The market impact of large trading orders: Correlated order flow, asymmetric liquidity and efficient prices

J. Doyne Farmer,[1,2,*] Austin Gerig,[1,†] Fabrizio Lillo,[3,1,‡] and Henri Waelbroeck[4,§]

[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*
[2]*LUISS Guido Carli, Viale Pola 12, 00198, Roma, Italy*
[3]*Dipartimento di Fisica e Tecnologie Relative, viale delle Scienze I-90128, Palermo, Italy*
[4]*Pipeline Financial Group, Inc., 60 E. 42nd St., Suite 624, New York NY 10165*

We study the price change associated with the incremental execution of large trading orders. The heavy tails of large order sizes leads to persistence in the signs of transactions: Buyer initiated transactions tend to be followed by buyer initiated transactions and seller initiated transactions tend to be followed by seller initiated transactions. The resulting predictability in order flow implies that to preserve market efficiency, liquidity must be asymmetric in the sense that trades of the same same sign as the large order generate smaller returns than returns of the opposite sign. The predictability of order flow increases during the execution of a large order, making returns smaller and causing the overall impact to be a concave function of order size. This depends on the information market participants have about order flow. Under assumptions described in the paper, the theory that we develop predicts the functional form of market impact, the degree to which impact is temporary or permanent, and its dependence on trading velocity. We perform empirical tests using data from the London Stock Exchange.

PRELIMINARY DRAFT
Comments appreciated

* jdf@santafe.edu
† gerig@santafe.edu
‡ lillo@santafe.edu
§ henri.waelbroeck@pipelinefinancial.com

Contents

I. INTRODUCTION

Market impact measures the expected price change induced by initiating a trade[1]. One of the puzzles of finance is that the market impact as a function of trading volume has a highly concave functional form, i.e. its derivative is a decreasing function of volume. This is widely assumed to reflect the diverse information present in trades of different sizes, e.g. because skilled professionals make intermediate to small trades while less skilled participants use an assortment of different sizes. In this paper we present an alternate point of view, arguing that the changes in the informativeness are driven by correlations in order flow, which are in turn driven by heavy tails in the size of large transactions. Under our view, while the concave form of impact does indeed depend on participant information, it has little or nothing to do with their identity.

Understanding market impact is important conceptually for what it tells us about the aggregation of information about trades into prices, and also for what it tells us about the underlying behavior of supply and demand. It is closely related to the demand elasticity of price, originally introduced by Alfred Marshall[2]. Knowing the market impact does not allow one to compute absolute price levels, but it does make it possible to answer the hypothetical question, "How much would I move the price if I were to make a trade?".

From an information point of view trades should be informative, in the sense that in a world in which each agent has different information, a trade and its incorporation into the price make this information public (Grossman and Stiglitz, 1980). One of our main accomplishments here is in developing a theory for how information about order flow is incorporated into prices. We explore two different models for predicting order flow and show that these lead to different functional forms for market impact. This has implications for market design, since our results imply that by controlling what information is made available during trading market designers can exert control over market impact.

Understanding market impact also has important practical implications. Practitioners care about understanding market impact because it reduces profits. Since market impact increases with trading size it places a limit on fund size. By adversely moving the prices at which transactions are made, impact can turn a profitable strategy into a losing strategy. This is the reason why savvy hedge funds close once they reach a critical size. The functional form matters: If market impact increases rapidly with volume, a fund's size is severely limited; if it increases slowly, the fund can grow much larger. Practitioners work hard to minimize market impact, and optimal algorithms for doing this are a topic of active research (Almgren and Chriss, 2000, 1999, Almgren, 2003). We give insight into the functional form that should be optimized and show that other properties such as the trading velocity are also important. Of course liquidity in markets varies enormously depending on context (see e.g. Gillemot, Farmer and Lillo, 2006), but what we are interested in here is how the predictability of order flow affects the market impact of transactions, averaged over different market conditions.

---

[1] "Initiation" refers to the party who is immediately responsible for a trade taking place. For example in a continuous double auction the initiating party is the one who places the order that causes an immediate transaction. This can also be defined in other market structures where actions are taken sequentially.

[2] For asynchronous market clearing, in which parties change their supply or demand functions one at a time, it is possible to show that the market impact for a series of trades is linearly proportional to the demand elasticity of price averaged over a sequence of trades (Farmer, Gerig and Lillo, in progress).

The functional form of market impact is one of the unexplained puzzles in finance. There have been many empirical studies of market impact[3]. All of them have observed a concave function of trading volume, i.e. the derivative is a decreasing function. The functional forms that have been reported to give a good fit to the data vary from study to study. We believe that much of this variation comes about because these studies in fact measure different things. Some of them measure the market impact of a single trade made in an order book, some measure the aggregate impact of sequential trades in an order book, some of them measure block trades, and many of them measure a mixture of all three. We believe that these are quite different and need to be analyzed separately. This paper is one of a series in which we develop theories for market impact for each of the above cases.

In this paper we focus on understanding how the predictability of order signs affects the total market impact of large trades that are executed sequentially in small pieces. Such trades are called *trading packages* or *hidden orders*, and the individual small pieces via which they are executed will be called *realized orders*. The strategic reasons for incremental execution were originally analyzed by Kyle (1985), who developed an idealized model with three types of traders: noise traders, who buy or sell a random quantity, a monopolist with inside information, and market makers who buy and sell at prices that keep the market efficient. He showed that in order to maximize profits the insider will split her trade into pieces and information will be gradually incorporated into the price as each piece is traded.

Our approach is complementary to Kyle's, and differs both in its goals and in the construction of the model. Our purpose is to compute the functional form of market impact based on empirically observable assumptions, taking the strategic motivations for order splitting as a given. The model contains only two types of traders, liquidity takers and liquidity providers. The liquidity takers are noise traders who decide to buy or sell at random, picking a random size V for their hidden orders from a known distribution $P(V)$. They split up V into small pieces $\bar{v}$ and trade incrementally at a constant rate, independent of V. Liquidity providers set prices so as to maintain market efficiency. To do this they have to compensate for the fact that, depending on $P(V)$, the incremental trading by noise traders induces positive autocorrelations in order flow. The need to maintain the price process as a martingale forces price responses to be asymmetric, i.e. the price response to buy orders must be different than that for sell orders. The asymmetry varies depending on the number of hidden orders, their signs, and their stage of execution. The predictability of order flow depends on $P(V)$, and as a result the concavity of the impact depends on the tail behavior of $P(V)$. The functional form of market impact also depends on the information available to liquidity providers for predicting order flow; we investigate two different information sets and show how they affect market impact.

Our theory stresses the consequences of the predictability of order flow for price formation. This was presaged by Hasbrouck's (1988, 1991) observation that only the unpredictable component of order flow can affect market impact[4]. The release of information enabling

[3] Empirical studies of market impact include Holthausen, Leftwich and Mayers (1987, 1990), Hasbrouck (1991), Hausman, Lo and MacKinlay (1992), Keim and Madhavan (1996), Torre (1997), Kempf and Korn (1999), Plerou et al. (2002), Evans and Lyons (2002), Lillo, Farmer, and Mantegna (2003), Potters and Bouchaud (2003), Gabaix et al. (2003, 2006), Chordia and Subrahmanyam 2004, Farmer, Patelli and Zovko (2005), Weber and Rosenow (2006), and Hopman (2006).

[4] Other price formation models that have made use of the need to enforce efficiency under predictable order flow include Garbade and Silber (1979), Roll (1984), Glosten and Milgrom (1985), Choi, Salandro and

its predictability should also affect stock prices, although this is more difficult to quantify. Order flow is highly predictable. Positive serial autocorrelation of signed order flow has been observed by many authors in many different markets, including the Paris Bourse (Biais, Hillion and Spratt, 1995), foreign exchange markets (Danielsson and Payne, 2001), and the NYSE [Ellul et al. (2005), Yeo (2006)]. These papers only studied the first autocorrelation, but this is only a small part of the story. For stocks in the London, Paris, and New York stock exchanges, the signs of trades in stock markets obey a long-memory process[5] [Bouchaud et al. (2004), Lillo and Farmer (2004)]. Typically *all* the coefficients of the autocorrelation function are positive at statistically significant levels out to lags of more than a thousand. The predictability of order flow is dramatic, and as stressed by Bouchaud et al. and Lillo and Farmer, would produce a serious violation of market efficiency if it weren't compensated for in price formation. Bouchaud et al. (2004, 2006) have shown how this affects the permanence of price impact; here we show how it also influences its dependence on trade size. We also show that the permanence of impact depends on the information available for predicting order flow.

We are not the first to model the price impact of large trades. Keim and Madhavan (1996) modeled block trades based on the cost of searching for counter-parties, and under the assumption that this increased as a power law in the number of counter-parties, derived a power law impact function. Evans and Lyons (2002) developed a theory for interdealer and public trading, and under the assumption that the public's demand function is linear derived a linear impact function. Gabaix et al. (2003, 2006) assumed that block traders have an additive risk aversion term in their utility function of the form $\mathcal{V}^{\delta/2}$, where $\mathcal{V}$ is the variance of their profits. Under the assumption of utility maximization and a random walk for prices, they showed that the impact of large trades should scale as $V^{\delta/2}$, where V is the trading volume. If $\delta = 1$, i.e. first order risk aversion, their theory predicts impact increases as the square root of volume, but with the more commonly used $\delta = 2$ it predicts a linear impact. These theories add insight, but in each case the functional form of the answer depends on an arbitrary functional form assumed in setting up the model. Our approach instead makes assumptions about the functional form of the volume distribution and the information available for predicting order flow. These have the important advantage that they can be inferred directly from empirical data.

We also provide preliminary empirical tests of our model. Empirical studies of the market impact of hidden orders are difficult to perform because linking realized trades together requires information about the identity of trading parties. Previous studies by Chan and Lakonishok (1993, 1995), and Torre (1997) were based on data from a brokerage that made it possible to explicitly track the orders of each client. Our approach follows the lead of Vaglica et al. (2007), who have developed a method for reconstructing large orders using information about brokerage codes. We use data from the London Stock Exchange containing codes identifying the member of the exchange submitting each trading order, and develop a simple algorithm for linking together a series of realized trades into underlying hidden orders. This approach has the disadvantage that it is impossible to classify hidden orders unambiguously,

Shastri (1988), and Stoll (1989) and Madhavan, Richardson and Roomans (1997).

[5] A standard example of a long-memory process is a fractionally integrated Brownian motion. We use the term in its more general sense to mean any process whose autocorrelation function is non-integrable (Beran, 1994). This can include processes with structure breaks, such as that studied by Ding, Engle and Granger (1993).

as there is always the possibility that two parties are trading simultaneously through the same member of the exchange. In fact we are able to estimate the probability that this happens, and so estimate the misclassification rate. Our approach has the advantage that it allows us to classify all trading orders for the whole market, and so gives us more data to work with than we would have if we were restricted to the clients of a single brokerage.

II. OUTLINE AND OVERVIEW

Because the theory developed here is complicated and has several interlocking pieces, we present an overview that can also act as a guide for understanding how the sections of the paper fit together.

In **Section III** we present the basic set up of the model. Under the assumption that the predictablity of order flow comes from hidden orders we derive a relationship connecting the persistence of hidden order flow to the predictability for realized order flow. We then impose market efficiency and derive the consequences for returns, showing that it means that expected returns are asymmetric, which is the basic reason that market impact is concave. As a hidden order develops its probability of continuing increases. This implies that the ratio of returns of the same sign (as the hidden order) to those of the opposite sign decreases. Under reasonable assumptions about scale, discussed in the next section, this makes the market impact concave.

In **Section IV** we discuss the problem of determining how the scale of returns changes during the development of a hidden order. The arguments given in Section III determine the ratio of positive and negative returns but do not determine their scale. We explore the consequences of assuming that the surprise in order flow is additive, i.e. that the amount by which the price moves depends on the difference between the observed order sign and the prediction. We show that this implies that the single transaction impact is linear and symmetric.

In **Section V** we address the fact that quantifying the predictability of order flow requires assumptions about the information and models of market participants. We develop two different models of predictability that in a certain sense bracket the space of possibilities. The first assumes that participants use a linear time series model to predict order flow based on public information about order signs; the second, which we call the "colored print" model, assumes more detailed information about hidden orders, including the underlying distribution of their size, the number of hidden orders present and the number of executions that each order has had so far.

Under assumptions about all the elements discussed above, namely volume distribution, scale, order flow prediction model, and neglecting quote adjustments resulting from updates in the order flow predictions, it is possible to derive the market impact function. We do this for each of the possibilities discussed above. To give an understanding of how the assumptions affect the results, we also do this for some assumptions that we believe are counterfactual; for example, we also consider the assumption that the hidden order volume distribution is a stretched exponential rather than a power law. So, for example, if we assume the color print model for order flow and constant volatility, an exponential distribution of hidden order volume leads to a linear market impact function, a stretched exponential leads to a power law market impact, and a power law hidden order distribution leads to a logarithmic impact function.

Under the colored print model one must also confront how participants detect the end of hidden orders. We show that this substantially modifies the impact and also modifies the response after the hidden order is finished. In particular this makes it possible to compute whether the impact is permanent or temporary.

In **Section VI** we present some preliminary empirical results testing these models. We first describe the method we use for approximately identifying hidden orders using only information about exchange membership and discuss the advantages and disadvantages of this method relative to other possibilities.

Finally in **Section VII** we summarize and provide a broader perspective on these results, and discuss future work.

III. MARKET EFFICIENCY AND RETURN ASYMMETRY

As we have already emphasized, the fact that hidden orders are only executed incrementally makes order flow extremely predictable. Enforcing market efficiency imposes strong constraints on the response of prices and leads to the conclusion that returns must be asymmetric. In this section we derive a relation between the structure of the hidden order process and the predictability of order flow. We then impose efficiency and derive a general relation between the predictability of hidden orders and returns.

For the sake of clarity we develop this theory at a microscopic level, i.e. we formulate the model at the level of individual transaction-to-transaction returns. This has the disadvantage of requiring the introduction of details that are not essential to understand the main ideas, but has the important advantage of making the microscopic mechanisms more explicit and enabling detailed empirical tests of the theory.

A. Basic set up of the model

Following Lillo, Mike, and Farmer (2005) we assume that all buying and selling decisions by liquidity takers correspond to hidden orders of size V that are executed through a series of incremental transactions of size $\bar{v}$. The signs of their hidden orders, $\eta = +1$ for buy and $\eta = -1$ for sell, are chosen at random in an IID manner. Their sizes are drawn from a distribution $P(V)$. For convenience we assume that all transactions have the same size $\bar{v}$ and that hidden orders have discrete sizes $V = N\bar{v}$, where N is a positive integer. (This simplifies the analysis and does not substantially affect any of the conclusions). Though we are most interested in long hidden orders, they may have lengths as short as $N = 1$, and indeed in the examples we will investigate here this is the most common length. All transactions come from hidden orders and have the same sign as the hidden order they come from. The sequence of transaction signs ϵ_t coming from different hidden orders is called *order flow*. We will say that a hidden order is *active* from the time it is created[6] until the time it is fully executed, i.e. until it has received N executions. $M(t)$ hidden orders can be active at the same time.

[6] For the purposes of this paper we will use empirical data and so don't need to worry about when new hidden orders are created – that is given to us from the data. For stochastic process models of hidden order creation see Lillo, Mike and Farmer.

While most of our results do not depend on this, for many purposes it is useful to treat the execution of hidden orders as a Poisson process in which a given active hidden order j has probability π_j of making an incremental transaction on any given timestep. (We will often drop the subscript j when it is obvious). We will call the Poisson rate π_j the *participation rate*. Under the Poisson execution model the average time between transactions of a given order is $\theta_j = 1/\pi_j$. Since we are using transaction time, $\theta_j - 1$ is the average number of intervening transactions coming from other hidden orders. Even though the signs of hidden orders are IID, depending on $P(V)$, the incremental transaction of large hidden orders can cause order flow to be persistent in the sense that it has positive autocorrelations.

As a simplifying assumption we assume that all hidden orders are executed with market orders. This assumption is valid for a class of algorithms known by practitioners as stealth algorithms, which are intended to avoid detection by not posting any limit orders, and almost entirely valid for volume participation strategies that aim to track market volume and therefore do not accept the uncertainty of execution that results from posting limit orders. Aggressive trading strategies aiming to capture over 20% of the market liquidity also execute mostly through market order transactions. The generalization of our arguments to passive limit order posting strategies lies beyond the scope of this paper.

Although the way in which we measure returns is arbitrary it is necessary to be consistent. We work in *transaction time* t, which is incremented according to $t \to t + 1$ every time a transaction occurs. Prices are measured in terms of the logarithmic midprice $p = \log(1/2(a+b))$, where a is the best offered selling price (the best ask) and b is the best offered buying price (the best bid). The logarithmic return $r_t = p_t - p_{t-1}$ is measured as the logarithmic price difference from immediately before a transaction is received until immediately before the next transaction is received. This means that each return includes the immediate impact of a transaction followed by an indeterminate number of quote driven price changes until the next transaction.

It is also useful to measure returns associated with hidden order executions. Let t_n be the time of the n^{th} execution, where $n = 1, \ldots, N$. In order to be able to define the last hidden order return ρ_N we will somewhat arbitrarily define $t_{N+1} = t_N + \theta$, where $\theta = (t_N - t_1)/(N-1)$ for $N > 1$ and $\theta = 1$ for $N = 1$. The hidden order return is then $\rho_n = p_{t_{n+1}} - p_{t_n}$. Our notation and our method for defining returns are illustrated in Figure 1.

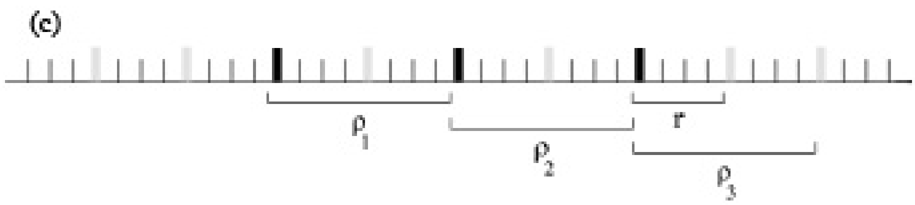

FIG. 1: Our method of defining returns. Solid bars are transactions of the hidden order and grey bars are transactions caused by other orders. The thin marks are quotes or cancellations. The return r_t of individual realized transactions is defined from just before a transaction to just before the next transaction, and the return ρ_n is defined from just before a transaction of a hidden order until just before its next transaction.

B. Price formation for individual incremental transactions

We model the total impact of hidden orders by summing the impacts of each incremental transaction. Our model for price formation for the incremental transactions is a generalization of the model of Madhavan, Richardson and Roomans (MRR, 1997),

$$p_t = p_{t-1} + K(\epsilon_t - E[\epsilon_t|\epsilon_{t-1}]) + \chi_t. \tag{1}$$

p_t is the midprice, K is a positive constant, ϵ_t is the sign of a transaction, $E[\epsilon_t|\epsilon_{t-1}]$ is the expected sign given the previous sign, and χ_t is a noise term that represents the arrival of new information that is not already contained in order flow. Since $-1 \leq E[\epsilon_t|\epsilon_{t-1}] \leq 1$, the contribution of a buyer-driven transaction is never negative, and of a seller-driven transaction is never positive.

We generalize and extend this model in several ways.

- We allow for a more general order flow model $\hat{\epsilon}_t = E_{t-1}[\epsilon_t|\Omega_{t-1}]$ to forecast the transaction sign ϵ_t based on information at time $t-1$, where Ω is an information set. As we make explicit in Section V, examples of possible information sets are a linear time series model based on order flow or the number of prior executions n for each hidden order j.

- Since we are interested in the expected impact of hidden orders, for each hidden order we measure time relative to the time $t_1^{(j)}$ of its first transaction. If t' sequentially labels all transactions in the temporal order which they occur, for hidden order j the relative time $t(j) = t' - t_1^{(j)} + 1$. (For convenience the argument j will usually be omitted).

- We split the noise term into two pieces, an idiosyncratic noise term $\xi_t(j)$ that is specific to hidden order j, and a noise term χ_t that describes innovations that survive after averaging over hidden orders.

- The sign of transactions ϵ_t is replaced by the variable $\varepsilon_t = \eta_j \epsilon_t$, which indicates whether the transaction sign agrees with the hidden order sign.

- We allow for a more general nonlinear function F describing the impact of individual transactions of the form $F(\varepsilon_t, \hat{\varepsilon}_t)$. F satisfies the conditions that $F(x,x) = 0$ and it is increasing in the first argument and decreasing in the second.

With these modifications our single transaction impact model is

$$p_t = p_{t-1} + \eta_j \left[F(\varepsilon_t, \hat{\varepsilon}_t) + \xi_t(j) + \chi_t \right]. \tag{2}$$

Expectations in our model are averages over hidden orders. The expectations of the idiosyncratic noise term $\xi_t(j)$ are by definition $E_{t-1}[\xi_t(j)] = E_t[\xi_t(j)] = 0$, where the subscript denotes the time when the average is taken. In contrast, the expectations of the systematic noise term χ_t are $E_{t-1}[\chi_t] = 0$ and $E_t[\chi_t] = \chi_t$. In the latter case the expectation is nonzero because χ_t is known at time t.

We have substantially generalized the allowed functional form of the impact in two different ways. First, we have allowed a general function F rather than the linear form assumed by MRR. Second, we have allowed it to depend on the observed transaction sign ε_t and the predicted transaction sign $\hat{\varepsilon}_t$ separately rather than on their difference. We will consider the latter simplification in Section IV.

For convenience we have implicitly assumed antisymmetry between buying and selling, i.e. that transactions from buy vs. sell hidden orders generate impacts of the same size but opposite sign. This is evident from that fact that if we flip the signs of all hidden orders, sending $\eta_j \to -\eta_j$, by definition from the setup of the model in the previous section this takes $\epsilon_t \to -\epsilon_t$, ε_t remains invariant, and under Eq. 2 we see that $r_t = p_t - p_{t-1} = -r_t$. This is a matter of convenience and parsimony – we could more generally have assumed one function F for buying and another F' for selling. This is easily tested empirically. This assumption should not be confused with the symmetry properties of F. For example, if F is a linear as assumed by MRR, then $F(-x) = -F(x)$ and the price response to order flow is the same but with a negative sign for surprises in order flow with the same sign as the hidden order as it is for surprises with the opposite sign. This is a much more important assumption that we will discuss in more detail in Section IV.

For computing the market impact of a given hidden order j it is useful to know the probability that next transaction will have the same sign or a different sign than the hidden order. These are defined as

$$\begin{aligned} \varphi_t^+ &= P(\epsilon_t = \eta_j | \Omega_{t-1}), \\ \varphi_t^- &= P(\epsilon_t \neq \eta_j | \Omega_{t-1}). \end{aligned} \tag{3}$$

This is simply related to the prediction $\hat{\varepsilon}_t$ of whether the order flow sign agrees with the hidden order sign as follows

$$\hat{\varepsilon}_t = \eta_j \hat{\epsilon}_t = E_{t-1}[\epsilon_t = \eta_j | \Omega_{t-1}] = \varphi_t^+ - \varphi_t^- = 2\varphi_t^+ - 1. \tag{4}$$

C. Consequences of market efficiency on returns

Our strategy is to use efficiency to compute the size of the expected *a priori* returns and use this *a posteriori* to compute the average impact. Using Eq. 2 we can compute the magnitude of the expected returns at $t-1$ conditioned on observing that the next order has a given sign,

$$\begin{aligned} \tilde{r}_t^+ &= E_{t-1}[r_t \eta_j | \epsilon_t = \eta_j \ \& \ \Omega_{t-1}] = F(1, \hat{\varepsilon}_t), \\ \tilde{r}_t^- &= -E_{t-1}[r_t \eta_j | \epsilon_t \neq \eta_j \ \& \ \Omega_{t-1}] = -F(-1, \hat{\varepsilon}_t). \end{aligned} \tag{5}$$

While the martingale condition above is naturally formulated in terms of the expected returns $\tilde{r}^+$ and $\tilde{r}^-$, they are not observable. This is because they represent expectations conditioned on information at time $t-1$ but the returns that can be measured from empirical data[7] also include information at time t. From Eq. 1 the average returns at time t are

$$\begin{aligned} r_t^+ &= E_t[r_t \eta_j | \epsilon_t = \eta_j \ \& \ \Omega_t] = F(1, \hat{\varepsilon}_t) + \chi_t, \\ r_t^- &= -E_t[r_t \eta_j | \epsilon_t \neq \eta_j \ \& \ \Omega_t] = -F(-1, \hat{\varepsilon}_t) - \chi_t. \end{aligned} \tag{6}$$

[7] The return at time t includes possible conditioning of order flow by liquidity takers, based on information at time t.

The expectations r^+ and r^- are observable but do not satisfy the efficiency condition exactly due to the information term χ_t. If the deviation from efficiency is $\varphi_t^+ \tilde{r}_t^+ - \varphi_t^- \tilde{r}_t^- = \tilde{\Delta}$, the *observable inefficiency* is

$$\Delta_t = \tilde{\Delta}_t + \chi_t = \varphi_t^+ r_t^+ - \varphi_t^- r_t^- . \tag{7}$$

We wish to emphasize that the observable inefficiency includes both the inefficiency of prices with respect to a given order flow model, as well as the informational terms χ_t, which are not an inefficiency but rather a direct response to information about order flow that is not reflected in the prediction model. We have defined the model in this way in order to make it directly comparable to observations.

When combined with the definition of volatility and the conservation of probability this gives a simple system of three linear equations. Let $\tilde{\nu}$ be a proxy of volatility, which we will take to be $\tilde{\nu} = \varphi^+ \tilde{r}^+ + \varphi^- \tilde{r}^-$. Providing $\tilde{r}^+$ and $\tilde{r}^-$ are both positive $\tilde{\nu}$ is the absolute value of returns.

$$\begin{aligned} \varphi^+ r^+ - \varphi^- r^- &= \Delta, \\ \varphi^+ r^+ + \varphi^- r^- &= \nu, \\ \varphi^+ + \varphi^- &= 1. \end{aligned} \tag{8}$$

We stress that all the quantities above can depend on t, so this defines a system of linear equations for each value of t. In particular the predictability of order flow as measured by φ_t^+ can vary within a hidden order due to variations in the information set Ω as the order develops.

We can now solve for the conditional expected returns r^+ and r^-, which gives

$$r^+ = \frac{\nu + \Delta}{2\varphi^+} \qquad r^- = \frac{\nu - \Delta}{2(1 - \varphi^+)}. \tag{9}$$

Note that the condition that ν corresponds to the absolute value of returns is met as long as $\nu > \Delta$, which guarantees that $r^- > 0$.

If the market is efficient and $\chi_t = 0$, then $\Delta = 0$ and the return asymmetry becomes

$$\frac{r^+}{r^-} = \frac{1 - \varphi^+}{\varphi^+}. \tag{10}$$

This implies that the response r^+ to buying orders differs from the response r^- to selling orders. When buy orders are more likely the response to buy orders is smaller, and vice versa. Solving for φ^+ and substituting into Eq. 10 gives

$$\frac{r^+}{r^-} = \frac{1 - \hat{\varepsilon}}{1 + \hat{\varepsilon}}. \tag{11}$$

This will be useful later for empirical testing.

There are several market microstructure effects that can produce an asymmetric price response. Possible causes are (1) a difference in the depth at the bid vs. the offer, (2) asymmetric liquidity taking, for example because buyers submit smaller market orders than sellers, or (3) differences between buy and sell quote driven price changes. We suspect that all of these contribute, but from the point of view of the theory we develop here their relative contribution is not important.

D. Predictability of order flow due to hidden orders

A key assumption that we make here is that the incremental execution of hidden orders is the dominant cause of the predictability of realized order flow, as proposed by Lillo, Mike and Farmer (2005). This is of course not the only possible cause of predictability. For example, Parlour (1998) developed a theory based on strategic considerations that predicts negative autocorrelations (which is the opposite of what is observed). Lebaron and Yamamoto (2007) have proposed a theory for positively correlated order flow based on the hypothesis that market participants imitate one another, which is also called *herding*. Nonetheless, as discussed in Section VI, in the stock markets that have been studied we believe the empirical evidence makes it clear that incremental hidden order execution is the dominant source of predictability. We will assume that it is the only cause.

We now derive a general relationship relating the predicted order flow imbalance φ^+ to the participation rate π and the probability $\mathcal{P}$ that the hidden order will continue. The continuation probability

$$\mathcal{P}(n) = P(N \leq n|\Omega_{t-1}) \tag{12}$$

is the probability that a hidden order will continue conditioned on information Ω_{t-1}. It is useful because it links the predictability of order flow to the hidden order process, and can be computed once we have chosen an information set Ω and fully specified the hidden order process.

By assumption the signs η_j and η_k of any two hidden orders j and k are independent. For convenience also assume that the unconditional probabilities of buy and sell orders are the same; the results are easily generalized to avoid this with an obvious trivial modification of the results. Suppressing the ubiquitous time indices, this allows us to write φ_t^+ in the form

$$\begin{aligned} \varphi^+ &= \left(\pi + \frac{1}{2}(1-\pi)\right)\mathcal{P} + \frac{1}{2}(1-\mathcal{P}) \\ &= \frac{1}{2}\left(1 + \pi\mathcal{P}\right). \end{aligned} \tag{13}$$

This can understood by examining the first expression, starting with the leftmost term: By definition the probability that a hidden order continues is $\mathcal{P}$. If it continues then there is a probability π that it has a transaction (which automatically has the same sign), and probability $1-\pi$ that another hidden order has a transaction, in which case there is a probability of one half of having the same sign. The probability that the order does not continue is $1-\mathcal{P}$, in which case there is probability one half of having the same sign.

The impact can be computed using Eqs. 13 and 8. Let J_t indicate that hidden order j is active at time t and $\bar{J}_t$ indicate it is inactive. If a hidden order is still active at time t the expected return is

$$E_t[r_t|J_t \ \& \ \Omega_t] = \eta_j\left[\pi r^+ + \frac{1-\pi}{2}(r^+ - r^-)\right] = \eta_j\left[\frac{\nu\pi(1-\mathcal{P}) + \Delta(1-\pi^2\mathcal{P})}{1-(\pi\mathcal{P})^2}\right]. \tag{14}$$

The middle expression can be interpreted as follows: Assuming hidden order j is a buy order there is probability π that the transaction at time t is an execution of hidden order j, in which case by definition the realized order has the same sign and the average return is r^+. There is probability $(1-\pi)$ that it is a transaction of some other hidden order, in which case since hidden order signs are independent there is an equal probability 1/2 for either generating a realized buy order with average return r^+ or a realized sell order with average return $-r^-$. If the hidden order is a sell order the factor η_j flips all the signs.

The expected return $E_t[r_t|J_t]$ is an *a posteriori* expected return, in the sense that it is based on information that the hidden order is still active. Thus it is generally not zero. The size of the expected return is determined by the imbalance between the conditional expectation r^+ for orders of the same sign and r^- for orders of the opposite sign.

E. Hidden order vs. transaction returns

In assessing the impact of a hidden order it is convenient to measure its impact in a natural time frame. We do this in terms of the number of executions n of the hidden order. We now derive a relation between the expected transaction by transaction returns $E_t[r_t|J_t \text{ \& } \Omega_t]$ and the expected hidden order returns $E_n[\rho_n|J_n \text{ \& } \Omega_n]$. For this purpose it is convenient to use the Poisson model of order placement. Letting the stochastic variable $\tau_n = t_{n+1} - t_n$ be the time interval between hidden order executions we assume that τ_n is independent of r_t, or alternatively that $E[r_t|J_t]$ is roughly constant from t_{n-1} to t_n. Under the Poisson approximation this gives

$$\begin{aligned} E_n[\rho_n|J_n \text{ \& } \Omega_n] &= \sum_{t=t_n+1}^{t_{n+1}+1} E_t[r_t|J_t \text{ \& } \Omega_t] = E[\tau_n]E_t[r_t|J_t \text{ \& } \Omega_t], \\ &= \sum_{\tau=1}^{\infty} \tau\pi(1-\pi)^{\tau-1} E_t[r_t|J_t \text{ \& } \Omega_t], \\ &= \theta E_t[r_t|J_t \text{ \& } \Omega_t], \end{aligned} \tag{15}$$

where $\theta = 1/\pi$ as defined in Section III A. This answer is intuitively pretty obvious – the return from θ steps is just θ times the return for one step. This is not very sensitive to the Poisson arrival hypothesis; for example, if we instead assume that hidden orders are executed at periodic intervals of constant length $\tau_n = \theta$ we get the same answer.

F. Predicted market impact

It is now possible to derive a general expression for market impact. We define the total impact of an hidden order in terms of the total logarithmic return $R = p_{t_N} - p_{t_1}$. Equation 14 is the expected single transaction return. The variables ν, $\mathcal{P}$ and Δ are all single transaction values, which in general will vary during the course of a hidden order. (We have already assumed that the participation rate π is a constant). Equation 15 is essentially a statement that we can regard all of these variables as constant in between executions of the hidden order. We can thus write $\nu(n)$, $\mathcal{P}(n)$ and $\Delta(n)$ to indicate the values of each of these variables after n executions of the order. We can then write the *a posteriori* expected impact $E[R|N]$ for a hidden order of length N as

$$\begin{aligned} E[R|N] &= \sum_{n=1}^{N} E_n[\rho_n|J_n \text{ \& } \Omega_n] = \theta \sum_{n=1}^{N} E_t[r_t|J_t \text{ \& } \Omega_t] \\ &= \eta_j \left[\sum_{n=1}^{N} \nu(n) \frac{1-\mathcal{P}(n)}{1-(\pi\mathcal{P}(n))^2} + \frac{1}{\pi} \sum_{n=1}^{N} \Delta(n) \frac{1-\pi^2\mathcal{P}(n)}{1-(\pi\mathcal{P}(n))^2} \right]. \end{aligned} \tag{16}$$

As we will demonstrate later, for the stocks we study typically $\pi < 0.1$, making it a good approximation to neglect terms of order π^2, since $\pi^2 \leq 0.01$ and all the terms with π^2 are subtracted from terms of order one. With this assumption $E[R|N]$ can be written in the simple form

$$E[R|N] \approx \eta_j \left[\sum_{n=1}^{N} \nu(n)(1 - \mathcal{P}(n)) + \theta \sum_{n=1}^{N} \Delta(n) \right]. \tag{17}$$

In order to actually compute the market impact we have to resolve the two unknown functions, $\nu(n)$ and $\mathcal{P}(n)$. The volatility proxy $\nu(n)$ can be thought of as the scale of the returns, and will be discussed in the next section. The continuation probability for hidden orders, $\mathcal{P}(n) = P(N \leq n|\Omega_{t-1})$, depends on the information set Ω and will be discussed in Section V. We also have to show that the Δ term is sufficiently smaller than the other terms so that it can be neglected. We investigate this question empirically in Section VI F.

IV. THE SCALE OF PRICE RESPONSES TO HIDDEN ORDERS

To compute the impact we have to fix the scale of the returns as the hidden order develops. This is controlled in our model by the volatility proxy $\nu(n)$. Market efficiency by itself is not sufficient to specify this – an additional assumption is needed.

One possible way to fix the scale is to make the *additive assumption* that the expected single transaction impact is of the form

$$F(\varepsilon_t, \hat{\varepsilon}_t) = F(\varepsilon_t - \hat{\varepsilon}_t). \tag{18}$$

This says that the price change caused by new information in order flow depends only on the difference between the observed order sign and its prediction. This was assumed by MRR and many other authors When combined with efficiency this is sufficient to specify F. From Eq. 5 the martingale condition can be written

$$\begin{aligned} E_{t-1}[r_t|\Omega_{t-1}] &= 0, \\ \varphi_t^+ \tilde{r}_t^+ - \varphi_t^- \tilde{r}_t^- &= 0, \\ \varphi_t^+ F(1, \hat{\varepsilon}_t) + \varphi_t^- F(-1, \hat{\varepsilon}_t) &= 0. \end{aligned} \tag{19}$$

Making use of the relation $\hat{\varepsilon}_t = 2\varphi_t^+ - 1$ and $\varphi^- = 1 - \varphi^+$ and simplifying the notation by letting $x = \varphi_t^+$ gives the following functional equation for F:

$$xF(2(1-x)) + (1-x)F(-2x) = 0. \tag{20}$$

A nonzero solution is $F(x) = \nu_0 x$, where ν_0 is a constant. We thus see that even though we assumed an apparently general form for the impact, efficiency plus the additivity assumption forces F to be linear. When this is substituted into the second relation of Eqs. 8 we can solve for the volatility, giving

$$\nu = \nu_0[1 - (\pi\mathcal{P})^2] + \Delta\pi\mathcal{P}. \tag{21}$$

By using Eq.s (6) we write

$$\begin{aligned} r_t^+ &= \nu_0 - \lambda_t \\ r_t^- &= \nu_0 + \lambda_t. \end{aligned}$$

where $\lambda_t = \nu_0 \hat{\varepsilon}_t - \chi_t$. This means that efficiency plus the additivity assumption implies that the impact is *symmetric*. This says that the change in the expected return for trades of the same sign is equal but opposite to the change in the expected return for trades of the opposite sign. As written above λ can be any increasing function, so this condition seems quite general. However, this also turns out to imply linearity. Eliminating λ using the condition for observed market efficiency, Eq. 11, and by assuming one can neglect the term χ, one gets

$$\begin{aligned} r^+ &= \nu_0(1-\hat{\varepsilon}) \\ r^- &= \nu_0(1+\hat{\varepsilon}). \end{aligned} \tag{22}$$

Notice that the equation above also predicts that the observed expected returns should obey linear relations with slope ν_0.

One of the interesting aspects of Eq. 21 is that the volatility proxy ν varies with n unless $\mathcal{P}$ is constant. This implies that the volatility is coupled to the development of the hidden order. In particular, if $\mathcal{P}$ is an increasing function of n the volatility will decrease as the hidden order is executed. However, the decrease goes as $(\pi P(n))^2$, which we have already said is small, and in the limit as $n \to \infty$ the volatility tends to a constant. Thus, one of the main predictions of the additive assumption is that the volatility should be relatively independent of the state of hidden orders.

Substituting Eq. 21 into Eq.16 we obtain the simple form

$$E[R|N] = \eta_j \left[\nu_0 \sum_{n=1}^{N} (1 - \mathcal{P}(n)) + \theta \sum_{n=1}^{N} \Delta(n) \right]. \tag{23}$$

The first term on the right hand side in Eq.(23)) shows how only the unpredictable component of order flow contributes to market impact at any given time along the execution. The second term allows for inefficiency and for informational effects as the market incorporates the information that makes order flow predictable. In this paper we will not attempt to separate out the inefficiency from informational effects; instead we will calculate the theoretical impact curve assuming $\Delta = 0$; in Sec. VI we will review empirical evidence for this assumption

V. PARTICIPANT MODELS OF ORDER FLOW

The participant model of order flow is needed to understand how its predictability affects market impact through Eq. (23)). The dependence on the participant model of order flow is evident through the dependence on $\mathcal{P}(n)$, the perceived probability that a hidden order will continue. In this section we introduce two possible participant order flow models. We will loosely refer to them as corresponding to two different "information sets" Ω, with the understanding that they involve both different information and different models based on this information[8]. The two models we will consider are:

1. *Linear time series model based on order signs.* Participants observe the historical sequence of order flow signs, which are public information, and use standard linear time series models to predict future signs.

[8] As discussed later, the linear time series model does not necessarily make optimal use of the information on which it is based.

2. *Colored print model of hidden order execution.* Participants see "colored prints" as each hidden order is executed, with unique colors for each active hidden order, telling them which transactions originate from the same order. They know how many hidden orders are active and know how much each has executed so far. They do not know the identities of the hidden orders or their true size, but they do know the unconditional distribution from which their sizes are drawn.

As discussed in more detail in Section V C, we view these participant models as extremes that in a certain sense bracket the range of possibilities. We now examine the models in more detail. In each case the calculation of market impact is reduced to understanding the behavior of $\mathcal{P}(n)$.

A. Linear time series model

Linear time series models are probably the most widely used forecasting tool. Here we analyze a linear time series model based on the signs of executed transactions, which in most continuous double auction markets are public information. Under the assumptions of our model this indirectly reflects the presence and persistence of hidden orders. We will assume a T^{th} order autoregressive AR model of the form

$$\epsilon_t = \text{sign}\left[\sum_{i=1}^{T} a_i \epsilon_{t-i} + \zeta_t\right], \tag{24}$$

as proposed by Lillo and Farmer (2004). ζ_t is uncorrelated noise, $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = -1$ if $x < 0$, and a_i are real numbers that can be estimated on historical data using standard methods. Taking expectations and using Eq. 4, this can be written

$$\varphi_t^+ = \frac{1}{2}\left[\sum_{k=1}^{\infty} a_k E_{t-1}[\epsilon_{t-k}] + 1\right]. \tag{25}$$

As discussed in Section VI we assume that the predictability of order flow is entirely due to the presence of hidden orders. For convenience consider a buy order; the calculation for a sell order can be done by simply flipping all the signs. Assume hidden order j has been active for $n\theta$ transactions with constant participation rate π. As before, hidden orders are IID, so that all other hidden orders are equally likely to be buy or sell. The expected sign $E_{t-1}[\epsilon_t] = 1\pi + 0(1-\pi) = \pi$ for $k \leq n\theta$ and $E_{t-1}[\epsilon_t] = 0$ for $k > n\theta$. Under the assumption[9] that $T > n\theta$, substituting into Eq. 25 and using Eq. 13 gives

$$\varphi^+ = \frac{1}{2}(1 + \pi\mathcal{P}) = \frac{1}{2}\left[\sum_{k=1}^{n\theta} a_k \pi + 1\right], \tag{26}$$

which implies

$$\mathcal{P}(n) = \sum_{k=1}^{n\theta} a_k. \tag{27}$$

[9] For the stocks we study here there is a great deal of data, and because the signs are a long-memory process the coefficients a_k can be reasonably accurately estimated up to very large values of T, e.g. $T \sim 500$.

Using Eq. 23 the total impact is

$$E[R|N] = \eta_j \left[\nu_0 \sum_{n=1}^{N} \left(1 - \sum_{k=1}^{n\theta} a_k \right) + \theta \sum_{n=1}^{N} \Delta(n) \right]. \tag{28}$$

To compute the impact we need to understand the behavior of the autoregressive coefficients a_k. We will distinguish two cases: (1) The best linear time series model is AR(1), as assumed by Madhavan, Richards and Rooman (1997). (2) The best linear model is a long-memory FARIMA process. We believe the empirical evidence for case (2) is overwhelming, but we solve for both of these to illustrate how this assumption effects the answer.

1. AR(1) order flow

As a first case suppose that the order flow is modeled by a zero-mean AR(1) model

$$\epsilon_t = \psi \epsilon_{t-1} + \zeta_t \tag{29}$$

where $a_1 = \psi$ and $a_k = 0$ for $k > 1$. Equation 27 gives $\mathcal{P}(n) = \psi$, i.e. the continuation probability is independent of n. The autocorrelation function of an AR(1) process decays exponentially, $C(\tau) \sim \exp[-\tau/\tau_c]$, where the time scale is $\tau_c = -1/\ln\psi$. Note that this time series model is the one implicitly assumed in the MRR model.

By using Eq.(28) and assuming $\Delta = 0$, the impact is

$$E[R|N] = \eta_j \nu_0 \left[1 + \sum_{i=1}^{N-1} (1-\psi) \right] = \eta_j \nu_0 [N(1-\psi) + \psi]. \tag{30}$$

For large N it is $E[R|N] \sim N$, i.e. the impact is linear. After the completion of the hidden order the price reverts in one transaction to the permanent impact value which is $\eta_j \nu_0 N(1-\psi)$.

2. FARIMA order flow

As argued in Section VI there is good evidence that for large times the autocorrelation $C(\tau)$ of order flow asymptotically decays[10] as a power law $C(\tau) \sim \tau^{-\gamma}$ for large τ. This implies the process has long-memory (Beran, 1994). 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

[10] The notation "$\sim$" means "asymptotically equivalent", and is widely used in extreme value theory, e.g. (Embrechts et al., 1997). $f(x) \sim g(x)$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = C, \tag{31}$$

where C is a constant.

satisfy $a_k \approx \phi k^{-\phi-1}$ where $\phi = (1-\gamma)/2$. For large k we can pass into the continuum limit and estimate the sum in Eq. 27 by an integral. This gives

$$\mathcal{P}(n) = 1 - (n\theta)^{-\phi}. \tag{32}$$

Assuming $\Delta = 0$ and under the symmetric assumption, from Eq. 23 the impact is

$$E[R|N] = \eta_j \nu_0 \left[1 + \sum_{i=1}^{N-1} \left(1 - \left(1 - (n\theta)^{-\phi}\right)\right)\right]. \tag{33}$$

In order to convert the sum to an integral we use the approximation

$$\sum_{n=1}^{N-1} f(n) \simeq \int_0^{N-1} f(x+1/2)\ dx, \tag{34}$$

which gives

$$E[R|N] \approx \nu_0 \left(1 + \frac{2^{\phi-1}\theta^{-\phi}}{1-\phi}[(2N-1)^{1-\phi} - 1]\right) \sim N^{1-\phi}. \tag{35}$$

Thus the market impact asymptotically increases with the length of the hidden order as $N^{1-\phi}$. 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 $\phi \approx 0.25$. This means that according to the linear time series model the impact should increase as roughly the 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 $\theta^{-\phi}$. This means that the slower 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.

Another interesting property of the linear time series model is that the impact is completely temporary. As was shown by Bouchaud et al. (2004, 2006) it asymptotically decays to zero as a power law τ^{ϕ}, where $\phi = (1-\gamma)/2$. This has to be true in order to preserve efficiency and compensate for the long-memory persistence of the order flow under the linear model.

We have derived these results based on the properties of a linear time series model to predict order flow, as originally suggested by Lillo and Farmer (2004), but it is also possible to derive the same result based on the hypothesis of a power law decay of purely temporary impact, as originally suggested by Bouchaud et al. (2004). Either approach gives the same result.

B. Colored print model of hidden order executions

We now investigate the possibility that market participants have more detailed information about hidden orders, including what would normally be considered private information. We assume that they know:

1. The number of active hidden orders $M(t)$ and the participation rate π_j of each order.

2. The correct distribution of hidden orders sizes $P(V)$, or under our assumpton $V = \bar{v}N$ this is equivalent to knowing the hidden order lengths $P(N)$.

3. The number of realized orders n_j that have been executed so far[11] for each hidden order j, where $1 \leq n_j \leq N_j$. One can imagine that each active hidden order has its own unique color, and that participants can see the color of each transaction as it happens. (This is the origin of the term "colored print").

4. Once the n^{th} colored print is observed the probability of the hidden order continuing is updated based on the probability $P(N)$. This update is held until further notice, i.e. the probability of hidden order continuation, $\mathcal{P}(n)$, is maintained until the next colored print is received regardless of the number of intervening transactions. We will return to relax this assumption and take into account the uncertainty of detecting the end of a hidden order in Section V B 3.

Participants *do not know* the true size of any particular hidden order j, and thus they do not know when the execution of order j will end. They also do not know who originates each order, so there is no distinction between informed and uninformed trades.

1. *Stretched exponential distribution of hidden order sizes*

To illustrate how the distribution of hidden order sizes determines the shape of the impact function we first consider a stretched exponential distribution, of the form

$$P(V) = \frac{\beta}{\Gamma(1/\beta, 1)} e^{-V^\beta}, \tag{36}$$

where the normalization factor $\Gamma(a, z)$ is the incomplete Gamma function and $\beta > 0$ is a parameter specifying whether the distribution decays for large V faster ($\beta > 1$) or slower ($\beta < 1$) than an exponential. For convenience we have chosen units so that $\bar{v} = 1$, so that $V \geq 1$. As already emphasized, we do not believe this is the correct functional form, as this distribution leads to an autocorrelation function for realized order signs that decays as $C(\tau) \sim \exp(-\tau^\beta)\tau^{\frac{2}{\beta}-1}$. The dominant term is exponential, which decays too fast to be compatible with the observed long-memory of order flow. Nonetheless, analysis of the stretched exponential is useful because it gives insight into how the tail behavior of the hidden order distribution causes the concavity of the market impact function and provides a useful null hypothesis.

Under our assumption $V = \bar{v}N$ we can equally well use N. The cumulative distribution is

$$P(N_j > N) = \int_N^\infty P(N)\, dN = \frac{\Gamma\left(\frac{1}{\beta}, N^\beta\right)}{\Gamma\left(\frac{1}{\beta}, 1\right)}. \tag{37}$$

Once an order has already had n executions, the probability that it will continue is

$$\mathcal{P}(n) = \frac{P(N > n+1)}{P(N_j > n)} = \frac{\Gamma\left(\frac{1}{\beta}, (n+1)^\beta\right)}{\Gamma\left(\frac{1}{\beta}, n^\beta\right)}. \tag{38}$$

[11] Under our convention about measuring returns from just before one transaction until just before the next information about hidden order arrival is necessarily split across returns. I.e. when the n^{th} hidden order transaction is made the immediate price change is based on liquidity as set by providers based on the $(n-1)^{th}$ hidden order arrival, but subsequent quote-driven transactions are potentially aware of the n^{th} arrival. This is obviously a small effect.

By using the asymptotic expansion of the Gamma function we get

$$\mathcal{P}(n) \sim \exp\left[-(n+1)^{\beta} + n^{\beta}\right]\left(1+\frac{1}{n}\right)^{1-\beta}. \tag{39}$$

The function $\mathcal{P}(n)$ is increasing and converging to 1 for $\beta < 1$, decreasing and converging to 0 for $\beta > 1$, and constant with the value e^{-1} for $\beta = 1$. This shows how the predictability of order flow depends on the distribution $P(V)$ of hidden order sizes: If $P(V)$ has tails that are heavier than those of an exponential, then the longer a hidden order goes on, the more likely it is to continue. If its tails are thinner than those of an exponential, the longer it goes on, the less likely it is to continue. The boundary case is when $P(V)$ is exponential, in which case the likelihood for a hidden order to continue is independent of its size.

The impact can be written

$$E[R|N] = \eta_j \nu_0 \left(1 + \sum_{n=1}^{N-1}(1-\mathcal{P}(n))\right) = \eta_j \nu_0 \left(1 + \sum_{n=1}^{N-1} \frac{P(n)}{P(x>n)}\right). \tag{40}$$

By using the above expressions

$$\begin{aligned} E[R|N] &= \eta_j \nu_0 \left(1 + \beta \sum_{n=1}^{N-1} \frac{e^{-n^{\beta}}}{\Gamma(1/\beta, n^{\beta})}\right) \simeq \eta_j \nu_0 \left(1 + \beta \int_0^{N-1} \frac{e^{-(n+1/2)^{\beta}}}{\Gamma(1/\beta,(n+1/2)^{\beta})} dn\right) \\ &= \eta_j \nu_0 \left(1 + \log[\Gamma(1/\beta, \frac{1}{2^{\beta}})] - \log[\Gamma(1/\beta, \left(N-\frac{1}{2}\right)^{\beta})]\right) \sim N^{\beta}. \end{aligned} \tag{41}$$

The impact is concave for $\beta < 1$, convex for $\beta > 1$, and linear for $\beta = 1$. Thus we see that once again the exponential distribution is the boundary case: If the distribution of hidden order sizes is exponential the impact is linear. If the tails are thinner than those of an exponential the impact is convex, and if the tails are fatter than those of an exponential, the impact is concave.

2. Power law distribution of hidden order sizes

A more realistic assumption is that hidden order sizes are drawn from a distribution that is asymptotically a power law for large V, of the form $P(V_j > V) \sim V^{-\alpha}$. As discussed in Section VI we believe there is good empirical evidence supporting this for several stock markets.

We now compute the impact. For convenience we assume a "pure" power law with tail exponent α, i.e. one with probability density function $P(N) = KN^{-(1+\alpha)}$. Given that a hidden order has already had n realized transactions, the probability $\mathcal{P}(n)$ that it will continue is

$$\mathcal{P}(n) = \frac{\sum_{t=n+1}^{\infty} K t^{-(1+\alpha)}}{\sum_{t=n}^{\infty} K t^{-(1+\alpha)}} = \frac{\zeta(1+\alpha, n+1)}{\zeta(1+\alpha, n)} \tag{42}$$

where ζ is the Riemann Zeta function. Using Eq. 23 (the symmetry assumption) we obtain

$$E[R|N] = \eta_j \left(\nu_0 \sum_{n=1}^{N}(1 - \frac{\zeta(1+\alpha,n+1)}{\zeta(1+\alpha,n)}) + \theta \sum_{n=1}^{N} \Delta(n)\right). \tag{43}$$

In the large N limit this can be written much more simply. For large n we can approximate $\mathcal{P}(n)$ as

$$\mathcal{P}(n) = \frac{\zeta(1+\alpha, n+1)}{\zeta(1+\alpha, n)} \approx \left(\frac{n}{n+1}\right)^{\alpha}. \tag{44}$$

Assuming $\Delta = 0$ and neglecting small terms, from Eq. 23 the impact in the large N limit is

$$E[R|N] = \eta_j \nu_0 \left(1 + \sum_{n=1}^{N} \left(1 - \left(\frac{n}{n+1}\right)^{\alpha}\right)\right) \approx \eta_j \nu_0 \left(1 + \alpha \sum_{n=1}^{N} \frac{1}{n}\right). \tag{45}$$

By replacing the sum with an integral according to Eq. 34 we have

$$E[R|N] = \eta_j \nu_0 (1 + \alpha \log(2N-1)) \sim \log N. \tag{46}$$

The predicted impact thus asymptotically grows as a logarithm, i.e. it grows much slower than under the linear time series model. This solution also has the interesting property that the impact is independent of the participation rate, i.e. it is the same no matter how quickly the trade is executed. We will see that once we take the need for detecting the end of the order into account this changes, as does the functional form of the impact.

3. Detecting the end of hidden orders

So far we have left unspecified how participants detect the end of hidden orders. As a result we have so far been unable to compute the behavior of the impact once the order ends. To do this we need to model how $\mathcal{P}$, the continuation probability, is changed both before and after the order ends. As we will now show, the change is large enough to significantly modify the functional form of the impact while the order is still active. Unlike the linear time series model, we find a permanent component of the impact.

In assumption (4) of the colored print model at the beginning of this section we assumed that $\mathcal{P}(n)$ is held constant between colored prints. Ultimately this assumption must be violated when participants discover the hidden order is over. The time it takes to discover the end of the hidden order will determine how persistent the impact is. For example, consider a buy hidden order that persists for N transactions. After the N^{th} transaction the asymmetric price response is $r^+(N)/r^-(N) < 1$. If participants do not immediately discover that the hidden order has ended, then there will be a temporary period of inefficiency during which the asymmetric price response persists even though there is no imbalance in order flow. During this period, since $\varphi^+ = \varphi^- = 1/2$, according to Eq. 14 the expected market impact will revert at a rate $E_t[r_t] = \epsilon_j (r^+ - r^-)/2$ per transaction. When the participants finally discover the hidden order is over the asymmetric price response will disappear so that $r^+ = r^-$, and any remaining impact will be persistent. The problem with this line of reasoning is that we assume that participants are confident the order is present as long as it is still present, and that after it ceases to be present for a period of unspecified length they do not discover it is over, during an arbitrary period of inefficiency.

Instead we now make a self-consistent model in which the liquidity providers use Bayesian reasoning to detect the end of hidden orders. As before we assume that liquidity takers execute their hidden orders with market orders according to a Poisson process. Assume the liquidity providers know the participation rate π, which is the rate of the Poisson process. Let m be the number of transactions since the last colored print at step n and let S indicate that the hidden order is still active, and $\bar{S}$ indicate that it is inactive. The probability that the order is alive is

$$\mathcal{Q} = P(S|m) = \frac{P(m|S)P(S)}{P(m)} \tag{47}$$

The terms on the right are easily computed as follows:

$$\begin{aligned}
P(m|S) &= (1-\pi)^m \\
P(S) &= \mathcal{P}(n) \\
P(m) &= P(m|S)P(S) + P(m|\bar{S})P(\bar{S}) \\
&= (1-\pi)^m \mathcal{P}(n) + 1(1-\mathcal{P}(n)) = 1 - \mathcal{P}(n)\left[1-(1-\pi)^m\right].
\end{aligned}$$

Solving for $\mathcal{Q}$ gives

$$\mathcal{Q}(n,m) = \frac{(1-\pi)^m \mathcal{P}(n)}{1-\mathcal{P}(n)\left[1-(1-\pi)^m\right]}. \tag{48}$$

Note that for $m = 0$ this reduces to $\mathcal{Q}(n,0) = \mathcal{P}(n)$, reflecting the fact that it is known with certainty that the order is still active.

The market impact can be computed by substituting $\mathcal{Q}$ for $\mathcal{P}$ in Eq.(14) and setting $\Delta = 0$. Because of the increased uncertainty about the continuation of the order, $\mathcal{Q}(n,m) \leq \mathcal{P}(n)$, which increases the impact. This is illustrated in Figure 2. We have run 5×10^5 simulations simulations of the hidden orders of different length and we have averaged the price profiles over the simulations. During the hidden order we have sampled the price any time a colored print is submitted whereas in the reversion part we have sampled the price every (uncolored) transaction. In order to make the time scales comparable we have rescaled the time during the reversion part by multiplying the time by π. The resulting average market impact function grows (probably) faster than a logarithm, thus illustrating how uncertainty about the end of the hidden order can alter the asymptotic scaling. The reason for this is evident in the inset, where we have plotted φ_t^+. When the fluctuations of the Poisson order placement process cause long intervals between colored prints the Bayesian liquidity provider starts to infer that the order has ended, decreasing $\mathcal{Q}$ and making the liquidity less asymmetric. This increases the expected impact.

It is possible to find an analytical solution of this model. We assume the additive assumption, which in this case becomes

$$r^+ = 2(1-p^+) = 1 - \pi\mathcal{Q} \tag{49}$$

$$r^- = 2p^+ = 1 + \pi\mathcal{Q}. \tag{50}$$

For convenience we have set $\nu_0 = 1$.

Between the n-th and the $n+1$-th colored print and after having observed m uncolored prints, the ex-post expected return of an uncolored transaction is

$$E[r] = \frac{r^+ - r^-}{2} = 1 - 2p^+ = -\pi\mathcal{Q}(n,m) \tag{51}$$

After M uncolored prints the $n+1$-th colored print is placed and its ex post return is $r^+ = 1 - \pi\mathcal{Q}(n,M)$. The probability that between the n-th and the $n+1$-th colored prints there are M uncolored prints is

$$q_M = (1-\pi)^M \pi \tag{52}$$

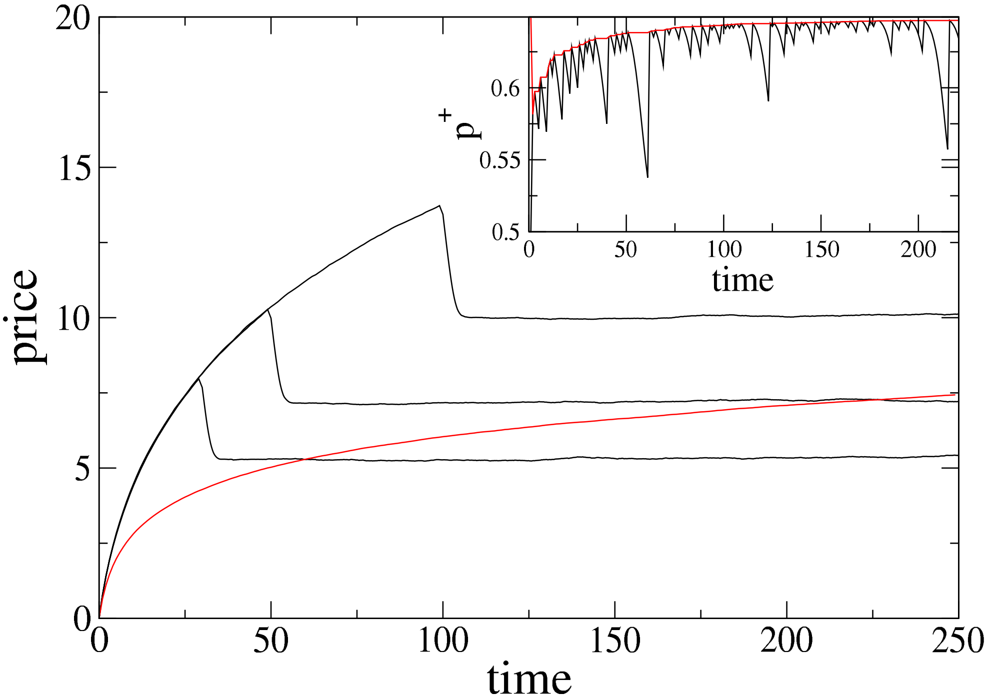

FIG. 2: Expected impact under the Poisson approximation under uncertainty about termination of the hidden order. The plot is in log-linear scale. The red line describes the simulations of the model with no updating of the continuation probability $\mathcal{P}$ between the n and the $n+1$ execution of the hidden order, whereas the blue line describes the continuous updating process $\mathcal{Q}$ with uncertainty about where the order stops. The black line is the exact result obtained from our theory (Eq. 45). The parameters are $\pi = 0.1$ and $\alpha = 1.5$. The simulated data are averaged over 5×10^5 simulations. The inset shows φ_t^+ for one specific realization. The red line refers to the model without updating and the black line refers to the model with continuous updating.

Therefore the ex post price impact between the n-th and the $n+1$-th colored print is on average

$$\rho_n = p_{n+1} - p_n = \sum_{M=0}^{\infty} q_M \left\{ -\pi \sum_{m=0}^{M-1} \mathcal{Q}(n,m) + 1 - \pi \mathcal{Q}(n,M) \right\} = \tag{53}$$

$$= 1 - \pi^2 \sum_{M=0}^{\infty} (1-\pi)^M \sum_{m=0}^{M} \frac{(1-\pi)^m \mathcal{P}(n)}{1 - \mathcal{P}(n)[1-(1-\pi)^m]} \tag{54}$$

Setting $a = 1 - \mathcal{P}(n)$ we get

$$\rho_n = a\pi^2 \sum_{M=0}^{\infty} (1-\pi)^M \sum_{m=0}^{M} \frac{1}{a + (1-a)(1-\pi)^m} \tag{55}$$

Unfortunately the sum cannot be performed analytically but only numerically. However we can get an approximate expression by converting the sums in integrals. We use the approximation that

$$\sum_{n=0}^{N} f(n) \simeq \int_0^{N+1} f(x-1/2)dx \tag{56}$$

and, after some calculations, we obtain

$$\rho_n \simeq -\pi^2 \frac{a \ln a + (1-a)\ln(1-a+a\sqrt{1-\pi})}{(1-a)\sqrt{1-\pi}(\ln(1-\pi))^2} \tag{57}$$

We are interested in the behavior for small a, i.e. values of $\mathcal{P}(n)$ close to 1. By expanding the above expression the leading term is

$$\rho_n \sim -\pi^2 \frac{a \ln a}{\sqrt{1-\pi}(\ln(1-\pi))^2} = \frac{\pi^2\alpha}{\sqrt{1-\pi}(\ln(1-\pi))^2} \frac{1}{n} \ln \frac{n}{\alpha} \sim \frac{\pi^2\alpha}{\sqrt{1-\pi}(\ln(1-\pi))^2} \frac{\ln n}{n} \tag{58}$$

where we have set $a = \alpha/n$. By integrating the leading term between 1 and N we get an approximated expression for the impact

$$E[R|N] \sim \frac{\alpha\pi^2}{2\sqrt{1-\pi}(\ln(1-\pi))^2}(\ln N)^2 \tag{59}$$

This shows that (i) the impact grows as the square of the logarithm of the hidden order size N and (ii) the impact is a decreasing function of the aggressiveness parameter π.

This example illustrates that in order to capture the true information set of market participants in the colored print model we need to take the detection of the end of a hidden order into account. This answer we have computed here depends on the Poisson process for realized order placement. It also requires taking the colored print model literally and assuming $\Delta = 0$, which we do not think are realistic assumptions Nonetheless, it illustrates how this aspect of the problem influences the size of the impact, and most importantly, how it affects the permanence of the impact.

C. Discussion of order flow models

The linear time series model and the colored print model represent computable extremes that in some sense bracket the level of information likely to be available in order flow models. Information about order flow varies from market to market[12], but in most modern financial markets signed order flow is public information. This is not, however, the only information available, and even if it is, linear time series models are not an optimal choice. In this section we review some of the possibilities and discuss their implications for order flow prediction.

Linear time series models serve as a lower bound for the predictability of order flow. Such models are easily to implement, even in real time, and one can expect that if there is anything to be gained from their use, they will be used. However, for order flow that

[12] In London, Paris, Spain, and many other electronic stock markets it is possible to observe order flow directly in real time. For the NYSE or NASDAQ one is only able to observe transactions and best quote changes in real time, but signed order flow can be inferred (with some mistakes) using the Lee and Ready algorithm (1991).

follows the model of Lillo, Mike, and Farmer (LMF, 2005), which forms the basis of our theory, linear time series models are not an optimal method of prediction. This is true even if the only information available is a time series of order flow signs. The reason for this is because linear time series models for long-memory processes average over a long window into the past, whereas the LMF model is strongly state dependent: After a hidden order stops the past behavior of that order ceases to have any predictive value. This is most apparent if there is only one hidden order active at a time, in which case it is often obvious when one hidden order stops and another starts. As a result the best model is not a linear time series model, but rather a regime switching or structural break model. This is similar to the situation for modeling long-memory for volatiity[13]. Preliminary results suggest that prediction methods that incorporate state dependence, such as hidden Markov models, can be more effective. We are currently investigating such models and hope to present some comparisons in the future.

In some cases participants can make use of 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. Such possibilities make it difficult to know the predictability of order flow *a priori*. When all such effects are taken into account, what information do participants actually have?

While it is implausible that the colored print model is strictly correct, we believe that an approximation of it may be plausible. This might involve a superior algorithm for exploiting publicly available order flow as well as the usage of other clues about order flow identity. Such an order flow would not be capable of discovering the initiation of hidden orders immediately, but it still might be effective enough for sufficiently long hidden orders to validate the asymptotic predictions of the colored print model.

D. Information revelation

Perhaps the most surprising aspect of these results concerns information revelation. The colored print model has more information revelation than the linear time series model yet asymptotically it has lower impact, since asymptotically a power law will always overtake a logarithm. This remains true even in the more accurate case where we require participants to detect the end of hidden orders.

This suggests that there are situations when it is to a liquidity taker's advantage to reveal information. If true this would be very surprising. It goes against the prevailing wisdom of market participants, who work hard to keep their intentions as secret as possible. What is not quite as clear is whether this is reasonable from a game-theoretic point of view: Under what situations might it be advantageous to provide opponents with information? In a two-play competitive zero-sum game, providing an opponent with additional information cannot hurt their performance, and can only improve it. The situation is not as clear in more general games.

[13] For example see the discussion in Poon and Granger (2003).

VI. EMPIRICAL RESULTS

In this section we present empirical results to test the theory that we have developed. These results are reported for 6 stocks. As we will show, assuming there is indeed a consistent functional form for market impact, this quantity of data we study here is not sufficient to determine it. Nonetheless, the data are sufficient to demonstrate that several aspects of our theory are correct, while calling some of the assumptions into question. The approach we present here provides a blue print for future studies with larger data sets.

A. Our data

We study six stocks traded on the London Stock Exchange AZN (AstraZeneca), BSY (British Sky Broadcasting Group), LLOY (Lloyds TSB Group), PRU (Prudential Plc), RTO (Rentokil Initial), and VOD (Vodafone Group). The choice of these six stocks is somewhat arbitrary, and is largely determined by the fact that we have carefully cleaned these data and believe that we have a reliable record of almost every order placement (see discussion below). AZN, LLOY, and VOD are among the most heavily traded names; see Table I for the trading volume for each stock. The data is from the on-book exchange (SETS) only, and is for the period from May 2, 2000 to December 31, 2002. There are time-stamp issues with orders traded off-book, and therefore it is difficult to determine the impact of these orders. Also, many off-book trades are eventually offloaded in the on-book exchange – if we included off-book data, many trades would be analyzed in duplicate. For these reasons, we have chosen to use only on-book data. The dataset contains a complete record of all order placements, so we are able to determine the signs of orders unambiguously. In the figures we consistently use AZN to illustrate our results, and present results from other stocks or make comments in the text in the few cases that the results are significantly different from those obtained with AZN.

These data have anonymized codes attached to each trading order indicating the LSE member firm through which it was submitted[14]. The number of member firms of the LSE is of order 100, but they vary in terms of total activity; a substantial portion of the trading volume for a stock can be focused within only a fraction of the firms. Membership in the exchange does not identify the individual trading accounts, and in most cases members are acting as brokerages, handling the trades for an unknown number of clients. Since we do not know who these are we will refer to such clients as “agents”.

While the data are generally quite reliable, the time stamps associated with orders are only accurate to the second, and the correct sequencing of orders posted in the same second is not guaranteed. This sometimes leads to inconsistencies such as execution against orders that do not yet exist when reconstructing the order book. Within any second intervals that create problems we have resequenced the data to avoid such inconsistencies. While this resequencing is not always unique, it is at least plausible. The consequent time rearrangements are small and we do not think this affects the results presented here.

---

[14] In the original data set these codes were randomly shuffled every month, but Tom O’Brien of the LSE has graciously provided us with a key that allowed us to unscramble the codes.

B. What causes the persistence of order flow?

One of the assumptions we have made throughout the development of our theory is that the persistence of order flow is caused by the incremental execution of hidden orders. In this section we review the empirical evidence that supports this hypothesis and present some new evidence. We also argue that in the markets where this has been studied the evidence indicates that the distribution of large orders is a power law and the resulting order flow has long-memory.

The evidence can be briefly summarized as follows:

1. The distribution of large orders in block markets is distributed as a power law $P(V > x) \sim x^{-\alpha}$, with $\alpha \approx 1.5$.

2. Order flow in order book markets shows long memory, with a decaying autocorrelation $C(\tau) \sim \tau^{-\gamma}$, with $\gamma \approx 0.5$.

3. A theory based on (1) and the incremental time and size-independent execution of hidden orders predicts $\alpha = 1 + \gamma$.

4. New evidence that we present here shows that transactions made under the same membership code show autocorrelations consistent with long-memroy, whereas transactions made under different order codes do not.

We now discuss this evidence in more detail.

A power law tail with tail exponent $\alpha \approx 1.5$ for the hidden order size distribution was originally observed for the NYSE by Gopikrishnan et al. (2000). Similar behavior was observed for the Paris and London stockmarkets by Gabaix et al. (2006), and under aggregation across many stocks the power law behavior becomes quite crisp, i.e. for large volumes the data fit a power law closely over several orders of magnitude. With these data, however, it is impossible to distinguish block trades and order book trades. Lillo, Mike and Farmer (2005) used data from the LSE in which it is possible to separate block trades from order book trades and study their distributions separately. An example is given in Figure 3. Using a Hill estimator on the largest one percent of the block trades gives a tail exponent $\alpha \approx 1.59$. In contrast, for the order book trades the tail exponent is $\alpha \approx 2.9$, showing that the order book trades have a much thinner tail. Since it is well known that incremental execution is widespread in order book markets while it is discouraged in block markets, the block market is assumed to be a better proxy for the distribution of intended trade sizes. When the data sets are mixed together, as they were in the studies of Gabaix et al., the much heavier tail of the block trades will dominate the tail of the aggregate distribution, justifying the assumption that the tail represents block trades. These estimates are fairly consistent from stock to stock.

It has been seen in several different studies that the first autocorrelation of order flow is positive[15]. Bouchaud et al. (2004) and Lillo and Farmer (2004) studied the tails of the autocorrelation function at large lags and observed long-memory for order flow in the Paris, London and NYSE stock exchanges. This means that the autocorrelation function $C(\tau)$ of the signs ϵ_t of transactions decays in time as $C(\tau) \sim \tau^{-\gamma}$, where $0 < \gamma < 1$. The autocorrelation function of a long-memory process is not integrable in the limit $\tau \to \infty$, and such

[15] See for example Biais (1995), Danielsson (2001), Ellul et al. (2005).

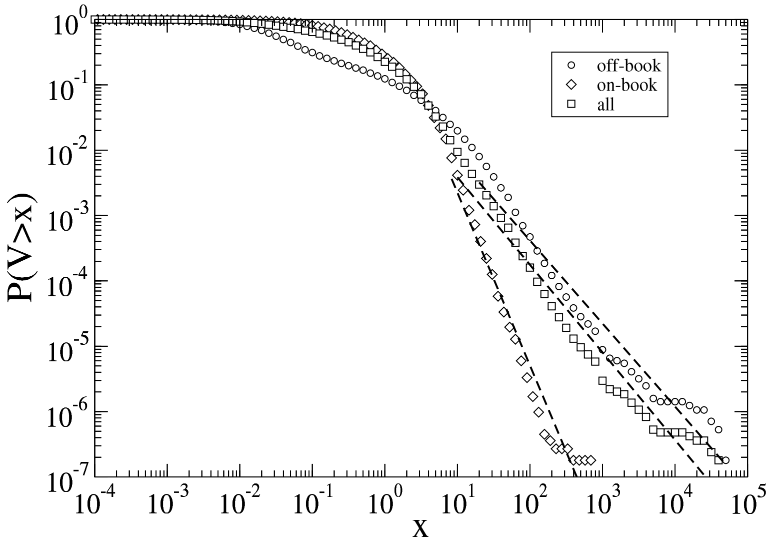

FIG. 3: Volume distributions of block trades (circles), order book trades (diamonds) and the aggregate of both (squares), from Lillo, Mike and Farmer (2005). This is for a collection of 20 different stocks, normalizing the volume of each by the mean volume before aggregating the data. The dashed black lines have the slope found by the Hill estimator (and are shown for the largest one percent of the data).

processes do not have a characteristic time scale, i.e. the integral of the autocorrelation function, $\int C(\tau)d\tau$, does not exist. Long-memory processes can be characterized by the exponent γ of the autocorrelation function or equivalently in terms of the Hurst exponent $H = 1-\gamma/2$. The observation of long-memory in stock markets is very robust; in London, for example, all of the twenty stocks examined showed long-memory at highly statistically significant levels under strict statistical tests (Lillo and Farmer, 2004). Whether long-memory also exists in other types of markets (FX, interest rates, commodities, etc.) is not known.

Under the assumptions of a power law distribution of hidden orders and that all active hidden orders are executed at the same rate, independent of their size or the number of previous executions, Lillo Mike and Farmer showed that in the limit $\tau \to \infty$ this leads to an autocorrelation function for order flow that decays as a power law $C(\tau) \sim \tau^{-\gamma}$. This prediction is testable by comparing 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[16]. For comparison[17] the

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

[17] 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 between γ and $\alpha - 1$ is probably fortuitious. The error analysis in the presence of long-memory is not trivial and we intend to refine it in the future.

average measured values of γ for these stocks was $\gamma = 0.57$, close to the predicted value $\hat{\gamma} = \alpha - 1 = 0.59$.

Further supporting evidence comes from a study of the Spanish stock exchange by Vaglica et al. (2007) who have reconstructed hidden orders using data with brokerage codes. They confirm directly that $V \sim N$, i.e. that uniform execution rate is a good assumption. Similarly, they find that N is distributed as a power law for large N with $\alpha \approx 1.5$. Using their methods, in an as yet unpublished study, this has been confirmed as well for the LSE.

We present additional direct evidence here for the hypothesis that the dominant cause of long-memory in revealed order flow is order splitting. This analysis takes advantage of the fact that we have the exchange membership code associated with each order book trade (see the discussion of the data in Section VI A). In Figure 4 we compare the autocorrelation function for trades with the same membership code to those with different membership codes for the stock AZN; the results for other stocks in our data set are essentially the same. The autocorrelation functions of the signs of trading orders from the same membership

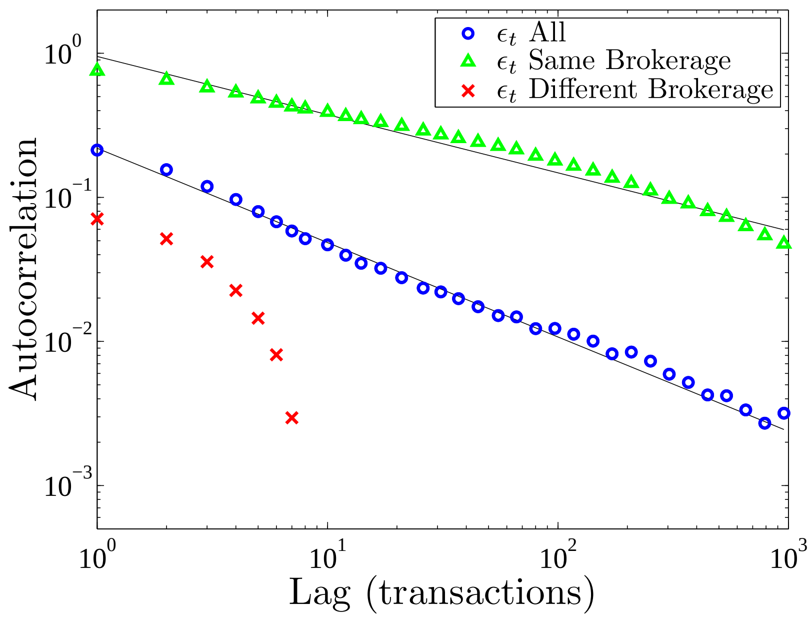

FIG. 4: Autocorrelation of trade signs as a function of transaction sequence for transactions with same brokerage code (green triangles), different brokerage code (red x), and all transactions irrespective of brokerage code (blue circles), plotted on double logarithmic scale. To suppress statistical fluctuations autocorrelation coefficients with nearby lags are averaged.

code and all membership codes both decay roughly as a power law, as indicated by their approximation to a straight line on log-log scale. More rigorous statistical tests based on the methods used in Lillo and Farmer (2004) confirm this. In contrast, the autocorrelation function for realized order signs from different brokerages decays rapidly, and is clearly not a power law. By lag 10 there are already negative values; because we are using a logarithmic scale we cannot even plot them. In contrast, for the data based on the same brokerage codes, all autocorrelations are positive out to lags of 1000.

To summarize, there are three empirical results supporting the hypothesis that order splitting is the primary cause of long-memory in trading signs. These are (1) observations of trading volume distributions consistent with power laws in block markets with exponent $\alpha \approx 1.5$; (2) agreement of observed long-memory of order-book transactions with the predicted relation $\alpha = \gamma + 1$; and (3) disappearance of long-memory for orders with different brokerage codes. While there are likely to be other factors that contribute to the predictability of order flow, the dominant cause appears to be order splitting of hidden orders whose volume is drawn from a distribution that in the limit $V \to \infty$ is a power law tail.

C. Detection of hidden orders

To test our theoretical predictions about market impact we need to identify individual hidden orders. Doing this accurately requires information about trading accounts. We do not have such information, but we do have anonymized membership codes making it possible to identify which orders are made by the same member of the exchange. This provides enough information to separate most hidden orders, particularly large ones. Although our method of doing this is deficient in several respects it is good enough to be very useful. An algorithm for identifying hidden orders with some advantages over ours was introduced by Vaglica et al. (2007); we will comment later on the pros and cons of the two approaches.

Our algorithm is very simple. We assume that all the transactions of a given hidden order are made by the same member of the exchange, that they are of the same sign, and that they are within $\theta_{max} = 100$ transactions of each other[18]. The algorithm proceeds by applying the criteria to the first 100 orders, labeling all orders with no matches as hidden orders of length one, and lumping together other transactions as being part of the same hidden order according to the criteria above. We then examine each new transaction successively, either adding it to any pre-existing hidden order (including those of length one) if it satisfies the criteria above or designating it as a new hidden order of length one. Under this procedure it is possible to have multiple hidden orders active at the same time. It is also possible that for the same membership code two hidden orders of opposite signs can be active at the same time.

We need to make several caveats, some of which are serious: (1) Our algorithm assumes that all agents submit their orders through a single member of the exchange. This is certainly not strictly true – agents are known to split their orders across several brokerages. Nonetheless, the fact that orders submitted through different brokerages do not have long-memory, as shown in Figure 4, indicates that agents submit most of their orders through the same member of the exchange. (2) Many different agents trade through the same member of the exchange, and more than one of them may be actively trading with that member at the same time. Thus the algorithm is in some cases lumping together two or more hidden orders. See our analysis below, which provides estimates of the error rate. (3) It is common to use a mixture of market and limit orders to execute a hidden order, but we are studying only hidden orders made up purely of market orders. This sometimes results in erroneous splitting of larger hidden orders. (4) The algorithm imposes an upper bound θ_{max} which will artificially split hidden orders if any two executions are separated by more than θ_{max} transactions. We analyze this effect quantitatively in the next section and argue that it works well for intermediate size hidden orders but splits large orders much too often. (5)

[18] The value $\theta_{max} = 100$ was chosen by varying θ_{max} and choosing the value that minimized the autocorrelation of the hidden order sequence, as described in Section VI E.

We are only considering the on-book market; the same agent may also trade in the off-book market. During this period roughly fifty percent of order volume is executed in the off-book market. The off-book market is not anonymous, and brokers strongly discourage order splitting, which suggests that off-book market trades are less relevant than on-book trades for understanding the behavior of hidden orders.

All of these effects introduce problems into our analysis that the reader should bear in mind. Problem (4) is particularly serious, and introduces problems that we will comment on later. In a future study we intend to do a careful comparison of several different methods for hidden order classification, including the method of Vaglica et al. (2007) and a method based on hidden Markov models.

D. Estimate of the error rate of the algorithm

As discussed above our reconstruction makes errors both in falsely merging hidden orders from different agents using the same member of the exchange to execute their order, and in falsely splitting large hidden orders. In this section we make estimates of both of these effects.

To get an estimate of false merging we compute the fraction of the time that each member has at least one active hidden order. For example, according to the results of our algorithm, for AZN the two most active members have active hidden orders in process about 20% of the time. Thus, under the hypothesis that each individual hidden order is executed according to a Poisson process, there is a 4% chance that two orders executed by one of these members will overlap. Relaxing the Poisson process assumption will lead to higher overlap rates, so this estimate is optimistic, but it suggests that the order of magnitude of the overlap effect is not prohibitively large.

Similarly, under the Poisson process assumption, since $(1 - \pi)$ is the probability that a hidden order with participation rate π does not have an execution on any given transaction, the probability that a very long hidden order will go for θ_{max} transactions without being falsely split by our algorithm is $p_s = (1 - \pi)^{\theta_{max}}$. The typical length at which splitting becomes likely is therefore $L_{max} = 1/p_s = 1/(1 - \pi)^{\theta_{max}}$. For AZN, for example, $\pi \approx 0.05$ implies that with $\theta_{max} = 100$, $L_{max} \approx 170$. We thus expect the tail of the distribution of hidden orders to be artificially truncated by our reconstruction algorithm at roughly this length.

E. Consistency checks

Several consistency checks give insight into the performance of our algorithm. One is the autocorrelation function of the signs of hidden orders. If the assumption that the signs of hidden orders are IID is correct and our reconstruction method is sufficiently accurate, then we should recover uncorrelated hidden order signs. We compute the autocorrelation function of reconstructed hidden order signs, putting them in sequence based on the time when each hidden order begins. For $\tau > 0$ the coefficients of the autocorrelation function are all close to zero, but with a slight negative bias. This can be seen for the stock AZN in Figure 5(a). The cumulative of the autocorrelation function is also plotted and appears to level out at around a lag of 400 transactions. We report the average of the first 100 autocorrelation coefficients for each stock in Table I. As seen in the table, the coefficients are negative but very close to zero. This should be compared with the strong positive autocorrelations observed for realized orders in Figure 4. This indicates that the assumption of IID hidden order signs is reasonable and that our algorithm is not making large systematic errors in signing and sequencing hidden orders.

	AZN	BSY	LLOY	PRU	RTO	VOD
Transactions $\times 10^5$	5.7	3.6	6.0	3.9	2.1	10
Avg Autocorr $\times 10^{-3}$	-2.0	-1.6	-.92	-2.1	-1.5	-.98
α	1.7	1.9	1.9	1.8	2.0	1.7
γ	.6	.6	.6	.6	.5	.7
Avg M	14	14	15	14	12	17
Avg θ	25	24	27	24	22	30
r^+ Slope Linear	-.7	-.62	-.96	-.42	-1.1	-1.1
r^- Slope Linear	1.3	1.3	.96	1.4	.61	.80
r^+ Slope Colored Print	-.4	-.05	-.37	-.44	-.54	-.64
r^- Slope Colored Print	.9	.67	.43	.42	1.3	.02

TABLE I: Summary Statistics. Rows from top to bottom, (1) number of transactions in the sample, (2) average autocorrelation for first 100 lags of hidden order size ϵN, (3) tail exponent, α, of the hidden order size distribution as measured using the Hill estimator, (4) power law exponent, γ, of the autocorrelation function for signed transactions, (5) average number of hidden orders active at one time, M, (6) average number of transactions between pieces of a hidden order θ, (7) slope of r^+/ν_0 when measured as a function of the linear model sign predictor $\hat{\epsilon}^{(ts)}(\tau = 10)$, (8) slope of r^-/ν_0 when measured as a function of the linear model sign predictor $\hat{\epsilon}^{(ts)}(\tau = 10)$, (9) slope of r^+/ν_0 when measured as a function of the colored print model sign predictor $\hat{\epsilon}^{(cp)}(\tau = 10)$, (10) slope of r^-/ν_0 when measured as a function of the colored print model sign predictor $\hat{\epsilon}^{(cp)}(\tau = 10)$

Even if the true sequence of hidden order signs is indeed IID, the problems of falsely merging or splitting hidden orders will affect the autocorrelation of hidden order signs observed in the reconstruction. Falsely merging orders induces a negative autocorrelation. Assuming an IID sequence, this is because nearby orders of the same sign are compressed into a single order, thus causing a tendency for the sequence of signs to alternate. Similarly, splitting orders induces a positive autocorrelation, due to the fact that a single order is replicated to become two or more nearby orders of the same sign. The net autocorrelation of the reconstructed hidden order sequence is thus a combination of these two effects, added to whatever autocorrelation might exist in the true sequence of hidden orders. The observed negative autocorrelation suggests that either the assumption of IID hidden order signs is slightly wrong or that there is a net tendency to falsely merge orders, or some combination of both. Nonetheless, the fact that the observed autocorrelation of order signs is not large indicates that overall these problems are not large.

Our primary goal here is to measure market impact, and the possibility that our algorithm artificially splits or merges orders potentially distorts the impact. Since the impact is a concave function of order size, artificially splitting a large order will tend to assign improperly small impacts to large orders and similarly merging orders tends to assign improperly large impacts to large orders. The above diagnostic suggests that this would not be a problem if this were size independent. However, as we will demonstrate later, large orders are much more likely to be split than small orders, which creates problems in measuring impact.

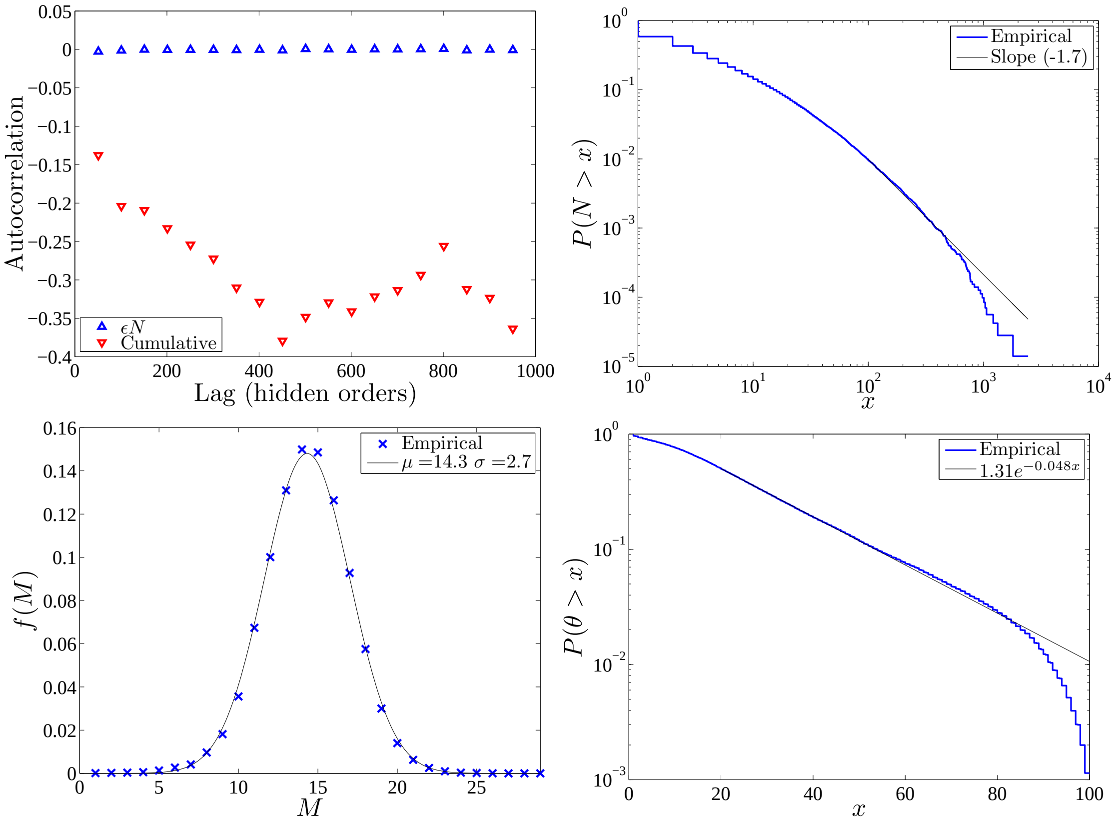

FIG. 5: Four diagnostics for our hidden order reconstruction algorithm applied to the stock AZN. (a) The autocorrelation function of signed hidden order size (ϵN) (blue up-triangles) and its cumulative (red down triangles). (b) $P(N > x)$, the probability that the length of a reconstructed hidden order is greater than a given value x, plotted on double logarithmic scale. For comparison a line of slope 1.7 is shown, corresponding to the asymptotic power law predicted from the relation $\alpha = \gamma + 1$ based on measurements of H from the realized order flow. (c) is a histogram of the number $M(t)$ of hidden orders that are active at the same time, compared to a normal distribution. (d) is the cumulative probability distribution $P(\tau > x)$ for the interval τ between successive executions of hidden orders, plotted on semi-logarithmic scale and compared to an exponential distribution.

We have shown that the distribution of hidden order volumes is an important determinant of the asymptotic market impact, and argued that the cumulative distribution of hidden order lengths is asymptotically a power law for large N with exponent $\alpha = \gamma + 1$. We can check whether our reconstruction is consistent with this hypothesis based on the empirical histogram of hidden order lengths, as shown in Figure 5(b). This figure is plotted in double logarithmic scale so that a power law appears as a straight line. The tail exponent α of the distribution is calculated using a Hill estimator[19], and the resulting estimate, $\alpha = 1.7$, is drawn as a straight line in the plot. In Table I we present the estimated values of α

[19] The Hill estimator is calculated as $\hat{\chi} = 1 + n / \sum_{i=1}^{n} \log(x_i / x_{min})$, where x_i represents empirical data and n is the number of observations $x_i \geq x_{min}$. n is set such that only the largest 1% of the data is included in the estimate.

for each stock and we show values of γ measured by computing the Hurst exponent of signs for the realized order flow[20]. As seen in the table, the agreement with the relation $\alpha = \gamma + 1$ is variable, but this relation is never wildly wrong. The estimates for α based on the reconstruction tend to be higher than $\gamma + 1$. In performing this reconstruction we have the problem that because this is an asymptotic relationship, for small values of N there is no reason to expect a power law, and for $N > 170$ we have predicted that the algorithm will truncate the distribution. This leaves very little dynamic range to test our original assumption. Thus it is impossible to strongly confirm this relation using the methods illustrated Figure 5(b), but in view of the problems mentioned above they are also not inconsistent with it. Note that the longest hidden orders in the reconstruction have lengths as long as $N = 2000$, corresponding to a timescale of roughly $N\theta = 40,000$ transactions, or roughly 40 trading days. The algorithm of Vaglica et al. (2007), in contrast, shows results consistent with a power law even for large sizes, recovers more long hidden orders, and for the Spanish stock market produces more consistent results.

Figure 5(c) is a normalized histogram of $M(t)$, the number of hidden orders that are active at a time. This is very well fit by a normal distribution with mean $E[M] = 14.3$ and standard deviation 2.7. Thus on average there are about 14 active hidden orders for AZN (the average number of orders for the other stocks is reported in Table I). For comparison, Lillo, Mike and Farmer estimated this by comparing the empirical autocorrelation function of realized orders to that predicted by their model, giving an estimate $E[M] \approx 21$. The reconstructed distribution and the mean value estimated from Figure 5(c) is sensitive to θ_{max}, but it is reassuring that these estimates are of the same order of magnitude.

Finally, Figure 5(d) shows the cumulative probability distribution $P(\tau > x)$ for the interval τ between successive executions of hidden orders for AZN, plotted on semi-logarithmic scale and compared to an exponential distribution. As expected the distribution is truncated for values of τ close to $\theta_{max} = 100$, but from roughly $\tau = 10$ to $\tau = 80$ there is reasonable agreement with an exponential. This is consistent with the Poisson hypothesis, and suggests that the probability of artificially splitting an order is about 1%, consistent with our previous estimates. The mean value of θ for each stock is reported in Table I.

F. Inefficiency and informational effects

In this section we present empirical tests of the assumption that $\Delta = 0$ under the order flow models of Section V. Recall that from Eq. 7 the observational inefficiency includes both real inefficiencies of the model (which might allow arbitrage) as well as the result of the direct informational effects χ_t (which will not allow arbitrage if all players have the same access to χ), so this is really a test of both inefficiencies and direct information. We look at how well the linear time series and colored print models predict order flow and explicitly demonstrate asymmetric price response.

To implement the linear time series model we use an autoregressive predictor of the form

$$\hat{\epsilon}_t^{(ts)} = \sum_{i=1}^{K} a_i \epsilon_{t-i}, \tag{60}$$

[20] The exact relation is $\gamma = 2(1-H)$. This procedure is more accurate than estimating the exponent directly from the autocorrelation function (see Lillo and Farmer, 2004).

and we set $a_i = \phi i^{-\phi-1}$ (as it would be for a FARIMA process). This relationship agrees well with fitting an autoregressive model to the order flow series with least squares. We use $K = 10,000$ in the results that follow.

For comparison we also predict order flow based on the colored print model. We first classify all the hidden orders using the algorithm described in Section VI C. A prediction algorithm for the probability of the next sign is then constructed by using Eqs. 4, 13 and 44 and averaging over all hidden orders that are active at time t. Let $A_{tj} = 1$ if order j is active at time t and $A_{jt} = 0$ otherwise, and let n_{jt} be the number of previous incremental transactions order j has experienced at time t. The expected sign imbalance $\hat{\epsilon}_t^{(cp)}$ is

$$\hat{\epsilon}_t^{(cp)} = \sum_j A_{tj}\eta_j\pi_j\mathcal{P}(n_{jt}) \approx \sum_j A_{tj}\eta_j\pi_j\left(\frac{n_{jt}}{n_{jt}+1}\right)^{\alpha}. \tag{61}$$

Both $\hat{\epsilon}_t^{(ts)}$ and $\hat{\epsilon}_t^{(cp)}$ can be used to predict the signs at the next time, and they can also be used to predict at any future time $t+\tau$, where τ is any positive integer. A natural performance measure for the quality of the predictions is the fraction of times the future sign matches the predicted sign at time $t+\tau$, i.e.

$$\hat{\varphi}^+(\tau) = P[\epsilon_{t+\tau} = \mathrm{sign}(\hat{\epsilon}_t)].$$

The probability for predicting the opposite sign is $\hat{\varphi}^-_{t+\tau} = 1 - \hat{\varphi}^+_{t+\tau}$. The "hat" superscript denotes that the prediction is associated with a particular time series model $\hat{\epsilon}_t$; while $\hat{\varphi}$ is related to φ_t^+ as previously defined in Eq. 3 the two quantities are not exactly the same. We use $\hat{\varphi}$ because it provides a convenient way to compare the predictive power of the two models.

Similarly the expected returns $\hat{r}^+_{t+\tau}$ and $\hat{r}^-_{t+\tau}$ associated with the prediction $\hat{\epsilon}_t$ are

$$\begin{aligned}\hat{r}^+(\tau) &= E_t[r_{t+\tau}\epsilon_t|\epsilon_{t+\tau} = \mathrm{sign}(\hat{\epsilon}_t)]\\ \hat{r}^-(\tau) &= -E_t[r_{t+\tau}\epsilon_t|\epsilon_{t+\tau} \neq \mathrm{sign}(\hat{\epsilon}_t)].\end{aligned}$$

A comparison of the predictive power of the two time series models is shown in Figure 6(a), where $\hat{\varphi}^+(\tau)$ is plotted as a function of τ for both the linear and colored print models. Not surprisingly, for all but the shortest values of τ the colored print model makes more accurate predictions[21]. The prediction accuracy is good for short values of τ, slowly decaying to zero as τ increases.

To allow us to see the asymptotic scaling of these results more clearly we compare the empirically measured predictability to the return asymmetry. For the market to be observationally efficient at time horizon τ it must satisfy $\hat{\varphi}^+(\tau)\hat{r}^+(\tau) - \hat{\varphi}^-(\tau)\hat{r}^-(\tau) = 0$, i.e. $\hat{\varphi}^+(\tau)/\hat{\varphi}^-(\tau) = \hat{r}^-(\tau)/\hat{r}^+(\tau)$. We can get an idea for how quickly Δ goes to zero by comparing the two ratios $\hat{\varphi}^+(\tau)/\hat{\varphi}^-(\tau)$ and $\hat{r}^-(\tau)/\hat{r}^+(\tau)$. Since in the limit as $\tau \to \infty$ both ratios approach 1, to use logarithmic scale it is convenient to subtract one, so that in the limit $\tau \to \infty$ the resulting quantity goes to zero and power law scaling (if it exists) will appear as a straight line). This is done in Fig. 6(b). The market is observationally inefficient with

[21] For $\tau < 3$ the linear time series model is actually more accurate than the colored print model. This is not evident in Figure 6(a) due to the fact that we are averaging together different values of τ, but it is apparent in Figure 6(b) where for small τ there is no averaging.

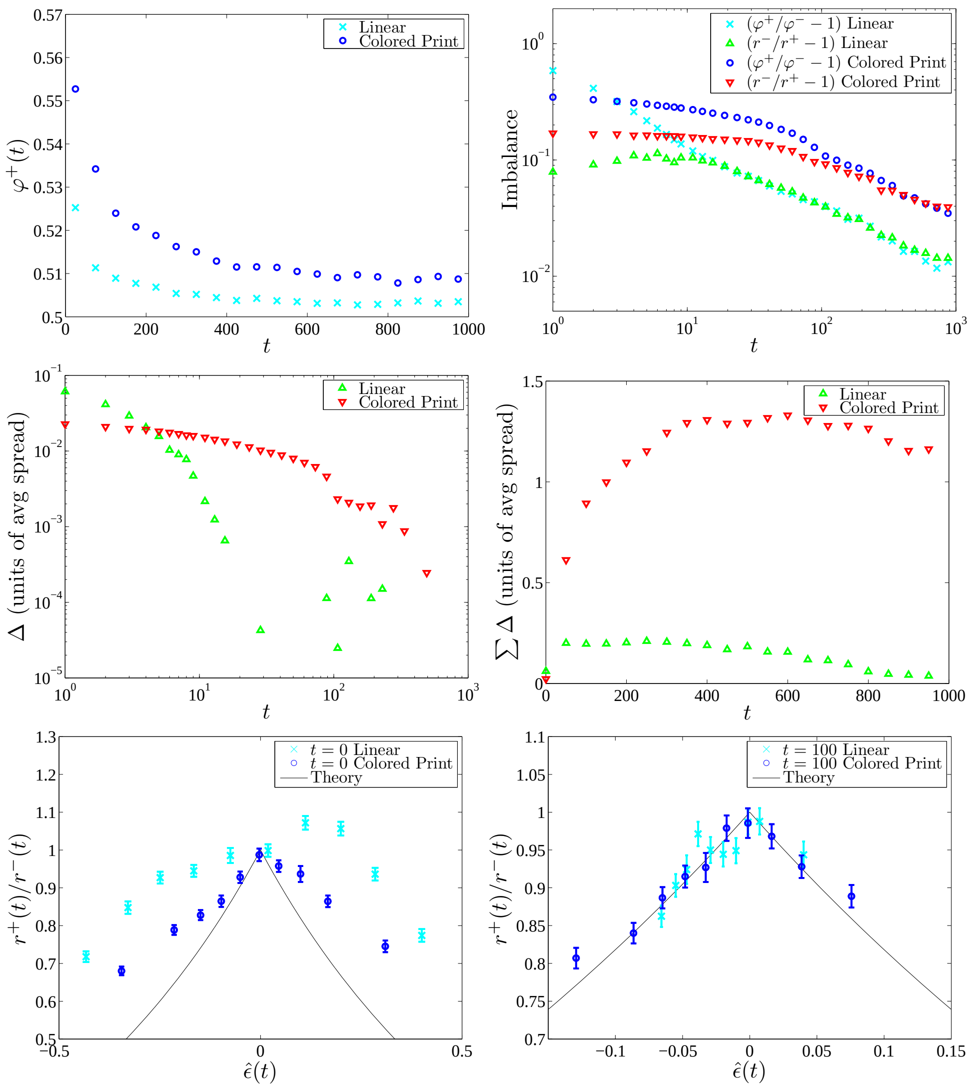

FIG. 6: Tests of efficiency and asymmetric liquidity for the linear times model and the colored print model. In each case the results showed are binned values plotting the average of the values in a given range of the argument on the horizontal axis. (a) The accuracy of the order flow prediction $\hat{\varphi}^+(\tau)$ as a function of the forecast horizon τ measured in number of transactions; (b) The quantity $\hat{\varphi}^+(\tau)/\hat{\varphi}^-(\tau) - 1$ is compared to $\hat{r}^-(\tau)/\hat{r}^+(\tau) - 1$, plotted on double logarithmic scale; to the extent that these don't agree the market is inefficient. (c) The inefficiency $\hat{\Delta}(\tau)$. (d) The cumulative inefficiency $\sum_{i=1}^{\tau} \hat{\Delta}(i)$ measured in units of the average bid-ask spread. (e) The ratio $\hat{r}^+/\hat{r}^-$ for the linear and colored print models as a function of the predicted sign $\hat{\epsilon}(\tau)$, compared to the predicted relationship under observed efficiency for $\tau = 1$ and $\tau = 100$. (f) Same as (e) for $\tau = 100$.

respect to the model $\hat{\epsilon}$ if $\hat{\varphi}^+(\tau)/\hat{\varphi}^-(\tau) - 1 \neq \hat{r}^-(\tau)/\hat{r}^+(\tau) - 1$. For the linear time series model we see that initially $\hat{\varphi}^+(\tau)/\hat{\varphi}^-(\tau) \gg \hat{r}^-(\tau)/\hat{r}^+(\tau)$, but by roughly $\tau = 20$ the two are essentially the same, and they remain that way for $\tau > 20$. In contrast, for the colored print model the initial inefficiency is not as large, but the convergence to observational efficiency is much slower, and is not complete until roughly $\tau = 400$.

In Figure 6(c) we plot $\hat{\Delta}(\tau) = \hat{\varphi}^+(\tau)\hat{r}^+(\tau) - \hat{\varphi}^-(\tau)\hat{r}^-(\tau)$. $\hat{\Delta}(\tau)$ decays quickly to zero for the linear model and much more slowly for the detailed model. To assess whether either of these inefficiencies are large enough to permit arbitrage, in Figure 6(d) we plot the cumulative observed inefficiency $\sum_{i=1}^{\tau} \hat{\Delta}(i)$ in units of the average bid-ask spread. The cumulative observed inefficiency for the linear model peaks at about 0.2. In contrast the inefficiency of the colored print model eventually exceeds the average spread, steadily growing until about $\tau = 400$, where it peaks at about 1.3 spread units. This varies quite a bit from stock to stock. For three of the stocks (LLOY, RTO and VOD) the cumulative inefficiency is always less than the spread, whereas for BSY, a lightly traded stock, at its maximum it rises to 2.8 in units of the average spread.

To measure the observable inefficiency in terms that are directly comparable to the impact, in Figure 7 we plot the cumulative Δ as a function of n, measuring it in units of the predicted impact (which also varies with n). For the linear time series model the cumulative

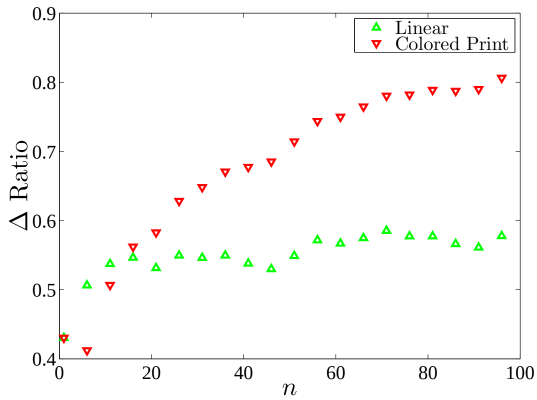

FIG. 7: The cumulative observable inefficiency $\sum_n \Delta(n)$ vs. n, the number of transactions, for both the linear time series model and the colored print model, for the stock AZN. Δ is measured as a fraction of the predicted impact under the assumption $\Delta = 0$, which also increases with n; note that it increases faster for the linear time series model than the colored print model.

observable inefficiency grows to roughly half the predicted impact and levels off, whereas for the colored print model it grows to almost 85% of the total. This suggests that the observable inefficiency is not negligible in either case, but it is particularly large for the colored print model. This is in part because for large n the predicted impact grows more slowly with the colored print model than for the linear time series model.

To explicitly test our assertion that efficiency in the face of predictable order flow is achieved through asymmetric price response, in Figure 6(e) we test this for the linear model by plotting the ratio $\hat{r}^{+}(\tau)/\hat{r}^{-}(\tau)$ as a function of the predicted sign $\hat{\epsilon}(\tau)$. We do the same for the detailed model in Figure 6(f). This is compared to the predicted relationship derived in Eq. 10. In both cases we observe that for small values of τ the ratio r^{+}/r^{-} does not satisfy the prediction, but by $\tau = 100$ the observed data are reasonably close to the prediction. This explicitly demonstrates that efficiency is achieved by an asymmetric price response, and for sufficiently large times satisfies the predicted efficiency condition.

G. Tests of the additive assumption

We now provide empirical tests of the additive assumption, Eq. 18, which states that price changes caused by new information in order flow only depend on the difference between the observed order sign and its prediction. As we showed in Section IV, the observable market efficiency condition implies that the normalized expected returns are linear and symmetric function of the prediction $\hat{\varepsilon}$ with slopes ± 1, i.e. $r^+/\nu_0 = 1 - \hat{\varepsilon}$ and $r^-/\nu_0 = 1 + \hat{\varepsilon}$. It is not hard to show that the time series predictions should show that when $\hat{\Delta} = 0$ the expected returns under the time series models show similar relations, i.e. $\hat{r}^+/\nu_0 = 1 - \hat{\epsilon}$ and $\hat{r}^-/\nu_0 = 1 + \hat{\epsilon}$. To test this in Figure 8 we plot $\hat{r}^+(\tau)/\nu_0$ and $\hat{r}^-(\tau)/\nu_0$ versus $\hat{\epsilon}^{(ts)}$ and $\hat{\epsilon}^{(cp)}$ for AZN.

Testing this relationship is difficult for two reasons. First, it is difficult to distinguish linear from nonlinear relationships for values close to the origin and this region is precisely the region where most of the empirical data is located. Second, as shown in Figure 6(a), returns are not immediately efficient, but become so after a certain length of time. This means the response of $\hat{r}^+$ and $\hat{r}^-$ to $\hat{\epsilon}^{(ts)}$ and $\hat{\epsilon}^{(cp)}$ as shown in Figure 8 is not sufficient to make market returns efficient – only when using lagged values of $\hat{\epsilon}^{(ts)}$ and $\hat{\epsilon}^{(cp)}$ is the market efficient. Using longer lags for these variables gives inconclusive results, however, as the the data become very noisy as τ gets large. We make a compromise and choose $\tau = 10$ for both order flow models. When we do this we observe linearity, consistent with our prediction (and indeed this holds true to a good approximation for all values of τ where there is enough data to measure it. However, the prediction that the slopes should be ± 1 is not met very precisely, as is evident in Table I. The agreement is better for the linear time series model, where linearity is a good assumption and the slopes are within 30% of 1, than for the colored print model, where linearity is also fairly well satisfied, but the slopes differ by a factor of two for r^+ and r^-.

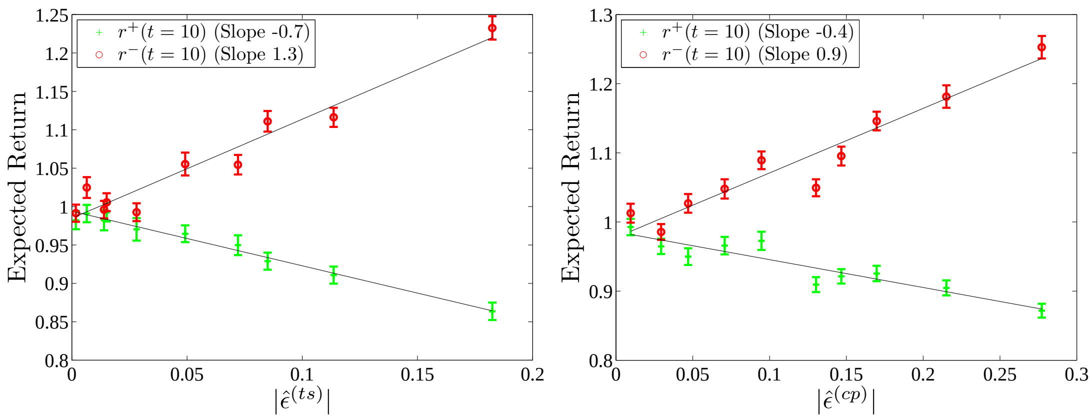

FIG. 8: Plot of r^+/ν_0 and r^-/ν_0 as a function of the predictor $\hat{\epsilon}^{(ts)}$ from the linear model and the predictor $\hat{\epsilon}^{(cp)}$ from the colored print model for the stock AZN, using $\tau = 10$. The prediction is that these should be lines with slopes ± 1.

The other prediction we made based on the additive assumption is that the volatility proxy $\nu(n)$ should obey Eq. 21. According to this prediction the volatility should decrease slightly and approach a constant. The size of the decrease is small, of order $(\pi \mathcal{P})^2$, which is typically less than 0.01. For comparison in Figure 9 we plot the average volatility proxy

$\nu(n)$ as a function of the number of executions n. We do this for each stock. As seen in the figure, in every case $\nu(n)$ is roughly constant, consistent with the prediction.

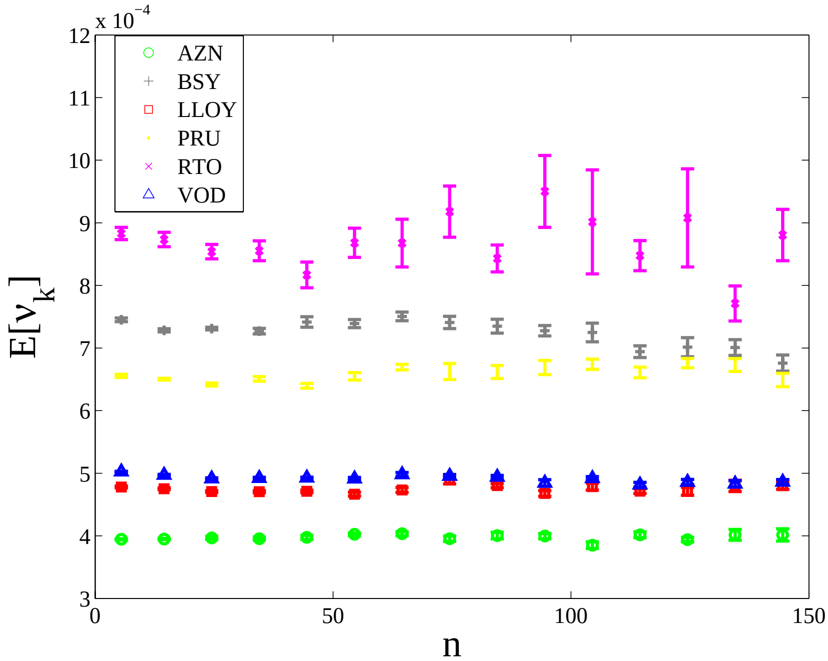

FIG. 9: ν as a function of n for all stocks in our sample.

H. Trading velocity

Both the linear time series model and the colored print model predict that there should be a dependence on the participation rate π, or in other words the impact depends on trading velocity. This means that the total impact depends on the rate of trading. In Figure 10 we plot the total impact as a function of participation rate for hidden orders in three different size ranges. In each case we see a similar result: Higher participation rate implies greater impact. This is in agreement with the linear time series model, and is opposite to what is predicted by the colored print model.

I. Returns of colored prints vs. other transactions

A clue about whether participants are able to distinguish colored prints from other transactions is to simply compare the returns from colored prints (the transactions associated with the same hidden order) to other transactions. This is shown as a function of n for the stocks AZN and VOD in Figure 11. The result is surprising. The returns r^+ from the hidden order decrease (as we would expect, but the returns r^+ for the other orders are essentially flat. This suggests that market participants are able to distinguish transactions coming from the same hidden order, and that the response for these orders is quite different than that for other orders. An alternate hypothesis is that there is something else different about these transactions, e.g. that they are of higher volume, but this is not the case – the average order

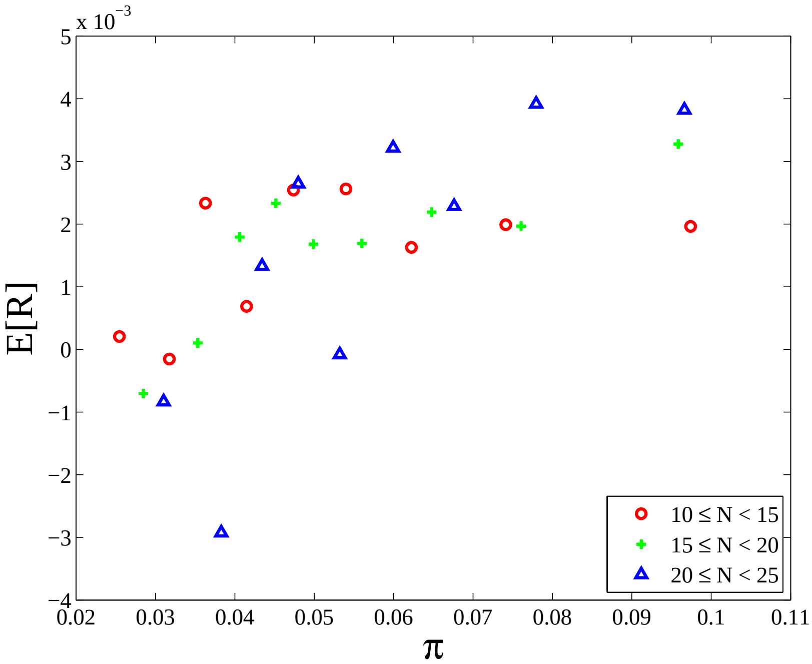

FIG. 10: The impact for AZN is plotted against the participation rate for hidden orders in three different size ranges. In each case the impact increases with participation rate.

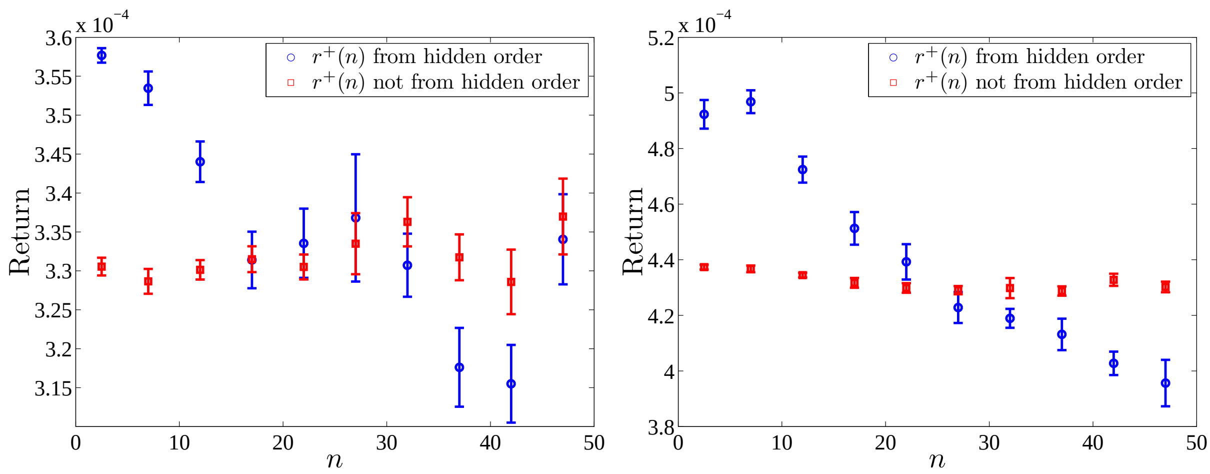

FIG. 11: The average returns for orders of the same sign as the hidden order are compared to those from other orders as a function of n, the number of transactions so far for that order. The data are binned and standard error bars are shown.

size is essentially the same, and changes very little with n. We have shown VOD because the effect is particularly striking, with a 25% decrease in the average return from $n = 1$ to $n = 50$.

J. Dependence on time since last colored print

One of the predictions of the colored print model is that the return asymmetry should be a decreasing function of the number of intervening trades m since the last colored print. We test this in Figure 12 by plotting r^+ and r^- as a function of m. There is indeed indeed

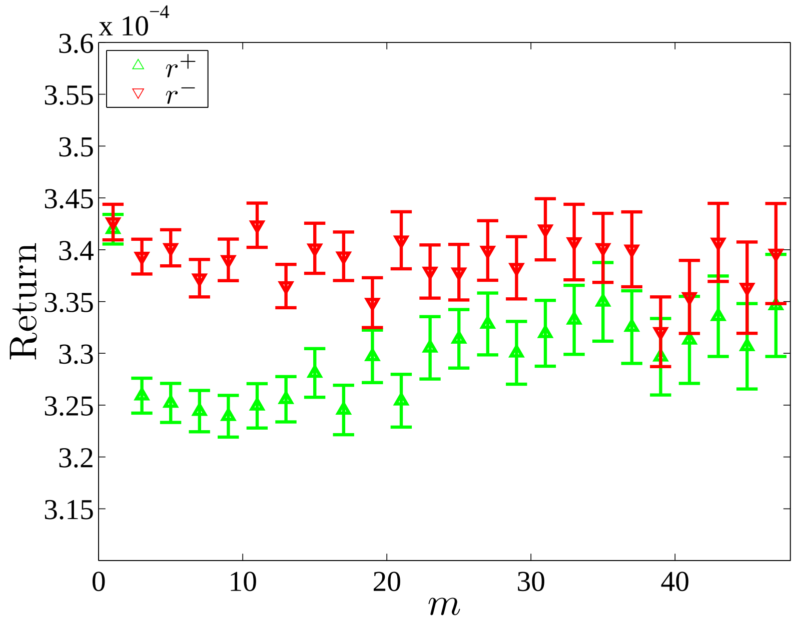

FIG. 12: The expected returns r^+ and r^- for AZN are plotted against m, the time since the last colored print.

a tendency for the return asymmetry to decrease. It is worth noting, though, that $m = 1$ and $m = 2$, which are not shown because they would not be on the same scale as the rest of the figure, the return asymmetry is actually opposite to what is predicted, i.e. $r^+ > r^-$. We believe this is because the information about a colored print takes time to be incorporated into the returns.

K. Tests of market impact

Our purpose in this paper was to derive equations for market impact. Unfortunately, as we show now, we are not able to make a definitive test at this point in time. There are two reasons for this. The first is that we do not have enough data, as we show in a moment. The second is that the excessive splitting of large orders by our reconstruction algorithm distorts the impacts of large orders. Nonetheless, we do have enough data to at least verify the concave behavior of the impact and to show that the predictions of our theory are not rejected by the data.

1. How much data is needed?

Based on the scalings that we have hypothesized and empirically observed it is possible to estimate the quantity of data that is needed to reasonably determine the functional form of the impact. Assume we have Y hidden orders, so that the number of hidden orders with length N is $k(N) = p(N)Y$, where $p(N)$ is the probability that an order has length N. For large N we have presented evidence that $p(N)$ scales as $p(N) \sim (1/\alpha)N^{-(\alpha+1)}$. Assume that the size of the impacts is a long-memory process with Hurst exponent H and single transaction volatility ν_0. Then under standard results for long-memory processes (Beran, 1994) the absolute error $E(N)/\nu_0$ for orders of length N scales as

$$\frac{E(N)}{\nu_0} \approx \frac{\sigma(N)}{\nu_0 k(N)^{1-H}} \approx \frac{N^{1/2}\alpha^{(1-H)}}{N^{-(1+\alpha)(1-H)}Y^{(1-H)}} = \left(\frac{\alpha}{Y}\right)^{(1-H)} N^{1/2+(1+\alpha)(1-H)} \tag{62}$$

For example, with $H = 0.75$ the absolute error increases as $N^{1.125}$. Even worse, the absolute error only decreases with the number of data points as $Y^{-0.25}$. So even with a data set containing a million hidden orders, the absolute error for a hidden order of length 100 is the order of one. Thus it takes a great deal of data to be able to accurately distinguish the functional form of the average market impact.

2. Empirical tests of market impact

The calculation above suggests that with the limited quantity of data we have here we are unlikely to be able to get a clear result concerning the functional form of the market impact. It is nonetheless interesting to at least make the attempt. For purposes of the theory we will assume that the distribution of order size volume scales as a power law with exponent α As reported in Section V A, under this assumption the linear time series model predicts that market impact should scale with hidden order size N as a power law with exponent $1 - \phi$, where $\phi = (2 - \alpha)/2$. In contrast, as reported in Section V B 3, the detailed model of hidden order executions predicts that market impact should scale with hidden order size N as $(\log N)^2$. To facilitate combining the data we divide each data point by the predicted scale factors given in Eqs. 35 and 45. In Figure 13 we plot the scaled impact of all hidden orders as a function of order size N and compare to the predictions. Note that when we do this we are not fitting any free parameters – all the parameters are determined based on independent measurements. From the result it appears that the linear model provides a worse fit. However, it should be born in mind that the error bars are quite large – the standard errors shown in the figure underestimate the errors. (A better error analysis will be presented in a later version of the paper).

VII. CONCLUSIONS

A. Summary

We have developed a theory for market impact that is based on the assumption of efficiency under predictable order flow. To fully describe the predictability of the order flow it is necessary to describe its intrinsic predictability, and to specify the information and models of participants. We have investigated several different options and shown how they affect the impact. We provide a summary of the assumptions and discuss some of the main conclusions below:

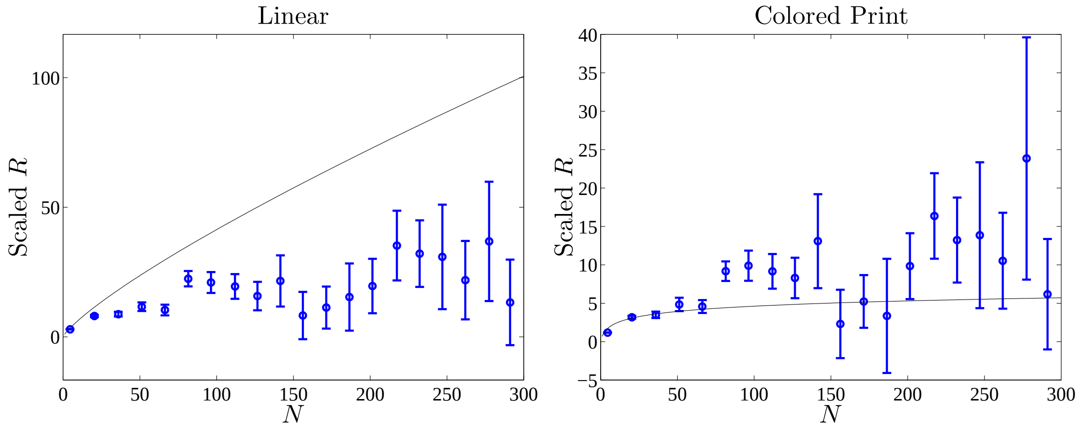

FIG. 13: (a) Scaled hidden order return, $(R\theta^{\phi}(1-\phi))/(\epsilon\nu_0)$, as a function of size, N, averaged over all stocks in our sample. This is compared to the predicted relation from the linear time series model. (b) Scaled hidden order return, $R/(\alpha\epsilon\nu_0)$, as a function of size, N, averaged over all stocks in our sample. This is compared to the predicted relation from the colored print model.

- *Origin of the predictability of order flow.* For the development of the theory we have assumed that order splitting of large hidden orders is the primary cause of the predictability of order flow. As discussed in Section VI, based on studies in several different markets, we think the empirical evidence for this assumption is strong.

- *Hidden order signs.* We assume the signs of hidden orders are IID. For the stocks in our sample the empirical support is good (see Figure 5 and Table I).

- *Volume distribution of hidden orders.* The empirical evidence that hidden order sizes are drawn from a power law distribution is good (see Section VI). We also investigate the alternative hypothesis of a stretched exponential distribution, which is not well supported by the data, and show that this makes a substantial difference in the results. Thus the heavy tails of the volume distribution are a key factor in determining the shape of the impact.

- *Uniform order splitting.* We assume that hidden orders are split into pieces of roughly the same size, independent of the position n within the hidden order. The evidence for this in the Spanish and London stock markets is good[22] Vaglica (2007). We also assume that large orders are executed at the same rate as small orders. There is a small observed empirical deviation from this (in Spain and London large hidden orders are executed more slowly than small orders). This deviation is not large, however, and we think that uniform order splitting is still a good assumption.

- *Participant order flow model.* We have presented two possible models, a linear time series model of order flow signs and a colored print model, in which participants are able to link the identities of transactions from hidden orders. At this stage it is not clear which of the two models is more appropriate; see the discussion in the next section.

[22] The evidence for London is based on work that is not yet published.

- *Observed inefficiency.* A nonzero contribution to the observed inefficiency Δ can come either from the information term χ or from a failure of participants to fully exploit arbitrage opportunities – we are unable here to distinguish these two causes. For the linear time series model, for lags of more than 20 transactions the observed inefficiency is reasonably small (see Section VI F). For the colored print model, in contrast, the observed inefficiency is much larger, and in several cases becomes larger than the spread, as shown in Fig. 6. The lags to achieve efficiency are much longer, the order of several hundred transactions, and in relative terms the cumulative observable inefficiency continues to grow with time, so that by $n = 100$ it is a significant fraction of the predicted impact, as seen in Figure 7.

- *Additivity assumption* The additivity assumption states that the price response is proportional to the new information in order flow, measured as the difference between the observed order sign and the expected order sign. This is fairly well supported for the linear time series model and somewhat less well supported for the colored print model. At this stage we cannot say with certainty whether the deviations are within statistical error; better data analysis is needed. However, based on standard errors it seems that linearity is well-supported, but the prediction from Eq. 22 that the slopes should be plus one and minus one are not well supported. This may be in part because this depends on the volatility proxy ν_0. For the colored print model, however, it seems that that the slope for r^+ is significantly different from the slope for r^-, so the additive assumption is not well satisfied.

- *Asymptotically constant volatility.* One of the predictions of the additivity assumption is that the volatility is asymptotically constant, i.e. that as a function of n the overall scale of the price responses should initially decrease and then approach a constant value. This has good empirical support (see Figure 9). Note, though, that the additivity assumption is not the only hypothesis that leads to asymptotically constant volatility.

- *Asymmetric liquidity.* In order for the strong temporal correlation of order flow to be compatible with observable efficiency, our model predicts the liquidity must be asymmetric, in the sense that transactions with the same sign as the hidden order produce a smaller price response than orders of the opposite sign, with this effect becoming stronger as the order progresses. The empirical support for this is very good, and is clearly observed for every stock in the sample (see Fig. 6(f)).

- *Velocity dependent impact.* Both participant order flow models predict a velocity dependent term in the impact function. The linear time series model predicts that the impact increases with trading velocity, whereas the colored print model predicts that it decreases with trading velocity. The former prediction is clearly more reasonable, in the sense that slower execution is less desirable than rapid execution, and one would naturally expect to pay a penalty for rapid execution. This is strongly born out in the data; see Figure 10.

- *Reversion of market impact.* The framework we have introduced makes it possible to predict the reversion of market impact when a trade is finished. The linear time series model predicts complete reversion, in line with the theory of Bouchaud et al. (2004). The colored print model predicts that the impact is partially temporary and partially permanent. We have not addressed this question empirically yet (there are some subtleties in the analysis and we did not feel our preliminary results were sufficiently reliable to present). This is an important question that needs to be more carefully resolved in the future.

To give insight into the effect of each assumption, we have made an effort whenever possible to analyze alternatives, even when we think they are not well supported by the data. Table II presents a summary of the predictions under different possibilities. All of these assume uniform order splitting and the additivity of information.

Volume distribution	*Order flow model*	*End detection*	*Predicted impact*	*Permanence*
Stretched exponential	Linear T.S.	N.A.	N	yes
Power law	Linear T.S. ($K = \infty$)	N.A.	$N^{1-\phi}$	no
Power law	Linear T.S. ($K < \infty$)	N.A.	N	yes
Stretched exponential	Colored print	Given	N^{β}	yes
Power law	Colored print	Given	$\log N$	yes
Power law	Colored print	Bayesian, Poisson	$(\log N)^2$	yes

TABLE II: Predicted impact under different combinations of the assumptions made in this paper as described in the text. *Volume distribution* refers to the underlying distribution of hidden order volumes; *Order flow model* refers to the method that participants use for predicting order flow; *End detection* refers to the method they use for determining when a hidden order has finished; *Predicted impact* is the scaling of the impact with the length N of the order for large N, and *Permanence* refers to whether or not the impact has a permanent component. *Linear T. S.* refers to the linear time series model, and K is the order of the model. All of the cases assume uniform order splitting and the additive assumption, and that the observable inefficiency is going to zero for large n. Note that the power law distribution for volume is empirically well-supported and the stretched exponential is not.

B. Discussion

In this paper we have attempted to lay out a framework for understanding the impact of large orders. The method we have used to do this can be compared to that of Black and Scholes for pricing options. That is, the two key elements of their method are an assumption about the random process followed by the underlying price and then the imposition of the requirement that there be no arbitrage. Similarly, we have made an assumption about the underlying distribution of volumes of hidden orders and then assumed that there should be no arbitrage between liquidity providers and liquidity takers. As summarized in the previous section, we have also had to make some other assumptions, such as uniform order splitting, additivity of new information, and an assumption about the participant model of order flow. Ultimately we think that it should be possible to derive most of these additional assumptions on more fundamental grounds. For example, the volume distribution of hidden orders is likely influenced by factors such as the distribution of fund sizes and the need for fund managers to minimize transactions costs, and can potentially be computed in an equilibrium model. Similarly, the assumption of uniform order splitting can potentially be derived as a rational choice. Of course, this needs to be done self-consistently with the model of market impact, and is not likely to be a trivial exercise.

In developing our theory we have attempted as much as possible to formulate everything in terms of quantities that are empirically measurable. For example, the distribution of trade volumes can be inferred from the correlations in order flow and by comparing block and order book markets, and thus we feel that this is a better basis for a theory that assumptions that depend on the functional form of utility, which are more or less arbitrary and difficult to verify empirically. This has allowed us to jump over several steps and to identify the elements of the theory that seem to be well-supported and those that are not, and as described in the previous paragraph, to indicate future topics of research that can further reduce the number of assumptions.

The most difficult component of the theory to observe is the participant model of order flow. We have made two extreme assumptions. At one end of the spectrum we assume the use of a linear time series model, and at the other we assume the colored print model. The colored print model is rational, but as we discuss later, the linear time series model is not. Under the colored print model participants are able to see which transactions come from the same hidden order, and to know how many previous transactions each hidden order has had. They must still infer when hidden orders end, which we have shown has several important consequences. We suspect that the truth lies somewhere in between. While many participants are probably very good at anonymizing their order flow, others are clearly not. We have examined the order flow on an anecdotal basis, and have found participants who leave clear marks for their presence, such as consistent use of limit orders of the same size and order placement at regular time intervals. Thus the order flow model may be diverse.

One of our surprising results is that even if we assume that the only available information is the time series of order signs, and even neglecting statistical estimation errors in setting parameters, the use of an infinite order time series model is not an optimal model, i.e. it is not rational in the usual sense. This is because the random process for order placement postulated by LMF (Lillo, Mike and Farmer, 2005) has nonlinear structure. When a long hidden order stops the optimal strategy is to reset the memory of the model, since the history of that order is no longer relevant. We have shown that in certain situations it is inherently better to reduce the order of the model. This has important consequences for both the functional form of the impact and in whether the impact is permanent or temporary (and is the reason for distinguishing the cases $K = \infty$ and $K < \infty$ in Table II). At this stage we cannot derive an optimal model for predicting order flow under the LMF framework.

One of the problems with the framework we have developed here is the need to assume that the direct information terms χ_t are zero. At this stage we have no way to measure these terms except in combination with the efficiency. The observable inefficiency Δ measures the combination of real inefficiencies and information terms. We can measure this combination but we cannot measure each of the components separately. We are working on methods of breaking these two terms apart.

In presenting our empirical analysis we have been able to gain considerable understanding of the factors influencing impact, but we still do not have the final story. There are two main problems. The first is that, as we have shown in Section VI K 1, because market impact becomes increasingly noisy for large N, the data requirements for an accurate measurement are considerable. The second is that the method we are using for detecting hidden orders is not optimal and has a tendency to unduly split large orders. We plan to redo all the analysis using a less biased method, but since developing such an algorithm is a significant project unto itself we have deferred this to a future paper[23]. It is also possible to use data in which there is detailed information at the level of trading accounts. Such data exists [Chan and Lakonishok, (1995), Odean et al. (2004), thus all the elements of this theory are empirically testable even without relying on participant identification algorithms.

In conclusion, we have introduced several new ideas about the causes of market impact. Once each of the elements of the theory are understood it should be possible to predict the functional form of market impact, including its dependence on trading velocity, and to quantitatively understand the extent to which impact is temporary or permanent. This is significant in providing a new approach to computing market impact that integrates the informational view of price formation with the more classic supply and demand approach.

Acknowledgments

We would like to thank J-P. Bouchaud for helpful comments. FL acknowledges support from the research project MIUR 449/97 "High frequency dynamics in financial markets" and from the European Union STREP project n. 012911 "Human behavior through dynamics of complex social networks: an interdisciplinary approach.". This work was supported by Barclays Bank and from NSF grant HSD-0624351. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or Pipeline Financial Group, Inc.

[1] R.F. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, 10(1):1–18, 2003.

[2] R.F. Almgren and N. Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3 (2):5–39, 2000.

[3] R.F. Almgren and N. Chriss. Value under liquidation. *Risk*, 12(12):61–63, 1999.

[4] Brad M. Barber, Yi-Tsung Lee, Yu-Jane Liu, and T. Odean. Do individual day traders make money? evidence from taiwan. Technical report, U.C. Davis, 2004.

[5] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall, New York, 1994.

[6] B. Biais, P. Hillion, and C. Spatt. An empirical analysis of the limit order book and order flow in the paris bourse. *The Journal of Finance*, 50(5):1655–1689, 1995.

[7] J-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and response in financial markets: The subtle nature of "random" price changes. *Quantitative Finance*, 4(2):176–190, 2004.

[8] J-P. Bouchaud, J. Kockelkoren, and M. Potters. Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance*, 6(2):115–123, 2006.

[23] We intend to develop an algorithm along the lines of that used by Vaglica et al. (2007). This algorithm has the disadvantage that it only uses transactions. This can result in a biased view of impact, since for limit orders impact is felt largely by selective execution due to adverse information effects. We intend to extend this algorithm to make use of limit orders as well as transactions, which would make it possible for us to use all the data in an unbiased manner.

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

[10] L. K.C. Chan and J. Lakonishok. The behavior of stock prices around institutional trades. *The Journal of Finance*, 50(4):1147–1174, 1995.

[11] Dmitriy Cherkashin, J. Doyne Farmer, and Seth Lloyd. A simple evolutionary game with feedback between perception and reality. Technical Report 2007-08-28, Santa Fe Institute, 2007.

[12] J.Y. Choi, D. Salandro, and K. Shastri. On the estimation of bid-ask spreads: Theory and evidence. *Journal of Financial and Quantitative Analysis*, 23:219–230, 1988.

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

[14] J. Danielsson and R. Payne. Measuring and explaining liquidity on the limit order book: Evidence from reuters d2000-2. Technical report, London School of Economics, 2001.

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

[16] A. Ellul, C. Holden, P. Jain, and R. Jennings. A comprehensive test of order choice theory: Recent evidence from the nyse. Technical report, Indiana University, 2005.

[17] J. P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modeling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, 1997.

[18] M. D. D. Evans and R. K. Lyons. Order flow and exchange rate dynamics. *Journal of Political Economy*, 110(1):170–180, 2002.

[19] J. D. Farmer, P. Patelli, and Ilija Zovko. The predictive power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):2254–2259, 2005.

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

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

[22] K. Garbade and W. Silber. Structural organization of secondary markets: Clearing frequency, dealer activity and liquidity risk. *Journal of Finance*, 34:577–593, 1979.

[23] Laszlo Gillemot, J. Doyne Farmer, and Fabrizio Lillo. There's more to volatility than volume. *Quantitative Finance*, 6(5):371–384, 2006.

[24] Lawrence R Glosten and Paul R Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.

[25] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley. Statistical properties of share volume traded in financial markets. *Physical Review E*, 62(4):R4493–R4496, 2000. Part A.

[26] S. J. Grossman and J. E. Stiglitz. On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408, 1980.

[27] J. Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46 (1):179–207, 1991.

[28] Joel Hasbrouck. Trades, quotes, inventories, and information. *Journal of Financial Economics*, 22(2):229–52, 1988.

[29] J. A. Hausman, A. W. Lo, and A. C. Mackinlay. An ordered probit analysis of transaction stock prices. *Journal of Financial Economics*, 31(3):319–379, 1992.

[30] R.W. Holthausen, R.W. Leftwich, and D. Mayers. The effect of large block transactions on

security prices: A cross-sectional analysis. *Journal of Financial Economics*, 19:237–267, 1987.

[31] R.W. Holthausen, R.W. Leftwich, and D. Mayers. Large block transactions, the speed of response, and temporary and permanent stock price effects. *Journal of Financial Economics*, 26:71–95, 1990.

[32] C. Hopman. Do supply and demand drive stock prices? *Quantitative Finance*, 2006. To appear.

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

[34] A. Kempf and O. Korn. Market depth and order size. *Journal of Financial Markets*, 2(1): 29–48, 1999.

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

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

[37] C. Lee and M. J. Ready. Inheriting trade direction from intraday data. *Journal of Finance*, 46:733–746, 1991.

[38] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 2004.

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

[40] F. Lillo, S. Mike, and J. D. Farmer. Theory for long memory in supply and demand. *Physical Review E*, 7106(6):066122, 2005.

[41] A. Madhavan, M. Richardson, and M. Roomans. Why do security prices change? a transaction-level analysis of nyse stocks. *The Review of Financial Studies*, 10(4):1035–1064, 1997.

[42] C. A. Parlour. Price dynamics in limit order markets. *The Review of Financial Studies*, 11 (4):789–816, 1998.

[43] V. Plerou, P. Gopikrishnan, X. Gabaix, and H. E. Stanley. Quantifying stock price response to demand fluctuations. *Physical Review E*, 66(2):article no. 027104, 2002.

[44] S-H. Poon and C. W. J. Granger. Forecasting volatility in financial markets: a review. *Journal of Economic Literature*, 41(2):478–539, 2003.

[45] M. Potters and J-P. Bouchaud. More statistical properties of order books and price impact. *Physica A*, 324:133–140, 2003.

[46] R. Roll. Orange juice and weather. *American Economic Review*, pages 861–880, 1984.

[47] H. Stoll. Inferring the components of the bid-ask spread: Theory and empirical tests. *Journal of Finance*, 44:115–134, 1989.

[48] N. Torre. *BARRA Market Impact Model Handbook*. BARRA Inc., Berkeley, 1997.

[49] G. Vaglica, F. Lillo, E. Moro, and R. Mantegna. Scaling laws of strategic behavior and size heterogeneity in agent dynamics. Technical report, U. of Palermo, 2007. URL `http://arxiv.org/abs/0704.2003`.

[50] P. Weber and B. Rosenow. Large stock price changes: volume or liquidity? *Quantitative Finance*, 6(1):7–14, 2006.

[51] Wee Yong Yeo. Persistence in limit order flow. Technical report, National University of Singapore, 2006.