as knowing the individual participant, since the member may either trade on its own account or act as a broker for other trades, or do both at once. Nonetheless, several recent papers have demonstrated that it is possible to extract useful information about the identity of individual traders using such information; for example showing that there are consistent behaviors that are persistent in time associated with particular membership codes, that such behaviors can be organized into a taxonomic tree, and that it is possible to detect the presence of large institutional trades (Lillo et al., 2008b; Zovko and Farmer, 2007; Vaglica et al., 2008).

Gerig et al. have used membership codes of the London Stock Exchange to test the hypothesis of the theory presented in Lillo et al. (2005). The autocorrelation function of market order signs is computed by considering realized orders placed by the same membership code or by different membership codes separately. Figure 2.2 shows the autocorrelation function of market order signs with the same membership code, different membership codes, and all transactions irrespective of membership code. The circles in the figure represent the autocorrelation function irrespective of the membership code and, as anticipated, it is well fitted by a power law. When only transactions with the same membership code are considered (the triangles), the autocorrelation is still power law with a slightly smaller exponent. Moreover, for a fixed lag τ, the autocorrelation function with the same membership code is one order of magnitude larger than the autocorrelation function irrespective of the membership code. Finally, when only transactions with different membership codes are considered, the autocorrelation function decays very rapidly to zero, and it is clearly not consistent with power-law

[12] The exception is that it is possible to split an order and trade with multiple brokers.

[13] The error bars in computing both γ and α are substantial, as can be seen by computing them for subsamples of the data, and the close agreement observed by Lillo et al. (2005) between γ and $\alpha - 1$ is probably fortuitious. Unfortunately, there still are not good statistical methods for assigning confidence intervals for exponents of power laws, particularly when the observations have long memory, but the errors can be roughly assessed by examining subsamples.

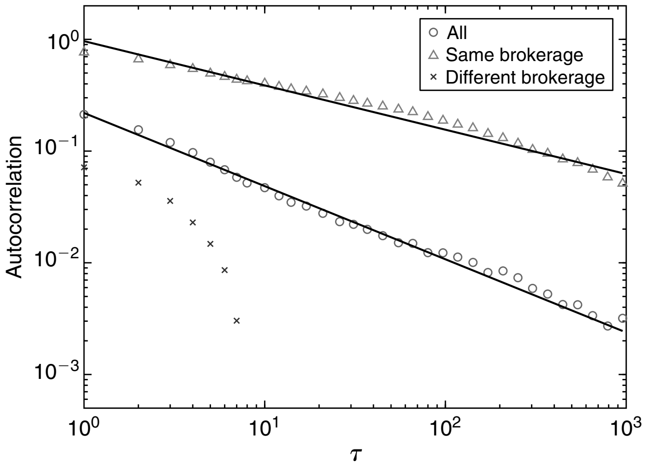


FIGURE 2.2 Autocorrelation of signs-versus-transaction lag for transactions with same membership code, different membership code, and all transactions irrespective of membership code, plotted on double logarithmic scale. The investigated stock is AstraZeneca (AZN) traded the LSE from 2000 to 2002.

behavior. Under the assumption that most investors use only a few brokers to execute a given hidden order, this plot strongly supports the hypothesis that the long memory of signs is due to the presence of investors that place many revealed orders of the same sign and that there is no clear sign of herding behavior among different investors. It is in principle possible that herding happens between investors using the same broker but not between investors with different brokers; however, the reasons why this would occur are unclear, and it seems implausible that it could explain such a dramatic difference.

2.4.5. Evidence for Heavy Tails in Volume

The theory we have developed makes it clear that the distribution of trading volume plays a key role in shaping many properties of the market, including the long memory of order flow, which, as we will show, in turn has important consequences for market impact. In recent years there has been a debate about the statistical properties of trading volume. This is partly due to the fact that markets have different structures and one should be careful in specifying which volume is considered in the analysis. Gopikrishnan et al. (2000) originally observed that volume of trades at the NYSE are asymptotically power-law distributed. Specifically, they claimed that for large volumes the probability distribution scales as

$$P(V > x) \sim x^{-3/2} \tag{2.7}$$

This law has been termed the "half cubic" law. The NYSE, like many other financial markets, employs two parallel markets that provide alternative methods of trading, called the on-book, or "downstairs," market and the off-book, or "upstairs," market. Orders in the on-book market are placed publicly but anonymously and execution is

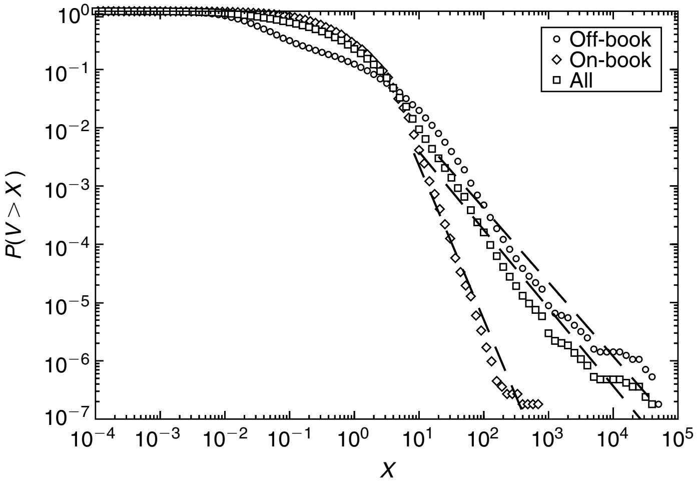


FIGURE 2.3 Volume distributions of off-book trades (*circles*), on-book trades (*diamonds*), and the aggregate of both (*squares*) for a collection of 20 different stocks, normalizing the volume of each by the mean volume before combining. The dashed black lines are for the slope found by the Hill estimator and are shown for the largest 1% of the data. (*Source:* Adapted from Lillo et al., 2005.)

completely automated. The off-book market, in contrast, operates through a bilateral exchange mechanism, via telephone calls or direct contact of the trading parties. The anonymous nature of the on-book market facilitates order splitting—that is, large orders are split into smaller pieces and traded incrementally. On the other hand, the off-book market is a block market, where large orders can be traded in a single transaction. The NYSE data used by Gopikrishnan et al. (2000) includes a mixture of order book trades and block trades. Since the typical size of block trades is much larger than the size of orders traded in the order book, the size of block trades dominates the tail of volume distribution.

This can be seen more clearly in a market (or database) where it is possible to separate block trades from order book trades. In Figure 2.3 (from Lillo et al., 2005) we show the cumulative distribution function of trading volume for off-book trades, on-book trades, and the aggregate of both for a collection of 20 LSE stocks. The distribution of block trades is consistent with the power-law hypothesis of Eq. 2.7 with an exponent close to 1.5, whereas the distribution of order book trades is not consistent with the half-cubic law and instead has a much thinner tail (see also Farmer and Lillo, 2004, and Plerou et al., 2004).

2.5. SUMMARY OF EMPIRICAL RESULTS FOR DIVERSE TYPES OF MARKET IMPACT

The relation between the transacted volume and the consequent expected price shift is called the *price impact*, or alternatively, the *market impact* function. Letting R be a price return associated with a trade of size V, the market impact a time l after the trade occurred is

$$\mathcal{I}(V,l) = E[R|V,l]$$

For many purposes it is useful to separate the dependence on volume from the dependence on time. One can make the hypothesis that the impact function can be written as a product of two functions, that is,

$$\mathcal{I}(V,t) = \mathcal{S}(V)\mathcal{R}(l)$$

In this section we primarily discuss the dependence on volume, saving the discussion of time dependence for Section 2.6.

At this stage we are intentionally being vague about the definition of the return R and the volume V; defining these more precisely is one of the main points of this chapter. The way in which the market impact behaves depends on the market structure as well as on what one means by "return" and "volume." Many studies have empirically investigated market impact with a range of different results; we argue that in many cases these differences stem from differences in what is being studied. The important distinctions that should be made are:

- First, one can consider market impact of an individual transaction vs. an aggregate of many transactions. Aggregation here means that the market impact is conditioned on a given number of trades or to a given interval of time. We discuss the volume dependence of individual impact in Section 2.5.1 and the time dependence in Sections 2.6.2–2.6.5, and we study the properties of aggregate impact in Sections 2.5.2 and 2.6.8.
- A second important aspect is the type of market exchange in which the transactions take place. As we said, most financial markets have upstairs or block markets as well as downstairs or order book markets. In the downstairs market, trades are made by placing orders in a limit order book, and it is quite common to aggressively split large trading orders into many small pieces. The upstairs market trades are arranged bilaterally between individuals. As a result of the varying market structures, the impacts can be quite different.
- A third factor that must be kept in mind is that large trading orders, which we will call *hidden orders*, are typically split into small pieces and executed incrementally. This is in contrast to *realized orders*, which are the actual orders that are traded—for example, the pieces into which hidden orders are split. For realized orders the impacts may be part of a larger process of order splitting that is invisible with the data that we have here. The impacts of hidden orders may be quite different than those of realized orders. The impacts of individual orders behave much like those of individual transactions, as described in the next bullet. We will discuss the impact of hidden orders in Section 2.6.7.
- Finally, even if we have discussed market impact in terms of transacted volume, other events in the market have an impact on price. Specifically, in double auction

market limit, orders and cancellations can have a market impact that is different from the impact of a market order.

In the following sections we discuss the empirical regularities in these different types of market impact.

2.5.1. Impact of Individual Transactions

We now discuss the impact of individual transactions in limit order book markets, whose volume we will denote by v. Many studies have examined the market impact for a single transaction, and all have observed a concave function of the transaction volume v, that is, one that increases rapidly for small v and more slowly for larger v. The detailed functional form, however, varies from market to market and even period to period. Early studies by Hasbrouck (1991) and Hausman, Lo, and MacKinlay (1992) found strongly concave functions but did not attempt to fit functional forms. Keim and Madhavan (1996) also observed a concave impact function for block trades.

Based on Trades and Quotes (TAQ) data for a set of 1000 NYSE stocks, the concavity of the market impact was interpreted by Lillo et al. (2003b) using the functional form

$$E[r|v] = \frac{\epsilon v^{\psi}}{\lambda} \tag{2.8}$$

The exponent $\psi(v)$ is approximately 0.5 for small volumes and 0.2 for large volumes. Even normalizing the volume v by daily volume, the liquidity parameter λ varies for different stocks; there is a clear dependence on market capitalization M that is well approximated by the functional form $\lambda \sim M^{\delta}$, with $\delta \approx 0.4$.

Potters and Bouchaud (2003) analyzed stocks traded at the Paris Bourse and NASDAQ and found that a logarithmic form gave the best fit to the data. For the London Stock Exchange, Farmer and Lillo (2004) and Farmer et al. (2005) found that for most stocks Eq. 2.8 was a good approximation with $\psi = 0.3$, independent of v. Hopman (2007) studied market impact on a 30-minute time scale in the Paris Bourse for individual orders and found $\psi \approx 0.4$, depending on the urgency of the order. Thus all the studies find strongly concave functions but report variations in functional form that depend on the market and possibly other factors as well. Figure 2.4 shows the price impact of buy market orders for five highly capitalized LSE stocks, that is, AZN, DGE, LLOY, SHEL, and VOD. The price impact is well fit by the relation $E[r|v] \propto v^{0.3}$.

2.5.2. Impact of Aggregate Transactions

Studies of aggregated market impact have produced variable results, reaching different conclusions that we will argue depend substantially on the time scale for aggregation. The BARRA market impact model, an industry standard, uses the TAQ data aggregated on a half-hour time scale (Torre, 1997). They compare fits using Eq. 2.8 and find $\psi \approx 0.5$; they obtain similar results using individual block data. Kempf and Korn (1999) studied data for futures on the DAX (the German stock index) on a five-minute time

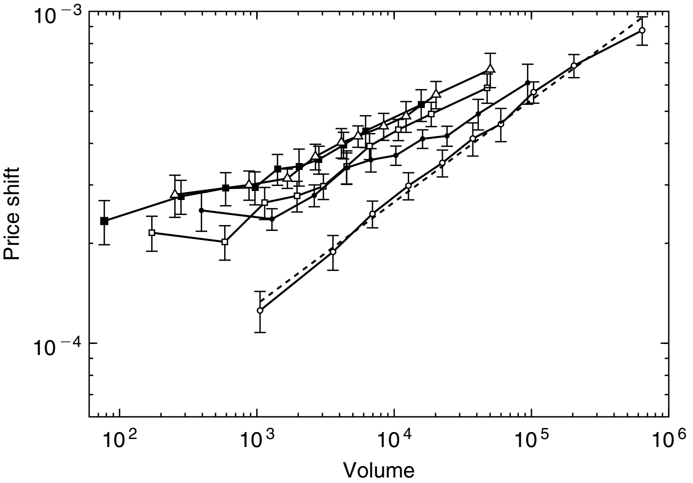


FIGURE 2.4 Market impact function of buy market orders for a set of five highly capitalized stocks traded in the LSE, specifically AZN (*filled squares*), DGE (*empty squares*), LLOY (*triangles*), SHEL (*filled circles*), and VOD (*empty circles*). Trades of different sizes are binned together, and the logarithmic price change's average size for each bin is shown on the vertical axis. The *dashed line* is the best fit of the market impact of VOD with a functional form as described in Eq. 2.8. The value of the fitted exponent for VOD is $\psi = 0.3$.

scale and found a very concave functional form. Plerou et al. (2002) studied data from the NYSE during 1994 and 1995 ranging from 5- to 195-minute time scales and fit the market impact function with a hyperbolic tangent. They noted that at shorter time scales this functional form did not work well for small v; $\tanh(v)$ is linear for small v, but at short time scales (e.g., 5 or 15 minutes) they observed a nonlinear impact function becoming more linear as they went toward longer time scales.

Evans and Lyons (2002) studied foreign exchange rate transactions data for DM and Yen against the dollar at the daily scale over a four-month period. They used the number of buyer-initiated transactions minus the number of seller-initiated transactions as a proxy for the signed order flow volume v and found a strong positive relationship to concurrent returns. Chordia and Subrahmanyam (2004) study impacts of stocks in the S&P 500 at a daily time scale and perform linear regressions but do not compare to other functional forms. For the Paris Bourse Hopman (2007) measures aggregate order flow as $\sum_i \epsilon_i v_i^{\psi}$, where the sum is taken over fixed time intervals. At a daily scale he finds that he gets the best linear regression against contemporary daily returns with $\psi \approx 0.5$. He also documents that the slope of the regression decreases with increasing time scale. Finally, as discussed in more detail later, Gabaix et al. (2003, 2006) have made extensive studies of data from the New York, London, and Paris stock markets on a 15-minute time scale and find exponents $\psi \approx 0.5$.

What is the origin of these differences in the observed functional form of the aggregate market impact? Part of the difference comes certainly from the fact that these studies consider different markets, different assets, and different time periods. However, another important difference across studies is the time scale of aggregation. There is no reason that the aggregate market impact over a ten-minute time interval should have the same functional form of that over a one-hour time interval or over an interval that is defined by 30 trades.

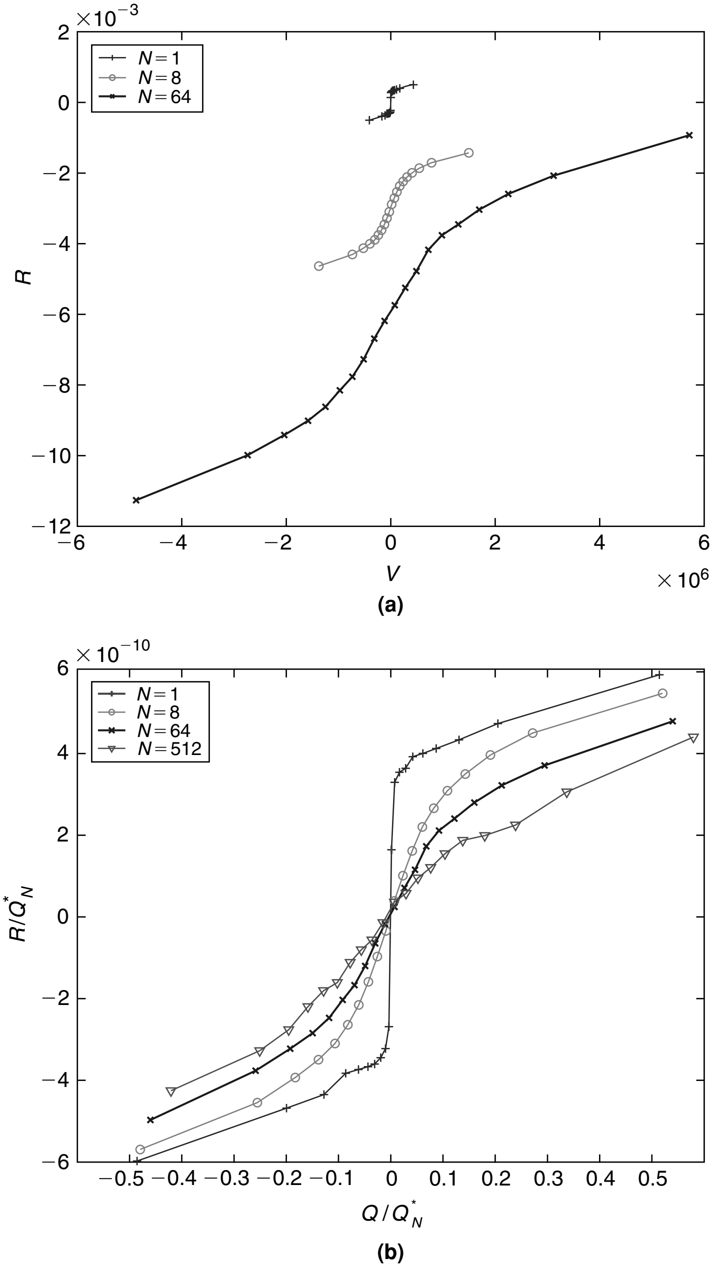
To get an idea of how the market impact changes its shape with aggregation scale, consider a specific example. Let v_t be the volume of transaction happening at time t (in event time). Let $r_t = \log(p_{t+1}/p_t)$ be the corresponding log-return, where p_t is the price of transaction t. For a sequence of N successive transactions beginning at time t, let $Q_N = \sum_{i=1}^{N} \epsilon_{t+i} v_{t+i}$ be the aggregate volume and $R_N = \sum_{i=1}^{N} r_{t+i}$ be the aggregate return. The average market impact conditioned on volume is

$$R(Q, N) = E[R_N | Q_N = Q] \tag{2.9}$$

that is, it is the expected return associated with a signed volume fluctuation Q. We write $R(Q, N)$ to emphasize that this can depend both on the signed trading volume imbalance Q and the number of transactions N. In Figure 2.5 we show empirical estimates for the market impact of the stock AZN, which is traded on the London Stock Exchange, from Lillo et al. (2008a). Figure 2.5 shows the market impact for different values of N with offsets added to the vertical axis to aid visualization. As one would expect, the scale increases with N. The shape of $R(Q, N)$ also changes, becoming more linear with increasing N. This is illustrated more clearly in Figure 2.5(b), where we rescale the horizontal and vertical axes using a rescaling factor based only on Q_N. The renormalization makes the increasing linearity clearer. As N increases, the market impact near $Q = 0$ becomes linear, and the size of the region that can be approximated as linear grows with increasing N. It also illustrates a surprising feature: The slope of the linear region decreases with N. These same basic features (increasing linearity and decreasing slopes) hold for all the stocks in our sample, in both the New York and London Stock Exchanges. This result shows that the shape and the scale of the aggregate market impact change with the aggregation scale. At short time scales the function is significantly nonlinear, but at large aggregation scales the market impact becomes close to linear, and the slope of the impact decays with the aggregation scale. For this reason it is in general misleading to compare aggregate impact curves with different scales unless one has a theory for how the market impact depends on aggregation scale. This also shows why the studies mentioned previously found different forms of the market impact. In Section 2.6.8 we present some models that help explain the behavior of aggregate impact observed in real data.

2.5.3. Hidden Order Impact

Because data for hidden orders, which are sometimes also called *trading packages*, are difficult to obtain, there are only a few studies (Chan and Lakonishok, 1993, 1995; Almgren et al., 2005; Vaglica et al., 2008). These studies show that hidden orders can be extremely long, involving thousands of realized trades spread over periods of many weeks or even months. As reviewed in Section 2.11.1, the most recent



(a)

(b)

FIGURE 2.5 Aggregate market impact $R(Q, N)$ for the LSE stock Astrazeneca for 2000–2002. **(a)** Plot of the shifted aggregate return $R(Q, N) + R_0$ vs. the aggregate signed volume Q for three values of N. The arbitrary constant R_0 is added to aid visualization; its values are $R_0 = \{0, -3 \times 10^{-3}, -6 \times 10^{-3}\}$ for $N = 1, 8$ and 64, respectively. **(b)** A rescaling for each N of both the horizontal and vertical axes by $Q^*_N = Q^{(95)}_N - Q^{(5)}_N$, where $Q^{(5)}_N$ is the 5% quantile and $Q^{(95)}_N$ is the 95% quantile of Q.

study by Vaglica et al. confirms that hidden orders obey a power-law distribution of size, which, as we argue in Section 2.6, plays an important role in determining their impact.

The theoretical considerations for treating hidden orders are quite different from those for individual orders, and they also very different from those of aggregated anonymous orders. The reason is that such orders come from the same agent, creating bursts of orders in the order flow which are all of the same sign. As we argued in Section 2.4.3, this generates strong correlations in order flow that have to be compensated for, as discussed in Section 2.6. The volume dependence of hidden order impact is intimately connected to the temporal aspects, and so we save the development of the theory for hidden order impact for the next section.

2.5.4. Upstairs Market Impact

Market impact in the upstairs market has been studied by Keim and Madhavan (1996). As in other cases, they find empirically that market impact is concave. They explain this based on a model for the difficulty of finding counterparties for trading. Ultimately, as pointed out by Gabaix et al. (2006), upstairs market impact should match hidden order impact, for the simple reason that the upstairs market is competing with the downstairs market, and if costs in the upstairs market are too high, they have the option of splitting up their trades in the downstairs market. This is convenient because it implies that a theory for either market automatically gives a theory for the other.

2.6. THEORY OF MARKET IMPACT

In this section we develop theoretical explanations for both the volume dependence and the temporal dependence of market impact. As stressed in the previous section, there are several distinct types of impact that require a different approach to their analysis. We begin in Section 2.6.1 by explaining why the impacts associated with individual trades are so concave, arguing that the dominant cause is selective liquidity taking.

Then, in Sections 2.6.2 through 2.6.5, we develop a theoretical approach to understanding the temporal behavior of impacts associated with individual trades. We show that the long memory of order flow and market efficiency play a crucial role, which one can take into account one of two ways. One can either assume a fixed impact, in which case the future contribution to the impact of each trade must decay to zero with time, or one can assume a varying but permanent impact, which implies asymmetry liquidity. We show that these two approaches are equivalent.

In Section 2.6.6 we present empirical results supporting these ideas. In Section 2.6.7 we develop a theory for the impact of hidden orders—that is, linked sets of trades made by large investors. Finally, in Section 2.6.8, we develop a theory for the aggregate impact of successive trades and show that it does a good job of explaining the empirical results of Section 2.5.2.

2.6.1. Why Is Individual Transaction Impact Concave?

Let us first consider the impact of individual transactions. Several different theories have been put forth to explain why market impact for single transactions is concave. These can be grouped into three classes: (1) size-dependent informativeness of trades (e.g., due to stealth trading, as postulated by Barclay and Warner, 1993), (2) average depth vs. price in the limit order book (Daniels et al., 2003), and (3) selective liquidity taking (Farmer et al., 2004).

The standard reason given for the concavity of market impact is that it reflects the informativeness of trades. If small trades carry almost as much information as large trades, the price changes caused by small trades should be nearly as big as those for large trades. For example, this could be due to "stealth trading,"; that is, because informed traders keep their orders small to avoid revealing their superior knowledge (Barclay and Warner, 1993). Hypothesis 2, due to Daniels et al. (2003), is that it reflects the accumulation of liquidity in the limit order book—that is, the depth in the order book as a function of the price will determine the market impact for a market order as a function of its size. Hypothesis 3 is that this is due to selective liquidity taking; that is, that liquidity takers submit large orders when liquidity is high and small orders when it is low (see Farmer et al., 2004; Weber and Rosenow, 2006; and Hopman, 2007).

Theory 2 is easily ruled out by computing the average virtual market impact as a function of volume. This is defined as the average price change that would instantaneously occur for an effective market order of size v (Weber and Rosenow, 2006; Farmer and Zamani, 2007). In Figure 2.6 we show the virtual impact for AZN, computed by

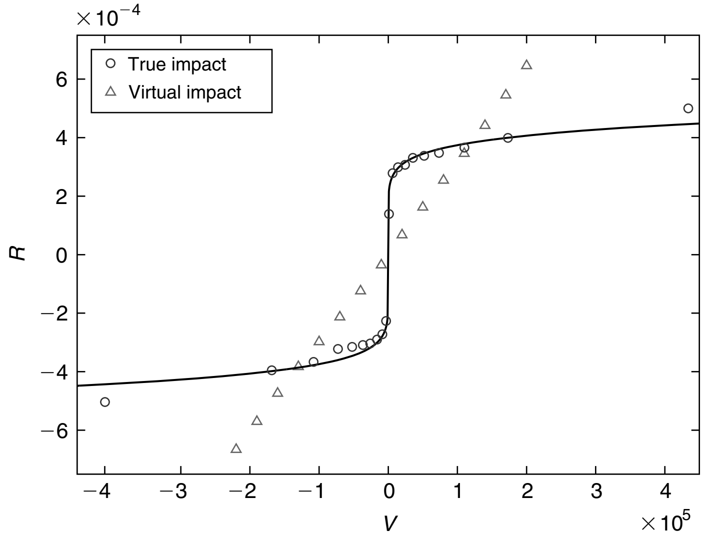


FIGURE 2.6 Comparison of virtual to true market impact: true impact (*circles*), virtual impact (triangles). The fitted curve for true impact (*black line*) is of the form $f(v) = Av^{\psi}$, with $\psi = 0.3$.

hypothetically submitting orders for a range of different values of v and measuring the immediate price response. This is done for each time when real effective market orders were submitted. The resulting price response is a direct probe of the depth of the limit order book. The fact that the mechanical impact is linear to very good degree of approximation makes it clear that this is not the cause of the concavity of the real market impact function.

The selective liquidity taking (Hypothesis 3) means that agents condition the size of their transactions on liquidity, making large transactions when liquidity is high and small transactions when it is low. As shown by Farmer et al. (2004), for LSE stocks it is rare that a trade penetrates more than one price level.[14] For example, for Astrazeneca, approximately 87% of the market orders creating an immediate price change have a volume equal to the volume at the opposite best. Moreover, approximately 97% of the market orders creating an immediate price change have a volume that is either equal to the opposite best or larger than this value but smaller than the sum of volume at the second best opposite price. This means that to a good approximation the market impact can be written in the very simple form

$$E[r|v] = P(+|v)E[r] \tag{2.10}$$

where $P(+|v)$ is the probability that a trade of size v generates a nonzero return—that is, the probability that $v \geq \Phi_b$, where Φ_b is the volume offered or bid at the opposite best price. $E[r]$ is the expected return given that there is a nonzero return, which is of the order of the bid–ask spread (see Section 2.7 for more precise statements). This demonstrates that trading orders that penetrate the opposite best are rare. This is because agents do not like to suffer price degradation more than the opposite best and so condition the size of their orders on what is being offered there.

We have now to explain why $P(+|v)$ is a concave function. An explanation in terms of selective liquidity taking is the following. Suppose that the volume at the best is drawn from a distribution $P_b(\Phi_b)$ and suppose that the liquidity taker draws the volume v she would like to trade from another distribution and independently from Φ_b. If $v < \Phi_b$ she places a market order of size v, whereas if $v > \Phi_b$ she places a market order of size Φ_b. What is the probability $P(+|v)$ under this simple model? A straightforward calculation shows that $P(+|v) = \int_0^v P_b(\Phi_b)d\Phi_b$; that is, it is equal to the cumulative distribution of the volume at the best. This is an increasing and concave function of v that could be used to fit the empirical $P(+|v)$. Under this model the shape of the market impact is explained by $P(+|v)$, that is, by the conditioning of trading orders on the liquidity that is offered. In other words, Theory 3 does a good job of explaining the data, at least qualitatively.

It is a matter of interpretation, however, whether this is also consistent with Theory 1; that is, that smaller trades are proportionately more informative than larger trades. From one point of view, one can simply say that the market impact *defines* the informativeness of trades. If so, then it is obviously consistent. However, if it means that price changes are a response to the new information contained in trades, the evidence presented here

[14] See Table 2 of Farmer et al.

is inconsistent with Theory 1. In the LSE the quoted volume is visible to all, and so, except for occasional latency problems, in which the quote changes just before a trade is placed, the trader is aware of the quote when she places the trade. The fact that the size of the trade is strongly correlated with the size of the best quote implies that the size of the trade carries little new information. This does not mean that the trade is based on inferior information; it merely means that other market participants do not learn much from its size when it occurs. It is the conditioning of trade size on best quotes that drives concavity and not because the smaller trades are nearly as "informed" as the larger trades.

2.6.2. A Fixed Permanent Impact Model

In the previous section we described how midquote prices react on average to market orders of a given volume v. The preceding discussion was restricted to the immediate impact, that is, the impact that is felt immediately after a trade is completed. In general this can have both temporary and permanent components. In this section we discuss the impact of individual transactions—that is, the average midquote price change between just before the n^{th} trade and just before the $n+1^{\text{th}}$ trade. It is an empirical fact that this immediate impact, defined as $E[r_n|\epsilon_n v_n]$, is nonzero and can be written as $E[r|\epsilon v] = \epsilon f(v)$, where f is a function that grows with v. Clearly, it is important to understand if and how this immediate impact evolves with time (which we will measure in terms of the sequence number of the trades). Is the impact of a trade permanent or transient? Is it fixed or is it variable? How does it depend on the past order flow history?

The simplest situation is that of a usual random walker, where position at any time is the sum over all past steps—however far in the past they might be. In financial language, this corresponds to the case where the impact of a transaction is permanent, which translates into the following equation for the midquote price m_n at time n:

$$r_n = m_{n+1} - m_n = \epsilon_n f(v_n; \Omega_n) + \eta_n, \tag{2.11}$$

where η_n is an additional random term describing price changes not directly attributed to trading itself—for example, the impact of news where quotes could instantaneously jump without any trade. We will assume here that η_n is independent of the order flow and we set $E[\eta] = 0$ and $E[\eta^2] = \Sigma^2$. We have included a possible dependence of the impact on the instantaneous state Ω_n of the order book. We expect such a dependence on general grounds: A market order of volume v_n, hitting a large queue of limit orders, will in general impact the price very little. On the other hand, one expects a very strong correlation between the state of the book Ω_n and size of the incoming market order: Large limit order volumes attract larger market orders.

The preceding equation can be written as:

$$m_n = \sum_{k<n} \epsilon_k f(v_k; \Omega_k) + \sum_{k<n} \eta_k \tag{2.12}$$

which makes explicit the nondecaying nature of the impact in this model: $\epsilon_k \partial m_n / \partial v_k$ (for $k < n$) does not decay as $n - k$ grows. This simple model makes the following predictions for the lagged impact function $\mathcal{R}_\ell$ and the lagged return variance $\mathcal{V}_\ell$:

$$\begin{aligned} \mathcal{R}_\ell &\equiv E[\epsilon_n \cdot (m_{n+\ell} - m_n)] = E[f]; \\ \mathcal{V}_\ell &\equiv E[(m_{n+\ell} - m_n)^2] = \left(E[f^2] + \Sigma^2\right) \ell \end{aligned} \tag{2.13}$$

that is, constant price impact and pure price diffusion, close to what is indeed observed empirically on small-tick, liquid contracts. However, if we consider the autocovariance of price returns within this model, we find that

$$E[r_n r_{n+\tau}] \propto E[\epsilon_n \epsilon_{n+\tau}] \sim \tau^{-\gamma} \tag{2.14}$$

which means that price returns are strongly autocorrelated in time. This fact would violate market efficiency because price returns would be easily predictable even with linear methods. We therefore come to the conclusion that the empirically observed long memory of order flow is incompatible with the previous random walk model if prices are efficient (Bouchaud et al., 2004; Lillo and Farmer, 2004; Challet, 2007). In other words one of the assumptions of the random walk model must be relaxed. Among the various possibilities we will relax either the assumption that price impact is permanent or the assumption that price impact is independent of the order flow. As we will see, these two possibilities are related one to each other, but for the sake of clarity we present them in two different subsections.

2.6.3. The MRR Model

To illustrate the preceding concepts, let us discuss a slight variant of a model due to Madhavan, Richardson, and Roomans (Madhavan et al., 1997) that helps define various quantities and hone in on relevant questions. The assumptions of the model are (1) that all trades have the same volume $v_n = v$ and (2) the ϵ_n's are generated by a Markov process with correlation ρ, which means that the expected value of ϵ_n conditioned on the past only depends on ϵ_{n-1} and is given by:

$$E\left[\epsilon_n | \epsilon_{n-1}\right] = \rho \epsilon_{n-1} \tag{2.15}$$

The case $\rho = 0$ corresponds to independent trade signs, whereas $\rho > 0$ describes positive autocorrelations of trade signs. Note that in this model, correlations decay exponentially fast, that is,

$$C_\ell = E[\epsilon_i \epsilon_{i+\ell}] = \rho^\ell \tag{2.16}$$

which, as we discussed in Section 2.4, does not conform to reality.

The MRR model postulates that the midpoint m_n evolves only because of unpredictable external shocks (or news) and because of the surprise component in the order flow. This postulate, of course, automatically removes any predictability in the price returns and ensures efficiency. Under the assumption that the surprise component of the order flow at the n^{th} trade is given by $\epsilon_n - \rho\epsilon_{n-1}$, one writes the following evolution equation for the price[15]

$$m_{n+1} - m_n = \eta_n + \theta[\epsilon_n - \rho\epsilon_{n-1}] \qquad \textbf{(2.17)}$$

where η is the shock component and the constant θ measures the size of trade impact.

These equations make it possible to compute several important quantities such as the lagged impact function defined earlier (Eq. 2.13). One may write:

$$m_{n+\ell} - m_n = \sum_{j=n}^{n+\ell-1} \eta_j + \theta \sum_{j=n}^{n+\ell-1} [\epsilon_j - \rho\epsilon_{j-1}] \qquad \textbf{(2.18)}$$

The full impact function is found to be constant, equal to:

$$\mathcal{R}_\ell = \theta(1-\rho^2), \quad \forall \ell \qquad \textbf{(2.19)}$$

We can also define the "bare" impact of a single trade $G_0(\ell)$, which measures the influence of a single trade at time $n - \ell$ on the the midpoint at time n. In terms of $G_0(\ell)$, the midpoint is therefore written as:

$$m_n = \sum_{j=-\infty}^{n-1} \eta_j + \sum_{j=-\infty}^{n-1} G_0(n-j-1)\,\epsilon_j \qquad \textbf{(2.20)}$$

here found to be given by $G_0(\ell = 0) = \theta$ and $G_0(\ell \geq 1) = \theta(1-\rho)$; a part $\theta\rho$ of the impact instantaneously decays to zero after the first trade, whereas the rest of the impact is permanent. The instantaneous drop of part of the impact compensates the sign correlation of the trades. Finally, the volatility, within this simplified version of the MRR model, reads:

$$\left[\theta^2(1-\rho^2) + \Sigma^2\right] \ell \qquad \textbf{(2.21)}$$

2.6.4. A Transient Impact Framework

Compared to the simplifying assumptions of the MRR model, the data shows that (1) the volumes v of the incoming market orders are very broadly distributed, with a power-law tail (see Section 2.4.5); (2) the sign time series ϵ_n has long-range correlations C_ℓ that decays again as a power-law $\sim c_0\ell^{-\gamma}$ with $\gamma < 1$, defining a long-memory process. The smallness of γ makes the correlation function C_ℓ nonsummable: The average relaxation time is infinite, whereas the correlation time of the Markovian sign process in the preceding MRR model is finite, equal to $(1-\rho)^{-1}$.

In this section we relax the assumption that impact of a single trade is permanent in time. Rather, we find that long-range correlations in trades imply that the impact

[15] The assumption that prices respond linearly to the order flow is a very strong assumption.

itself has to decay slowly with time. In the next section, we discuss an alternative but equivalent model, where the impact is permanent but asymmetric and history dependent.

Transient Impact and Mean Reversion

What would happen if the impact of each trade was purely transient—for example, an exponential decay in time? Eq. 2.11 would now read:

$$m_n = \sum_{k<n} \alpha^{n-k-1} \epsilon_k f(v_k; \Omega_k) + \sum_{k<n} \eta_k, \quad (0 \leq \alpha < 1) \tag{2.22}$$

The lagged impact and the return variance would then be given by:

$$\mathcal{R}_\ell = \alpha^{\ell-1} E[f] \quad \mathcal{V}_\ell = 2E[f^2]\frac{1-\alpha^\ell}{1-\alpha^2} + \Sigma^2 \ell \tag{2.23}$$

That is, a short-time volatility $\approx E[f^2] + \Sigma^2$ larger than its long-time value Σ^2, in which only the "news" component survives. The price would exhibit significant high-frequency mean reversion: Impact kicks it temporarily up and down, but the long-term wandering of the price is unrelated to trading. Of course, one could be in a mixed situation where the impact decays exponentially but toward a positive value, in which case the long-term volatility still involves an impact component. This conforms with conventional wisdom about efficient markets: an increased value of high-frequency volatility driven by the *tâtonnement* process and a long-term volatility made up both of unexpected news and long-term impact of market orders, which translates private information into prices. However, recall that this does not conform to observations, which show volatility very nearly constant across all time scales (see Section 2.3.9).

What is the relation between the average $\mathcal{R}_\ell$ and the impact of a single trade that we call $G_0(\ell)$ henceforth? If trades were uncorrelated, the two quantities would be identical, but trade correlations, as we shall see, change the picture in a rather interesting way.

Mathematical Theory of Long-Term Resilience

The long-term memory of trades is *a priori* paradoxical and hints of a nontrivial property of financial markets, which can be called *long-term resilience*. Take again Eq. 2.20 with the assumption that single trade impact is lag independent, $G_0(\ell) = G_0$, and that volume fluctuations can still be neglected. The midprice variance is easily computed to be:

$$\mathcal{V}_\ell \equiv \langle (m_{n+\ell} - m_n)^2 \rangle = [\Sigma^2 + G_0^2]\ell + 2G_0 \sum_{j=1}^{\ell} (\ell - j) C_j \tag{2.24}$$

When $\gamma < 1$, the second term of the *rhs* can be approximated, when $\ell \gg 1$, by $2c_0 G_0 \ell^{2-\gamma}/(1-\gamma)(2-\gamma)$, which grows faster than the first term. In other words, the price would *superdiffuse*, or trend, at long times, with a volatility diverging with the lag ℓ. This, of course, does not occur: The market reacts to trade correlations so as to prevent the occurence of such trends. In fact, within the present linear model, the impact to single trades must be transient rather than permanent. Before explaining why and how this occurs in practice, let us first express mathematically how the efficiency of prices imposes strong constraints on the shape of the single trade impact function. For an arbitrary function $G_0(\ell)$, the lagged price variance can be computed explicitly and reads:

$$\mathcal{V}_\ell = \sum_{0\le j<\ell} G_0^2(\ell-j) + \sum_{j>0} [G_0(\ell+j) - G_0(j)]^2 + 2\Delta(\ell) + \Sigma^2\ell \qquad \textbf{(2.25)}$$

where $\Delta(\ell)$ is the correlation-induced contribution:

$$\begin{aligned}\Delta(\ell) = &\sum_{0\le j<k<\ell} G_0(\ell-j)G_0(\ell-k)C_{k-j}\\ &+ \sum_{0<j<k} [G_0(\ell+j) - G_0(j)]\,[G_0(\ell+k) - G_0(k)]\,C_{k-j}\\ &+ \sum_{0\le j<\ell}\sum_{k>0} G_0(\ell-j)\,[G_0(\ell+k) - G_0(k)]\,C_{k+j}\end{aligned} \qquad \textbf{(2.26)}$$

Assume that $G_0(\ell)$ itself decays at large ℓ as a power law, $\Gamma_0\ell^{-\beta}$. When $\beta, \gamma < 1$, the asymptotic analysis of $\Delta(\ell)$ yields:

$$\Delta(\ell) \approx \Gamma_0^2 c_0 I(\gamma, \beta)\ell^{2-2\beta-\gamma} \qquad \textbf{(2.27)}$$

where $I > 0$ is a certain numerical integral. If the single trade impact does not decay ($\beta = 0$), we recover the above superdiffusive result. But as the impact decays faster, superdiffusion is reduced, until $\beta = \beta_c = (1-\gamma)/2$, for which $\Delta(\ell)$ grows exactly linearly with ℓ and contributes to the long-term value of the volatility. However, as soon as β exceeds β_c, $\Delta(\ell)$ grows sublinearly with ℓ, and impact only enhances the high-frequency value of the volatility compared to its long-term value Σ^2, dominated by "news." We therefore reach the conclusion that the long-range correlation in order flow does not induce long-term correlations nor anticorrelations in the price returns if and only if the impact of single trades is transient ($\beta > 0$) but itself nonsummable ($\beta < 1$). This is a rather odd situation in which the impact is not permanent (since the long-time limit of G_0 is zero) but is not transient either because the decay is extremely slow. The convolution of this semipermanent impact with the slow decay of trade correlations gives only a finite contribution to the long-term volatility. The mathematical constraint $\beta = \beta_c$ will be given more financial flesh later.

Within this framework, one can also compute the average impact function $\mathcal{R}_\ell$. From Eq. 2.13 one readily obtains:

$$\mathcal{R}_\ell = G_0(\ell) + \sum_{0<j<\ell} G_0(\ell-j)C_j + \sum_{j>0} [G_0(\ell+j) - G_0(j)]\,C_j \qquad \textbf{(2.28)}$$

This equation can be understood as a way to extract the impact of single trades G_0 from directly measurable quantities, such as $\mathcal{R}_\ell$ and C_n; see Section 2.6.6 and Appendix 2.2. From a mathematical point of view, the asymptotic analysis can again be done when $G_0(\ell)$ decays as $\Gamma_0 \ell^{-\beta}$. When $\beta + \gamma < 1$, one finds:

$$\mathcal{R}_\ell \approx_{\ell \gg 1} \Gamma_0 c_0 \frac{\Gamma(1-\gamma)}{\Gamma(\beta)\Gamma(2-\beta-\gamma)} \left[\frac{\pi}{\sin \pi\beta} - \frac{\pi}{\sin \pi(1-\beta-\gamma)} \right] \ell^{1-\beta-\gamma} \tag{2.29}$$

where we have explicitly given the numerical prefactor to show that it exactly vanishes when $\beta = \beta_c$, which means that in this particular case one cannot satisfy oneself with the leading term. When $\beta < \beta_c$, one finds that $\mathcal{R}_\ell$ diverges to $+\infty$ for large ℓ, whereas for $\beta > \beta_c$, $\mathcal{R}_\ell$ diverges to $-\infty$, which is perhaps counterintuitive but means that when the decay of single trade impact is too fast, the accumulation of mean reverting effects leads to a negative long-term average impact—see Figure 2.7. When β is precisely equal to β_c, $\mathcal{R}_\ell$ tends to a finite positive value $\mathcal{R}_\infty$: The decay of single trade impact precisely offsets the positive correlation of the trades.

In this framework, volume fluctuations have been neglected. An extended version of the model, which is directly related to the discussion in the next section, is presented in Appendix 2.2 (see also Bouchaud et al., 2004).

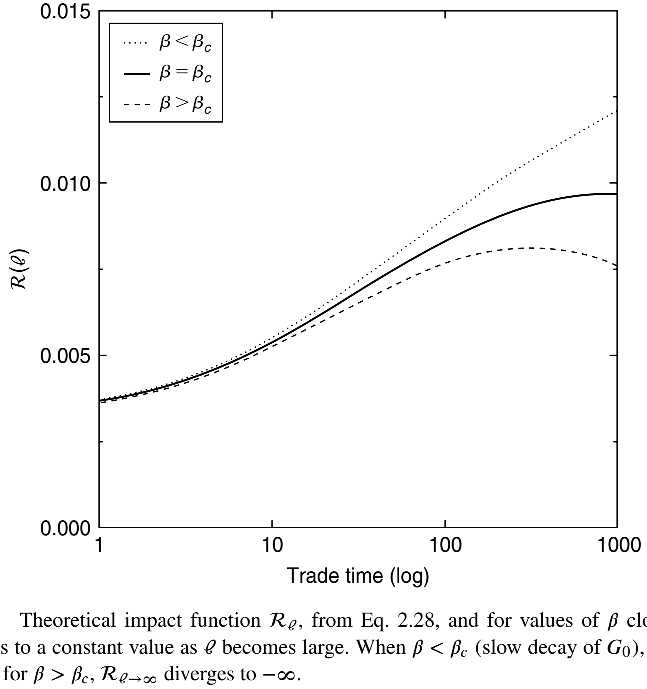


FIGURE 2.7 Theoretical impact function $\mathcal{R}_\ell$, from Eq. 2.28, and for values of β close to β_c. When $\beta = \beta_c$, $\mathcal{R}_\ell$ tends to a constant value as ℓ becomes large. When $\beta < \beta_c$ (slow decay of G_0), $\mathcal{R}_{\ell \to \infty}$ diverges to $+\infty$, whereas for $\beta > \beta_c$, $\mathcal{R}_{\ell \to \infty}$ diverges to $-\infty$.

2.6.5. History Dependent, Permanent Impact

An alternative interpretation of the preceding formalism is to assume that price impact is permanent, but history dependent as to ensure statistical efficiency of prices (Lillo and Farmer, 2004; Farmer et al., 2006; Gerig, 2007).

Predictable Order Flow and Statistical Efficiency

Let us consider a generalized MRR model:

$$r_n = m_{n+1} - m_n = \eta_n + \theta(\epsilon_n - \hat{\epsilon}_n), \quad \hat{\epsilon}_n = E_n[\epsilon_{n+1}|I] \tag{2.30}$$

where I is the information set available at time n. In line with our discussion in Section 2.3.8, we assume that in the market there are three types of traders. First, there are directional traders (liquidity takers) that have large hidden orders to unload and, by placing many consecutive orders with the same sign, create a correlated order flow. The second group of agents consists of the liquidity providers, who post bids and offers and attempt to earn the bid–ask spread. The third group is made by noise traders—that is, traders placing uncorrelated order flow. Anticipating the discussion in Section 2.7.3, it is indeed reasonable to assume that the strategies of the first two types of agents will adjust in such a way as to remove any predictability of the midpoint change; in other words, that $E_{n-1}[r_n|I] = 0$ as implied by Eq. 2.30. This is a plausible first approximation, although one can expect (and indeed observe) deviations from strict unpredictability at high frequencies.

Within the preceding simplified model, in which we have neglected volume fluctuations (see Appendix 2.2 for an attempt to include them), there are only two possible outcomes. Either the sign of the n^+ transaction matches the sign of the predictor $E_n[\epsilon_{n+1}|I]$ or they are opposite. Let us call r_n^+ and r_n^- the expected ex-post absolute value of the return of the n^{th} transaction, given that ϵ_n either matches or does not match the predictor. If we indicate with φ_n^+ and (φ_n^-) the *ex ante* probability that the sign of the n^{th} transaction matches (or disagrees) with the predictor ϵ_n, we can rewrite $E_{n-1}[r_n|I] = 0$ as:

$$\varphi_n^+ r_n^+ - \varphi_n^- r_n^- = 0 \tag{2.31}$$

Within the MRR model as described previously, this means

$$r_n^+ = \theta(1 - \hat{\epsilon}_n) \tag{2.32}$$

$$r_n^- = \theta(1 + \hat{\epsilon}_n) \tag{2.33}$$

This result shows that the most likely outcome has the smallest impact. We call this mechanism *asymmetric liquidity*: Each transaction has a permanent impact, but the impact depends on the past order flow and its predictability. The price dynamics and the impact of orders therefore depend on (1) the order flow process, (2) the information set I available to the liquidity provider, and (3) the predictor used by the liquidity provider to forecast the order flow.

Equivalence with the Transient Impact Model

In the following we consider the case where the information set available to liquidity providers is restricted to the past order flow. We call this information set *anonymous* because liquidity providers do not know the identity of the liquidity takers and are unable to establish whether or not two different orders come from the same trader. We assume also that the predictor used by liquidity takers to forecast future order flow comes from a linear model. In some cases, such as for an order flow generated according to the model presented in Section 2.4.3, this may not be an optimal predictor. However, linear time series models are probably the most widely used forecasting tools. Here we analyze a linear time series model based on the signs of executed transactions, and we assume a K^{th} order autoregressive AR model of the form

$$\hat{\epsilon}_n = \sum_{i=1}^{K} a_i \epsilon_{n-i} \qquad \textbf{(2.34)}$$

where a_i are real numbers that can be estimated on historical data using standard methods (see Lillo and Farmer, 2004; Bouchaud et al., 2004; and Appendix 2.2). The MRR model corresponds to an AR(1) order flow, with $a_1 = \rho$ and $a_k = 0$ for $k > 1$, with an exponential decay of the correlation.

The resulting impact model, Eq. 2.30 with a general linear forecast of the order flow, is in fact *equivalent*, when $K \to \infty$, to the temporary impact model of the previous section (see the appendix of Bouchaud et al., 2004). It is easy to show that one can rewrite the generalized MRR model in terms of a propagator as

$$m_n = m_{n-1} + \theta \epsilon_n + \sum_{i=1}^{\infty} [G(i+1) - G(i)] \epsilon_{n-i} + \eta_n, \quad \theta = G(1) \qquad \textbf{(2.35)}$$

The equivalence is obtained with the relation:

$$\theta a_i = G(i+1) - G(i) \quad \text{or} \quad G(i) = \theta[1 - \sum_{j=1}^{i-1} a_j] \qquad \textbf{(2.36)}$$

More General Information Models

In the previous section we saw that the fixed/temporary impact model is equivalent to the variable/permanent impact model under the additional assumptions that (1) the information set available to the liquidity provider is the set of the past order flow and (2) that liquidity providers use a linear forecast model to predict the future order flow from the past and to adjust price response. These two assumptions of the variable/permanent impact model are far from general. In the following we discuss the more general situations in which a different information set and forecast model can arise.

In most financial markets order flow is available in real time to all market participants; thus it is clear that any liquidity provider could use the past order flow time series to trade efficiently. However, in some cases participants can use information other than the time series of order flow signs. There are often indirect clues about the identity of orders such as the consistent use of particular round lots for orders that arrive at regular intervals. Activity in block markets can also provide clues about the activity of large orders. Another case is when a trader is trying to execute his large order by a so-called "slicing and dicing" algorithm. The liquidity provider could be able to detect the presence of this trader, and therefore the liquidity provider has additional information to add to his information set.

The algorithm used by the liquidity provider to forecast the future order flow depends on the information set and on the degree of sophistication of the liquidity provider. Even if linear forecasting methods are widespread, they can lead to suboptimal predictions if the time series one is trying to forecast is strongly nonlinear. For example, in Section 2.4.3 we discussed a microscopically based order flow model that reproduces the correlation properties observed in the real order flow. This model (Lillo, Mike, and Farmer, 2005) is clearly nonlinear. Despite the fact that an optimal forecast method for this order flow model is not easily available, one can find suboptimal nonlinear forecast models that outperform the linear forecast method. When one incorporates nonlinear forecast models in the variable/permanent impact model, the price dynamics will not be equivalent to the fixed/temporary model.

In conclusion, the variable/permanent model sets a general framework for describing the interaction between order flow and price dynamics. In a paper in progress, Gerig et al. (2008) show how different assumptions on the information set and on the forecast method lead to different functional forms of the impact of hidden orders and on the dynamical properties of prices.

Mechanisms for Asymmetric Liquidity

Let us rephrase in more intuitive terms the results established earlier. Due to the small outstanding liquidity, order flow must develop temporal correlations. This is such an obvious empirical fact that high-frequency traders/market makers quickly come to learn about it and adapt to it. In the simple MRR model where signs are exponentially correlated, the probability that a buy follows a buy is $p_+ = (1+\rho)/2$. The unconditional impact of a buy is θ (see Eq. 2.60); however, a second buy immediately following the first has a reduced impact equal to $\mathcal{R}_1^+ = \theta(1-\rho)$. The second buy is not as surprising as the first and therefore should impact the price less. A sell immediately following a buy, on the other hand, has an *enhanced* impact equal to $\mathcal{R}_1^- = \theta(1+\rho)$, in such a way that the conditional average impact of the next trade is zero: $p_+\mathcal{R}_1^+ + (1-p_+)\mathcal{R}_1^- \equiv 0$ Gerig (2007). This is the "asymmetric liquidity" effect explained previously (Lillo and Farmer, 2004; Farmer et al., 2006; and Gerig, 2007; see also Bouchaud et al., 2006, where it is called "liquidity molasses"). This mechanism is expected to be present in general; because of the positive correlation in order flow, the impact of a buy following a buy should be less than the impact of a sell following a buy—otherwise, trends would appear.

But what are the mechanisms responsible for asymmetric liquidity, and how can they fail (in which case markets cease to be efficient)? This is still an open empirical question that started to be investigated only recently. For example, Lillo and Farmer (2004) showed that when the order flow becomes more predictable, the probability that a market order triggers a price change is larger for market orders with the unexpected sign than for those with the expected one. Moreover, the same authors showed that the ratio between the volume of the market order and the volume at the opposite best is lower (higher) for market orders with an expected (unexpected) sign.

Another related basic mechanism is "stimulated refill": Buy market orders trigger an opposing flow of sell limit orders, and vice versa (Bouchaud et al., 2006). This rising wall of limit orders decreases the probability of further upward moves of the price, which is equivalent to saying that $\mathcal{R}_1^+ < \mathcal{R}_1^-$, or else that the initial impact of the first trade reverts at the second trade. This dynamical feedback between market orders and limit orders is therefore fundamental for the stability of markets and for enforcing efficiency. It can be directly tested on empirical data. For example, Weber and Rosenow (2005) have found strong evidence for an increased limit order flow compensating market orders.

Since such a dynamical feedback is so important to reconcile correlation in order flow with the diffusive nature of price changes, it is worth detailing its intimate mechanism a little further and insisting on cases where this feedback may break down. Recall our discussion of the market ecology in Section 2.3.8: Market participants can be, in a first approximation, classified as a function of their trading frequencies. Large latent demand arises from low-frequency participants; the decision to buy or sell can be considered as fixed over a time scale of a few hours or a few days, much longer than the average time between trades. These participants create long-term correlations in the sign of the trades. Higher-frequency traders try to profit from microstructural effects and short time predictability. Even if institutionally designated market makers are no longer present in most electronic markets, these high-frequency strategies are in fact akin to market making—they make money from providing liquidity to lower-frequency traders. This is why we often (incorrectly) call this category of participants *market makers*.[16] So, one should think of two rather large latent supply and offer quantities that await favorable conditions, in terms of both price and quantity, to be executed on the market. Then begins a kind of hide-and-seek game, where each side attempts to guess the available liquidity on the other side. A "tit-for-tat" process then starts, whereby market orders trigger limit orders and limit orders attract market orders. A buy trade at the ask (say) is a signal that an investor is indeed willing to trade at that particular price. But the seller who placed a limit order at the ask is also, by definition, willing to trade at that price. The natural consequence is that a flow of refill orders is expected to occur at the ask immediately after a buy trade (and at the bid after a sell).

[16] Of course, the preceding distinction between participants must be taken with a grain of salt: Low-frequency decisions may be executed using smart high-frequency algorithmic trading. In this case, the same participant is at the same time a low-frequency trader and a market maker.

In other words, optimized execution strategies that look for micro-opportunities impose strong correlations between market order flow of one sign and limit order flow of the opposite sign. Imagine a case where buy market orders eat up sell limit orders at the ask, with no refill. The ask then moves up one tick. By making the price more expensive, the flow of buy market orders slows and the probability that a sell limit order reappears at the previous ask increases. Imagine now that the refill process is too intense; sell limit orders at the ask now pile up. This has two effects: (1) the probability of a large market order that executes a large volume in one shot increases; (2) the large volume at the ask decreases the probability of further sell limit orders joining the queue because the priority of these new orders is low. Both cases (no refill or intense refill) therefore induce a clear feedback mechanism ensuring local stability of the order book.

The previous mechanism can be thought of as a dynamical version of the supply-demand equilibrium, in the following sense: Incipient up trends quickly dwindle because as the ask moves up, the buy pressure goes down, while the sell pressure increases. Conversely, liquidity induced mean reversion—keeping the price low—attracts more buyers and soon gives way. Such a balance between liquidity taking and liquidity providing is at the origin of the subtle compensation between correlation and impact explained previously. It is interesting to notice that several other dynamical systems operate similarly, with a competition between two antagonist systems; heartbeats is an interesting example: The sympathetic and parasympathetic system act in opposition to speed/slow the cardiac rhythm.

One easily envisions that such a subtle dynamical equilibrium can quickly break down; for example, an upward fluctuation in buy order flow might trigger a momentary panic, with the opposing side failing to respond immediately. These liquidity micro-crises are probably responsible for the large number of price jumps; if the feedback mechanism changes sign, this can even lead to crashes. The tug of war is a vivid illustration of this phenomenon. A major challenge of microstructure theory is to turn the previous qualitative story into a quantitative model for heavy-tailed return distributions and volatility clustering, with interesting potential ideas on how to limit the occurence of these liquidity micro-crises. We are convinced that a consistent theory of hidden liquidity and stimulated refill is well within reach at this stage.

2.6.6. Empirical Results

The section reviews how the preceding ideas can be directly tested and measured on high-frequency data.

We start with the full impact function, defined by Eq. 2.13, which is easily measured, at least when the lag ℓ is not too large. When ℓ becomes of the order of the number of daily trades or more, the error bar on $\mathcal{R}_\ell$ quickly becomes large. The main features of $\mathcal{R}_\ell$ are, however, quite robust from stock to stock and also across different markets. For example, $\mathcal{R}_\ell$ for France Telecom in 2002 is shown in Figure 2.8. One sees a mild increase by a factor $\lambda \sim 2$ between $\ell = 1$ and $\ell = 1000$ before a saturation or maybe a decline for larger lags. This behavior is quite typical, in particular the

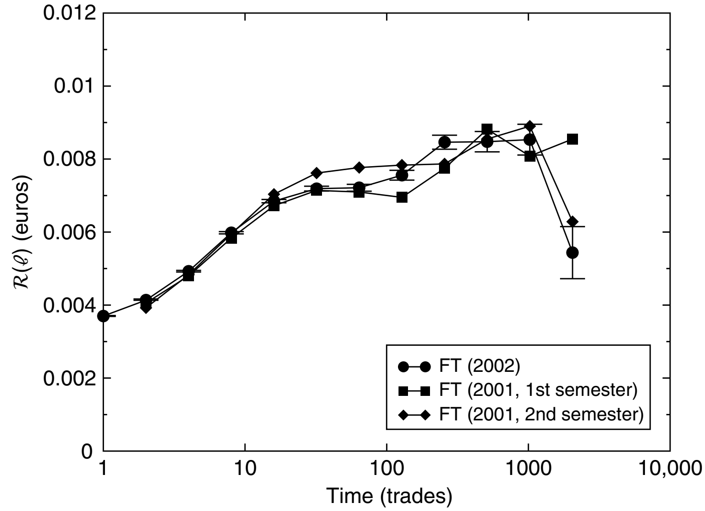


FIGURE 2.8 Average empirical response function $\mathcal{R}_\ell$ for FT during three different periods (1st and 2nd semester of 2001 and 2002); error bars are shown for the 2002 data. For the 2001 data, the y axis has been rescaled such that $\mathcal{R}_1$ coincides with the 2002 result. $\mathcal{R}_\ell$ is seen to increase by a factor ~2 between $\ell = 1$ and $\ell = 100$.

roughly twofold increase between small lags and large lags. So $\mathcal{R}_\ell$ reveals some non-trivial temporal structure; recall that $\mathcal{R}_\ell$ is constant within models where the midpoint reacts to surprise in order flow. In an MRR setting, the amplification factor λ should be $1/(1 - C_1)$, which is found to be in the range of 1.2 to 1.4, still too small to explain $\lambda \sim 2$.

As noted, one can in fact extract the theoretical impact of single trades $G_0(\ell)$ from the empirically measured impact $\mathcal{R}_\ell$ and the correlation between the sign of the trades C_ℓ, using Eq. 2.28. This was done in Bouchaud et al. (2006) and indeed produces nice, power-law decaying $G_0(\ell)$'s; see Figure 2.9 for a few examples. Within the previous restrictive theoretical framework, this provides a direct proof of the transient nature of the impact of single market orders and the long-term resilience of markets. This is quite important as far as execution strategies are concerned; see Section 2.10.

We should, however, list a number of caveats. One is the assumption that the impact is time translation invariant—that is, only the lag ℓ is relevant. This is clearly questionable, since strong intraday seasonality effects are expected. For example, there are indications that the trade sign correlation function C_ℓ for a given lag ℓ is quite different intraday and from one day to the next (Eisler et al., 2008). Similarly, we expect that the single trade impact should decay differently intraday and overnight. Second, we have to a large extent discarded the interesting correlations between the state of the order book Ω_n, the incoming volume v_n and the resulting impact (see Eq. 2.11). All this complexity was replaced by an average description: $\epsilon_n f(v_n; \Omega_n) \longrightarrow \epsilon_n \ln v_n$. Certainly, a refined version is needed, in particular because the fluctuations of $f(v_n; \Omega_n)$ will contribute to

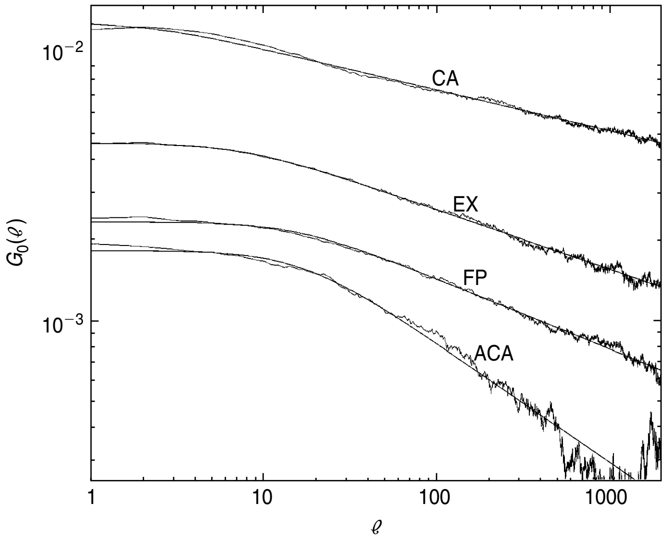


FIGURE 2.9 Comparison between the empirically determined $G_0(\ell)$, extracted from $\mathcal{R}$ and $\mathcal{C}$ using Eq. 2.28, and the power-law fit $G_0^f(\ell) = \Gamma_0/(\ell_0^2 + \ell^2)^{\beta/2}$ for a selection of four stocks: ACA, CA, EX, and FP.

the diffusion properties (see Eq. 2.25). Finally, we have chosen from the start to give a special role to market orders, as if only those impact the price. But this is not true: Obviously, limit orders also impact the price. In fact, it is precisely the impact of limit orders that offsets that of market orders and leads to a decay of the single trade impact $G_0(\ell)$. In other words, we have studied an effective model in terms of market orders only, dumping into $G_0(\ell)$ the counteracting effect of limit orders. A more symmetric version of the model, which treats market and limit orders on an equal footing, would be quite enticing (Eisler et al., 2008).

We now consider some empirical evidence for asymmetric liquidity. Figure 2.10 shows the behavior of the conditional returns r^+ and r^- defined in Eq. 2.33 as a function of the sign predictor $\hat{\epsilon}$. The data we show in Figure 2.10 is for Astrazeneca, a stock traded at the LSE. The sign predictor is the linear predictor defined in Eq. 2.33. The larger the absolute value of $\hat{\epsilon}$, the stronger the predictability of the next market order sign. We have plotted the average value of the return conditioned to be in the direction of the predictor, r^+, and the average return when the sign of the predictor is wrong, r^-. We see that r^- is indeed larger than r^+ and this difference increases with the predictability of the order flow. This is a clear evidence for asymmetric liquidity. Note also that both r^+ and r^- are approximately described by a linear function of the predictor $\hat{\epsilon}$. This is expected under the model described in the "Predictable Order Flow and Statistical Efficiency" section (see Eq. 2.33). However, the slopes of r^+ and r^- vs. $\hat{\epsilon}$ are different, challenging the implicit symmetric assumption (Eq. 2.33) in the MRR model. Other evidence for the buildup of the "liquidity molasses" accompanying the flow of market order can be found in Bouchaud et al. (2006) and Weber and Rosenow (2005).

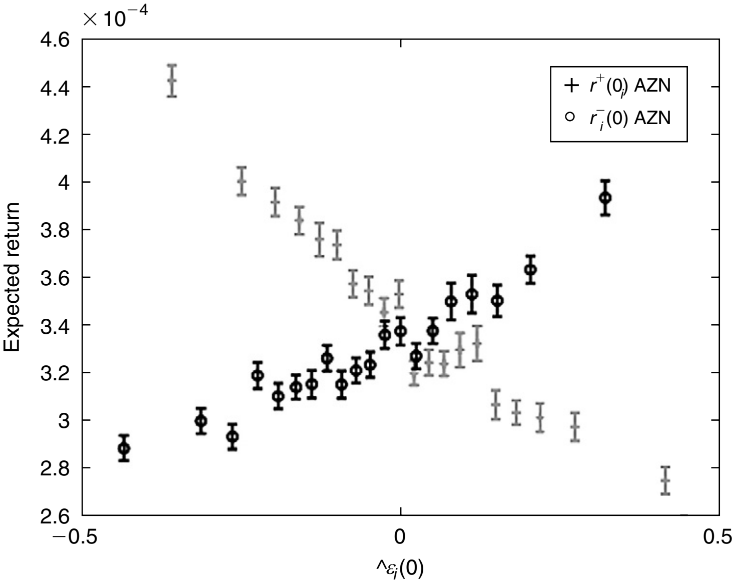


FIGURE 2.10 Expected return as a function of the sign predictor $\hat{\epsilon}$. The quantity r^+ (r^-) refers to trades with a sign that is equal (opposite) to the one of the predictor. Data are binned in such a way that each point contains an equal number of observations. Error bars are standard errors. (*Source:* Adapted from Gerig, 2007.)

2.6.7. Impact of a Large Hidden Order

We now want to calculate, within the stated theoretical framework, the impact of a hidden order of size N. For simplicity, let us first assume that the hidden order is made of N consecutive trades made by the same institution, though this remains "hidden" if trades are anonymous. Let us call m_0 the price at the beginning of the hidden order and compute the average price m_{N+t} observed t transactions after the completion of the hidden order. Within the generalized MRR model with a linear predictor of the order flow, a straightforward calculation shows that

$$E[m_{N+t}] - m_0 = \epsilon\theta \sum_{i=t+1}^{t+N} [1 - \sum_{j=1}^{i-1} a_j] \tag{2.37}$$

For $t = 0$ this expression gives the (temporary) total impact of the hidden order, whereas for $t > 0$ we can calculate the price reversion after the completion of the hidden order, and the permanent impact (if any) for $t \to \infty$.

This result can be generalized to the case where there is only one hidden order active at a given time, which mixes with a flow of uncorrelated orders with a constant participation rate π. The total time needed to execute the hidden order is then $T = N/\pi$. It is possible to show in this case that (Farmer, Gerig, Lillo, and Waelbroeck, 2008):

$$E[m_N] - m_0 = \epsilon\theta \sum_{i=1}^{N} \left(1 - \sum_{k=1}^{i/\pi} a_k\right) \tag{2.38}$$

Let us estimate this formula in the case where the autocorrelation C_τ of order flow asymptotically decays as a power law $C_\tau \sim \tau^{-\gamma}$ for large τ. There are several different ways of generating and forecasting long-memory processes. Here we assume that the participants observing public information model the time series with a FARIMA process. It is known (Beran, 1994) that for large k the best linear predictor coefficients of a FARIMA process satisfy $a_k \approx k^{-\beta-1}$, where $\beta = (1-\gamma)/2$. For large k we can go into the continuum limit, and from Eq. 2.38 the impact is:

$$E[m_N] - m_0 = \epsilon\theta \left[1 + \sum_{i=1}^{N-1} \left(1 - \left(1 - (n/\pi)^{-\beta}\right)\right)\right] \tag{2.39}$$

Converting the sum to an integral gives:

$$E[m_N] - m_0 \approx \epsilon\theta \left(1 + \frac{2^{\beta-1}\pi^{\beta}}{1-\beta}[(2N-1)^{1-\beta} - 1]\right) \sim \pi^{\beta} N^{1-\beta} \tag{2.40}$$

Thus, for a fixed participation rate, the market impact asymptotically increases with the length of the hidden order as $N^{1-\beta}$. A typical decay exponent for the autocorrelation of order signs is $\gamma \approx 0.5$ (Lillo and Farmer, 2004; Bouchaud et al., 2004), which means that $\beta \approx 0.25$. This means that according to the linear time series model, the impact should increase as roughly the $\frac{3}{4}$ power of the order size. An interesting property of this solution is that it depends on the speed of execution. The size of the impact varies as π^{β}. This means that the more slowly an order is executed, the less impact it has, and in the limit as the order is executed infinitely slowly, the impact goes to zero. Note, however, that if the execution time $T = N/\pi$ is *fixed*, the impact becomes linear with N but decays as $T^{-\beta}$.

To investigate the reversion dynamics, we again make use of the Eq. 2.37. We assume that the liquidity provider uses a FARIMA model to forecast order signs, and for the sake of simplicity in the following, we assume that $\pi = 1$—that is, that there are no noise traders. Realistically, the regression made by the liquidity provider on past signs will use a finite lag K, leading to:

$$\hat{\epsilon}_n = \sum_{i=1}^{K} a_i^{(K)} \epsilon_{n-i} \tag{2.41}$$

where (Beran, 1994):

$$a_i^{(K)} = -\binom{K}{i} \frac{\Gamma(i - H + 1/2)\Gamma(K - H - i + 3/2)}{\Gamma(1/2 - H)\Gamma(K - H + 3/2)} \tag{2.42}$$

and $H = 1/2 - \beta$ is the Hurst exponent of the FARIMA process. It is possible to derive an analytical exact result for the permanent impact. In fact, from Eq. 2.37, one can obtain:

$$E[m_\infty] - m_0 = \epsilon\theta N(1 - \sum_{j=1}^{K} a_j^{(K)}) = \epsilon\theta N \frac{4^{H-1}\sqrt{\pi}\,\Gamma[H]\,\sec[(K-H)\pi]}{\Gamma(3/2 + K - H]\Gamma[2H - 1 - K]} \quad \textbf{(2.43)}$$

Using the Stirling formula and the reflection formula for the Gamma function, one can show that for large K, the permanent impact scales as:

$$E[m_\infty] - m_0 \sim \epsilon\theta \frac{N}{K^\beta} \quad \textbf{(2.44)}$$

If K is infinite, $E[m_\infty] - m_0 = 0$; that is, the impact is completely temporary. This can be shown in the mathematically equivalent propagator model Bouchaud et al. (2004, 2006). For a FARIMA forecast model with finite K (or equivalently, if the sign autocorrelation function decays fast beyond time scale K), the permanent impact is nonzero and is linear in N. Even if for large K the permanent impact is small, the convergence to zero with the memory K is very slow.

Another interesting issue that can be discussed within the model is the decay of the impact immediately after the end of the hidden order (defined by Eq. 2.37). One finds that the initial drop for $t \ll N$ is in fact very sharp for $\beta < 1$: $m_{N+t} - m_N \propto -t^{1-\beta}$, such that the slope of the decay is infinite when $t \to 0$ (in the continuous limit).

2.6.8. Aggregated Impact

Impact is often measured not on a trade-by-trade level but rather on a coarse-grained time scale, say five minutes or a day. One then speaks of positive correlations between signed order flow and price returns. At the level of single trades, impact is strongly concave in volume and decays in time. How does this translate at a coarse-grained level? In Section 2.5.2 we have discussed this from an empirical point of view. Here we show how the impact theories we have developed so far make predictions about the impact function, following the approach of Lillo et al. (2008a).

Suppose one aggregates the returns and volumes of N consecutive trades (not necessarily from the same hidden order). Using the same notation as in Section 2.5.2, the total volume imbalance is $Q_N = \sum_{n=0}^{N-1} \epsilon_n v_n$. Conditioned to a particular value $Q_N = Q$, what is the average price return $R(Q)$? The answer to this question depends on the order flow and the properties of the impact function. In the following discussion, we consider two extreme cases. In the first case we consider an unrealistic model where the order flow is described by an independent identically distributed random process, and the impact is fixed and permanent. In the second case we consider a correlated order flow and a fixed/temporary impact model.

Independent Identically Distributed Order Flow

If the unconditional distribution of market order volume and the functional form of the impact function are known, it is possible to find a closed expression for the impact $R(Q)$. Consider a series of N transactions with signed[17] volumes v_i corresponding to total return $R = \sum_{i=1}^{N} r_i$ and total signed volume $Q = \sum_{i=1}^{N} v_i$. The expected return given Q can be written:

$$\begin{aligned} R(Q, N) \equiv E[R|Q] &= \int RP(R|Q, N)\, dR \\ &= \frac{1}{P_N(Q)} \int RP(R, Q, N)\, dR \end{aligned} \tag{2.45}$$

where $P_N(Q)$ is the probability density for Q. We assume that the N individual price impacts r_i due to the IID signed volumes v_i are given by a deterministic function[18] $r_i = f(v_i)$. Let the distribution of individual v_i be $p(v_i)$. Then the joint distribution of v_i is $P(v_1, \ldots, v_N) = p(v_1) \ldots p(v_N)$. The integral above becomes:

$$\int RP(R, Q, N)\, dR = \int dv_1 \ldots dv_N p(v_1) \ldots p(v_N) \sum_{i=1}^{N} f(v_i)\delta(Q - \sum_{i=1}^{N} v_i) \tag{2.46}$$

where we introduced the Dirac delta function.

By making use of the integral representation of the Dirac delta function, after some manipulations it is possible to rewrite $R(Q, N)$ as:

$$R(Q, N) = \frac{N}{2\pi} \frac{1}{P_N(Q)} \int d\lambda e^{(N-1)h(\lambda)} g(\lambda) e^{-i\lambda Q} \tag{2.47}$$

where $h(\lambda)$ is the logarithm of the Fourier transform of the volume distribution and $g(\lambda)$ is the Fourier transform of the product of the volume distribution and the impact function. Moreover, $P_N(Q)$ is the probability density that the total signed volume in the N trades is Q.

The functional form of the aggregate impact $R(Q, N)$ can be calculated by integrating this expression. It is possible to show that many of the properties of the solution are robust, independent of the details of the model. For small values of Q the aggregate impact $R(Q)$ is always linear with a slope that depends on N and on the details of the volume distribution and the impact function. For example, if the impact function is a power-law function $\epsilon|v|^{\psi}$ and the volume distribution decays asymptotically as $P(V) \sim V^{-\alpha-1}$, for large N the aggregate impact behaves for small Q as:

$$R(Q, N) \sim \frac{Q}{N^{\kappa}} \tag{2.48}$$

where κ depends in a nontrivial way on α and ψ (see Lillo et al., 2008a). For example, if volumes have a finite second moment and the impact function is concave, then

[17] Only in this subsection we indicate with v_i the signed and not the absolute value of volume.

[18] The results remain the same if a noise term is added to the impact function.

$\kappa = 0$; in contrast, if the second moment of the volume does not exist and the impact function is sufficiently concave, then $\kappa > 0$. The latter case agrees with what is seen in Figure 2.5, where the slope of the aggregate impact decreases with N. Thus theories for the aggregate impact make falsifiable predictions connecting volumes, order flow, and impact.

Transient Impact Model

Within the model of Section 2.6.4, the aggregate impact reads:

$$R(Q,N) = \sum_{n=0}^{N-1} G_0(N-n)E[q_n|Q] + \sum_{m<0}[G_0(N-m) - G_0(-m)]E[q_m|Q] \tag{2.49}$$

where $q_n = \epsilon_n \ln v_n$, and we assume that volumes are lognormally distributed (see Appendix 2.2). Because trades are long ranged correlated, the second term is nonzero. But one can show it is subdominant when $N \gg 1$, so we discard it in a first approximation. In the first term, one can compute $E[q_n|Q] = x$ using the fact that the q_ns are, within the model, Gaussian with $rms = s$. Noting also that typical values of Q are of order $N^{1-\gamma/2} \ll N$, one finds:

$$x \approx \frac{sQ}{\mathcal{I}N}, \quad \mathcal{I} = 2\int_0^\infty \mathrm{d}u\, u\, e^{us-u^2/2} \tag{2.50}$$

With $R(Q,N) \approx \Gamma_0 N^{1-\beta} x/(1-\beta)$ and the previous relation between $\beta = (1-\gamma)/2$, we finally find the following result, written in a suggestive scaling form:

$$R(Q,N) = \sqrt{N}\frac{s\Gamma_0}{\mathcal{I}(1-\beta)}\left(\frac{Q}{N^{1-\gamma/2}}\right) \tag{2.51}$$

This means that by rescaling the return and the signed volume by their respective root mean square value, one obtains at large N a limiting curve that is a straight line. Whereas for small N impact is strongly concave, impact becomes linear when $N \gg 1$. One can go one step further and compute the leading nonlinear correction in Q when N is large. One finds that it is negative, as a remnant of the small N concavity, and becomes noticeable at increasingly larger values of $Q \sim N$, as seen in empirical data; see Figure 2.5.

The important conclusion of this model is that although the impact of individual trades is concave and decays in time, the compensating effect of correlated trades leads to a well-defined *linear relation* between order imbalance and returns at an aggregated level. This is important because such a relation is often interpreted as a manifestation of the permanent component of the impact.

Is this linear relation telling us that part of the trades have indeed correctly *predicted* the aggregated return (in Hasbrouck's words); see Hasbrouck (2007)? In light of all the previous results, it looks much more plausible to us that anonymous trades in fact statistically *induce* price changes, although in a quite nontrivial and perhaps unexpected fashion.

2.7. THE DETERMINANTS OF THE BID–ASK SPREAD

In modern electronic markets, liquidity is *self-organized* in the sense that any agent can choose, at any instant of time, to either provide liquidity or consume liquidity. The liquidity of the market is partially characterized by the bid–ask spread S, which sets the cost of an instantaneous round trip of one share (a buy instantaneously followed by a sell, or vice versa).[19] A liquid market is such that this cost is small. A question of both theoretical and practical crucial importance is to know what fixes the magnitude of the spread in the self-organized setup of electronic markets and the relative merit of limit vs. market orders. In the economics literature (O'Hara, 1995; Biais et al., 1997; Madhavan, 2000; Glosten and Milgrom, 1985), the existence of the bid–ask spread is often attributed to three types of liquidity providing costs (Stoll, 1978):

- Order processing costs (this includes the profit of the market maker)
- Adverse selection costs—liquidity takers may have superior information on the future price of the stock, in which case the market maker loses money
- Inventory risk—market makers may temporarily accumulate large long or short positions that are risky; if agents are risk sensitive and have to limit their exposure, this may add extra costs

A somewhat surprising conclusion of early econometric studies is that order processing costs account for a large fraction of the spread. This may make sense in illiquid markets where market makers exploit a monopolistic situation to open large spreads but cannot be the correct picture in highly liquid, electronic markets in which market making is highly competitive. What we argue is that the main determinant of the spread is in fact impact.

2.7.1. The Basic Economics of Spread and Impact

What are the basic economics behind a trade—that is, the encounter between a liquidity taker and one (or several) liquidity provider(s)?

The Average Gain of Market Makers

Consider the sequence of all trades (not necessarily coming from the same hidden order). Let the n^{th} trade have volume v_n and sign ϵ_n. The profit collectively made by

[19] Other determinants of liquidity discussed in the literature are the depth of the order book and market resiliency, see Black (1971), Kyle (1985).

liquidity providers on that given trade, marked to market at time $n + \ell$, is given by:

$$\mathcal{G}_L(n, n + \ell) = v_n \epsilon_n \left[(m_n + \epsilon_n \frac{S_n}{2}) - m_{n+\ell} \right] \quad \textbf{(2.52)}$$

where S_n is the value of the spread at that moment in time. Think of a buy trade $\epsilon_n = +1$. This equation compares the money received by the liquidity provider when the trade occurs $(v_n(m_n + \frac{S_n}{2}))$ to its mark-to-market (midpoint) price at time $n + \ell$. Symmetrically, the profit made by the liquidity taker using market orders is $\mathcal{G}_L(n, n + \ell) = -\mathcal{G}_M(n, n + \ell)$. This equation clearly shows that the profitability of market making comes from the spread $(+S_n/2)$, whereas the losses are induced by market impact $(-\epsilon_n(m_{n+\ell} - m_n))$, which may or may not come from more informed traders (see the following discussion).

Neglecting for simplicity volume fluctuations at this stage $(v_n \equiv v)$ and using Eq. 2.13, we see that the average gain of the market maker in the absence of extra costs is given by:

$$E[\mathcal{G}_L](\ell) = v \left(E[\frac{S}{2}] - \mathcal{R}_\ell \right) \quad \textbf{(2.53)}$$

which shows explicitly that for a given total market impact $\mathcal{R}_\ell$, the spread S should be larger than a minimum value for market-making strategies to be at all profitable on a time scale ℓ—or else, for a given value of S, the impact function $\mathcal{R}_\ell$ should be as small as possible. We recover here the idea that it is in the interest of liquidity providers to control the growth of $\mathcal{R}_\ell$ by tuning the liquidity asymmetry.

In fact, this reasoning neglects the cost of unwinding the market-maker position, and a better estimate will be provided later in this article. But the main message of the preceding simple computation is that the spread compensates for the impact of market orders. In the microstructure literature, this is refered to as *adverse selection*; as alluded to previously, this implies that market orders originate from better informed traders, with an information on the future price on average worth $\mathcal{R}_\ell$. But the same result would hold if impact was purely statistical, with no information content whatsoever. In fact, one could even revert the logic and claim that it is the spread that determines the impact: If *some* traders accepted to pay $m_n + S_n/2$ for the stock, it is natural that the market as a whole revises its fair-price estimate from m_n to $m_n + \alpha S_n/2$, where $\alpha \geq 0$ is a number measuring how trades influence the participants beliefs, leading to $\mathcal{R}_\infty = \alpha S/2$. The MRR model with spread (see The MRR Model with a Bid–Ask Spread section), in this context, assumes that market participants believe that the last traded price is indeed the correct price $(\alpha = 1)$. Clearly, in that model, the cost of a market order or the gain of a limit order are exactly zero. This leaves us, by the way, in the familiar but uncomfortable situation of the "no trade theorem": If the spread is such that the information content of a market order is compensated, why would the informed trader trade at all?

How Informed Are the Trades?

So, are some market orders informed? Can one find convincing ex-post signatures of informed trades? A minimal definition of an informed trade is a trade that earns a profit significantly larger than the transaction costs (including both brokerage fees and market slippage). Introducing the signed return $r(n, n+\ell) \equiv \epsilon_n(m_{n+\ell} - m_n)$, the profit of the n^{th} market order on time scale ℓ is:

$$\mathcal{G}_M(n, n+\ell) = v_n \left[r(n, n+\ell) - \frac{S_n}{2} \right] \qquad \textbf{(2.54)}$$

Note that by definition the average of $r(n, n+\ell)$ is equal to the total impact $\mathcal{R}_\ell$, which is positive. If one averages this equation over *all* trades, one in fact finds that $E[\mathcal{G}_M]$ is close to zero, which means that the spread compensates for the average impact, at least measured on short time scales ℓ (between a few seconds to a few days). More precisely, on liquid NYSE stocks in 2005 (when market makers were still present), one finds that $E[\mathcal{G}_M]$ is zero within error bars, which means that, after transaction costs, market orders lose money, on average. The situation is slightly better for liquid PSE stocks in 2002, where one finds $E[\mathcal{G}_M] = gE(S)/2$ with $g \approx 0.3$ (see Figure 2.14 later). This amounts to 3 to 5 bp per trade, close to the transaction costs. So, on average, and although market orders do impact prices, there does not seem to be much *short-term* information in these orders, at least judging from their ex-post profitability. The question of longer-term information is, of course, left open here, simply because the statistics are not sufficient to judge the average profitability of trades on long time scales and because long-term drift effects cannot be neglected. (On average, buy trades are profitable in the long run!)

We can look in more detail at the full distribution of $r_\ell \equiv r(n, n+\ell)$, $P(r_\ell)$, which contains much more information. Note that its second moment $E[r^2|\ell]$ is very close to the volatility on scale ℓ, which soon becomes much larger than $\mathcal{R}^2$ when ℓ increases. Concerning the shape of $P(r_\ell)$, two extreme scenarios could occur (see Figure 2.11 for an illustration):

- A small proportion of well-informed trades *predict* the future price while a majority of trades are uninformed and do not impact the price at all. The distribution of r_ℓ should then be composed of a broad blob, symmetric around $r_\ell = 0$, corresponding to uninformed trades, plus a hump (or more plausibly, a broad shoulder) on the positive side, corresponding to well-informed trades. The nonzero value of $E[r_\ell]$ comes from these informed trades. This is the scenario behind, for example, the Kyle model or the Glosten-Milgrom model.
- All trades are equally weakly informed or even not informed, but *all* statistically impact prices. In this case one expects a symmetric broad blob but around the average impact $E[r_\ell]$.

Empirically, the distribution of r_ℓ is found to be very close to the second picture for ℓ corresponding to intraday time scales. In particular, no noticeable asymmetry

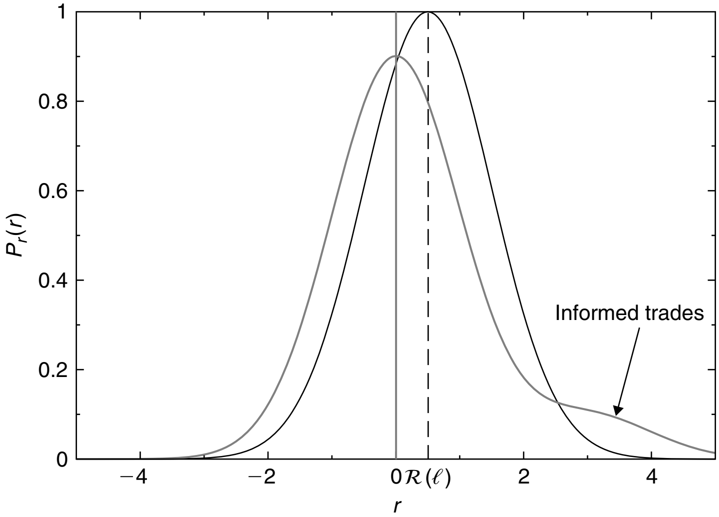


FIGURE 2.11 Two extreme cases for the distribution $P(r_\ell)$ of signed returns r_ℓ. Thick curve means nearly all trades are uninformed but impact prices, leading to a symmetric $P(r_\ell)$ around a nonzero average impact. Narrow curve means most trades are uninformed and do not impact prices, while some trades are informed and predict correctly the future return, leading to a thick tail in the $r_\ell > 0$ region.

(beyond the existence of a nonzero value of $E[r_\ell]$) is observed on liquid stocks; see Figure 2.12 for an example. This suggests that trades, on average, impact prices but do not seem to "predict" future prices—at least not on short time scales. The strong relation between order imbalance and price returns would then be a tautological consequence of this impact (see Section 2.6) and not a signature of "true" information revelation.

2.7.2. Models for the Bid–Ask Spread

The Glosten-Milgrom Model

One of the earliest theories of the spread that makes the preceding discussion is the sequential trade model of Glosten and Milgrom (Glosten and Milgrom, 1985). One assumes that market orders are either due (with some probability q) to informed traders, who know the end-of-day price p_f, or (with probability $1 - q$) to noise traders. The value of q is assumed to be known by the market maker, which is not necessarily very realistic (a similar assumption is made within the Kyle model). The end-of-day price p_f can either be above ($p_>$) or below ($p_<$) the open price. The probabilities for either outcome at the start of the day are $\delta_+ = \delta_- = 1/2$ for simplicity. But as trading occurs, either at the bid or at the ask, the market maker updates in a Bayesian way the value of $\delta_+ = 1 - \delta_-$: trades at the ask increase the value of δ_+, whereas trades at the bid increase δ_-.

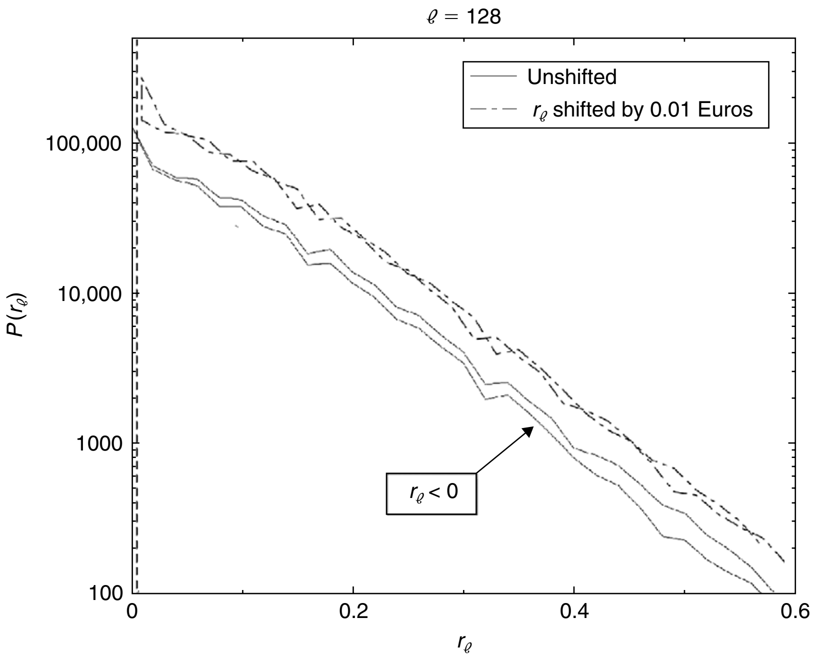


FIGURE 2.12 Probability distribution $P(r_\ell)$ of the quantity $r = (m_{n+\ell} - m_n).\varepsilon_n$ (in Euros) for $\ell = 128$; data are again for France Telecom during 2002. The negative part of the distribution has been folded back to positive r to highlight the small positive assymetry of the distribution. The average value $\mathcal{R}_\ell = E[r] \approx 0.01$ is shown by the vertical dashed line. The dashed-dotted line corresponds to the distribution of $r - 0.01$, for which no asymmetry of the type shown in Figure 2.11 can be detected. This curve has been shifted upwards for clarity.

This leads to a certain update rule for δ_+ as a function of the sign of the next trade, which we do not write here explicitly. Anticipating the value of $\delta_\pm$ after the next trade allows the market maker to position his quotes in such a way as not to have *ex-post* regrets. More precisely:

$$a = \delta_+(+)p_> + \delta_-(+)p_<, \quad b = \delta_+(-)p_> + \delta_-(-)p_< \tag{2.55}$$

where (±) refers to the sign of the next trade. This leads to the following prediction for the bid–ask S_n after the n^{th} trade:

$$S_n = 4q\delta_+^{(n)}\delta_-^{(n)}(p_> - p_<) \tag{2.56}$$

where $\delta_\pm^{(n)}$ is the updated value of $\delta_\pm$ after n trades (with $\delta_\pm^{(0)} = 1/2$), and we have neglected terms of order q^2, which must be small if this model is to be realistic. This model is by construction compatible with a random walk for the midpoint, with a volatility per trade σ_1 proportional to the bid–ask spread, as reported later in this chapter. It also predicts that the bid–ask spread declines on average throughout the day, since the update rule drives δ_+ either to zero or to one: As trading occurs, the market maker discovers more accurately which outcome is more likely.

A detailed comparison of this model with empirical data is given in Wiesinger et al. (2008). Here we simply note that as far as order of magnitude goes, the spread at the beginning of the day (when $\delta_\pm = 1/2$) is typically 0.1%, whereas the daily volatility fixes the order of magnitude of $p_> - p_<$ to typically 2%, leading to $q \sim 0.05$. Within this framework, one finds again that the fraction of short-time "informed trades" must be small. One also finds that in this model the spread decays exponentially fast with time, at variance with the slow, power-law relaxation that has been observed (see Section 2.7.4).

The MRR Model with a Bid–Ask Spread

The original MRR model is in fact slightly different from the model described in Section 2.6.3. MRR model rather assumes that it is the "true" fundamental price p_n rather than the midpoint m_n, which is impacted by the surprise in order flow, and hence:

$$p_{n+1} - p_n = \eta_n + \theta[\epsilon_n - \rho\epsilon_{n-1}] \qquad \textbf{(2.57)}$$

MRR then specifies a rule for the bid and ask price, which in turn allows one to *compute* the midpoint m_n. Since market makers cannot guess the surprise of the next trade, they post a bid price b_n and an ask price a_n given by:

$$a_n = p_n + \theta[1 - \rho\epsilon_{n-1}] + \phi, \quad b_n = p_n + \theta[-1 - \rho\epsilon_{n-1}] - \phi \qquad \textbf{(2.58)}$$

where ϕ is the extra compensation claimed by the market maker, covering processing costs and the shock component risk. This rule ensures no *ex-post* regrets for the market maker: Whatever the sign of the trade, the traded price is always the "right" one. The midpoint $m \equiv (a + b)/2$ immediately before the n^{th} trade is now given by:

$$m_n = p_n - \theta\rho\epsilon_{n-1} \qquad \textbf{(2.59)}$$

whereas the spread is given by $S = a - b = 2(\theta + \phi)$.

More generally, assuming that only the sign surprise matters, one can write, for arbitrary correlations between signs:

$$m_{n+\ell} - m_n = \sum_{j=n}^{n+\ell-1} \eta_n + \theta \sum_{j=n}^{n+\ell-1} \left\{\epsilon_j - E_j\left[\epsilon_{j+1}\right]\right\} \qquad \textbf{(2.60)}$$

where the last term is the conditional expectation of the next sign. In the Markovian case, $E_j[\epsilon_{j+1}] = \rho\epsilon_j$, and we recover the previous result. The impact function, in the general case, reads:

$$\mathcal{R}_\ell = \theta\,[1 - C_\ell] \qquad \textbf{(2.61)}$$

Using Eq. 2.53, one sees that the long-term profit of market makers is zero. However, due to correlations between trades, the longtime impact is enhanced compared to the short-term impact by a factor:

$$\lambda = \frac{1}{1 - C_1} > 1 \tag{2.62}$$

As we've discussed very generally, spread and impact are two sides of the same coin. This is particularly clear within the MRR model, where the half-spread $S/2$ is set to be equal to the long-term impact $\mathcal{R}_\infty = \theta$. This means that the profit of market makers is exactly zero (provided $\phi = 0$), but also, as noted previously, that the profit of putatively informed market orders is zero. The spread in the MRR model is

$$S = 2(\theta + \phi) = 2(\mathcal{R}_\infty + \phi) = 2\lambda\mathcal{R}_1 + 2\phi \tag{2.63}$$

where $\lambda = (1 - \rho)^{-1}$. Appendix 2.3 provides an alternative, enlightening derivation.

One computes the midpoint volatility on scale ℓ, defined as

$$\sigma_\ell^2 = \frac{1}{\ell}\langle(m_{\ell+i} - m_i)^2\rangle \tag{2.64}$$

One finds a sum of a trade-induced volatility $\theta^2(1-\rho)^2$ and a "news"-induced volatility Σ^2:

$$\sigma_1^2 = \langle(m_{n+1} - m_n)^2\rangle = \Sigma^2 + \theta^2(1-\rho)^2 \tag{2.65}$$

and

$$\sigma_\infty^2 = \Sigma^2 + \theta^2(1-\rho)^2(1 + 2\frac{\rho}{1-\rho}) = \Sigma^2 + \theta^2(1-\rho^2) \geq \sigma_1^2 \tag{2.66}$$

The MRR model therefore leads to two simple relations among spread, impact, and volatility per trade

$$S = 2\lambda\mathcal{R}_1 + 2\phi \quad \sigma_1^2 = \mathcal{R}_1^2 + \Sigma^2 \tag{2.67}$$

where $\lambda = (1-\rho)^{-1}$ and ϕ is any extra compensation claimed by market makers. These relations are generalized to more realistic assumptions and tested empirically in the next two sections.

2.7.3. Limit vs. Market Orders: The Microstructure Phase Diagram

Market Order Strategies

As we have mentioned, the gain (or cost) of a given market order can be defined as $v_n[r(n, n+\ell) - \frac{S_n}{2}]$. This definition in fact marks the trade to market after ℓ trades and is often referred to as the *realized spread* (Bessembinder, 2003; Stoll, 2000). The volume-weighted averaged gain (over a large number of trades) of market orders over a long horizon $\ell \gg 1$ is therefore:[20]

$$E[\mathcal{G}_M] \approx \lambda \frac{E[v\mathcal{R}_1(v)]}{E[v]} - \frac{E[vS]}{2E[v]} \tag{2.68}$$

In this expression we have introduced the volume-dependent lagged impact function

$$\mathcal{R}_\ell(v) = E[\epsilon_n(m_{n+\ell} - m_n)|v_n = v] \tag{2.69}$$

and we have used the previous definition of the amplification factor λ: $\mathcal{R}_{\ell \gg 1} = \lambda \mathcal{R}_1$. In the plane $x = E[v\mathcal{R}_1(v)]/E[v]$, $y = E[vS]/E[v]$ (which will repeatedly be used later), the condition $E[\mathcal{G}_M] = 0$ defines a straight line of slope 2λ separating an upper region where market orders are on average costly from a region where single market orders are favored: see Line a in Figure 2.13. For large spreads, the positive average cost of the market orders would deter their use; limit orders would then pile up and reduce the spread.

Below Line a of slope 2λ, market orders have a negative cost, and one might be able to devise profitable strategies based solely on market orders. The idea would be to try to benefit from the impact term $\mathcal{R}_\infty$ in the previous balance equation. The growth of $\mathcal{R}_\ell$ ultimately comes from the correlation between trades—that is, the succession of buy (sell) trades that typically follow a given buy (sell) market order. The simplest "copycat" strategy one can rigorously test on empirical data is to imagine placing a market order with vanishing volume fraction (so as not to affect the subsequent history of quotes and trades), immediately following another market order. This strategy suffers on average from the impact of the initial trade, used as a guide to guess the direction of the market. Therefore, the profit $\mathcal{G}_{CC}$ of such a copycat strategy, marked to market after a long time and neglecting further unwinding costs, is reduced to:

$$\mathcal{G}_{CC} = [\lambda - 1]\frac{E[v\mathcal{R}_1(v)]}{E[v]} - \frac{E[vS]}{2E[v]} \tag{2.70}$$

By requiring that this gain is nonpositive, one obtains a lower line in the plane x, y, of slope $2(\lambda - 1)$. Only below Line d can the preceding infinitesimal copycat strategy be profitable. We therefore expect markets to operate above this line and below Line a of slope 2λ.

Note also that the longtime impact of an isolated market order, uncorrelated with the order flow, is given by $G_0(\ell \gg 1)$, which is small (see Section 2.6.2). These isolated market orders thus also have a positive cost equal to half the spread. The only way to benefit from the average impact $\mathcal{R}_\ell$ is to free-ride on a wave of orders launched by others, as in the copycat strategy. Let us now take the complementary point of view of limit orders and determine the region of profitable market-making strategies.

[20] Note that this definition neglects the fact that one single large market order may trigger transactions at several different prices, up the order book ladder, and pay more than the nominal spread. Nevertheless, this situation is empirically quite rare in the markets we are concerned with and corresponds to only a few percent of all cases; see Farmer et al. (2004).

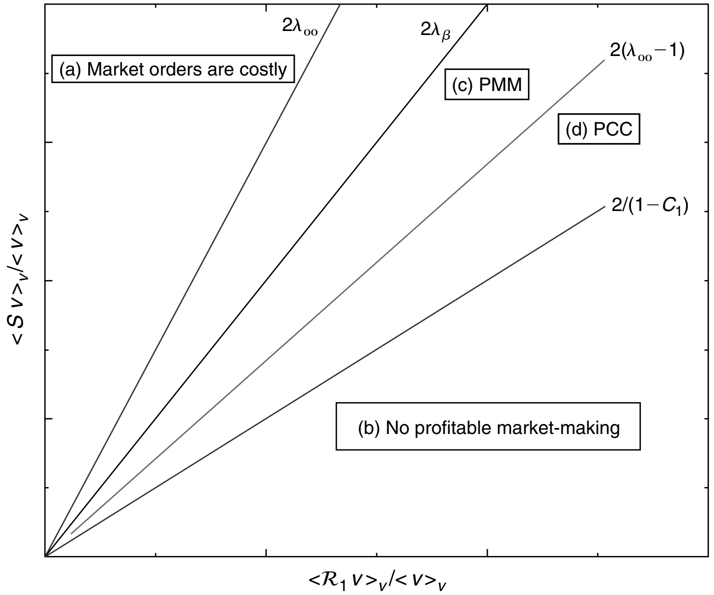


FIGURE 2.13 General "phase diagram" in the plane $x = E[v\mathcal{R}_1(v)]/E[v]$, $y = E[vS]/E[v]$, showing several regions: **(a)** above the line of slope 2λ, market orders are costly (on average) and market-making is profitable; **(b)** below the line of slope $\approx 2/(1 - C_1)$, limit orders are costly and no market-making strategy is profitable; **(c)** above the thick line of slope $2\overline{\lambda}_\beta$, market-making on time scale β^{-1} (or faster) is profitable (PMM); **(d)** below the fine line of slope $2(\lambda - 1)$, copycat strategies can be profitable (PCC). Since neither market orders nor liquidity providing should be systematically penalized to ensure steady trading, we expect that markets should operate in the "neutral wedge" in between the (b) and (a) lines. Competition between liquidity providers should push the market toward the (b). Since copycat strategies should not be profitable either, the PCC (d) line cannot lie above (b). Note that the (b), (a), and (c) all coincide within the MRR model.

An Infinitesimal Market-Making Strategy

We now compute the gain of a simple market-making strategy, which amounts to participating in a vanishing fraction of all trades through limit orders. The simplest strategy is to consider a market maker with a certain time horizon who provides an infinitesimal fraction φ of the total available liquidity. As illustrated by Eq. 2.68, the cost incurred by the market maker comes from market impact: The price move is anticorrelated with the accumulated position. When the crowd buys, the price goes up while the market-making strategy accumulates a short position, which would be costly to buy back later, and vice versa.

We consider a *steady-state* market-making strategy that avoids explicit unwinding costs. The strategy is such that tendered volume dynamically depends on the accumulated position, which ensures that the inventory is always bounded. We choose the tendered fraction φ to be given by $\varphi_i = \varphi_0(1 + \alpha V_i \epsilon)$, where V_i is the (signed) position accumulated up to time i^-, and $\epsilon = +1$ for orders placed at the ask and $\epsilon = -1$ for orders placed at the bid. This mean-reverting strategy ensures that the typical position is always bounded. One can now use this strategy for an arbitrary long time T; its profit and loss is simply given by

$$\mathcal{G}_L = \sum_{i=0}^{T-1} \varphi_i \epsilon_i v_i \left(m_i + \epsilon_i \frac{S_i}{2} \right) \tag{2.71}$$

For large T one can replace this expression by its average:

$$\mathcal{G}_L = TE \left[\varphi_i \epsilon_i v_i \left(m_i + \epsilon_i \frac{S_i}{2} \right) \right] \tag{2.72}$$

with $O(T^0)$ corrections due to the residual position at T. This quantity has been computed in Wyart et al. (2008) and depends on the value of $\beta = 1 - \alpha\varphi_0 E[v]$ that fixes the typical time scale $\ell^* = (1-\beta)^{-1}$ of the market-making strategy. When $\beta \to 0$ (fast market-making), the gain per unit time and unit volume reduces to

$$\frac{\mathcal{G}_L(\beta \to 0)}{T\varphi_0 E[v]} \approx \frac{E[vS]}{2E[v]} [1 - C_1] - \frac{E[v\mathcal{R}_1(v)]}{E[v]} \tag{2.73}$$

whereas $\beta \to 1$, corresponding to slow market making, yields:

$$\frac{\mathcal{G}_L(\beta \to 1)}{T\varphi_0 E[v]} = \frac{E[vS]}{2E[v]} - \frac{E[v\mathcal{R}_1(v)]}{E[v]} \tag{2.74}$$

The competition between impact and spread is more favorable to limit orders when the strategy is fast ($\beta = 0$) than when it is slow ($\beta = 1$). Imposing that there is a certain frequency β such that the gain of market-making strategies is zero leads to a linear relation between spread and impact, generalizing the prerious MRR relation Eq. 2.67:

$$\frac{E[vS]}{E[v]} = 2\lambda_\beta \frac{E[v\mathcal{R}_1(v)]}{E[v]} \tag{2.75}$$

Using the empirical shape of $\mathcal{R}_\ell$ and C_ℓ, the slope $2\lambda_\beta$ is found to increase between $\approx 2/(1-C_1)$ and 2λ when β increases from zero to one. When $\beta \to 1$, $\lambda_\beta \to \lambda$ and the lower limit of profitability of very slow market making is precisely Line a of Figure 2.13, where market orders become profitable. Faster strategies correspond to smaller values of λ_β, closer to $1/(1-C_1)$, leading to an extended region of profitability for market making.

From the assumption that the preceding market-making strategy for any value of β should be at best marginally profitable (since one might find more sophisticated strategies that take full advantage of the correlations between signs and volumes), we finally obtain the following bound between spread and impact:

$$\frac{E[vS]}{E[v]} \leq \frac{2}{1-C_1}\frac{E[v\mathcal{R}_1(v)]}{E[v]} \tag{2.76}$$

defining Line b of slope $2/(1-C_1)$ in the x, y plane of Figure 2.13. Consistent with the MRR model, when $\lambda = 1/(1-C_1)$, the Lines a and b of Figure 2.13 exactly coincide. Using that fact that $\mathcal{R}_1^{n+} \leq \mathcal{R}_1^{(n-1)+}$, a simple generalization of the argument presented in Appendix 2.3 allows one to show directly that the cost of limit orders is indeed negative above Line b.

Eqs. 2.68 and 2.76 and the resulting microstructural "phase diagram" of Figure 2.13 are the central results of this section. The preceding analysis delineates, in the impact-spread plane, a central wedge bounded from above by a slope 2λ and from below by a slope $\approx 2/(1-C_1)$, within which both market orders and limit orders are viable. In the upper wedge, market orders would always be costly and would be substituted by limit orders. In the lower wedge, market-making strategies, even at high frequencies, would never eke out any profit. Such a market would not be sustainable in the absence of any incentive to provide liquidity. But if the spread happened to fall in this region, the enhanced flow of market orders would soon reopen the gap between bid and ask. In the MRR model, this wedge reduces to a single line.

Comparison with Empirical Data

In conclusion of the preceding theoretical section, one expects electronic markets to operate in the vicinity of Line b of Figure 2.13, that is, there should be a linear relation between spread and market impact with a slope close to $2/(1-C_1)$. This prediction has been tested on empirical data in Wyart et al. (2008), where different markets were considered. The prediction can be tested in two different ways: for a given stock across time and across all different stocks. In both cases, a rather convincing agreement with the theory is obtained. We show, for example, in Figure 2.14 the cross-sectional test of Eq. 2.75 over 68 different stocks of the PSE in 2002. The relative values of the spread and the average impact vary by a factor of five between the various stocks, which makes it possible to test the linear relations Eqs. 2.70 and 2.76. A linear fit with zero intercept gives a slope of 2.86,[21] while the average of $2/(1-C_1)$ over all stocks is found to be ≈ 2.64.

However, the situation appears to be different on the NYSE, where specialists are present. Plotting the data corresponding to the 155 most actively traded stocks on the NYSE in 2005 in the spread-impact plane, one now finds that the empirical results cluster around the upper Line a limit where market orders become costly; see Figure 2.15. The regression has a significantly larger slope of 3.3, larger than $2/(1-C_1) \approx 2.78$, and a positive intercept $2\phi \approx 1.3$ basis points.[22] This suggests the existence of monopoly rents on NYSE; even if there is some competition to provide liquidity with other

[21] The intercept of a two-parameter regression is in fact found to be slightly negative.

[22] This is five times smaller than the average spread, leading to $\phi/\theta \sim 0.25$, much smaller than the result $\phi/\theta \sim 1-2$ found within the MRR model in 1990 or a similar value reported in Stoll (2000).

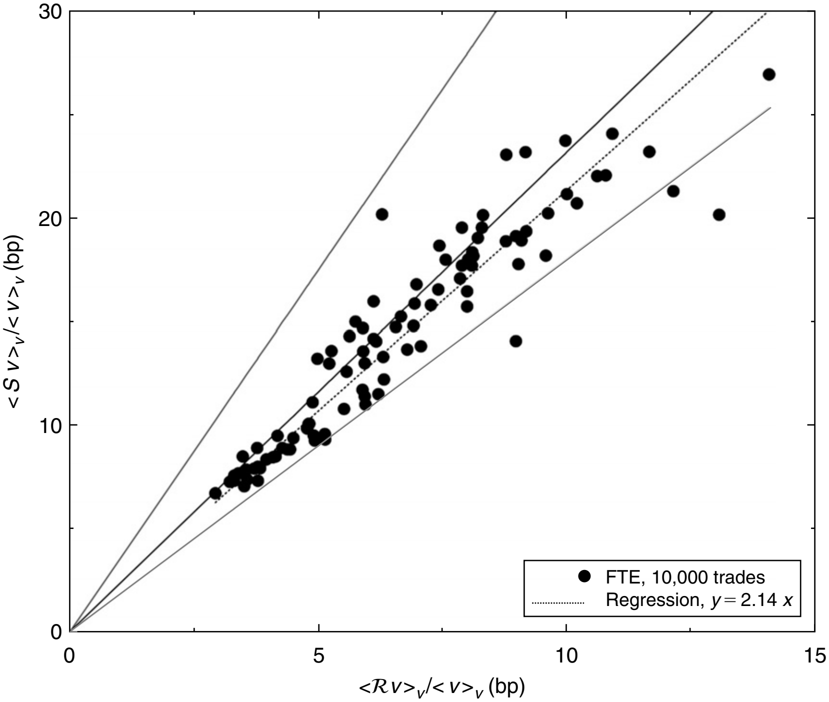


FIGURE 2.14 Plot for 68 stocks on the Paris Stock Exchange in 2002. Each point corresponds to a pair $(y = \langle vS \rangle / \langle v \rangle, x = \langle v\mathcal{R}_1 \rangle / \langle v \rangle)$, computed by averaging over the year. Both quantities are expressed in basis points. We also show the different bounds, from largest to smallest slope: Eqs. 2.68, 2.76, and 2.70, and a linear fit that gives a slope of 2.86, while $\langle 2/(1 - C_1) \rangle \approx 2.64$. The correlation is $R^2 = 0.90$.

market participants. Market makers post spreads that are systematically overestimated compared to the situation in electronic markets, with a nonzero extrapolated spread 2ϕ for zero market impact. This result is in agreement with older studies on the NYSE: Harris and Hasbrouck (1996) used data from the early 1990s to show that limit orders were more favorable than market orders; and Handa and Schwartz (1996) showed that pure limit order strategies were indeed profitable. We refer to Wyart et al. (2008) for more discussion.

The empirical analysis therefore shows that for liquid markets, an approximate symmetry between limit and market orders holds, in the sense that neither market orders nor limit orders are systematically unfavorable. Markets operate in the "neutral wedge" of Figure 2.13. In fully electronic markets, competition for providing liquidity is efficient in keeping the spread close to its lowest value. For markets with specialists, such as the NYSE, spreads appear to be significantly larger and market orders are now marginally costly on average.

Note that the preceding analysis does not require any model-specific assumptions such as the nature of order flow correlations or the fraction of informed trades. In fact,

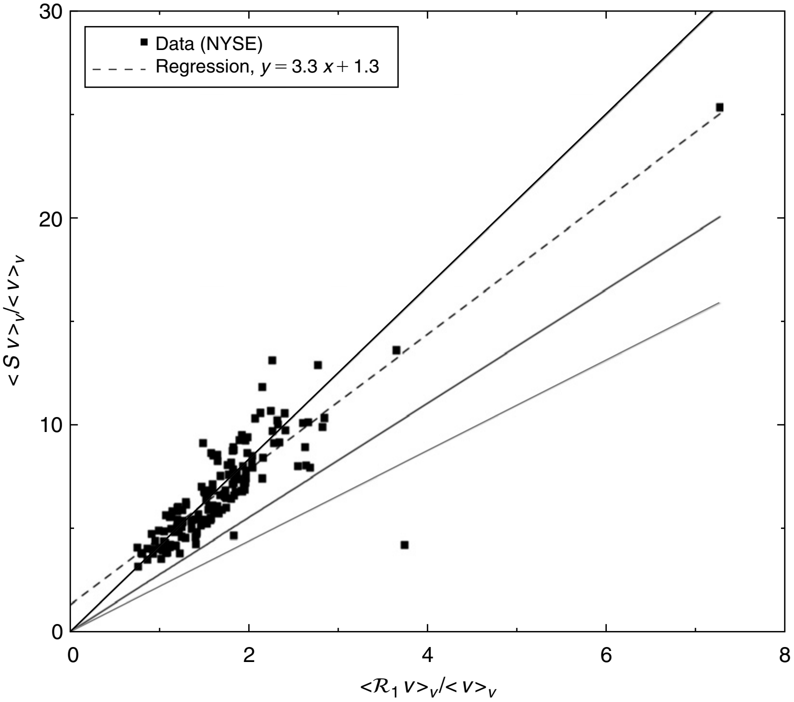


FIGURE 2.15 Plot of 155 stocks on the NYSE in 2005. Each point corresponds to a pair ($y = \langle vS\rangle/\langle v\rangle$, $x = \langle v\mathcal{R}_1\rangle/\langle v\rangle$), computed by averaging over the year. Both quantities are expressed in basis points; also shown are bounds, from largest to smallest slope: Eqs. 2.68, 2.76, and 2.70. Data clearly show that market orders are less favorable than in the electronic Paris Bourse. The regression now has a positive intercept of 1.3 bp with an $R^2 = 0.87$.

the preceding results hold even if trades are all uninformed but still mechanically impact the price.

2.7.4. Spread Dynamics After a Temporary Liquidity Crisis

The preceding analysis has shown the existence of relations between market impact and the unconditional value of the spread. The spread, however, is a variable with interesting temporal dynamics. Several studies have characterized the statistical properties of spread. Generally these studies have found that the spread distribution is fat tailed and the time correlation properties are consistent with a long-memory process (Plerou et al., 2005; Mike and Farmer, 2008; Gu, Chen, and Zhou, 2007).

It is also interesting to ask how the spread responds after a temporary liquidity crisis. As we describe in more detail in Section 2.8.1, even at the scale of individual transactions, price returns are heavy tailed; that is, it is not infrequent to observe individual transactions triggering large price changes. This often happens because a market order removes all the volume at the best, and the next-to-best occupied price level has a price very different from the price at the best (Farmer et al., 2004). As a consequence, even a small order can create a large price change, creating a very large spread. A large spread is what we mean here by a "temporary liquidity crisis."

We now describe the average dynamics followed by the spread as it converges to its "typical" value. First, a large spread is a strong incentive for limit orders inside the spread and a strong disincentive for market orders. Direct measurements of the order flow conditional on the spread value confirm this intuition (Mike and Farmer, 2008; Ponzi et al., 2008). The limit order flow inside the spread has a limit price distribution that is roughly independent of spread size and monotonically decreasing when one moves from the same best toward the opposite best. This suggests that the typical spread dynamics is not a fast reversion to its typical value, but rather it is a slow process where each liquidity provider competes with the others to close the spread. Each player tries to do this as slowly as possible to get a more favorable price from the incoming market orders, but at the same time competition prevents this process from being too slow. Empirically this slow decay has been measured in Zawadowski et al. (2006) and Ponzi et al. (2008). One way of quantifying the average dynamics is by computing the quantity, Ponzi et al. (2008),

$$G(\tau|\Delta) = E(S_{t+\tau}|S_t - S_{t-1} = \Delta) - E(S_t) \qquad \textbf{(2.77)}$$

where S_t is the spread at time t (in seconds). This quantity is the expected value of the spread at time $t + \tau$ conditional to the fact that at time zero there is a spread change of size Δ. Figure 2.16 shows this quantity for the stock AZN traded at the LSE as a

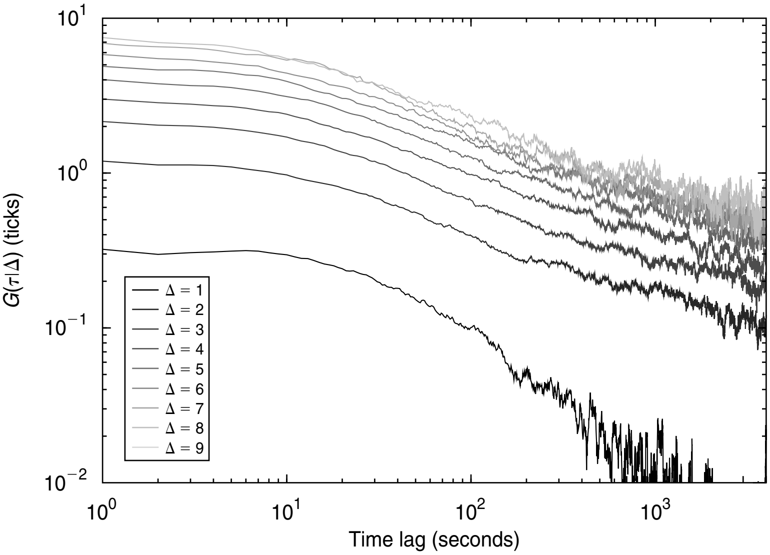


FIGURE 2.16 Conditional spread decay $G(\tau|\Delta)$ defined in Eq. 2.77 for the stock AZN, showing $G(\tau|\Delta)$ for different positive values of Δ (in ticks) corresponding to an opening of the spread at time lag $\tau = 0$. (*Source*: Adapted from Ponzi et al., 2008.)

function of τ for different positive and negative values of Δ. The decay of $G(\tau|\Delta)$ as a function of τ is very slow and for large values of τ is compatible with a power-law decay with a fitted exponent in the range 0.4–0.5. A similar slow decay of the volatility after a shock has been reported in Lillo and Mantegna (2003), Zawadowski et al. (2006), and Joulin et al. (2008).

2.8. LIQUIDITY AND VOLATILITY

One of the best-known statistical regularities of financial time series is the fact that the empirical distribution of asset price changes is heavy tailed; that is, there is a higher probability of extreme events than in a Gaussian distribution. This property has been verified by many authors on many different financial time series (e.g., Mandelbrot, 1963; Lux, 1996; Gopikrishnan et al., 1998). Extensive empirical analyses have shown that the distribution of price change over time intervals ranging from a few minutes to one or a few trading days is asymptotically distributed in a way that is approximately independent of the time interval size.

2.8.1. Liquidity and Large Price Changes

Many estimates indicate that the part of the distribution describing large price changes is a power law. For larger time intervals, the tail behavior of the return distribution becomes slowly consistent with a Gaussian tail in accordance with the central limit theorem. The heavy-tailed property of large price change is important for financial risk, since it means that large price fluctuations are much more common than one might expect under a Gaussian hypothesis.

There have been several conjectures about the origin of heavy tails in prices. Two theories that make testable hypotheses about the detailed underlying mechanism are the subordinated random process theory of Clark (1973) and the recent theory of Gabaix et al. (2003). The first model has its origins in a proposal of Mandelbrot and Taylor (1967) that was developed by Clark. Mandelbrot and Taylor proposed that prices could be modeled as a subordinated random process $Y(t) = X(\tau(t))$, where Y is the random process generating returns, X is Brownian motion, and $\tau(t)$ is a stochastic time clock whose increments are independent and identically distributed and uncorrelated with X. Clark hypothesized that the time clock $\tau(t)$ is the cumulative trading volume in time t. In simple terms, the subordination hypothesis states that price changes would be Gaussian if one measured them in equal intervals of volume (or number of trades) rather than in real time intervals.

Gabaix et al.'s proposal, in contrast, is that high-volume orders cause large price movements. They argue that the distribution of large trade size scales as $P(V > x) \sim x^{-\alpha}$, where v is the volume of the trade and $\alpha \approx 1.5$. Based on the assumption that agents maximize a first-order utility function, with a risk penalty term that is proportional to standard deviation rather than variance, they claim that the average market impact function has the form $\Delta p \propto V^{\psi}$, where $\psi \approx 0.5$. From this follows that large price changes have a power-law distribution with exponent $\alpha/\psi \approx 3$. For a critique of the empirical results and a rebuttal, see Farmer and Lillo (2004) and Plerou et al. (2004).

Both the Clark and Gabaix theories emphasize the role of trading volume as the determinant of large price changes. Even if it is clear that volume has some role in determining price changes, recent studies show that trading volume could not be the key factor. In a recent paper, Farmer et al. (2004) considered the distribution of returns generated by individual market orders. They showed that even at this microscopic time scale, price returns are heavily tailed, and more important, the size of price moves is essentially independent of the volume of the orders; see also Joulin et al. (2008). Both these facts seriously challenge the explanation of fat tails based on volume fluctuations. In that paper Farmer et al. showed that price returns associated with individual transactions are driven by liquidity fluctuations. The authors proposed and tested a mechanism for explaining how liquidity fluctuations determine large price changes. Even for the most liquid stocks in the London Stock Exchange, the limit order book often contains large gaps, corresponding to a block of adjacent price levels containing no quotes. When such a gap exists next to the best price, a new market order can remove the best quote and generate a large price change. At this time scale the distribution of large price changes merely reflects the distribution of gap sizes in the order book. The LSE data indicate that approximately 85% of the trades having a nonzero price impact have a volume equal to the volume at the best. Moreover, 97% of the trades having a nonzero price impact generate a price change equal to the first gap. In summary, the fluctuations of the gap sizes in the book are a key determinant of large price changes. The gap size is a measure of the liquidity available in the market as limit orders. Thus fluctuations of liquidity—that is, in the market's ability to absorb new market orders—are the origin of large price changes, whereas the trading volume plays a minor role.

The previously proposed mechanism raises the question of the importance of temporary liquidity crises, evidenced by large gaps in the book, for price changes over long time intervals. Although a definite answer is not available, there are three indications that short time scale and long time scale price fluctuations may be related. First, the gap size displays long-memory properties in time; see Lillo and Farmer (2005). This means that the gap size—that is, the liquidity availability—is strongly correlated in time. Periods when the typical gap size is large are likely to be followed by periods of large gaps; that is, liquidity availability is a persistent quantity. Second, it has been shown that the permanent component of the price impact is roughly proportional to the immediate impact caused by the trade (Ponzi et al., 2008). Thus the distribution of permanent price impacts, which is closely related to the distribution of price changes over relatively long time intervals, is approximately the same as the distribution of temporary price impacts, that is, of gaps in the order book. The third indication concerns the relative importance of volume and liquidity in explaining aggregate price changes, as discussed in more detail in the next section.

2.8.2. Volume vs. Liquidity Fluctuations as Proximate Causes of Volatility

The existence of a relation between volume and volatility has been known for a long time. This relation has been often interpreted as a causal relation, suggesting that volume (or number of transactions) is the driving factor determining volatility (Ane and Geman, 2000). In the previous section we discussed the subordination hypothesis, which states that returns would be Gaussian if measured in equal intervals of volume rather than in equal intervals of real time. The theory by Gabaix et al. (2003, 2006) reaches the same conclusion. Here we present some evidence challenging this view and indicating that liquidity fluctuations may be more important than volume in explaining volatility fluctuations. The question can be posed in terms of Eq. 2.1—that is, $\Delta p = \mathcal{T}(I)/\lambda$: Which is more important in determining the size of price movements, $\mathcal{T}(I)$ or λ?

In a recent paper, Gillemot et al. (2006) have presented evidence based on several different tests involving comparisons of long memory and regressions of the volatility in specific time intervals, showing that liquidity is a more important determinant of volume. Even when one aggregates returns over a fixed number of transactions (or volume), the return probability density function remains heavy tailed with properties very similar to those in fixed intervals of time. A simple way to see this effect is given in Figure 2.17, which shows the empirical probability $P(|r| > x)$ as a function of x for

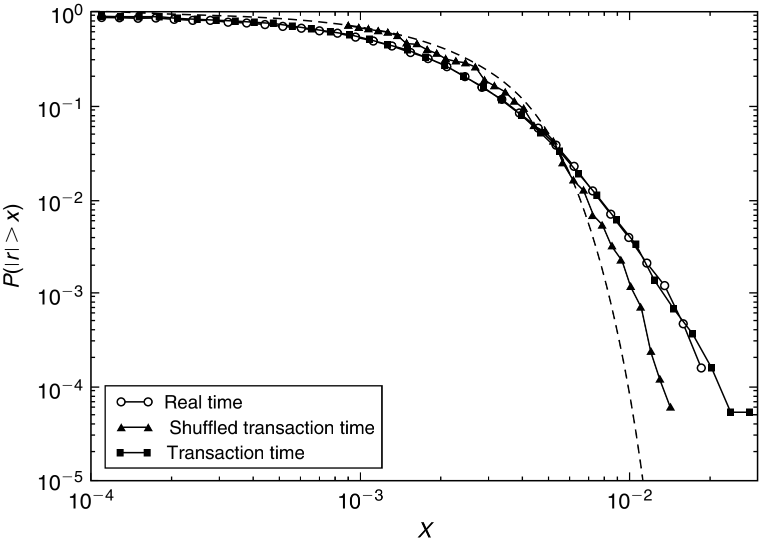


FIGURE 2.17 Cumulative distribution of absolute (log) returns $P(|r| > x)$ for the NYSE Procter & Gamble stock under different time clocks, plotted on double logarithmic scale. The circles refer to 15-minute returns, the squares refer to returns aggregated with a fixed number of transactions, and the triangles show the cumulative distribution obtained by randomly shuffling individual transaction returns and then aggregating them in a way that matches the number of transactions in each real-time interval. The dashed line corresponds to a normal distribution.

the NYSE stock Procter & Gamble. Here r is the price return over a 15-minute time interval. Suppose returns are measured in transaction time; that is, every 87 transactions rather than every 15 minutes, where 87 is chosen because it is the average number of transactions in 15 minutes (during the period from January 29, 2001, to December 31, 2003). The empirical distribution of transaction time returns matches that of real-time returns very well. Since in this case the number of transaction is held constant, this shows that the heavy tail of the return distribution is not due to variations in the number of transactions. The same effect is seen by aggregating transactions with volume rather than the number of transactions fixed (see Gillemot et al., 2006, for details).

This result shows that the fluctuation in number of trades or volume associated with a fluctuating trading activity is not the main determinant of the heavy tails of the return distribution. To highlight this effect, Figure 2.17 also shows the distribution of returns obtained from a surrogate distribution, constructed by randomly shuffling the returns of individual transactions and by aggregating them in a way that matches the number of transactions in each real-time interval. In doing so the unconditional distribution of returns of individual transactions is preserved, as are the fluctuation properties of trading frequency, but any temporal correlations of individual trade returns are destroyed. The figure shows that the tail of the surrogate distribution is less heavy than the real one, indicating that fluctuations and the time correlation properties of the reaction of prices to trades—that is, liquidity—are more important than fluctuations in trading frequency.

More supporting evidence for the importance of liquidity in determining volatility comes from a recent paper testing the microscopic random walk hypothesis against real data (Laspada et al., 2008). The price dynamics can be described as a random walk in which the increments are due to individual transactions. Under the assumption that the sign and the size of the price increments are mutually independent stochastic processes, it is possible to derive an exact expression for the volatility expected in a time interval with a given number of transactions. When one tests this expression on real data, it is found that for one-hour intervals the model consistently overpredicts the volatility of real price by about 70% and that this effect becomes stronger as the length of the time interval increases. This fact suggests that the assumption of independence of size and sign of price changes is wrong. However, data show that the contemporaneous correlation between size and sign of returns is nonstatistically significant. By performing a series of shuffling experiments, Laspada et al. (2008) show that the discrepancy between the volatility of the model and of the data is caused by a subtle but long-memory noncontemporaneous correlation between the signs and sizes of individual returns. Therefore, even after controlling for the number of transactions and the order imbalance in a given time interval, the random walk model has a strong bias in predicting the volatility, which is caused by the long memory of liquidity. This once again indicates that volume is not the key factor in explaining volatility. The neglected subtle relation between return signs and sizes shows that fluctuating liquidity is an important factor in explaining volatility.[23]

[23] In Section 2.6 we discuss how such a correlation is a consequence of the long memory of order flow and of market efficiency. The asymmetric liquidity models described in Section 2.6 predict a reduction of volatility relative to what one would expect under an unconditional permanent impact model.

Finally, the correlation between large volumes and large returns was directly studied in Joulin et al. (2008), both for trade-by-trade data and for one-minute bins, with the conclusion that such a correlation is totally absent from the data.

2.8.3. Spread vs. Volatility

It is worth investigating the relation between spread and volatility in the framework of the MRR model discussed previously. In fact, this model predicts a simple relation between volatility and impact, as shown in Eq. 2.67. Together with the relation between spread and impact we've discussed at length, this suggests a direct link between volatility per trade and spread, which we motivate and test in this section.

By definition of the volatility per trade $\sigma_1^2 = E[(m_{\varrho+1} - m_{\varrho})^2]$ and of the instantaneous impact $r_{i,i+1} \equiv (m_{i+1} - m_i) \cdot \epsilon_i$, one has as an identity:[24]

$$\sigma_1^2 \equiv E\left[r_{i,i+1}^2\right] \tag{2.78}$$

The instantaneous impact $r_{i,i+1}$ is expected to fluctuate over time for several reasons. First, the volume of the trade, the volume in the book, and the spread strongly fluctuate with time (Mike and Farmer, 2008; Wyart et al., 2008). Large impact fluctuations may also arise from quote revisions due to addition or cancellation of limit orders. Second, there might also be important news affecting the "fundamental price" of the stock. These may result in large, instantaneous jumps of the midpoint with virtually no trade at all. In order to account for both effects, one may generalize the previous MRR relation (Eq. 2.67), as in Bouchaud et al. (2004), Rosenow (2002), and Wyart et al. (2008):

$$\sigma_1^2 = A\mathcal{R}_1^2 + \Sigma^2 \tag{2.79}$$

where $\mathcal{R}_1 \equiv E[\mathcal{R}_1(v)]$ is the average impact after one trade, A is a coefficient accounting for the variance of impact fluctuations, and Σ^2 is the news component of the volatility (see Section 2.6.2). This relation holds quite precisely across different stocks of the PSE, with a correlation of $R^2 = 0.96$ (see Figure 2.18). Perhaps surprisingly, the exogenous "news volatility" contribution Σ^2 is found to be small. (The intercept of the best affine regression is even found to be slightly negative.) This could be related to the observation made in Farmer et al. (2004)—and discussed earlier—that for most price jumps, some limit orders are cancelled too slowly and get "grabbed" by fast market orders. This means that most of these events also contribute to the impact component $\mathcal{R}_1$.[25] We can neglect Σ^2 in the preceding equation; in this sense the volatility of the stocks can be mostly attributed to market activity and trade impact. This is in agreement with the conclusions of Evans and Lyons on currency markets (Evans and Lyons, 2002); see also the discussion in Bouchaud et al. (2004) and Hopman (2007).

[24] Neglecting the extremely small drift contribution.

[25] One could argue that our results simply show that the news volatility Σ itself is proportional to $\overline{\mathcal{R}}_1$ and thus to the spread S. However, there is no reason that this should *a priori* be the case. For example, a model where rare jumps of typical amplitude J and probability per trade $p \ll 1$ lead to $\Sigma = \sqrt{p}J$, whereas the cost of such jumps, contributing to S, is $pJ \ll \Sigma$.

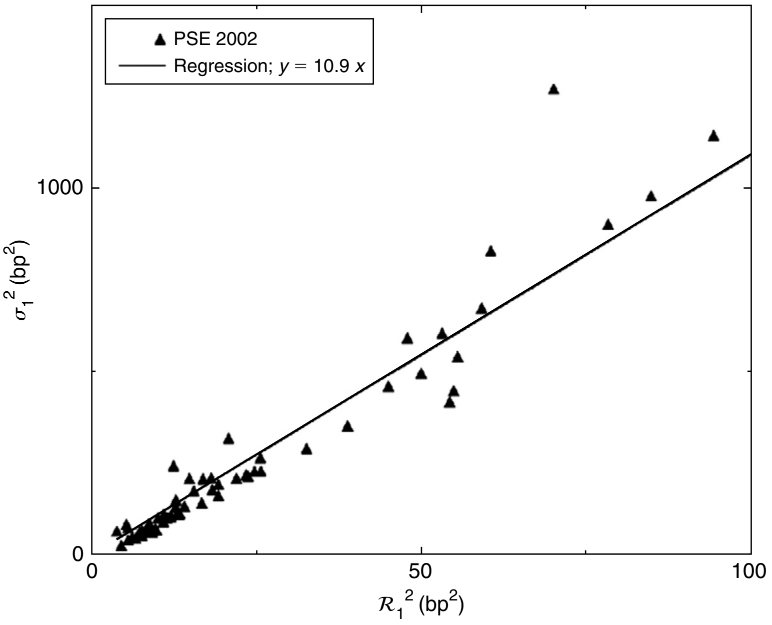


FIGURE 2.18 Plot of σ_1^2 vs. $\overline{\mathcal{R}}_1^2$, showing that the linear relation Eq. 2.79 holds quite precisely with $\Sigma^2 = 0$ and $a \approx 10.9$. (The intercept of the best affine regression is even found to be slightly negative.) Data here correspond to the 68 stocks of the PSE in 2002. The correlation is very high: $R^2 = 0.96$.

A final important assumption is that of *universality*. When the tick size is small enough and the typical number of shares traded is large enough, all stocks within the same market should behave identically up to a rescaling of the average spread and the average volume. In particular we assume that the statistics of (1) the volume of market orders (2) the spread S, and (3) the impact $\mathcal{R}_1$ and the various correlations between these quantities are independent of the stock when these quantities are normalized by their average value. Empirical evidence for (at least approximate) universality can be found in Lillo et al. (2003b) and Bouchaud et al. (2002). However, one expects that universality holds only for large-cap, small-tick stocks; large-tick stocks are not covered by the following analysis.

Universality then implies that:

$$E[vS] = BE[v]E[S], \quad E[v\mathcal{R}_1(v)] = B'E[v]\mathcal{R}_1 \qquad \textbf{(2.80)}$$

where B, B' are stock independent numbers. Equation 2.80 accounts well for the Paris Stock Exchange data studied in Wyart et al. (2008), where it was found that $B \approx 1.02$ and $B' \approx 1.80$: The incoming volume and the spread are nearly uncorrelated, whereas the volume traded and the impact are correlated ($B' > 1$), as expected.

Therefore, using Eq. 2.76 as an equality and Eqs. 2.79 and 2.80 with $\Sigma^2 = 0$, we obtain the main result of this section:

$$E[S] = C\,\sigma_1 \tag{2.81}$$

where C is a stock-independent numerical constant, which can be expressed using the constants introduced as $C = 2\lambda B'/\sqrt{AB}$. This very simple relation between volatility *per trade* and average spread was noted in Bouchaud et al. (2004), Zumbach (2004), and Wyart et al. (2008), and we present further data to support this conjecture. Therefore, the fact that the cost of limit and market orders should be nearly equal on average Eqs. 2.68 and 2.76 and the absence of a specific contribution of news to the volatility lead to a particularly simple relation between liquidity and volatility. As an important remark, we note that the preceding relation is not expected to hold for the volatility *per unit time* σ, since it involves an extra stock-dependent and time-dependent quantity, namely the trading frequency f, through:

$$\sigma = \sigma_1\sqrt{f} \tag{2.82}$$

The predicted linear relation between spread and volatility per trade was tested empirically in Wyart et al. (2008) on small-tick stocks. For example, the results for the Paris Stock Exchange are shown in Figure 2.19. One finds that Eq. 2.81 describes the data very well, with R^2 values over 0.9. One can also check that there is an average intraday pattern that is followed in close correspondence both by $E[S]$ and σ_1: Spreads

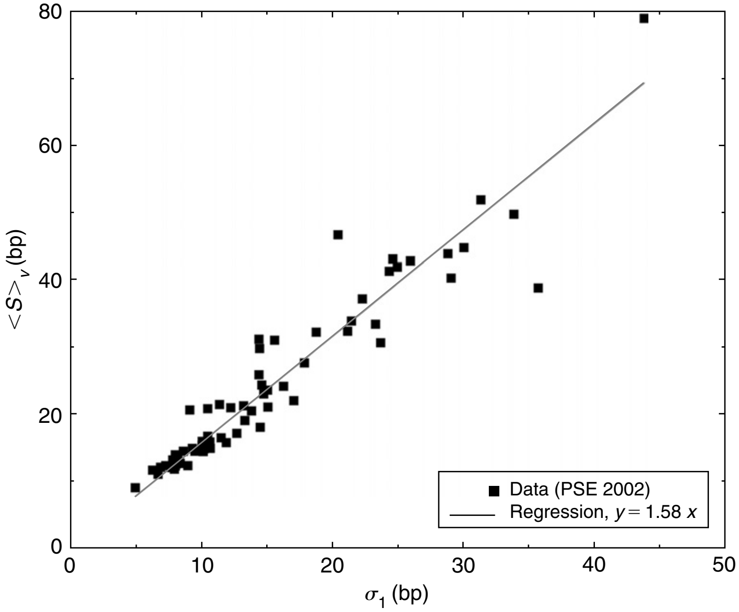


FIGURE 2.19 Test of Eq. 2.81 for 68 stocks from the Paris Stock Exchange in 2002, averaged over the entire year. The value of the linear regression slope is $c \approx 1.58$, with $R^2 = 0.96$.

are larger at the opening of the market and decline throughout the day. Note that the trading frequency f increases as time elapses, which, using Eq. 2.82, explains the familiar U-shaped pattern of the volatility per unit time.

Note that there are two complementary economic interpretations of the relation $\sigma_1 \sim S$ in small-tick markets:

- Since the typical available liquidity in the order book is quite small, market orders tend to grab a significant fraction of the volume at the best price; furthermore, the size of the "gap" above the ask or below the bid is observed to be on the same order of magnitude as the bid–ask spread itself, which therefore sets a natural scale for price variations. Hence both the impact and the volatility per trade are expected to be on the order of S, as observed.
- The relation can also be read backward as $S \sim \sigma_1$: When the volatility per trade is large, the risk of placing limit orders is large and therefore the spread widens until limit orders become favorable.

Therefore, there is a clear two-way feedback that imposes the relation $\sigma_1 \sim S$, and that can in fact lead to liquidity instabilities: Large spreads create large volatilities, which in turn may open the spread more. A detailed study of such effects would be highly valuable. On average, however, any deviation from the balance between spread and volatility tends to be corrected by the resulting relative flow of limit and market orders.

The result $\sigma_1 \sim S$ therefore appears as a fundamental property of the market organization, which should be satisfied within any theoretical description of the microstructure. This is an important constraint on models of order flow; however, none of the simple models studied in the past (zero intelligence models, Daniels et al., 2003; bounded-range models, Foucault et al., 2005, Luckock, 2003, and Rosu, 2008; or diffusion-reaction models, Slanina, 2001) is able to predict the preceding structural relation between S and σ_1 (see, however, Mike and Farmer, 2008, for recent developments using a "low intelligence" model, as discussed in the "A Simple Empirical Agent-Based Model for Liquidity Fluctuations" section).

2.8.4. Market Cap Effects

It is interesting to study the systematic dependence of the volatility and spread as a function of market capitalization M. Across stocks, the volatility per unit time shows a systematic slow decrease with M, $\sigma \propto M^{-\varphi}$, where φ is small. The trading frequency f, on the other hand, increases with M as $f \propto M^{\zeta}$. For stocks belonging to the FTSE-100, Zumbach finds $\zeta \approx 0.44$ (Zumbach, 2004), whereas for U.S. stocks the scaling for f is less clear (Eisler and Kertecz, 2006), with apparently two regimes, one for $M > 10$ B\$, where $\zeta \approx 0.44$, and the other for $M < 10$ B\$, for which $\zeta \approx 0.86$. The average amount per trade v_m, on the other hand, also increases with M in such a way that $f \times v_m$ is directly proportional to M. This last scaling holds with rather good accuracy and merely states that the total volume of transactions is proportional to market capitalization, which is somewhat expected a priori. What is interesting is that this is insured by having both the frequency of trades *and* the volume per trade increase with M, and not, for example, the transaction frequency at fixed amount per trade. The constant of proportionality is such that $\sim 10^{-3}$ of the total market cap is exchanged per day, on average, both in London and in New York (Zumbach, 2004; Eisler and Kertecz, 2006).

Combining the above two relations for the volatility per trade $\sigma_1 = \sigma/\sqrt{f}$ results in the following scaling law for the spread S,

$$S \sim \sigma_1 \propto M^{-\omega} \quad \omega = \varphi - \frac{\zeta}{2} \approx 0.22 \qquad \textbf{(2.83)}$$

The average spread therefore decreases with market capitalization. This result is in good agreement with data from the LSE, Zumbach (2004), and from the PSE, Wyart et al. (2008). It can also be directly compared with the impact data of Lillo et al. (2003b) in the NYSE, where it was established that

$$\mathcal{R}_1(v) \approx M^{-0.3} F\left(M^{0.3} \frac{v}{\overline{v}}\right) \qquad \textbf{(2.84)}$$

where $\overline{v}$ is the average volume per trade for a given stock and F a master curve that behaves approximately as a power law with exponent ψ. Since spread and impact are proportional, this last result is directly comparable to Eq. 2.83. The average over v of the preceding result then leads to $E[\mathcal{R}_1] \sim M^{-\omega}$ with $\omega \approx 0.3(1-\psi)$, which is in the range 0.15–0.25 (see Section 2.5.1 for a discussion of the value of ψ).

2.9. ORDER BOOK DYNAMICS

The previous section stresses the key role that liquidity plays in price formation. In double auction markets, prices are formed in the limit order book. Thus one obvious approach to understanding liquidity is to investigate the causes of liquidity fluctuations in the limit order book. Although the dynamics of liquidity are still very much an open question, several studies have identified statistical regularities in the behavior of limit order books and give some insight into the relationship between order flow and liquidity.

2.9.1. Heavy Tails in Order Placement and the Shape of the Order Book

There are several statistical regularities of limit orders placement. First, as mentioned, limit order signs are also well described by a long-memory process with a Hurst exponent very close to the one for market order signs. Lillo and Farmer (2004) reported a value of $H = 0.69$ for market orders and of $H = 0.71$ for limit orders.

Limit orders are characterized also by the limit price. The absolute value of the difference between the limit price and the best available price is a measure of the patience of the trader. Patient (impatient) traders submit limit orders very far from (close to) the spread. One of the statistical regularities recently observed in the microstructure of financial markets is the power-law distribution of limit order price in continuous double auction financial markets (Bouchaud et al., 2002; Zovko and Farmer, 2002). Let

$b(t) - \Delta$ denote the price of a new buy limit order and $a(t) + \Delta$ the price of a new sell limit order. Here $a(t)$ is the best ask price and $b(t)$ is the best sell price. The Δ is measured at the time when the limit order is placed. It is found that $\rho(\Delta)$ is very similar for buy and sell orders. Moreover, for large values of Δ the probability density function is well fitted by a single power law:

$$\rho(\Delta) \sim \frac{1}{\Delta^{1+\mu}} \tag{2.85}$$

There is no consensus on the value of the exponent μ. Zovko and Farmer (2002) estimated the value $\mu = 1.5$ for stocks traded at the London Stock Exchange, whereas Bouchaud et al. (2002) estimated the value $\mu = 0.6$ for stocks traded at the Paris Stock Exchange. More recently, Mike and Farmer (2008) fitted the limit order distribution for LSE stocks with a Student distribution, with 1.3 degrees corresponding to a value $\mu = 1.3$. This power law extends from 1 tick to over 100 ticks (sometimes even 1000 ticks), corresponding to a relative change of price of 5% to 50%. Such a broad distribution of limit order prices tells us that the opinion of market participants about the price of the stock in the near future could be anything from its present value to 50% above or below this value, with all intermediate possibilities. This means that market participants, quite oddly, anticipate the existence of large price jumps that would lead to trading opportunities.

A heavy tail in the distribution of relative limit price Δ indicates that there is a large heterogeneity in the limit price—that is, in the patience associated with each limit order. Patience is in turn related to the time scale the investor is willing to wait before her order is filled. The typical time to fill[26] of a limit order grows with Δ. In a recent study Lillo (2007) suggested that the origin of the heavy tails in the distribution of the relative limit price Δ can be attributed to a heterogeneity of time scales characterizing the trading behavior of individual utility maximizer investors and tested this theory using brokerage data from the LSE.

The order flow and the interaction of orders determine the instantaneous state of the book Ω_t. By averaging over time, empirical studies consistently show that the average shape of the order book is roughly symmetric between the bid and offer side of the book and is consistent across various stocks (Bouchaud et al., 2002; Zovko and Farmer, 2002; Mike and Farmer, 2008). They show that the maximum of the averaged book is not the best price, as shown in the top panel of Figure 2.20, even though this is the most likely place for an order to be placed. In Section 2.9.3 we present statistical models explaining this fact.

It is important to stress that the average shape of the book is very different from the "typical" shape of the book. As Farmer et al. (2004) showed, for most LSE stocks the typical shape of the book is extremely sparse (see the bottom panel of Figure 2.20). This occurs when the ratio between tick size and price is small so that there are often

[26] The mean time to fill of a limit order is infinite if the price process can be approximated by a random walk. "Typical" means some other measure such as the median time to fill.

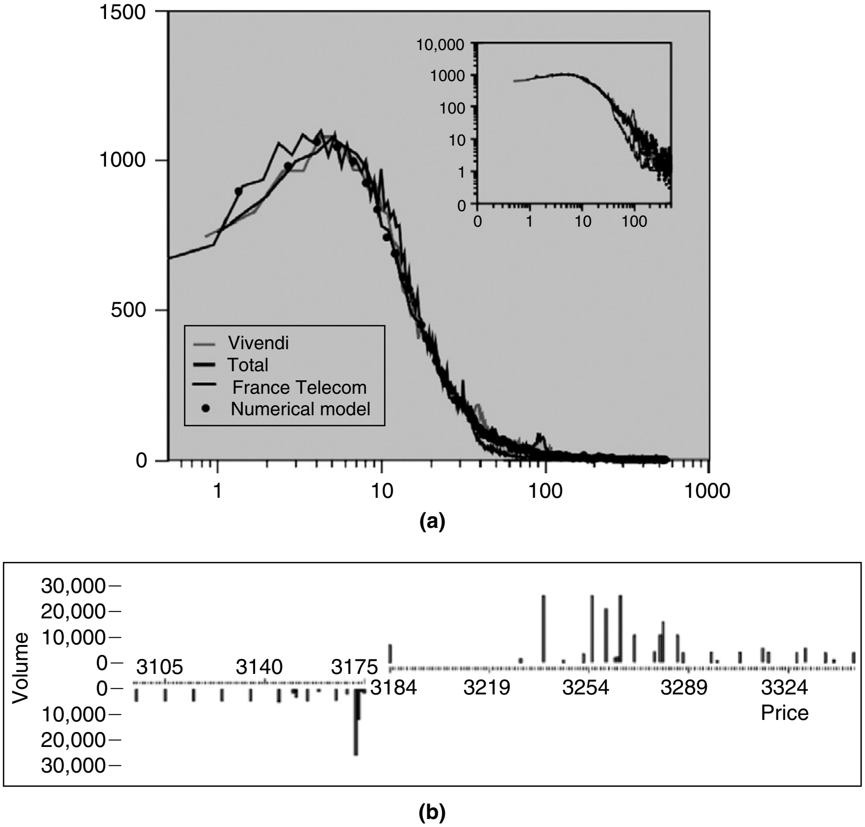


FIGURE 2.20 (**a**) Average shape of the order book. (**b**) Instantaneous shape of the order book.

many unoccupied price levels. As we discussed in Section 2.8.1, this fact has important consequences for the price impact of individual transactions and on the origin of large price fluctuations.

2.9.2. Volume at Best Prices: The Glosten-Sandas Model

The distribution of available volumes at best can be fitted by a gamma distribution with an exponent less than unity, meaning that the most probable value of the volume is much smaller than the average value. Both the value of the spread S and the quantity available at the bid and the ask, Φ_b, Φ_a, give information on the willingness of liquidity providers to enter a trade. One would like to understand the relation between these quantities—intuitively, large spreads are more favorable to liquidity providers and should attract larger volumes. More generally, it would be extremely interesting to have a theory for the shape of the whole order book, that is, the relation between the available volume and the distance from the best price.

The approach of Glosten and Sandas attempts to answer these questions, within a framework where market orders are informed trades (Glosten, 1994; Sandas, 2001). The idea is now that information is time dependent and modeled as a random variable that gives the predicted future variation of the midpoint, which we call (in conformity with the previous notation) $\epsilon_n r(n, n+\ell)$. Just before the n^{th} trade, a liquidity provider considers the volume of the queue at the ask, $\Phi_{a,n}$ and decides to add an extra (infinitesimal) limit order if its expected gain, conditional on execution, is greater than some minimum value $\mathcal{G}_{\min} \geq 0$. This reads:

$$E[m_{n+\ell} - m_n | \epsilon_n = 1, v_n \geq \Phi_{a,n}] \leq \frac{S_n}{2} - \mathcal{G}_{\min} \tag{2.86}$$

At this stage, Glosten and Sandas add several questionable assumptions. A crucial one is that the volume that the informed trader chooses to trade is proportional to the information he has: $v_n = \alpha r(n, n+\ell)$, *independently* of the shape of the book at that moment, and in particular of the available volume at the ask. He is prepared to walk up the book if necessary, which occurs with only a very small probability in practice: As discussed in Section 2.6.1, trading is, in fact, discretionary. Introducing the probability of information content $P_\ell(r)$ and dropping the index n for convenience, the previous conditional expectation inequality reads:

$$\int_{\Phi_a/\alpha}^{+\infty} r P_\ell(r)\mathrm{d}r \leq \left[\frac{S}{2} - \mathcal{G}_{\min}\right] \int_{\Phi_a/\alpha}^{+\infty} P_\ell(r)\mathrm{d}r \tag{2.87}$$

where we have used the fact that information is assumed to be reliable, that is, the expected midpoint change is indeed given by the informed trader prediction. To achieve a quantitative prediction, Sandas further assumes that $P_\ell(r)$ has an exponential shape:[27]

$$P_\ell(r) = \beta e^{-\beta r} \longrightarrow \frac{\Phi_a}{\alpha} + \frac{1}{\beta} \leq \frac{S}{2} + \mathcal{G}_{\min} \tag{2.88}$$

In fact, this calculation can be reinterpreted to give the total volume of orders available $\Phi_<$ at a price less or equal to $p = m + S/2$ and therefore makes a prediction for the shape of the order book:

$$\Phi_<(p) = \alpha(p - m) - \alpha\mathcal{G}_{\min} - \frac{\alpha}{\beta} \tag{2.89}$$

that is, a linear order book with slope α and, in principle, a *negative* intercept. (The prediction for the buy side of the book is obvious by symmetry.) Note that within this framework, the volume-dependent impact of market orders is by assumption linear: $\mathcal{R}_\ell(v) = v/\alpha$, which we already know is quite a bad representation of real data, where impact is always strongly sublinear (see Section 2.5.1). Altogether, this model fares quite badly compared with empirical data:

[27] This exponential assumption is in fact not so important. For example, a pure power-law distribution $P_\ell(r) \propto r^{-1-\mu}$ when $r > r_0$ would lead to the following result instead: $\Phi_a/\alpha \leq (1 - \mu^{-1})[S/2 + \mathcal{G}_{\min}]$ $(\mu > 1)$.

- The order book intercept, which should be negative according to the model, is found to be positive when the model is fitted to empirical data, suggesting negative costs for placing limit orders.
- The slope α, when obtained from the slope of the order book, is found to be ten times larger than when obtained from direct impact estimates.
- As mentioned, the empirical shape of order books is nonmonotonic, exhibiting a maximum away from the best price. This is not accounted for by the model.

The reason for such a failure is essentially that, as discussed in Section 2.6.1, as shown by Farmer et al. (2004), the volume of the incoming market order is in fact strongly correlated with the available volume at the best price. This is in fact why impact is sublinear in volume and is at the heart of the liquidity game we have been detailing in the previous pages. One cannot consider that the market order flow is an exogenous process to which the limit order flow must adapt; rather, the two coevolve in a strongly intertwined manner.

One can, however, directly test Eq. 2.86 on empirical data, without any further theoretical assumptions, much as we did in the previous section. We choose $\ell = 1, 10, 100$ and identify a "neutral line" in the S, Φ plane separating the region (above that line) where executed limit orders are profitable from a region where they are costly (see Figure 2.21 and Eisler et al., 2008). One sees that after the $\ell = 1$ trade the separation line is flat and is located around the value of the average spread. This means that the value of the spread is such that limit orders and market orders break even on average at high frequencies, as discussed in Section 2.7.3. However, judged on longer time scales, the profitability of a limit order behind a large preexisting order only becomes positive for spreads significantly larger than the average. In other words, correlations between spread and volume, of the type predicted by the Glosten-Sandas model (Eq. 2.86), indeed appear on longer time scales.

2.9.3. Statistical Models of Order Flow and Order Books

An alternative point of view is to model the order flow directly as a stochastic process, decomposed into three components: market orders, addition of limit orders, and cancellation of limit orders.

Zero-Intelligence Models

There is a long list of literature that develops models of this type.[28] We will describe the approach of Daniels et al. (2003) (see also Smith et al., 2003), which has the advantage that it makes predictions that can be tested against real data. They assume that each elementary event is independent and concerns a fixed "quantum" of volume v. Buy

[28] Examples of stochastic process models of limit order books include Mendelson (1982), Cohen et al. (1985), Domowitz and Wang (1994), Bak et al. (1997), Bollerslev et al. (1997), Eliezer and Kogan (1998), Maslov (2000), Slanina (2001), Challet and Stinchcombe (2001), Daniels et al. (2003), Chiarella and Iori (2002), Bouchaud et al. (2002), Smith et al. (2003), Farmer et al. (2005), and Mike and Farmer (2008). See Smith et al. (2003) for a more detailed survey.

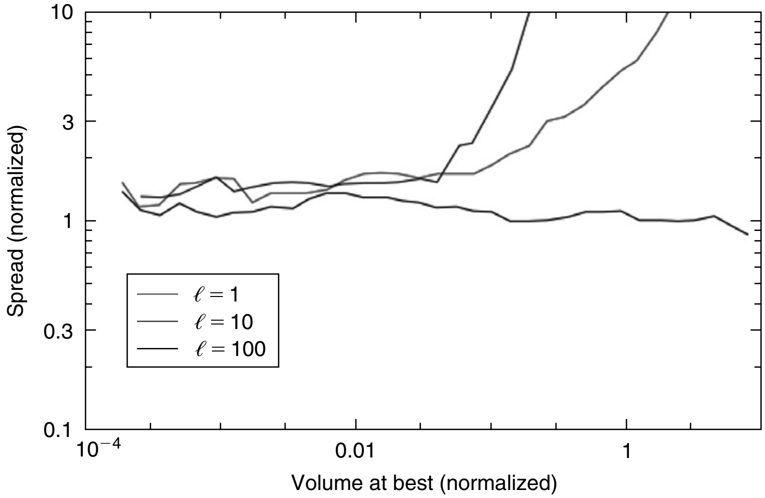


FIGURE 2.21 Neutral line for the profitability of adding a new limit order at the best price, for three different values of the horizon ℓ. The profit is positive above and negative below the curves. The indicated time horizons are given in number of transactions. The curves were gained by averaging for the ten most liquid stocks traded in the Paris Stock Exchange during 2002. Both volume and spread were normalized by their means for each stock before averaging. (*Source*: From Eisler et al., 2008).

and sell market orders are described by two Poisson processes of rate μ. Limit orders have a constant probability per unit time ρ to land anywhere they will not generate an immediate transaction, and existing limit orders have a probability ν to be canceled. This model is of course highly schematic, since it neglects all correlations between market and limit orders, in particular, the "stimulated refill" effect that we argued to be so important. Another important effect neglected is the dependence of the canceling rate on the size of the queue: One can actually observe that the probability of cancellation decreases as the number of orders at that price increases.

A simple self-consistent argument makes it possible to estimate the size of the spread S in this model. The total flux of limit orders between the midpoint and $S/2$ is by definition $\int_0^{S/2} d\Delta \rho(\Delta)$, where Δ is the distance from the midpoint and we are allowing here for the possibility that ρ might depend on Δ. If we assume that S is sufficiently small so that ρ is approximately constant, one finds that this incoming flux is $\approx \rho(0)S/2$. Whenever $\mu > \rho(0)S/2$, the rate of market order eats up the limit orders that appear within the spread completely, and the average volume present is close to zero. The cancellation term can be safely neglected if removal by market orders is more efficient—that is, when $\mu \gg \nu(0)$. But the argument breaks down when $S \sim 2\mu/\rho(0)$, which sets the typical position of the best price, provided the tick size is small compared to S. The spread is therefore larger for larger market order rates and smaller when the flow of limit orders is larger, as expected intuitively. The above "scaling" result for the spread has been derived more quantitatively when ρ and ν are independent of Δ. One finds for the average spread:

$$E[S] = \frac{\mu}{\rho} F\left(\frac{\nu}{\mu}\right) \tag{2.90}$$

where $F(u)$ is a monotonically increasing function that can be approximated as $F(u) \approx 0.28 + 1.86u^{3/4}$. Therefore, in the limit where cancellation can be neglected, one recovers the previous result $S \approx 0.28\mu/\rho(0)$. This prediction can be compared with empirical data by independently measuring the spread and the rates of the various processes (Farmer et al., 2005). In view of the simplicity of the model, the agreement with data is quite good, but systematic deviations remain. In view of the importance of feedback mechanisms that are neglected, this is hardly surprising.

These results are interesting because they demonstrate that some properties of the limit order book are dictated more or less automatically by the structure of the continuous double auction itself. In particular, Eq. 2.90 is an "equation of state" relating statistical properties of price formation to those of order flow. This equation of state is clearly inaccurate due to the extreme assumptions that must be made to derive it. However, it has some reasonable level of empirical validity, suggesting that such a relationship indeed exists for real markets. See the discussion concerning attempts to find a more realistic equation of state in the section "A Simple Empirical Agent-Based Model for Liquidity Fluctuations."

Statistical Model of Order Book

The preceding model can also explain the hump shape of the average order book. From a theoretical point of view, however, the problem is difficult to handle: If one chooses a fixed reference frame, the rates of incoming orders and cancellations change with the midpoint, whereas if one chooses the reference frame where the midpoint is fixed, limit orders that are already present get shifted around. The main difficulty comes from the fact that the motion of the midpoint is dictated by the order flow. To make progress, one can artificially decouple the motion of the midpoint and impose that it follows a random walk. An approximate quantitative theory of the volume in the book $\Phi(\Delta)$ can then be written as follows: Sell orders at distance Δ from the current midpoint at time t are those that were placed there at a time $t' < t$ and have survived until time t. These orders (1) have not been cancelled, and (2) have not been crossed by the ask at any intermediate time t'' between t' and t.

An order at distance Δ at time t in the reference frame of the midpoint $m(t)$ appeared in the order book at time t' at a distance $\Delta + m(t) - m(t')$. The average order book can thus be written, in the long time limit $t \to \infty$, as

$$\Phi_{\mathrm{st}}(\Delta) = \lim_{t\to\infty} \Phi(\Delta, t) = \int_{-\infty}^{t} \mathrm{d}t' \int du \rho\,(\Delta + u)\,\mathcal{P}\left(u|C(t,t')\right) \mathrm{e}^{-\nu(t-t')} \tag{2.91}$$

where $\mathcal{P}\,(u|C(t,t'))$ is the conditional probability that the time evolution of the price produces a given value of the midpoint difference $u = m(t) - m(t')$, given the condition that the path always satisfies $\Delta + m(t) - m(t'') \geq 0$ at all intermediate times $t'' \in [t', t]$.[29] The evaluation of $\mathcal{P}$ requires the knowledge of the statistics of the price process, which we assume to be purely diffusive. In this case, $\mathcal{P}$ can be calculated using the method of images. One finds:

$$\mathcal{P}\left(u|C(t,t')\right) = \frac{1}{\sqrt{2\pi D\tau}}\left[\exp\left(-\frac{u^2}{2D\tau}\right) - \exp\left(-\frac{(2\Delta+u)^2}{2D\tau}\right)\right] \qquad \textbf{(2.92)}$$

where $\tau = t - t'$ and D is the diffusion constant of the price process.

After a simple computation, one finally finds, up to a multiplicative constant that only affects the overall normalization,

$$\Phi_{\mathrm{st}}(\Delta) = \Phi(\Delta, t\to\infty) = \mathrm{e}^{-\alpha\Delta}\int_0^{\Delta} \mathrm{d}u\rho(u)\sinh(\alpha u) + \sinh(\alpha\Delta)\int_{\Delta}^{\infty} \mathrm{d}u\rho(u)\mathrm{e}^{-\alpha u} \qquad \textbf{(2.93)}$$

where $\alpha^{-1} = \sqrt{D/2\nu}$ measures the typical variation of price during the lifetime of an order ν^{-1}.

The preceding formula depends on the statistics of the incoming limit order flow, modeled by $\rho(u)$. When $\rho(u) = e^{-\beta u}$, all integrals can be perfomed explicitly and one finds:

$$\Phi_{\mathrm{st}}(\Delta) = \Phi_0\frac{\alpha\beta}{\alpha-\beta}\left[e^{-\beta\Delta} - e^{-\alpha\Delta}\right] \qquad \textbf{(2.94)}$$

which can easily be seen to be zero for $\Delta = 0$, reach a maximum and decay back to zero exponentially at large Δ. Here Φ_0 is the total volume in the sell side of the book.

We have seen that the limit order price distribution is characterized by a power law with exponent μ (see Eq. 2.85). When $\mu < 1$, the parameter α in the preceding formula can be rescaled away in the limit where α^{-1} is much larger than the tick size (this is relevant for small-tick stocks, where $\alpha^{-1} \sim 10$ ticks). In this case, the shape of the average order book only depends on μ. In rescaled units $\delta = \alpha\Delta$, it is given by the following convergent integral:

$$\Phi_{\mathrm{st}}(\Delta) = \mathrm{e}^{-\delta}\int_0^{\delta} \mathrm{d}u\, u^{-1-\mu}\sinh(u) + \sinh(\delta)\int_{\delta}^{\infty} \mathrm{d}u\, u^{-1-\mu}\mathrm{e}^{-u} \qquad \textbf{(2.95)}$$

For $\Delta \to 0$, the average available volume vanishes in a singular way, as $\Phi_{\mathrm{st}}(\Delta) \propto \Delta^{1-\mu}$, whereas for $\Delta \to \infty$, the average volume simply reflects the incoming flow of orders: $\Phi_{\mathrm{st}}(\Delta) \propto \Delta^{-1-\mu}$. We have shown in Figure 2.22 the average order book obtained numerically from the previous Poisson model with a power-law order flow and compared it with Eq. 2.95, for various choices of parameters and $\mu = 0.6$, as found for various stocks of the Paris Stock Exchange. After rescaling the two axes, the numerical models lead to very similar average order books, and the analytic approximation, although crude, appears rather effective. The average shape of the order book therefore

[29] We neglect here the fluctuations of the spread. The condition should in fact read $\Delta + a(t) - a(t'') = \Delta + m(t) - m(t\text{''}) + (S(t) - S(t\text{''}))/2 \geq 0$.

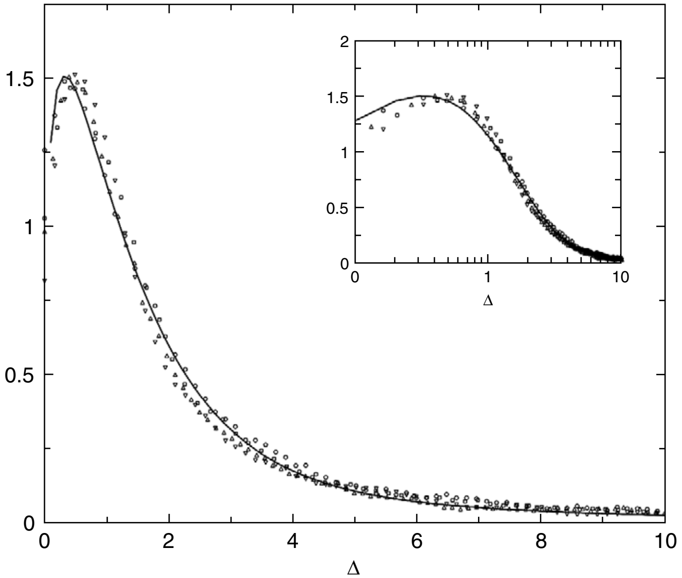


FIGURE 2.22 Average order book for a Poisson rate model with various choices of parameters (see Bouchaud et al., 2002) and $\mu = .6$. After rescaling the axes, the various results roughly fall on the same curve, which is well reproduced by the simple analytic approximation leading to Eq. 2.95, shown as the *solid line*.

reflects the competition between a power-law flow of limit orders with a finite lifetime and the price dynamics that remove the orders close to the current price.

A Simple Empirical Agent-Based Model for Liquidity Fluctuations

We now return to discuss the problem of the relationship between order flow and liquidity. The pure zero intelligence model of Daniels et al. (2003) was limited by its extreme assumptions of Poisson processes and the use of a highly stylized simplified model for order placement. A model based on more realistic assumptions was made by Mike and Farmer (2008). They made simple econometric models for order placement and cancellation and showed that by simulating this model it was possible to reproduce many of the empirical features of prices, including a quantitative match for the distribution of returns and the distribution of the spread.

In Section 2.9.1 we discussed the remarkable heavy tails in order placement. This result applied only to orders placed inside the same best.[30] Mike and Farmer also studied the distribution of order prices for orders placed inside the spread or crossing the opposite best (i.e., those generating immediate transactions). Remarkably they found, in a certain sense, that the same power-law behavior applied there as well. The frequency

[30] For example, for buy orders the same best is the best bid; the power law applies to orders placed at prices less than the best bid. The "opposite best" for buy orders is the best ask.

of order placement peaks at the same best and dies out on either side and can be reasonably well fit by a Student distribution (which has a power-law tail). Under the rule that orders that cross the opposite best price are executed, this simple rule does a reasonably good job of explaining execution frequency. One of the predictions that emerges automatically is that when the spread is small it is more likely for an order to cross the opposite best; that is, market orders become more likely. This at least partially explains the "stimulated refill" process mentioned earlier, since when the spread is large, orders chosen at random are more likely to fall inside the spread (and therefore accumulate in the limit order book), whereas when the spread is small executions are more likely. In fact, the model based on this process relied on this effect to preserve stability in the number of orders accumulating in the order book.

In this model the rate of cancellation was empirically found to depend on factors such as the number of orders in the order book, the imbalance in the order book, and the position of a given order relative to the opposite best price. Finally, it takes as an exogenous input the long memory of order signs discussed in Section 2.4. When these three elements (order placement, order sign, and cancellation) are put together, it is possible to simulate this model, generating a time series of order books with the corresponding prices. Note that it is critical that there is feedback between price formation and the order placement process. The resulting series of prices are not efficient, which is not surprising given that no effort was made to make them so and there are no agents who can take advantage of inefficiencies.

Nonetheless, for a subset of stocks with properties similar to those that were used to build the model, which were called "Type I" stocks, it does a good job of reproducing many of the properties of real prices.[31] In particular, it provides a good quantitative match with the distribution of returns and the distribution of the spread. This match includes not just the shape of the distributions but their scale, including the absolute level of volatility. That is, for Type I stocks a simulation of prices based on the measured parameters of the order flow produces forecasts of volatility that make a good match in absolute terms; in other words the predictions and measured values lie along the identity line. This provides further evidence for the existence of an equation of state relating order flow and prices. (It remains to extend this model so that it also works well for Types II and III.)

To summarize, the interesting point about this model is that it suggests that volatility is directly related to fairly simple properties of the order flow.

2.10. IMPACT AND OPTIMIZED EXECUTION STRATEGIES

The fact that trades impact prices is obviously detrimental to trading strategies: Since, again, liquidity is so small, trades must typically be divided into small chunks and spread

[31] Type I stocks are those with reasonably low volatility and small tick size. Type II stocks are those with high volatility, and Type III are stocks with large tick sizes. At this stage the model only performs well for Type I stocks.

throughout the day. But because of impact, the price paid for the last lot is on average higher than the price for the first lot. This poses a well-defined problem: What is the optimal trading profile as a function of time of day, such that the average execution price is as low as possible compared to the decision price (a quantity often called "implementation shortfall")?

Assume that a trader has a total volume V to execute; he decides to cut his order into N trades, each of size v, with $nv = V$. His trading profile $\phi(t)$ is such that the number of trades between t and $t + dt$ is $\phi(t)dt$. His own impact on the price of the stock at time $t' \geq t$ is modeled as

$$p(t') - p_0(t') = P(0) \int_0^{t'} dt \phi(t) G_0(t' - t) \ln v \tag{2.96}$$

where G_0 is the continuous time version of the single trade impact discussed in Section 2.6.4. Using all the results obtained previously, one has

$$G_0(t - t') = \frac{g_0 S}{f^\beta |t' - t|^\beta} \tag{2.97}$$

where g_0 is a number of order unity (since impact and spread are proportional) and f the number of trades per unit of time. We neglect here the possible dependence of the spread S and of f on time of day.

The total extra cost due to impact for a given profile $\phi(t)$ is therefore given by:

$$\int_0^T dt \int_0^t dt' \phi(t) G_0(t - t') \phi(t') \equiv \frac{1}{2} \int_0^T dt \int_0^T dt' \phi(t) G_0(|t - t'|) \phi(t') \tag{2.98}$$

where T is the trading period (say, one day). The previous quantity should be minimized with the constraint that the total trading volume is fixed, that is:

$$\int_0^T dt \phi(t) v = V \tag{2.99}$$

This problem can be easily solved using the method of functional derivatives with a Lagrange multiplier z to enforce the constraint. This leads to the following linear equation for the profile:

$$\int_0^T dt' G_0(|t - t'|) \phi(t') = z \tag{2.100}$$

where z is such that Eq. 2.99 is satisfied.

As a pedagogical example, let us assume that the impact decays exponentially as:

$$G_0(\tau) = G_0 \exp(-\alpha\tau) + G_\infty \tag{2.101}$$

Thanks to the constraint Eq. 2.99, the value of G_∞ can be reabsorbed in z and drops out of the computation; the permanent part of the impact is irrelevant to the optimization of execution costs (although the resulting implementation shortfall, of course, depends on G_∞). The solution of the optimization problem then reads:

$$G_0\phi^*(t) = z\delta(t) + z\delta(T-t) + \frac{z\alpha}{2} \qquad \textbf{(2.102)}$$

and the constraint is:[32]

$$\frac{1}{G_0}\left[2\frac{z}{2} + \frac{z\alpha T}{2}\right] = V \longrightarrow z = \frac{G_0 V}{1+\alpha T/2} \qquad \textbf{(2.103)}$$

so finally:

$$\phi^*(t) = \frac{V}{1+\alpha T/2}\left[\delta(t) + \delta(T-t) + \frac{\alpha}{2}\right] \qquad \textbf{(2.104)}$$

the optimal profile is composed of two peaks at the open and at the close of the day and a flat profile in between. The ratio of the volume traded within the day to the volume traded at the open and at the close is $\alpha T/2$; for a fast-decaying impact (αT large), most of the volume should be spread out evenly intraday, whereas for a slowly decaying impact, trading should mostly concentrate at the open and at the close.

More generally, it can be shown that the solution to Eq. 2.100 is symmetric around $T/2$ and U-shaped (this is also mentioned in Hasbrouck, 2007, Chapter 15). In particular, when $G_0(\tau)$ is given by Eq. 2.97, one finds that the optimal profile diverges at both $t = 0$ and $t = T$, respectively, as $t^{\beta-1}$ and $(T-t)^{\beta-1}$. An approximate solution to Eq. 2.100 in that case reads:

$$\phi^*(t) \approx V\frac{\Gamma[2\beta]}{T^{2\beta-1}\Gamma^2[\beta]}t^{\beta-1}(T-t)^{\beta-1} \qquad \textbf{(2.105)}$$

It is interesting to note that none of the parameters g_0, S, f entering in the numerical evaluation of G_0 appear in the shape of the profile, since again these can be reabsorbed in the definition of z at an early stage of the computation.

A generic U-shape solution for the optimized execution profile suggests an interesting interpretation of the observed U-shaped total traded volume as a function of the time of day.

2.11. TOWARD AN EMPIRICAL CHARACTERIZATION OF A MARKET ECOLOGY

The description of financial markets we have depicted is based on the assumption of the existence of different degrees of heterogeneity among market participants. The first level

[32] Note that the two delta functions only contribute half of their "area" to the total volume, since they are at the edge of the integration range.

of heterogeneity is due to the existence of a broad distribution of scales among market participants. Here scale refers to any quantity that measures the typical size of the trades of an investor. Moreover, the size of the hidden order determines the time horizon over which the order is worked and the number of transactions needed to complete the order.

As described in Section 2.3.8, the second degree of heterogeneity is due to the existence of (at least) two classes of agents acting systematically on opposite sides of the market. One group corresponds to liquidity providers and the other to liquidity takers. It would be extremely valuable to have a comprehensive empirical study that connects the heterogeneity of market participants with their strategy and with the properties of price dynamics. Unfortunately, it is not easy to obtain databases containing this level of information. Some data providers are starting to release datasets containing information about the financial institutions involved in the transaction and/or submission or cancellation of orders from the book.

It is important to stress that such financial institutions are not individual traders or agents but rather are usually credit entities and investment firms that are members of the stock exchange and are entitled to trade at the exchange. Very often these institutions are acting both as brokers for other clients and trading for their own account. Although an institution may act on behalf of many individuals and institutions with different strategies, recent findings show that in most cases it is possible to characterize an institution with an overall strategy, corresponding to that of the bulk of their trades. In the following two sections we present the results of two recent papers investigating the behavior of institutions in the Spanish Stock Exchange.

2.11.1. Identifying Hidden Orders

In a recent paper, Vaglica et al. (2008) used brokerage data on transactions in the Spanish Stock Exchange to identify hidden orders and to characterize their statistical properties. The identification of hidden orders is done using an algorithm designed to identify segments of the inventory time series of an institution characterized by an approximately constant and statistically significant drift term. The working hypothesis is that these segments are associated with hidden orders. A hidden order is characterized by the traded volume V, the number of transactions N, and the (real) time period T needed to complete the order.[33] It is found that the distribution of these quantities scales asymptotically for large values as

$$P(V > x) \sim \frac{1}{x^2} \quad P(N > x) \sim \frac{1}{x^{1.8}} \quad P(T > x) \sim \frac{1}{x^{1.3}} \qquad \textbf{(2.106)}$$

These relations show that the size of the hidden orders is asymptotically Pareto distributed in accordance with the hidden order model described in Section 2.4.3. It should be noted that the value of the exponent for V and for N is slightly larger than the

[33] In Vaglica et al. (2008), the investigated variables are the volume and the number of trades associated with those transactions characterizing the hidden order as a buy or sell hidden order.

value 1.5 expected by the theory described in Section 2.4.3 and a more careful testing of the theory is needed. The low value of the exponents indicates that the size of hidden orders is a very heterogeneous quantity, probably reflecting the heterogeneity of market participants. To test this hypothesis, Vaglica et al. (2008) have considered the distributional properties of hidden orders of individual brokerage codes. It is found that the distribution of hidden order size of individual brokers is consistent with a lognormal distribution, whereas the pool of the hidden orders of all brokers is not consistent with a lognormal. This indicates that investor size heterogeneity is at the origin of the power-law distribution of hidden order size.

The size variables of hidden orders are clearly related to each other. If the volume V is large, we expect that the number of transactions N and the time needed to complete the orders will also be large. The relation between the size variables reflects the strategic behavior chosen by the trader to work the order. By performing a principal component analysis of the hidden orders, Vaglica et al. find that

$$N \sim V^{1.1} \quad T \sim V^{1.9} \quad N \sim T^{0.66} \tag{2.107}$$

The fraction of variance explained by the first eigenvalue is on the order of 88%, so these characterizations are reasonably sharp.

The first relation indicates that the number of transactions of a hidden order is roughly proportional to its size. This means that even if a trader needs to trade a large hidden order, she will not split the order into larger chunks. This observation is consistent with the empirical finding that it is rare that the size of market orders is larger than the volume available at the opposite best (see Section 2.8.1 and Farmer et al., 2004). The other two relations indicate that the larger the volume of the hidden order, the slower the trading rate. This result has also been verified by using other statistical hidden order detection algorithms and still needs to be properly understood. Finally, it is worth noting that the relations of Eq. 2.107 also hold approximately true when one considers hidden orders belonging to a single brokerage code. In other words, the scaling relations of Eq. 2.107 are not the effect of heterogeneity among traders.

2.11.2. Specialization of Strategies

The presence of distinct classes of institutions and their mutual interaction has been investigated in a recent work by Lillo et al. (2008b). This study clearly identifies classes of institutions that are characterized by a similar trading behavior. Specifically, the study has focused on the cross-correlation between the inventory variation of different institutions. In general it is found that the cross-correlation of the inventory variation of different institutions is often statistically significant, for both positive and negative values. Principal component analysis reveals that the first eigenvalue of the correlation matrix is associated with a factor that is strongly correlated with price return. To give an idea of the level of correlation of the activity between different institutions, in Figure 2.23 we show the contour plot of the correlation matrix of daily inventory variation of the institutions trading the stock BBVA in the Spanish Stock Exchange in 2003.

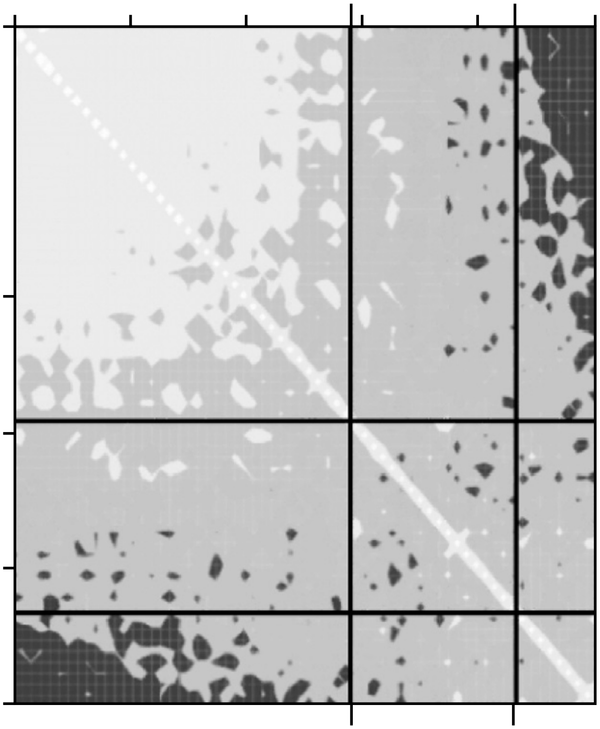


FIGURE 2.23 Contour plot of the correlation matrix of daily inventory variation of institutions trading the stock BBVA in 2003. This is plotted by sorting the firms in the rows and columns according to the strength of the correlation of their inventory variation with the return of the price of BBVA during the same period. Shades are chosen to highlight positive or negative firm daily inventory cross-correlation values above a given significance level. Specifically, *light gray* (black) indicates positive (negative) cross-correlation with a significance of 2σ, whereas *medium gray* indicates positive (negative) cross-correlation below 2σ. The *thick lines* in the matrix are obtained from the bottom panel by partitioning the firms into three groups according to the value of the correlation between returns and inventory variation (smaller than -2σ, between -2σ and 2σ, and larger than 2σ). (*Source*: Adapted from Lillo et al., 2008b).

The various shades of gray refer to different levels of correlation (see the figure caption). The institutions are sorted according to the value of the correlation of their inventory variation with the price return of BBVA. Two groups of firms are shown, one on the top left corner and the other on the bottom right corner.

The figure shows a clear block structure that makes it possible to identify communities of institutions characterized by a similar trading behavior. Specifically, the trading institutions can be partitioned into three subsets. The first (see the bottom right corner in Figure 2.23) is composed of institutions with an inventory variation positively correlated with price return—that is, these institutions buy when the price goes up and sell when the price goes down. Moreover, they are typically large institutions and have strongly autocorrelated order flow, probably because of order splitting. The second subset (see the top left corner in Figure 2.23) is composed of institutions whose inventory variation is negatively correlated with price return; these institutions buy when the price goes down and sell when the price goes up. The size of these institutions is very heterogeneous, as is the autocorrelation of their order flow. Finally, the third group is made up of uncategorized firms. As Figure 2.23 shows, the cross-correlation between the inventory variation of an institution belonging to the first group and an institution belonging to the second group is typically negative (dark areas in the top right and bottom left corners). This and other more direct evidence suggests that institutions belonging to these two groups are often trading counterparties.

2.12. CONCLUSION

In this review, we discussed market impact on two different, but overarching, levels. The first level deals with ultra high-frequency scales: that of elementary transactions (a level that in physics is called "microscopic"). It is concerned with the phenomenological description and mathematical modeling of empirical observations on order flow, impact, order book, bid–ask spread, and so on, which are of direct interest for high-frequency trading, execution, and slippage control. Results on that front are surprisingly rich and to some extent unexpected. Among the most salient points, one finds that impact of individual trades is nonlinear (concave) in volume and has a nontrivial lag dependence that can be thought of alternatively as a history-dependent impact. This is at variance with many simple models, including the famous Kyle model, where impact is assumed to be linear and permanent. The subtle temporal structure of impact can be traced back to the long memory in the fluctuations of supply and demand. The compatibility of the long memory in order flow and the absence of predictability of asset returns has profound consequences on price formation.

The second level deals with phenomena on a longer "coarse-grained" time scale and is more in line with the questions economists like to ask about markets and prices, such as: Are prices in equilibrium? What is the information content of these prices? or Why is the volatility so high? Much as in physics, where the detailed understanding of the microscopic world provides invaluable insight into macroscopic phenomena, we believe that a consistent picture of the microstructure mechanisms will help put in perspective some of these traditional questions about markets and prices. From the set of results presented previously, two concepts seem to emerge with possible far-reaching theoretical consequences:

- Because the outstanding liquidity of markets is always very small, trading is inherently an incremental process, and prices cannot be instantaneously in equilibrium and cannot instantaneouly reflect all available information. There is nearly always a substantial offset between latent offer and latent demand, which only slowly gets incorporated in prices. Only on an aggregated level does one recover an approximately linear impact with a permanent component.
- On anonymous, electronic markets, there cannot be any distinction between "informed" trades and "uninformed" trades. The average impact of all trades must

be the same, which means that impact must have a mechanical origin: If everything is otherwise held constant, the appearance of an extra buyer (seller) must on average move the price up (down).

The theory of rational expectations and efficient markets has increasingly emphasized information and belittled the role of supply and demand, in contradiction with the intuition of traders and of the layman.[34] The work we have reviewed here underlines the role of fluctuations in supply and demand, which may or may not be exogenous and may or may not be informed in a traditional sense—it does not really matter. Attempts to estimate *ex-post* the fraction of truly informed trades leads to very small numbers, at least judged on a short time basis, meaning that the concept of informed trades is not very useful to understand what is going on in markets at high frequencies. But still, prices manage to be almost perfectly unpredictable, even on very short time scales. The conclusion is that any useful notion of information must be internal to the market; trades, order flow, and cancellations *are* information, whatever the final cause of these events.

We are aware that some of these ideas go strongly against the prevailing view in market microstructure theory and entail a rather abrupt change of paradigm. We hope that this work will help clarify the issues and motivate further work to reconstruct a fully rigourous modeling framework, deeply rooted in the empirical data. Such data is now widely available and so abundant that it should be possible to raise the achievements of microstructure theory to the level of precision achieved in the physical sciences.

[34] On this point, see the lucid discussion in Lyons (2001), from which we reproduce the following excerpt: *Consider an example, one that clarifies how economist and practitioner worldviews differ. The example is the timeworn reasoning used by practitioners to account for price movements. In the case of a price increase, practitioners will assert, "there were more buyers than sellers." Like other economists, I smile when I hear this. I smile because in my mind the expression is tantamount to "price had to rise to balance demand and supply."*

APPENDIX 2.1: MECHANICAL VS. NONMECHANICAL IMPACT

As we summarized in Section 2.3.4, there are two very different views of what causes impact. The standard view is that it is essentially driven by information; the arrival of a trade signals new information, which causes market participants to update their valuations. But suppose a trade arrives that is really not based on any information. Does such a trade have a purely mechanical effect on prices? If so, what is the nature of that effect?

In Section 2.3.4 we introduced one such notion of mechanical impact, imagining a standard market clearing framework in which agents randomly alter their excess demand functions asynchronously. As each agent alters her demand function, she makes trades that generate market impact. Whether or not these are permanent depends on whether the alternations are permanent. Insofar as such alternations are permanent, the effect on prices will also be permanent.

In this appendix we examine another notion of mechanical impact for continuous double auctions. We define a mechanical impact as what happens if someone places an order in the order book if this order has no effect on any future orders. We are essentially asking the question of what happens to the price if an order is injected into the order book at random but no one pays any attention to it. We describe a method for analyzing order book data to answer this question (Farmer and Zamani, 2007). The essential result is that though there is a significant instantaneous mechanical impact due to the simple fact that such an order can consume the best quotes and move the midprice, this impact decays to zero. This decay seems to follow a power law, decaying very fast initially and very slowly later on. The reason for this decay is that orders are continually being removed from the order book, and as this happens the mechanical impact decays away. The mechanical impact as defined in this sense largely reflects the rate at which orders are flushed out of the order book.

A2.1.1. Definition of Mechanical Impact for Order Books

The following definition of mechanical impact makes the convenient simplifying assumption that the market framework is a continuous double auction. Consider the order flow $\Omega = (\omega_1, \omega_2, \ldots, \omega_t, \ldots)$ consisting of individual orders ω_t, which can be either new trading orders or cancellations of existing trading orders. Each individual order could be originated because of information relating to the value of the asset, or it could be originated "at random,"—for example due a demand for liquidity driven by events having no bearing on the asset being traded.

Under the rules of the continuous double auction, any initial limit order book and subsequent order flow generates a unique sequence of limit order books, which corresponds to a unique sequence of midprices. Auction A can be regarded as a deterministic function

$$b_{t+1} = A(b_t, \omega_{t+1})$$

that maps an order ω_t and a limit order book b_t onto a new limit order book b_{t+1}. For a given order flow $\Omega_t^{t+\tau} = \{\omega_t, \omega_{t+1}, \ldots, \omega_{t+\tau}\}$, Auction A is applied to each successive order to generate the limit order book $b_{t+\tau}$ at time $t + \tau$,

$$b_{t+\tau} = A^\tau(b_t, \Omega_t^{t+\tau})$$

The continuous double auction can thus be thought of as a deterministic dynamical system with initial condition b_0 and exogenous input Ω.

Each limit order book b_t defines a unique logarithmic midprice $p_t = p(b_t)$. The midpoint price at time $t + 1$ can be written in terms of the composition of the price operator p and the auction operator A as $p_{t+1} = p \circ A(b_t, \omega_{t+1})$. Thus, any initial limit order book b_t and subsequent order flow $\Omega_t^{t+\tau}$ will generate a series of future prices $p_{t+1}, p_{t+2}, \ldots, p_{t+\tau}$, where, for example, the last price $p_{t+\tau}$ is

$$p_{t+\tau} = p \circ A^\tau(b_t, \Omega_t^{t+\tau}) \qquad \textbf{(2.108)}$$

To give a precise meaning to mechanical impact, suppose we modify a particular order ω_t and replace it with a new order ω'_t while leaving the rest of the order flow unaltered. Since by assumption this modification does not affect the rest of the order flow, we can freely assume that it occurred for purely mechanical reasons. We can then compare the future stream of prices generated by the order flow $\Omega_t^{t+\tau} = \{\omega_t, \omega_{t+1}, \ldots, \omega_{t+\tau}\}$ to that generated by the altered order flow $\Omega'^{t+\tau}_t = \{\omega_t, \omega_{t+1}, \ldots, \omega_{t+\tau}\}$, for example, for time $t + \tau$, $p'_{t+\tau} = p \circ A^\tau(b_t, \Omega'^{t+\tau}_t)$.

This can be used to measure the mechanical impact of any existing order ω_t by comparing the prices that are generated when ω_t is present to those that would have been generated if were were absent. We thus replace $\Omega_t^{t+\tau}$ by $\Omega'^{t+\tau}_t = \{0, \omega_{t+1}, \ldots, \omega_{t+\tau}\}$, where 0 in this case represents a null order, that is, one that does not change the order book. We can then define the *mechanical impact* $\Delta p_\tau^M(t)$ of the order ω_t as

$$\Delta p_\tau^M(t) = p \circ A^\tau(b_t, \Omega_t^{t+\tau}) - p \circ A^\tau(b_t, \Omega'^{t+\tau}_{t+1}) \qquad \textbf{(2.109)}$$

The real price p contains both the informational and mechanical impact of order ω, while in the hypothetical price p' the mechanical impact is missing; that is, it contains only the informational impact. Under subtraction, only the mechanical impact remains. This isolates the part of the price impact that is "purely mechanical," in the sense that it is generated solely by the effect of placing an order in the book and observing its effect under the deterministic operation of the continuous double auction. The *information impact* can be defined as the portion of total impact that cannot be explained by mechanical impact, that is,

$$\Delta p_\tau^I = \Delta p_\tau^T - \Delta p_\tau^M$$

where Δp_τ^T is the total impact. Whatever components of the total impact not explained by mechanical impact must be due to correlations between the order ω_t and other events. With the data we have it is impossible to say whether the placement of the order ω_t causes changes in future events Ω_{t+1} or whether the properties of Ω_{t+1} are correlated with those of ω_t due to a common cause. In either case, changes in price that are not caused mechanically must be due to information—either the information contained in ω_t affecting Ω_{t+1} or external information affecting both ω_t and Ω_{t+1}. These ideas can be extended to apply to arbitrary modifications of the order stream—for example, infinitesimal modification of order ω_t, and to define mechanical generalization of elasticity (Zamani and Farmer, 2008).

A2.1.2. Empirical Results

This definition has been applied to several stocks in the London Stock Exchange and studied as a function of the order sequence number t (which as previously simply labels the temporal sequence in which the orders are received). It is clear that the mechanical impact is highly variable. In some cases there is an initial burst of mechanical impact, which dies to zero and then remains there. In some cases there are long gaps in which the impact remains at zero and then takes on nonzero values after more than 1000 transactions. In other cases there is no mechanical impact at all.

Despite the extreme variability, when an average is taken over time, a consistent pattern emerges. By definition for $\tau = 1$ the impact is entirely mechanical, since the only order that can affect the price is the reference event ω_t. After $\tau = 1$, however, the mechanical impact decays so that on average by the time of the next transaction it is roughly half its initial value; that is, it is half the total impact. In the limit as $\tau \to \infty$, for the stocks investigated in the LSE, to a good approximation for large τ the mechanical impact decays to zero as a power law $\Delta p_\tau^M(t) \sim \tau_M^\alpha$, with exponent $\alpha_M \approx 1.6$. See the example given in Figure A2.1.1. For event time the total impact and mechanical impact are by definition the same at $\tau = 1$. This is because in moving from $\tau = 0$ to $\tau = 1$ the only event that affects the price is the reference event ω_t; alterations in Ω_{t+1} cannot effect Δp_1^T. For larger values of τ the mechanical impact decreases and the informational impact increases. As the example shows, over the time scale shown here (100 orders), when measured in units of the average spread, the mechanical impact is initially about 0.17 and then decays monotonically toward zero. In contrast, the total impact increases toward what appears to be an asymptotically constant value slightly greater than 0.3. This is the source of our statement that the initial value of mechanical impact is about half the asymptotic value of the total impact.

In thinking about this, we should stress a few points. By requiring that any associated alterations of orders be considered information, we have taken a very strict definition of mechanical impact. Within our definition of "informational" there are two fundamentally different ways in which placing or removing an order can be correlated with the placement or removal of other orders. One is that placing or removal of an order causes a change in the placement or removal of another order. The alternative, however, is that the placement or removal of the two orders are caused by the same external event and are therefore correlated. From this point of view it can appear as though one order causes the other, simply because it happens to occur a bit earlier.

Though it might be surprising that the mechanical effect of order placement is completely temporary, in fact, this has a trivial explanation: Once all the orders were

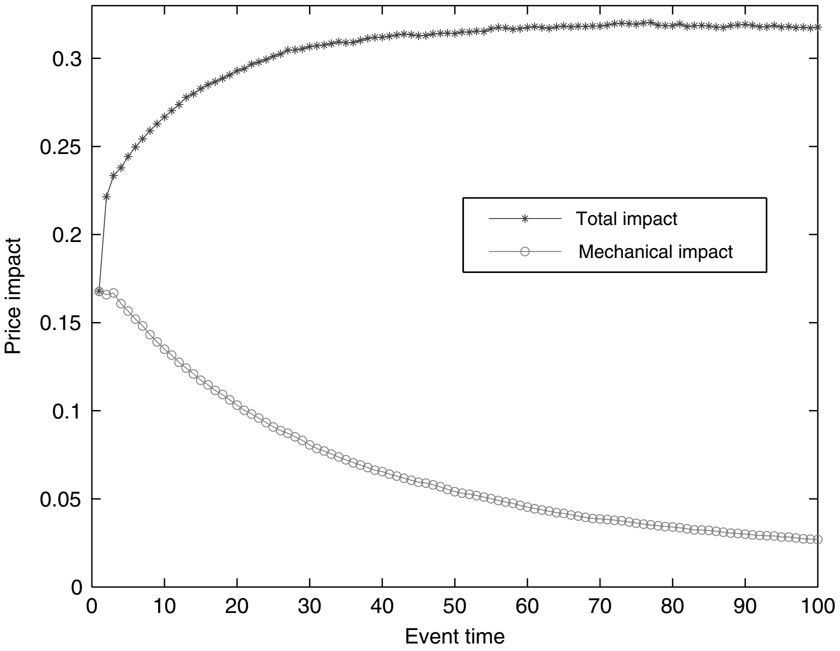


FIGURE A2.1.1 Average mechanical impact $E_t[\epsilon_t \Delta p_\tau^M(t)]$ (*squares*) and total impact $E_t[\epsilon_t \Delta p_\tau^T(t)]$ (*stars*) in units of the average spread, plotted as a function of number of the order sequence for the LSE stock AZN.

originally in the book when the reference order was placed, by definition all trace of the original order's presence is gone and so the mechanical impact is zero. Thus the power-law decay of market impact that is observed for mechanical impact is just a reflection of the rate at which orders turn over in the order book and is not related to the decay of the total impact discussed in Section 2.6.

APPENDIX 2.2: VOLUME FLUCTUATIONS

How should one take into account volume fluctuations in the formalism developed in Section 2.6.4? Since the volume of trades v is rather broadly distributed, the impact of trades could itself be highly fluctuating as well. This is not so, because large trade volumes mostly occur when a comparable volume is available at the opposite best price, in such a way that the impact of large trades is in fact quite similar to that of small trades. Mathematically, we have seen that the average impact is a slow power-law function v^ψ or even a logarithm $\log v$. As a simplifying limit, we postulate a logarithmic impact and a broad, lognormal distribution of v.

The resulting impact of the n^{th} trade $q_n = \epsilon_n \ln v_n$ is then a (zero mean) Gaussian random variable, which inherits long-range correlations from the sign process. Suppose that, as in the MRR model, only the surprise in q_n moves the price; this ensures *by* *construction* that the price returns are uncorrelated. An elegant way to write this down mathematically is to express the (correlated) Gaussian variables q_n in terms of a set of auxiliary uncorrelated Gaussian variables η_m, through:

$$q_n = \sum_{m \leq n} K(n-m)\eta_m \quad E[\eta_m \eta_{m+\ell}] = \delta_{\ell,0} \tag{2.110}$$

where $K(n)$ is a certain kernel such that the q_n have the required correlations:[35]

$$C_\ell = E[q_n q_{n+\ell}] \equiv \sum_{m \geq 0} K(m+n)K(m) \tag{2.111}$$

In the case where C decays as $c_0 \ell^{-\gamma}$ with $0 < \gamma < 1$, it is easy to show that the asymptotic decay of $K(n)$ should also be a power-law $k_0 n^{-\delta}$ with $2\delta - 1 = \gamma$ and $k_0^2 = c_0 \Gamma(\delta)/\Gamma(\gamma)\Gamma(1-\delta)$. Note that $1/2 < \delta < 1$. Inverting Eq. 2.110 leads to:

$$\eta_n = \sum_{m \leq n} Q(n-m) q_m \tag{2.112}$$

where Q is the matrix inverse of K such that $\sum_{m=0}^{\ell} K(\ell - m)Q(m) = \delta_{\ell,0}$. For a power-law kernel $K(n) \sim k_0 n^{-\delta}$, one obtains $Q(n) \sim (\delta - 1)\sin \pi\delta/(\pi k_0) n^{\delta-2} < 0$ for large n. Note that whenever $\delta < 1$, one can show that $\sum_{m=0}^{\infty} Q(m) \equiv 0$.

When all $q_m, m \leq n-1$ are known, the corresponding $\xi_m, m \leq n-1$ can be computed; the predicted value of the yet unobserved q_n is then given by:

$$E_{n-1}[q_n] = \sum_{m<n} K(n-m)\eta_m \tag{2.113}$$

and the surprise in q_n is simply:

$$q_n - E_{n-1}[q_n] = K(0)\eta_n \tag{2.114}$$

The generalization of the price equation of motion (Eq. 2.60) is therefore:

$$m_{n+1} - m_n = \xi_n + \theta K(0)\eta_n \tag{2.115}$$

which, again by construction, removes any predictability in the price returns. From this equation of motion one can derive $G_0(\ell)$ and $\mathcal{R}_\ell$.[36] From the expression of the η_n in terms of the q_n, one finds:

$$G_0(\ell) \equiv \theta K(0) \sum_{m=0}^{\ell-1} Q(m) = -\theta K(0) \sum_{m=\ell}^{\infty} Q(m) \tag{2.116}$$

[35] The following equation can be uniquely solved to extract $K(\ell)$ from C_ℓ using the so-called Levinson-Durbin algorithm for solving Toepliz systems (see, e.g., Percival, 1992).

[36] We now define G_0 as the impact of the q_n on the price.

Using the previous asymptotic estimate of the sum of matrices $Q(m)$, we finally obtain

$$G_0(\ell) \sim_{\ell \gg 1} \theta \frac{\sin(\pi\delta)K(0)}{\pi k_0} \ell^{\delta-1} \equiv \Gamma_0 \ell^{-\beta} \qquad \textbf{(2.117)}$$

Identifying the exponents leads to $\beta = 1 - \delta = 1 - \gamma/2$, recovering the above equality. The quantity θ, relating surprise in order flow to price changes, measures the so-called "information content" of the trades. It can be measured from empirical data using the preceding relation between prefactors.

Finally, from Eq. 2.115, one finds the full impact function:

$$\mathcal{R}_\ell = E[(m_{n+\ell} - m_n)q_n] = \theta K(0)^2 \quad \forall \ell \qquad \textbf{(2.118)}$$

that is, a completely *flat* impact function, independent of ℓ, as in the simplified MRR model decribed previously. However, if we assume with MRR that the fundamental price, rather than the midpoint, is impacted by the surprise in q_n, we find that the full impact function is again given by Eq. 2.61: $\mathcal{R}_\ell = \theta[1 - \mathcal{C}_\ell]$, which increases with ℓ.

APPENDIX 2.3: THE BID–ASK SPREAD IN THE MRR MODEL

A complementary point of view to that given in the main text is to analyze the cost of limit orders within the MRR model. The following argument is interesting because it can be, in essence, generalized to more complex cases as well. Suppose one wants to trade at a random instant in time. Compared to the initial midpoint value, the average execution cost of an infinitesimal buy limit order is given by:

$$\mathcal{C}_L = \frac{1}{2}\left(-\frac{S}{2}\right) + \frac{1}{2}\left(\mathcal{R}_1 + \mathcal{C}_L^+\right) \qquad \textbf{(2.119)}$$

With probability $1/2$, the order is executed right away, $S/2$ below the midpoint; otherwise, the midpoint moves on average by a quantity $\mathcal{R}_1$, to which must be added the cost of a limit order conditioned to the last trade being a buy, $\mathcal{C}_L^+$, for which a similar equation can be obtained:

$$\mathcal{C}_L^+ = \frac{1-\rho}{2}\left(-\frac{S}{2}\right) + \frac{1+\rho}{2}\left(\mathcal{R}_1^+ + \mathcal{C}_L^{++}\right) \qquad \textbf{(2.120)}$$

with obvious notations. Since the MRR model is Markovian, one has $\mathcal{R}_1^+ = \mathcal{R}_1$ and $\mathcal{C}_L^{++} = \mathcal{C}_L^+$, so that:

$$\mathcal{C}_L^+ = -\frac{S}{2} + \frac{1+\rho}{1-\rho}\mathcal{R}_1 \qquad \textbf{(2.121)}$$

Plugging this last relation in Eq. 2.119, we finally find:

$$\mathcal{C}_L = -\frac{S}{2} + \frac{1}{1-\rho}\mathcal{R}_1 \qquad \textbf{(2.122)}$$

Imposing that $C_L \equiv 0$, one recovers the MRR relation between the spread and the asymptotic impact (see Eq. 2.67).

References

Almgren, R., C. Thum, H. L. Hauptmann, and H. Li, Equity market impact, *Risk*, July 2005.

Ane, T., and H. Geman, Order flow, transaction clock, and normality of asset returns, *Journal of Finance*, 2000, 55, 2259–2284.

Bak, P., M. Paczuski, and M. Shubik, Price variations in a stock market with many agents. *Physica A*, 1997, 246, 430–453.

Barber, B. M., Y. -T. Lee, Y. -J. Liu, and T. Odean, Do individual day traders make money? Evidence from Taiwan, technical report, U.C. Davis, 2004.

Barclay, M. J., and J. B. Warner, Stealth trading and volatility, *Journal of Financial Economics*, 1993, 34, 281–305.

Beran, J., *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.

Bessembinder, H., Issues in assessing trade execution costs, *Journal of Financial Markets*, 2003, 6, 233–257.

Biais, B., T. Foucault, and P. Hillion, *Microstructure des marches financiers*, Presses Universitaires de France, 1997.

Black, F., Towards a fully automated exchange, *Review of Financial Analysts*, 1971, 27, 29–36.

Black, F., Noise, *Journal of Finance*, 1986, 41, 529–543.

Bollerslev, T., I. Domowitz, and J. Wang, Order flow and the bid–ask spread: An empirical probability model of screen-based trading, *Journal of Economic Dynamics and Control*, 1997, 21, 1471–1491.

Bouchaud, J. -P., and R. Cont, A Langevin approach to stock market fluctuations and crashes, *European Physics Journal B*, 1998, 6, 543–550.

Bouchaud, J. -P., Y. Gefen, M. Potters, and M. Wyart, Fluctuations and response in financial markets: The subtle nature of "random" price changes, *Quantitative Finance*, 2004, 4, 176–190.

Bouchaud, J. -P., J. Kockelkoren, and M. Potters, Random walks, liquidity molasses and critical response in financial markets, *Quantitative Finance*, 2006, 6, 115–123.

Bouchaud, J. -P., M. Mezard, and M. Potters, Statistical properties of the stock order books: Empirical results and models, *Quantitative Finance*, 2002, 2, 251–256.

Campbell, J. Y., and R. J. Shiller, The dividend-price ratio and expectations of future dividends and discount factors, *Review of Financial Studies*, 1989, 1, 195–228.

Casdagli, M., S. Eubank, J. D. Farmer, and J. Gibson, State space reconstruction in the presence of noise, *Physica D*, 1991, 51, 52–98.

Challet, D., So you are making money in financial markets. Should you tell your friends how?, technical report, 2007.

Challet, D., and R. Stinchcombe, Analyzing and modeling 1+1d markets, *Physica A*, 2001, 300, 285–299.

Chan, L. K., and J. Lakonishok, Institutional trades and intraday stock price behavior, *Journal of Financial Economics*, 1993, 33, 173–199.

Chan, L. K., and J. Lakonishok, The behavior of stock prices around institutional trades, *Journal of Finance*, 1995, 50, 1147–1174.

Chiarella, C., and G. Iori, A simulation analysis of the microstructure of double auction markets, *Quantitative Finance*, 2002, 2, 346–353.

Chordia, T., and A. Subrahmanyam, Order imbalance and individual stock returns: Theory and evidence, *Journal of Financial Markets*, 2004, 72, 485–518.

Clark, P. K., Subordinated stochastic process model with finite variance for speculative prices, *Econometrica*, 1973, 41, 135–155.

Cohen, K. J., R. M. Conroy, and S. F. Maier, Order flow and the quality of the market, in Y. Amihud, T. Ho, and R. A. Schwartz (eds.), *Market Making and the Changing Structure of the Securities Industry*, pp. 93–110, Rowman & Littlefield, Lanham, 1985.

Curty, P., and M. Marsili, Phase coexistence in a forecasting game, *Journal of Statistical Mechanics*, P03013, 2006.

Cutler, D. M., J. M. Poterba, and L. H. Summers, What moves stock prices? *The Journal of Portfolio Management*, 1989, 15, 4–12.

Daniels, M. G., J. D. Farmer, L. Gillemot, G. Iori, and D. E. Smith, Quantitative model of price diffusion and market friction based on trading as a mechanistic random process, *Physical Review Letters*, 2003, 90, 108102–108104.

DeLong, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann, Positive feedback and destabilizing rational speculation, *Journal of Finance*, 1990, 45, 379–395.

Ding, Z., C. W. J. Granger, and R. F. Engle, A long memory property of stock returns and a new model, *Journal of Empirical Finance*, 1993, 1, 83–106.

Domowitz, I., and J. Wang, Auctions as algorithms: computerized trade execution and price discovery, *Journal of Economic Dynamics and Control*, 1994, 18, 29–60.

Eisler, Z., J. -P. Bouchaud, and J. Kockelkoren, technical report, in preparation, 2008.

Eisler, Z., and J. Kertecz, Size matters, some stylized facts of the market revisited, *European Journal of Physics*, 2006, B51, 145–154.

Eliezer, D., and I. I. Kogan, Scaling laws for the market microstructure of the interdealer broker markets, technical report, www.arxiv.org/abs/cond-mat/9808240, 1998.

Embrechts, P., C. Kluppelberg, and T. Mikosch, *Modelling Extremal Events* Springer Verlag, 1997.

Engle, R., and J. Rangel, The spline GARCH model for unconditional volatility and its global macroeconomic causes, technical report, NYU and UCSD, 2005.

Evans, M. D. D., and R. K. Lyons, Order flow and exchange rate dynamics, *Journal of Political Economy*, 2002, 110, 170–180.

Farmer, J., A. Gerig, F. Lillo, and S. Mike, Market efficiency and the long memory of supply and demand: Is price impact variable and permanent or fixed and temporary?, *Quantitative Finance*, 2006, 6, 107–112.

Farmer, J. D., Market force, ecology and evolution, *Industrial and Corporate Change*, 2002, 11, 895–953.

Farmer, J. D., L. Gillemot, F. Lillo, S. Mike, and A. Sen, What really causes large price changes?, *Quantitative Finance*, 2004, 4, 383–397.

Farmer, J. D., and F. Lillo, On the origin of power laws in financial markets, *Quantitative Finance*, 2004, 4, c7–c11.

Farmer, J. D., P. Patelli, and I. Zovko, The predictive power of zero intelligence in financial markets, *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102, 2254–2259.

Farmer, J. D., and N. Zamani, Mechanical vs. informational components of price impact, *European Physical Journal B*, 2007, 55, 1899–2000.

Fisher, F. M., *Disequilibrium Foundations of Equilibrium Economics*, Cambridge University Press, 1983.

Foucault, T., O. Kadan, and E. Kandel, Limit order book as a market for liquidity, *The Review of Financial Studies*, 2005, 18, 1171–1217.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley, Institutional investors and stock market volatility, *Quarterly Journal of Economics*, 2006, 121, 461–504.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley, A theory of power-law distributions in financial market fluctuations, *Nature*, 2003, 423, 267–270.

Gallagher, D., and A. Looi. Trading behavior and the performance of daily institutional trades, *Accounting and Finance*, 2006, 46, 125–147.

Gerig, A., *A Theory for Market Impact: How Order Flow Affects Stock Price*, Ph.D. thesis, University of Illinois, 2007.

Gillemot, L., J. D. Farmer, and F. Lillo, There's more to volatility than volume, *Quantitative Finance*, 2006, 6, 371–384.

Glosten, L., Is the electronic limit order book inevitable?, *Journal of Finance*, 1994, 49, 1127–1161.

Glosten, L. R., and P. R. Milgrom, Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics*, 1985, 14, 71–100.

Gopikrishnan, P., M. Meyer, L. Amaral, and H. E. Stanley, Inverse cubic law for the probability distribution of stock price variations, *European Physical Journal B.*, 1998, 3, 139–140.

Gopikrishnan, P., V. Plerou, X. Gabaix, and H. E. Stanley, Statistical properties of share volume traded in financial markets, *Physical Review E*, 2000, 62, R4493–R4496. Part A.

Granger, C. W. J., and R. Joyeux, An introduction to long-range time series models and fractional differencing, *Journal of Time Series Analysis*, 1980, 1, 15–30.

Grossman, S. J., *The Informational Role of Prices*. MIT Press, 1989.

Grossman, S. J., and J. E. Stiglitz, On the impossibility of informationally efficient markets, *American Economic Review*, 1980, 70, 393–408.

Gu, G. -F., W. Chen, and W. -X. Zhou, Quantifying bid–ask spreads in the Chinese stock market using limit-order book data: Intraday pattern, probability distribution, long memory, and multifractal nature, *European Physical Journal B*, 2007, 57, 81–87.

Guedj, O., and J.-P. Bouchaud, Expert's earnings forecasts: Bias, herding and gossamer information, *International Journal of Theoretical and Applied Finance*, 2005, 8, 933–946.

Handa, P., and R. A. Schwartz, Limit order trading, *Journal of Finance*, 1996, 51, 1835–61.

Harris, L., and J. Hasbrouck, Market vs. limit orders: The superdot evidence and order submission strategy, *Journal of Financial and Quantitative Financial Analysis*, 1996, 31, 213–231.

Hasbrouck, J., Measuring the information content of stock trades, *Journal of Finance*, 1991, 46, 179–187.

Hasbrouck, J., Trades, quotes, inventories, and information, *Journal of Financial Economics*, 1988, 22, 229–52.

Hasbrouck, J., *Empirical Market Microstructure: The Institutions, Economics and Econometrics of Securities Trading*, Oxford University Press, 2007.

Hausman, J., A. W. Lo, and A. C. MacKinlay, An ordered profit: Analysis of transaction stock prices, *Journal of Financial Economics*, 1992, 31, 319–379.

Hopman, C., Do supply and demand drive stock prices?, *Quantitative Finance*, 2007, 7, 37–53.

Hosking, J. R. M., Fractional differencing, *Biometrika*, 1981, 68, 165–176.

Joulin, A., A. Lefevre, D. Grunberg, and J. -P. Bouchaud, Stock price jumps: News and volume play a minor role, Wilmott *Magazine*, 2008, 46.

Keim, D. B., and A. Madhavan, The upstairs market for large-block transactions: Analysis and measurement of price effects, *Review of Financial Studies*, 1996, 9, 1–36.

Kempf, A., and O. Korn, Market depth and order size, *Journal of Financial Markets*, 1999, 2, 29–48.

Kyle, A. S., Continuous auctions and insider trading, *Econometrica*, 1985, 53, 1315–1335.

La Spada, G., J. D. Farmer, and F. Lillo, The nontrivial random walk of stock prices, *European Journal of Physics B*, 2008, 64, 607–614.

LeBaron, B., and R. Yamamoto, Long-memory in an order-driven market, *Physica A*, 2007, 383, 85–89.

Lillo, F., Limit order placement as an utility maximization problem and the origin of power law distribution of limit order prices, *European Journal of Physics*, 2007, 55, 453–459.

Lillo, F., and J. D. Farmer, The long memory of the efficient market, *Studies in Nonlinear Dynamics & Econometrics*, 2004, 8, Article 1.

Lillo, F., and J. D. Farmer, The key role of liquidity fluctuations in determining large price fluctuations, *Fluctuations and Noise Letters*, 2005, 5, L209–L216.

Lillo, F., J. D. Farmer, and A. Gerig, A theory for aggregate market impact, technical report, Santa Fe Institute, unpublished research, 2008a.

Lillo, F., J. D. Farmer, and R. N. Mantegna, Master curve for price impact function, *Nature*, 2003b, 421, 129–130.

Lillo, F., and R. N. Mantegna, Power-law relaxation in a complex system: Omori law after a financial market crash, *Physical Review E*, 2003, 016119.

Lillo, F., S. Mike, and J. D. Farmer, Theory for long memory in supply and demand, *Physical Review E*, 2005, 7106, 066122.

Lillo, F., E. Moro, G. Vaglica, and R. Mantegna, Specialization and herding behavior of trading firms in a financial market, *New Journal of Physics*, 2008b, 10, 043019.

Lo, A. W., Long-term memory in stock market prices, *Econometrica*, 1991, 59, 1279–1313.

Lobato, I. N., and C. Velasco, Long memory in stock-market trading volume, *Journal of Business & Economic Statistics*, 2000, 18, 410–427.

Luckock, H., A steady-state model of the continuous double auction, *Quantitative Finance*, 2003, 39, 385–404.

Lux, T., The stable Paretian hypothesis and the frequency of large returns: An examination of major german stocks, *Applied Financial Economics*, 1996, 6, 463–475.

Lyons, R., *The Microstructure Approach to Foreign Exchange Rates*, MIT Press, 2001.

Madhavan, A., Market microstructure: A survey, *Journal of Financial Markets*, 2000, 3, 205–258.

Madhavan, A., M. Richardson, and M. Roomans, Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies*, 1997, 10, 1035–1064.

Mandelbrot, B. B., The variation of certain speculative prices, *Journal of Business*, 1963, 36, 394–419.

Mandelbrot, B. B., and J. W. van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Review*, 1968, 10, 422–437.

Mandelbrot, B. B., and H. M. Taylor, On distribution of stock price differences, *Operations Research*, 1967, 15, 1057–1062.

Maslov, S., Simple model of a limit order-driven market, *Physica A*, 2000, 278, 571–578.

Mendelson, H., Market behavior in a clearing house, *Econometrica*, 1982, 50, 1505–1524.

Mike, S., and J. D. Farmer, An empirical behavioral model of liquidity and volatility, *Journal of Economic Dynamics and Control*, 2008, 32, 200–234.

Milgrom, P., and N. Stokey, Information trade and common knowledge, *Journal of Economic Theory*, 1982, 26, 17–27.

Odean, T., Do investors trade too much?, *American Economic Review*, 1999, 89, 1279–1298.

O'Hara, M., *Market Microstructure Theory*, Blackwell, 1995.

Packard, N. H., J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Geometry from a time series, *Physical Review Letters*, 1980, 45, 712–716.

Percival, D., Simulating gaussian random processes with specified spectra, *Computing Science and Statistics*, 1992, 24, 534.

Plerou, V., P. Gopikrishnan, X. Gabaix, and H. E. Stanley, Quantifying stock price response to demand fluctuations, *Physical Review E*, 2002, 66, article no. 027104.

Plerou, V., P. Gopikrishnan, X. Gabaix, and H. E. Stanley, On the origin of power laws in financial markets, *Quantitative Finance*, 2004, 4, 11–15.

Plerou, V., P. Gopikrishnan, and H. E. Stanley, Quantifying fluctuations in market liquidity: Analysis of the bid–ask spread, *Physical Review E*, 2005, 71, 046131–9.

Ponzi, A., F. Lillo, and R. Mantegna, Market reaction to temporary liquidity crises and the permanent market impact, preprint at arXiv:physics/0608032v1.

Roll, R., Orange juice and weather, *American Economic Review*, 1984, pp. 861–880.

Rosenow, B., Fluctuations and market friction in financial trading, *International Journal of Modern Physics C*, 2002, 13, 419–425.

Ross, S., *Neoclassical Finance*. Princeton University Press, 2004.

Rosu, I., A dynamic model of the limit order book, 2008, *Review of Financial Studies*.

Sandas, P., Adverse selection and comparative market making: Empirical evidence from a limit order market, *Review of Financial Studies*, 2001, 14, 705–734.

Sebenius, J. K., and J. Geanakoplos, Don't bet on it: Contingent agreements with asymmetric information, *Journal of the American Statistical Association*, 1983, 78, 424–426.

Shiller, R. J., Do stock prices move too much to be justified by subsequent changes in dividends?, *American Economic Review*, 1981, 71, 421–436.

Shleifer, A., *Clarendon Lectures: Inefficient Markets*, Oxford University Press, 2000.

Slanina, F., Mean-field approximation for a limit order-driven market model, *Physical Review E*, 2001, 64, Article no. 056136, Part 2.

Smith, E., J. D. Farmer, L. Gillemot, and S. Krishnamurthy, Statistical theory of the continuous double auction, *Quantitative Finance*, 2003, 3, 481–514.

Stoll, H., Friction, *Journal of Finance*, 2000, 55, 1479.

Stoll, H. R., The supply of dealer services in securities markets, *Journal of Finance*, 1978, 33, 1133–51.

Takens, F., Detecting strange attractors in turbulence, in D. Rand and L.-S. Young (eds.), *Dynamical Systems and Turbulence*, vol. 898, pp. 366–381, Springer-Verlag, 1981.

Torre, N., *BARRA Market Impact Model Handbook*, BARRA Inc., 1997.

Vaglica, G., F. Lillo, E. Moro, and R. Mantegna, Scaling laws of strategic behavior and size heterogeneity in agent dynamics, *Physical Review E*, 2008, 77, 0036110.

Weber, P., and B. Rosenow, Order book approach to price impact, *Quantitative Finance*, 2005, 5, 357–364.

Weber, P., and B. Rosenow, Large stock price changes: Volume or liquidity?, *Quantitative Finance*, 2006, 6, 7–14.

Wiesinger, J., Z. Eisler, A. Joulin, and J. -P. Bouchaud, Dynamics of the bid–ask spread in the Glosten-Milgrom model: analytical results, technical report, 2008.

Wyart, M., J. -P. Bouchaud, J. Kockelkoren, M. Potters, and M. Vettorazzo, Relation between bid–ask spread, impact and volatility in double auction markets, *Quantitative Finance*, 8, 41–57, 2008.

Zamani, N., and J. D. Farmer, Decomposition of mechanical and informational components of orderflow, technical report, 2008.

Zawadowski, A., G. Andor, and J. Kertecz, Short-term market reaction after extreme price changes of liquid stocks, *Quantitative Finance*, 2006, 4, 283–295.

Zovko, I., and J. D. Farmer, The power of patience: A behavioral regularity in limit order placement, *Quantitative Finance*, 2002, 2, 387–392.

Zovko, I., and J. D. Farmer, Correlations and clustering in the trading of members of the London Stock Exchange, in S. Abe, T. Mie, H. Herrmann, P. Quarati, A. Rapisarda, and C. Tsallis (eds.), *Complexity, Metastability and Nonextensivity; An International Conference*, AIP Conference Proceedings, Springer, 2007.

Zumbach, G., How the trading activity scales with the company sizes in the FTSE 100, *Quantitative Finance*, 2004, 4, 441.