How Prices Respond to Worked Orders[*]

Austin Gerig
University of Technology, Sydney
austin.gerig@uts.edu.au

J. Doyne Farmer
Santa Fe Institute and
LUISS Guido Carli
jdf@santafe.edu

Fabrizio Lillo
Santa Fe Institute, Università di Palermo,
and Scuola Normale Superiore di Pisa
lillo@unipa.it

Very preliminary. Please do not quote.

January 2011

Using a structural model of price formation, we analyze how prices respond to worked orders, i.e., to orders that are split into pieces and transacted over time. Our results replicate several empirical findings that have otherwise been difficult to explain. In the model, (1) prices respond more to the earlier pieces of a worked order than the later pieces so that price impact is a nonlinear and concave function of total order size, (2) prices revert, at least partially, after an order completes—this holds true even when disregarding order processing and inventory costs, and (3) under most circumstances, the speed of execution affects the price response, with faster execution causing a larger immediate impact than slower execution. Given the current prevalence of worked orders in markets (and the dramatic effects these orders have had on order flow variables), our findings should be of interest to investors as well as market designers and regulators.

Keywords: algorithmic trading; market impact; price impact; worked orders.
JEL Classification: G19.

[*]This work was supported by the U.S. National Science Foundation Grant No. HSD-0624351

How Prices Respond to Worked Orders

Abstract

Using a structural model of price formation, we analyze how prices respond to worked orders, i.e., to orders that are split into pieces and transacted over time. Our results replicate several empirical findings that have otherwise been difficult to explain. In the model, (1) prices respond more to the earlier pieces of a worked order than the later pieces so that price impact is a nonlinear and concave function of total order size, (2) prices revert, at least partially, after an order completes—this holds true even when disregarding order processing and inventory costs, and (3) under most circumstances, the speed of execution affects the price response, with faster execution causing a larger immediate impact than slower execution. Given the current prevalence of worked orders in markets (and the dramatic effects these orders have had on order flow variables), our findings should be of interest to investors as well as market designers and regulators.

Keywords: algorithmic trading; market impact; price impact; worked orders.
JEL Classification: G19.

I. INTRODUCTION

In financial markets, large orders are often "worked" when transacted, i.e., they are split into small pieces and executed over a period of time. Working an order has become standard practice in recent years—even for modest sized orders—due to the widespread adoption of electronic trading.[1] It is now cheap, fast, and relatively easy to program a computer to slice and execute an order automatically, whereas in the past, this task would have required the services of someone positioned on the exchange floor.[2] Despite their current prominence in markets, little is understood theoretically about how worked orders should influence prices. This is mainly because the original microstructure models of price formation were developed before working an order was common practice.

In this paper, we analyze how prices respond to worked orders by applying a structural model of price formation to an order that is incrementally executed. Qualitatively, our results match nicely with the following empirical findings:[3]

(1) Prices respond more to the earlier pieces of the order than to the later pieces, so that price impact is a nonlinear and concave function of the total order size.

(2) Prices revert, at least partially, after the order completes.

(3) Slowly transacting the order results in lower transaction costs.

A plot of the typical price response is shown in Fig. 1.

Unfortunately, few previous papers have focused on explaining the origin of these results, which has led to some confusion in the literature. Many papers that model or estimate price response assume, contrary to (1), that price impact must be linear.[4] In addition, it is largely assumed and accepted that

[1] The average transaction size on the NYSE has dropped by a factor of three in the past five years to approximately 300 shares per trade (Angel, Harris, and Spatt (2010)). This result suggests that most orders above 300 shares (approx. $\$10,000$) are now worked. Fig. 3 in Chordia, Roll, and Subrahmanyam (2010) reports that the dollar fraction of small transactions on the NYSE (transactions less than $\$10,000$) has exploded from 5% to over 60% in the past ten years.

[2] Automated order execution is often called *algorithmic trading*; although algorithmic trading is sometimes used to refer to any type of automated trading, including automated liquidity provision.

[3] See Torre (1997), Almgren et al. (2005), Gerig (2007), Yuen (2007), and Moro et al. (2009)

[4] The linear assumption is often made for simplicity—only a standard regression is needed to estimate impact—and is often justified by reference to Kyle (1985) and Huberman and

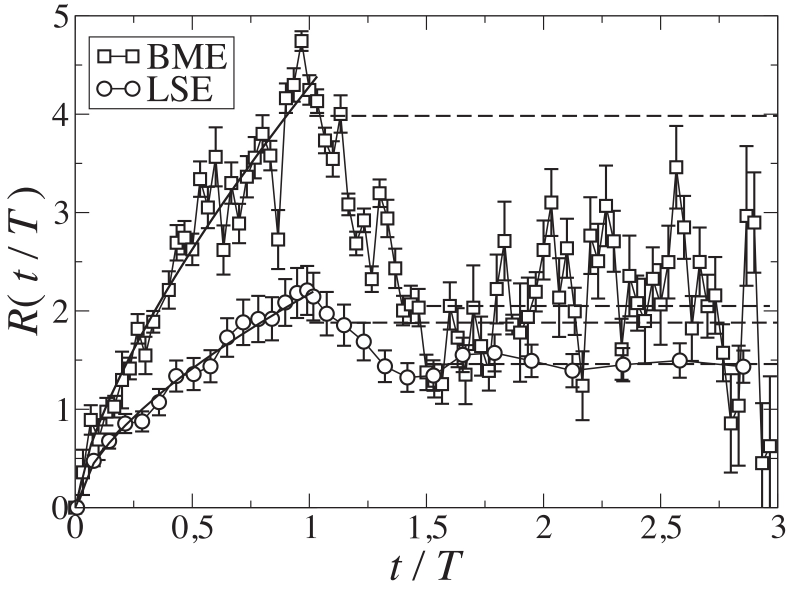

Figure 1: Plot of the price response to a worked order in the Bolsa de Madrid (BME) and the London Stock Exchange (LSE) (from Moro et al. (2009)). At time $t/T = 0$ the order starts and at time $t/T = 1$ the order completes. See Moro et al. (2009) for details.

the price reversion described in (2) results solely from compensation costs to liquidity providers or to market imperfections,[5] but as we show, it also can result when liquidity providers receive no revenue. Finally, the result in (3), although widely known amongst practitioners, has been largely ignored in the microstructure literature.

The intuition behind our results is straightforward: When a large order is split and transacted over time, it produces predictability in order flow that the market observes and uses to anticipate further transactions from the order (from hereon we refer to the unsplit order as an *order* and to the small transacted

Stanzl (2004). Unlike the Kyle (1985) model, our model is based on order size distributions that are non-Gaussian. The linear impact result of Huberman and Stanzl (2004) depends on the assumption that impact is completely permanent and that the market cannot discern split orders from complete orders. The two forms of impact we derive in this paper violate these assumptions.

[5] For a review article that includes discussion on the temporary component of impact see Stoll (2000).

pieces as *transactions*[6]). If the later transactions are more predictable than the earlier ones (we describe the conditions where this holds below), they impact the price less and the price response is a nonlinear and concave function of total order size. In addition, prices revert after an order completes due to the market's anticipation of further transactions from the order that do not materialize. Finally, if liquidity providers cannot precisely determine which transactions come from which orders, then an order that is executed quickly is indistinguishable from several orders transacting in the same direction at the same time, causing prices to react more abruptly than they would have, had the order been transacted slowly.

The model we use for worked order execution is based on a simple strategy where an active order attempts to transact a constant fraction of volume in the market.[7] Because the order transacts over time, it leaves a footprint that is observed and used to modify future transaction prices. The exact way that prices are modified depends on the information available to liquidity providers, which we bracket between two extremes in our analysis. First, we assume that liquidity providers cannot determine which transactions come from which orders, so that they use a simple autoregressive model to predict future order flow. This is similar to Hasbrouck (1988) and Madhavan, Richardson, and Roomans (1997). We call this the *autoregressive model.* Second, we assume that liquidity providers can discern each order separately, as if each transaction in the market had a 'color' that can be used to associate it with it's corresponding parent order. We call this the *colored print model.* In the colored print model, liquidity providers predict future order flow by calculating the probability that an order continues transacting conditioned on observing that the order has transacted the observed number of pieces so far. Pricing is therefore similar to price schedules that are set using 'tail expectations' in a limit order market[8], although here the analysis is in a dynamic setting. The colored print model is also similar to the 'fast execution' of a worked order in Back and Baruch (2007).

The model we use for price response is based on previous structural models that assume trades convey information to the market (Glosten and Harris (1988), Hasbrouck (1988), Madhavan, Richardson and Roomans (1997)). As is common in many of these models (see Hasbrouck (2007)), our model subtracts off the predictable component of order flow when determining price impact.

[6] Practitioners sometimes refer to these as *parent orders* and *child orders* respectively.

[7] The strategy is similar to a VWAP or "volume weighted average price" strategy, which is a standard execution strategy for worked orders.

[8] See Glosten (1994).

We therefore make an implicit assumption that prices respond symmetrically for a 'predicted' vs. an 'unpredicted' transaction. We discuss this assumption and its significance below. The assumption is relaxed in Farmer, Gerig, Lillo, and Waelbroeck (2010), where it is replaced by a 'fair pricing' condition that states the average execution price of an order should equal the final market price after the order completes.

This work was originally inspired by two empirical findings that are closely related to each other. First, that the size distribution of orders is asymptotically Pareto, and second that order flow in markets is highly autocorrelated and predictable.[9] The Pareto tail of the order size distribution was originally reported for the NYSE by Gopikrishnan et al. (2000) and has been verified for several other markets (Plerou et al. (2004)). The extreme predictability of order flow was reported independently by two groups of researchers who showed that buying and selling in markets exhibits long-term persistence (Lillo and Farmer (2004) and Bouchaud et al. (2004)). Lillo, Mike, and Farmer (2005) develop a model where these two observations are not independent, and where the long memory of order flow results from the splitting of orders that are drawn from a size distribution that is Pareto. Here, we explore how orders from a Lillo, Mike, and Farmer (2005) model are priced in an efficient market.

Finally, our work is related to the growing literature on the optimal execution of orders.[10] These papers take as given the price impact function and then optimize order placement according to the preferences of the individual transacting the order. Here, we are interested in the more fundamental question of why prices respond in the way that they do.

II. MODEL

A. *Model of Worked Order Execution*

We model the situation where an investor wishes to buy or sell a certain quantity of a security, but where the market only allows one unit at a time to be bought or sold. Orders, therefore, must be *worked*—they must be split

[9] This predictability is so long-lived that, for many securities, if you were to observe a buyer initiated transaction on a certain day, there is a greater than 50% chance that an observed transaction *several weeks later* will also be buyer initiated (see Lillo and Farmer (2004)).

[10] See Bertsimas and Lo (1998), Almgren and Chriss (2000), and Obizhaeva and Wang (2006)

into unit-sized pieces that are transacted one at a time. To reduce confusion, we refer to full-sized orders as *orders* and to the unit-sized pieces of an order as *transactions*. An order, indexed by i, is defined by three parameters, $\Psi_i = \{\varepsilon_i, \pi_i, N_i\}$. ε_i is the sign of the order—whether it is a buy ($\varepsilon_i = +1$) or a sell ($\varepsilon_i = -1$), π_i is the participation rate of the order—the probability per unit time that the order transacts, and N_i is the size, or the number of units, in the order. We assume that π_i is constant (although possibly different for different orders), which means that investors use a type of VWAP (or volume weighted average price) strategy, where an order attempts to participate in a certain constant fraction of all market transactions. These strategies are common in financial markets, and we leave more complicated strategies for future analysis. Time is denoted t and is measured in units of trades, i.e., it is updated by one whenever a trade occurs in the market. The times when each transaction from an order are submitted are denoted $t_{(1,i)}, t_{(2,i)}, \ldots, t_{(N,i)}$, and the total number of transactions that have been submitted by an order at time t is labeled $n_i(t)$.

B. Structural Model of Price Formation

Each transaction that is submitted by an investor is priced centrally at the market by a group of competitive, risk-neutral liquidity providers. Transaction prices, therefore, are set to the expected value of the security given that the transaction has occurred, and they follow a martingale (see Glosten and Milgrom (1985)). If Ω_{t-1} is the information available to liquidity providers at time $t-1$ and p_t is the post-trade expected value of the security at time t, then,

$$E[p_t | \Omega_{t-1}] = p_{t-1}. \tag{1}$$

Notice that p_t is both the post-trade expected value of the security and also the transaction price. These values are the same because we do not consider any revenue for liquidity providers. Thus, the results we derive for impact (and specifically the temporary effects that we describe later) are independent of any liquidity provider costs or profits.

We assume that liquidity providers update the estimated value of a security using two sources of information: (1) public news events and (2) order flow in the market, i.e., whether transactions are buys or sells.[11] The update in

[11] Order flow causes liquidity providers to update their estimated value because we assume there is a positive probability that transactions originate from informed individuals.

estimated value due to public news events is denoted ϵ_t; it covers all news released between times $t-1$ and t. We assume ϵ_t is an independent and identically distributed random variable with mean zero. We assume that a transaction, by itself, causes a value revision of $x_t\lambda$, where x_t is the sign of the transaction ($x_t = +1$ for a buy and $x_t = -1$ for a sell) and $\lambda > 0$ is a scale parameter that measures the degree of information asymmetry for the security. If transactions were not predictable, then we would write our structural model as, $p_t = p_{t-1} + x_t\lambda + \epsilon_t$.[12] However, transactions are highly autocorrelated and predictable in real markets due to the working of orders (this means $E[x_t|\Omega_{t-1}] \neq 0$ and our structural model would violate Eq. 1). To fix this problem, we introduce a term, γ_t, that compensates for the predictability of order flow,

$$p_t = p_{t-1} + (x_t - \gamma_t)\lambda + \epsilon_t. \tag{2}$$

Enforcing Eq. 1, we find that $E[\gamma_t|\Omega_{t-1}] = E[x_t|\Omega_{t-1}]$. In general, γ_t can depend on x_t, but we will assume that prices respond symmetrically for a 'predicted' transaction, $x_t = Sign(E[x_t|\Omega_{t-1}])$, vs. an 'unpredicted' transaction, $x_t \neq Sign(E[x_t|\Omega_{t-1}])$, so that γ_t is the same in either case.[13] This assumption forces $\gamma_t = E[x_t|\Omega_{t-1}]$. Defining $\hat{x}_t \equiv E[x_t|\Omega_{t-1}]$, our final structural model for price formation is,

$$p_t = p_{t-1} + (x_t - \hat{x}_t)\lambda + \epsilon_t. \tag{3}$$

As in Hasbrouck (1988) and Madhavan, Richardson and Roomans (1997), it is the *innovation* in order flow, $(x_t - \hat{x}_t)$, that causes liquidity providers to update their beliefs.[14]

C. Predictability of Order Flow

The specific form of $\hat{x}_t$ is dependent on the structure of the market and the prediction model used by liquidity providers. Markets can have different rules for broadcasting order flow information, so that information available to liquidity providers, Ω_t, is not necessarily standard across markets. Even with a specific

[12] This equation is similar to the structural model in Glosten and Harris (1988) when the revenue of liquidity providers is assumed zero.

[13] Farmer, Gerig, Lillo, and Waelbroeck (2010) use an alternative assumption.

[14] Our model is very similar to the structural model used in Madhavan, Richardson, and Roomans (1997), except here we assume no revenue for liquidity providers and we allow the predicted component of x_t to be influenced by the entire information set of the liquidity provider and not just x_{t-1}.

information set, the model that is used by liquidity providers to predict order flow may or may not optimally utilize this information. In this section, we first present a baseline autoregressive model for $\hat{x}_t$ that is similar to what has been proposed elsewhere (see Hasbrouck (2007)).[15] This model assumes that only the signs of transactions are used to determine $\hat{x}_t$. Second, we present a specific extreme case where $\hat{x}_t$ is determined using precise order information, specifically $n_i(t)$ and π_i for all orders in the market. In this scenario, the only information that is unknown to liquidity providers is the total size of an order and when it will complete.

Assuming that liquidity providers only use information about past trade signs, x_t, and that an autoregressive model is used to predict current order flow, then

$$\hat{x}_t = \sum_{k>0} a_k x_{t-k}. \tag{4}$$

We call this model for the predictability of order flow the *autoregressive model.*

If liquidity providers can discern each order separately (as if each transaction in the market had a 'color' that associated it with its corresponding parent order), then x_t can be estimated more precisely than with an autoregressive model. Assuming transaction prints are 'colored' so that order information $n_i(t)$ and π_i is available to liquidity providers,[16]

$$\hat{x}_t = \sum_i \varepsilon_i \pi_i \mathcal{P}(n_i(t)), \tag{5}$$

where $\mathcal{P}(n_i(t))$ is the probability that the i^{th} order is still actively transacting, given that it has so far transacted n pieces at time t. This equation sums the probability that each order will transact based on complete information about the transactions of that order. To complete the setup, we assume that the end of order i is signaled or realized with probability π_i at each time step after it has completed and that $\mathcal{P} = 0$ for that order from then on. This ensures that the time between the transactions of an order does not influence the calculation of $\mathcal{P}(n_i(t))$.[17] We call this model for the predictability of order flow the *colored print model.*

[15] See also Bouchuad et al. (2004) and Bouchaud, Kockelkoren, and Potters (2006). The impact functions in these papers can be shown to result from an autoregressive model for order flow (see Gerig (2007)).

[16] Technically, π_i, would need to be estimated, but we assume it is given.

[17] For simplicity, in this paper we do not consider the more complicated case when the end of an order is determined, on average, at an interval larger or smaller than $1/\pi_i$.

The motivation for focusing on these two scenarios is the following. In most markets, transaction information is broadcast to all participants directly after the transaction, and in most cases, the initiator of the transaction, x_t, can be determined. Therefore, the use of an autoregressive model to predict order flow should be uncontroversial. We treat the autoregressive model as a baseline model for order flow predictability because of this. For several reasons, order flow might be more predictable than the autoregressive model. In general, the counterparties involved in a transaction are anonymous, so that transactions do not have a 'color'. However, in some markets this information is available to a select number of liquidity providing individuals, such as specialists at the NYSE. Other markets have experimented with broadcasting this information, e.g., NASDAQ Pathfinders.[18] Even if not broadcast, there might be ways to determine the likely presence of a large order that continually transacts in a predictable way. In a market where prices are determined by individuals with privileged information about split orders or in a market where order information is broadcast or is discernable, the predictability of transaction sign is better modeled by Eq. 5 than Eq. 4. In reality, we would expect markets to operate somewhere between these two extreme models.

III. PRICE RESPONSE TO WORKED ORDERS

In this section, we derive how prices respond to worked orders using the structural model of the previous section. We define a price impact function, $R(t|\Psi)$ that measures the average price response at time t due to an order with given parameters, Ψ. This function measures the price response due to a trade in the direction of the trade, so that it is positive both when buy transactions increase the price and when sell transactions lower the price. Based on the empirical results of Gopikrishnan et al. (2000), Plerou et al. (2004), Vaglica et al. (2008), and Moro et al. (2009), we assume that order sizes are Pareto distributed, $g(N) = \alpha N^{-(\alpha+1)}$, with $\alpha < 2$ and that order signs, ε_i are uncorrelated. The following set of propositions hold (proofs are given in the appendix):

PROPOSITION 1. *For the autoregressive model, the price response to an order while it is transacting is concave and is given by the following approximate equation:*

[18] See https://data.nasdaq.com/Path.aspx

$$R\left(t_n|\Psi\right) \approx \frac{2\lambda}{\alpha}\pi^{\left(1-\frac{\alpha}{2}\right)}n^{\frac{\alpha}{2}}\left(1+\mathcal{O}\left[\frac{1}{n}\right]\right). \tag{6}$$

The autoregressive coefficients are positive at all lags k, which makes the later transactions of an order more predictable than the earlier transactions (the sign of the earlier transactions are included in the sum of Eq. 4 and therefore increase $\hat{x}_t$). Because the later pieces are more predictable, they impact the price less and total impact is a concave function of the amount traded.

COROLLARY TO PROPOSITION 1. *For the autoregressive model, impact is larger for quickly transacted orders and smaller for slowly transacted orders.*

This follows immediately from the equation for impact in Proposition 1. In the autoregressive model, the price response to a single transaction decays in time as the transaction influences later prices through the autoregressive coefficients in Eq. 4. An order that transacts at a higher participation rate, π, does not allow as much time for the price decay of the individual transactions to occur, and therefore causes a larger overall price response.

PROPOSITION 2. *For the colored print model, the price response to an order while it is transacting is concave and is given by the following approximate equation:*

$$R\left(t_n|\Psi\right) \approx \lambda[1+\alpha\log(n)]+\mathcal{O}\left[\frac{1}{n}\right]. \tag{7}$$

In the colored print model, the later transactions of an order are more predictable than the earlier transactions. This occurs because the probability that an order continues, $\mathcal{P}(n)$, increases with n for Pareto distributed order sizes (which means $\hat{x}$ in Eq. 5 increases as n increases). Because the later pieces are more predictable, they impact the price less and total impact is a concave function of the amount traded.

There are two notable differences between the price impact of a worked order in the autoregressive vs. the colored print model. (1) The speed of execution affects the impact in the autoregressive model but does not in the colored print model. When prints are colored, liquidity providers can discern the difference between one order transacting quickly and two orders of the same sign transacting slowly but simultaneously. These two scenarios are indistinguishable in the autoregressive model. Quickly worked orders, therefore, receive a worse price in the autoregressive model. Notice that adding a small amount of noise to the 'color' of a transaction would cause quicker orders to also impact the price more than slower orders in the colored print model. (2) Impact is asymptotically larger in the autoregressive model than in the colored print model. This result is not surprising considering that liquidity providers use more information in the colored print model than they do in the autoregressive model, and therefore can make better predictions about order flow. When order flow is more predictable, it impacts the price less. This doesn't mean that impact is always less if transaction prints are colored, instead it means that impact is shifted to the beginning of an order because the later transactions are better 'announced' by the earlier transactions.[19]

PROPOSITION 3. *For the autoregressive model, when an order completes, prices revert as an inverse power of the time since completion. The reversion is given by the following approximate equation:*

$$R(t_N + \tau|\Psi) \approx \lambda N \frac{1}{\tau^{\left(1-\frac{\alpha}{2}\right)}} \left(1 - \mathcal{O}\left[\frac{1}{\tau}\right]\right). \tag{8}$$

In the autoregressive model, prices revert after the order completes because the transactions of the order continue to influence prices through the autoregressive coefficients in Eq. 4. This influence is in the opposite direction of the order, which means prices revert.

As $\tau \to \infty$, the order's impact reverts completely.[20] In reality, liquidity providers would have some upper bound for the lag k in Eq. 4, which would result in at least a portion of the impact being permanent.

[19] The scale of the price response to a single transaction, λ, should be different in the two models.

[20] Full reversion also occurs in Bouchuad et al. (2004) and Bouchaud, Kockelkoren, and Potters (2006).

PROPOSITION 4. *For the colored print model, when an order completes, prices revert exponentially in time to an amount $\lambda\mathcal{P}(N)$ less than the original impact. The reversion is given by the following approximate equation:*

$$R(t_N + \tau|\Psi) \approx \lambda \left\{1 + \alpha \log(N) - \mathcal{P}(N)\left[1 - (1-\pi)^{\tau}\right]\right\} + \mathcal{O}\left[\frac{1}{N}\right]. \qquad (9)$$

In the colored print model, prices revert because it takes time for liquidity providers to become aware that the order has finished. $\hat{x}_t$ is therefore valued and continues to influence prices (in the opposite direction of the order) even after the order completes. Another way to demonstrate the reversion is the following: because liquidity providers assign some probability to the order continuing on and further influencing prices, then prices must respond in the other direction if the order does not continue. Otherwise prices would not follow a martingale.

In comparison to the autoregressive model, the price reversion in the colored print model is faster but not as pronounced: orders have permanent impact on the colored print model, but do not in the autoregressive model.

IV. CONCLUSIONS

Due to recent changes in the structure of financial markets, it has become relatively easy to work an order—to transact the order piecemeal over time. As a result, many orders that previously would have transacted in one lot are now worked. This has had dramatic effects on order flow variables, with average transaction sizes plummeting and the number of transactions skyrocketing. There exists little previous work that analyzes how these changes affect price formation. This paper attempts to fill that void.

By applying a structural model to incrementally transacted orders, we analyze how prices fluctuate in a market where orders are worked. As demonstrated, our results replicate several empirical findings that have otherwise been difficult to theoretically explain: (1) we find that prices respond in a nonlinear and concave way as an order is transacted, (2) that prices revert, at least partially, after an order completes, and (3) that slowly transacting an order, in general, causes a smaller price response.

The details of our results depend on the particular information available to liquidity providers, which we bracket between two extremes. When liquidity providers only use past transaction data to predict order flow, then impact increases as a power of time. When transaction prints are 'colored', so that they can be associated with their parent orders, then order flow predictability is greatly increased and impact is asymptotically smaller: it increases logarithmically in time. In addition, the reversion is different for the two scenarios; it is quicker but less pronounced when prints are colored. These differences should be of interest to investors, regulators, and market designers; all of whom have power in deciding what information is broadcast to liquidity providers.

The differences in price response for the two models are also theoretically interesting. If transaction prints are colored, then liquidity provision reduces to a dynamic version of Glosten (1994) and prices are set to 'tail-expectations'. This is the usual method used to predict the impact of orders with various sizes. It is highly unlikely that liquidity providers can precisely determine transaction color in real markets. A more likely scenario is that colors are noisy and that price response is somewhere between the two scenarios we analyze.

APPENDIX

Proof of Proposition 1

The expected price impact of a hidden order with parameters, $\Psi = \{\varepsilon, \pi, N\}$ measured from when the order starts to the time when it completes, $T = t_N$, is,

$$\begin{aligned} R\left(t_N|\Psi\right) &= \varepsilon E_\Psi\left[\sum_{t=t_1}^{t_N} p_t - p_{t-1}\right], \\ &= \lambda N - \varepsilon\lambda E_\Psi\left[\sum_{t=t_1}^{t_N} \hat{x}_t\right]. \end{aligned} \tag{10}$$

For the autoregressive model, the coefficients a_k can be estimated as follows. Given that order sizes are power law distributed, $g(N) = \alpha N^{-(1+\alpha)}$, and that orders are split into pieces and transacted, then the resulting order flow time series, x_t, exhibits long memory with Hurst exponent $H = 3/2 - \alpha/2$ (see Lillo, Mike, and Farmer (2005)). The coefficients a_k can be determined by modelling x_t as a FARIMA$(0, H - 1/2, 0)$ process, where in the large k limit, $a_k \sim k^{-H-1/2}$.

$$\begin{aligned} \varepsilon\lambda E_\Psi\left[\sum_{t=t_1}^{t_N} \hat{x}_t\right] &= \lambda E_\Psi\left[\sum_{n=1}^{N}\sum_{k=1}^{t_N-t_n} a_k\right], \\ &\approx \lambda\sum_{n=1}^{N}\left(1 - E_\Psi\left[(t_N - t_n)^{-\left(1-\frac{\alpha}{2}\right)}\right]\right), \\ &\approx \lambda\sum_{n=1}^{N}\left(1 - \pi^{\left(1-\frac{\alpha}{2}\right)}\frac{\Gamma[N-n-(1-\alpha/2)]}{\Gamma[N-n]}\right), \\ &\approx \lambda\sum_{n=1}^{N}\left[1 - \frac{\pi^{\left(1-\frac{\alpha}{2}\right)}}{(N-n)^{\left(1-\frac{\alpha}{2}\right)}}\left(1 + \mathcal{O}\left[\frac{1}{N-n}\right]\right)\right], \\ &\approx \lambda\left[N - \frac{2}{\alpha}\pi^{\left(1-\frac{\alpha}{2}\right)}N^{\frac{\alpha}{2}}\left(1 + \mathcal{O}\left[\frac{1}{N}\right]\right)\right]. \end{aligned} \tag{11}$$

This uses $a_k \approx (1-\alpha/2)k^{\alpha/2-2}$ for a FARIMA process with $H = 3/2 - \alpha/2$ and that $t_N - t_n$ is gamma distributed. The sums are approximated by converting to integrals over the intervals $[1, t_N - t_n]$ and $[0, N]$ respectively. The total impact for the autoregressive model is therefore,

$$R\left(t_N|\Psi\right) \approx \frac{2\lambda}{\alpha}\pi^{\left(1-\frac{\alpha}{2}\right)}N^{\frac{\alpha}{2}}\left(1+\mathcal{O}\left[\frac{1}{N}\right]\right). \tag{12}$$

Proof of Proposition 2

For the colored print model,

$$\begin{aligned}
\varepsilon\lambda E_\Psi\left[\sum_{t=t_1}^{t_N}\hat{x}_t\right] &= \pi\lambda E_\Psi\left[\sum_{n=1}^{N-1}\sum_{t=t_n+1}^{t_{n+1}}\mathcal{P}(n(t))\right], \\
&= \pi\lambda\sum_{n=1}^{N-1}\mathcal{P}(n)E_\Psi\left[\sum_{t=t_n+1}^{t_{n+1}}1\right], \\
&= \pi\lambda\sum_{n=1}^{N-1}\mathcal{P}(n)E_\Psi\left[t_{n+1}-t_n\right], \\
&= \lambda\sum_{n=1}^{N-1}\mathcal{P}(n), \\
&\approx \lambda\left[N-1-\alpha\log(N)+\mathcal{O}\left[\frac{1}{N}\right]\right]. \qquad (13)
\end{aligned}$$

This uses that $\mathcal{P}(n) \approx 1-\alpha/n+\mathcal{O}[1/n]^2$ for the given order size distribution and that $t_{n+1}-t_n$ is exponentially distributed. The sum is approximated with an integral over the interval $[1, N]$. The total impact for the colored print model is therefore,

$$R\left(t_N|\Psi\right) \approx \lambda[1+\alpha\log(N)]+\mathcal{O}\left[\frac{1}{N}\right]. \tag{14}$$

Proof of Proposition 3

Consider the total impact of an order measured at some time τ after the order has completed,

$$\begin{aligned} R(t_N + \tau|\Psi) &= \varepsilon E_\Psi \left[\sum_{t=1}^{t_N+\tau} p_t - p_{t-1} \right], \\ &= \lambda N - \varepsilon \lambda E_\Psi \left[\sum_{t=1}^{t_N+\tau} \hat{x}_t \right], \end{aligned} \tag{15}$$

For the autoregressive model,

$$\begin{aligned} \varepsilon \lambda E_\Psi \left[\sum_{t=1}^{t_N+\tau} \hat{x}_t \right] &= \lambda E_\Psi \left[\sum_{n=1}^{N} \sum_{k=1}^{t_N+\tau-t_n} a_k, \right], \\ &= \lambda \sum_{n=1}^{N} \left(1 - E_\Psi \left[(t_N + \tau - t_n)^{-\left(1-\frac{\alpha}{2}\right)} \right] \right), \\ &\approx \lambda \sum_{n=1}^{N} \left[1 - \frac{1}{\tau^{\left(1-\frac{\alpha}{2}\right)}} \left(1 - \mathcal{O}\left[\frac{1}{\tau} \right] \right) \right], \\ &\approx \lambda N \left[1 - \frac{1}{\tau^{\left(1-\frac{\alpha}{2}\right)}} \left(1 - \mathcal{O}\left[\frac{1}{\tau} \right] \right) \right]. \end{aligned} \tag{16}$$

The impact after the order completes in the autoregressive model is therefore,

$$R(t_N + \tau|\Psi) \approx \lambda N \frac{1}{\tau^{\left(1-\frac{\alpha}{2}\right)}} \left(1 - \mathcal{O}\left[\frac{1}{\tau} \right] \right). \tag{17}$$

Proof of Proposition 4

For the colored print model,

$$\varepsilon \lambda E_\Psi \left[\sum_{t=t_1}^{t_N+\tau} \hat{x}_t \right] = \varepsilon \lambda E_\Psi \left[\left(\sum_{t=t_1}^{t_N} \hat{x}_t + \sum_{t=t_N+1}^{t_N+\tau} \hat{x}_t \right) \right], \tag{18}$$

We have already calculated the expectation of the first sum in parantheses, which resulted in Eq. 13. Focusing on the second sum,

$$\begin{aligned}
\varepsilon\lambda E_\Psi\left[\sum_{t=t_N+1}^{t_N+\tau}\hat{x}_t\right] &= \pi\lambda E_\Psi\left[\sum_{t=t_1}^{t_N+\tau}\mathcal{P}(n(t))\right], \\
&= \pi\lambda\mathcal{P}(N)E_\Psi\left[\sum_{k=1}^{\tau}(1-\pi)^{(k-1)}\right], \\
&= \lambda\mathcal{P}(N)\left[1-(1-\pi)^\tau\right]. \qquad (19)
\end{aligned}$$

Combining these together,

$$\varepsilon\lambda E_\Psi\left[\sum_{t=t_1}^{t_N+\tau}\hat{x}_t\right] \approx \lambda\left(N-1-\alpha\log(N)+\mathcal{P}(N)\left[1-(1-\pi)^\tau\right]+\mathcal{O}\left[\frac{1}{N}\right]\right). \qquad (20)$$

The impact after the order completes in the colored print model is therefore,

$$R_H(t_N+\tau|\Psi) \approx \lambda\left(1+\alpha\log(N)-\mathcal{P}(N)\left[1-(1-\pi)^\tau\right]\right)+\mathcal{O}\left[\frac{1}{N}\right]. \qquad (21)$$

REFERENCES

Almgren, Robert, and Neil Chriss, 2000, Optimal execution of portfolio transactions, *Journal of Risk*, 3, 5-39.

Almgren, Robert, Chee Thum, Emmanuel Hauptmann, and Hong Li, 2005, Direct estimation of equity market impact, *Risk*, 21-28.

Angel, James, Lawrence Harris, and Chester S. Spatt, 2010, Equity trading in the 21st century, working paper, Marshall School of Business.

Bertsimas, Dimitris, and Andrew W. Lo, 1998, Optimal control of execution costs, *Journal of Financial Markets*, 1, 1-50.

Bouchaud, Jean-Philippe, Yuval Gefen, Marc Potters, and Matthieu Wyart, 2004, Fluctuations and response in financial markets: the subtle nature of 'random' price changes, *Quantitative Finance*, 4, 176-190.

Bouchaud, Jean-Philippe, Julien Kockelkoren, and Marc Potters, 2006, Random walks, liquidity molasses and critical response in financial markets, *Quantitative Finance*, 6, 115-123.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2010, Recent trends in trading activity and market quality, working paper, Emory University and University of California, Los Angeles.

Farmer, J. Doyne, Austin Gerig, Fabrizio Lillo, and Henri Waelbroeck, 2010, How efficiency shapes market impact, working paper, Santa Fe Institute.

Gerig, Austin, 2007, A theory for market impact: How order flow affects stock prices, Ph. D. Thesis, University of Illinois.

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

Glosten, Lawrence R., and Lawrence E. Harris, 1988, Estimating the components of the bid/ask spread, *Journal of Financial Economics*, 21, 123-142.

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable? *Journal of Finance*, 49, 1127-1161.

Gopikrishnan, Parameswaran, Vasiliki Plerou, Xavier Gabaix, and H. Eugene Stanley, 2000, Statistical properties of share volume traded in financial markets, *Physical Review E*, 62, R4493.

Hasbrouck, Joel, 1988, Trades, quotes, inventories, and information, *Journal of Financial Economics*, 22, 229-252.

Hasbrouck, Joel, 2007, *Empirical Market Microstructure*, (Oxford University Press, New York).

Huberman, Gur, and Werner Stanzl, 2004, Price manipulation and quasi-arbitrage, *Econometrica*, 72, 1247-1275.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica*, 53, 1315-1335.

Lillo, Fabrizio, and J. Doyne Farmer, 2004, The long memory of the efficient market, *Studies in Nonlinear Dynamics & Econometrics*, 8, 1-33.

Lillo, Fabrizio, Szabolcs Mike, and J. Doyne Farmer, 2005, Theory for long memory in supply and demand, *Physical Review E*, 71, 066122.

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

Moro, Esteban, Javier Vicente, Luis G. Moyano, Austin Gerig, J. Doyne Farmer, Gabriella Vaglica, Fabrizio Lillo, and Rosario N. Mantegna, 2009, Market impact and trading profile of hidden orders in stock markets, *Physical Review E*, 80, 066102.

Obizhaeva, Anna, and Jiang Wang, 2006, Optimal trading strategy and supply/demand dynamics, working paper, Massachusetts Institute of Technology.

Plerou, Vasiliki, Parameswaran Gopikrishnan, Xavier Gabaix, and H. Eugene Stanley, 2004, On the origin of power-law fluctuations in stock prices, *Quantitative Finance*, 4, C11-C15.

Stoll, Hans R., 2000, Friction, *Journal of Finance*, 55, 1479-1514.

Torre, Nicolo, 2007, *BARRA Market Impact Model Handbook*, (BARRA Inc).

Yuen, Peter, 2007, Introduction to execution costs, What is implementation shortfall?, Credit Suisse.

Vaglica, Gabriella, Fabrizio Lillo, Esteban Moro, and Rosario N. Mantegna, 2008, Scaling laws of strategic behavior and size heterogeneity in agent dynamics, *Physical Review E*, 77, 036110.