A theory of aggregate market impact

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

[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*
[2]*Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL, 61801*
[3]*Dipartimento di Fisica e Tecnologie Relative, viale delle Scienze I-90128, Palermo, Italy*

(Dated: January 14, 2008)

Aggregate market impact is the net price change corresponding to an imbalance in the net signed volume of a sequence of trades. Trades that are initiated by buyers have a positive signed volume and tend to induce positive price impacts, and trades initiated by sellers have a negative signed volume and tend to induce negative price impacts. We develop a theory for aggregate market impact in terms of the competing random walks of net volume and net returns. Under the assumptions that individual impacts are permanent and IID we show that the aggregate impact R for N trades of net signed volume V scales as $R \sim VN^{-\kappa}$. If the distributions of volume fluctuations and return fluctuations are sufficiently thin tailed, $\kappa = 0$, but if either of them are sufficiently heavy tailed, $\kappa \neq 0$, with $\kappa > 0$ for realistic parameter values. We show that the same result holds under numerical extensions of the theory to more realistic assumptions of long-memory and temporary impacts, and demonstrate that the theory is in good agreement with data from the London Stock Exchanges. From a practical point of view these results are important for trade optimization, implying that market impact can be made arbitrarily small by trading sufficiently slowly. From a theoretical point of view they are significant because market impact is closely related to excess demand. Our theory suggests that structural considerations, in this case understanding the process of aggregation, dominate strategic considerations (i.e. utility maximization).

PRELIMINARY DRAFT.
PLEASE DO NOT DISTRIBUTE.

*jdf@santafe.edu
†gerig@santafe.edu
‡lillo@lagash.dft.unipa.it

Contents

I. Introduction 3
A. Motivation 3
B. Different kinds of market impact 5
C. Previous work 6
1. Empirical studies of market impact for single transactions 6
2. Empirical studies of aggregate market impact 6
3. Empirical studies of hidden orders 7
4. Why is market impact concave? 7
5. Contrast to Gabaix et al. 7

II. Relation of price impact to supply and demand 9
A. Market clearing with continuous excess demand 9
B. Revealed excess demand in the limit order book 10

III. Aggregation theory assuming permanent IID impact 11
A. General theory 12
B. Theory for power-law impact and volume distribution 13
C. Width of the region of linear impact 16
D. Behavior of aggregate impact for large volumes 16
E. Limitations of the model 16
1. Noisy impact 16
2. Finite size effect 18
3. Real time vs. transaction time 19

IV. Aggregation theory with temporary, long-memory impact 19
A. Failure of IID model in describing aggregate impact of financial data 19
B. Theory for the origin of long memory of signed order flow 20
C. Reconciling efficiency and long-memory: Temporary price propagator vs. asymmetric liquidity 20
D. Aggregation theory for the propagator model 21
E. Aggregation model for the asymmetric liquidity model 22

V. Comparison of theory to empirical data 24
A. Data 24
B. Testing the theory 24
C. Determinants of deviations from linearity in aggregate impact 25

VI. Conclusions 25

Acknowledgments 25

A. Saddle point approximation and asymptotic expansion 26

References and Notes 27

I. INTRODUCTION

A. Motivation

Understanding the nature of supply and demand is one of the oldest problems in economics. Supply and demand curves have two distinctly different functions. The first is that the intersection of the supply and demand curves determines the equilibrium value of the price, and the second is that their slopes make it possible to estimate how the price will respond to a change in excess demand, defined as demand minus supply. Here we study the aggregate market impact function, which is closely related to the slope of excess demand, and thus performs the second function. We choose to study aggregate market impact rather than excess demand for two reasons: (1) Unlike excess demand, it is observable in standard financial markets, and so a theory for it is falsifiable. (2) The functional form of the aggregate market impact can be understood in terms of a theory based on a generalization of the central limit theorem for competing random walks.

To make our goals in this paper clear we immediately present some empirical results. Let v_t be a signed transaction, where t is an index labeling the time sequencing of the transactions, which we will loosely refer to as "time". $|v_t|$ is the size of each transaction, measured either in shares or monetary units; buyer initiated transactions have positive signs and seller initiated transactions have negative signs. Let $r_t = \log(p_{t+1}/p_t)$ be the corresponding log-return, where p_t is the price of transaction t. For a sequence of N successive transactions beginning at time t, let $V_{t,N} = \sum_{i=1}^{N} v_{t+i}$ be the aggregate volume and $R_{t,N} = \sum_{i=1}^{N} r_{t+i}$ be the aggregate return. The average market impact conditioned on volume is

$$R(V, N) = E_t[R_{t,N}|V_{t,N} = V], \tag{1}$$

i.e. it is the expected return associated with a signed volume fluctuation V. The expectation E_t is taken by averaging over the transaction time index t. We write $R(V, N)$ to emphasize that this can depend both on the signed trading volume imbalance V and the number of transactions N.

In Figure 1 we show empirical estimates for the market impact for the company Astrazeneca, which is traded on the London Stock Exchange. $R(V, N)$ is estimated by recording $V_{t,N}$ and $R_{t,N}$ over a three year period for all t, sorting $V_{t,N}$ into bins containing roughly equal numbers of points, and computing the mean values of the pairs $(E_t[V_{t,N}], E_t[R_{t,N}])$ for each bin. In Figure 1 we show the market impact for different values of N with offsets added to the vertical axis to aid visualization. As one would expect, the scale increases with N. The shape of $R(V, N)$ also changes, becoming more linear with increasing N. This is illustrated more clearly in Figure 1(b), where we rescale the horizontal and vertical axes using a rescaling factor based only on $V_{t,N}$. The renormalization makes the increasing linearity clearer. As N increases the market impact near $V = 0$ becomes linear, and the size of the region that can be approximated as linear grows with increasing N. It also illustrates a surprising feature: The slope of the linear region decreases with N. These same basic features (increasing linearity and decreasing slopes) hold for all the stocks of the London Stock Exchanges in our sample.

Our main goal in this article is to explain the shape of market impact, and in particular, its increasing linearity and decreasing slopes. Our explanation is based on a generalization of the central limit theorem to double random walks, which may have heavy tails. From an intuitive point of view, these two facts can be qualitatively understand as follows: The

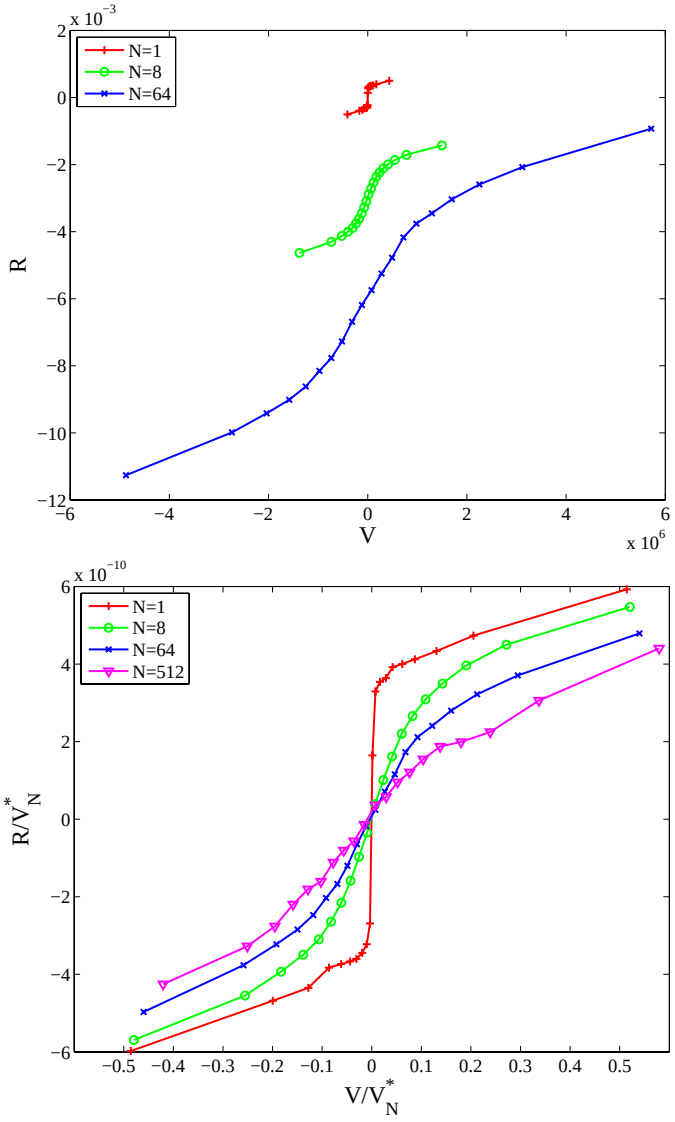

FIG. 1: Aggregate market impact $R(V,N)$ for the LSE stock Astrazeneca for 2000-2002. In (a) we plot the shifted aggregate return $R(V,N)+R_0$ vs. the aggregate signed volume V for three values of N. The arbitrary constant R_0 is added to aid visualization; its values are $R_0 = \{0, -3\times10^{-3}, -6\times10^{-3}\}$ for $N = 1, 8$ and 64 respectively. In (b) for each N we rescale both the horizontal and vertical axes by $V_N^* = V_N^{(95)} - V_N^{(5)}$, where $V_N^{(5)}$ is the 5% quantile and $V_N^{(95)}$ is the 95% quantile of $V_{t,N}$.

market impact is determined by a competition between two random walks for V and R, which have the same number of steps. For large N each random walk approaches a smooth limiting distribution near $V = 0$ and $R = 0$, so that for any fixed value of N, $R(V,N)$ becomes increasingly linear in V. If the increments of R and V have similar behavior in their tails, then the slope is constant as a function of N. However, if the increments of V are sufficiently heavy tailed compared to those of R, as is typically the case in financial markets, the distribution of V spreads with N faster than it does with R, and the slope decreases

with N. We make these arguments more precise in terms of a nested set of theories, starting with a simple theory in which we can compute everything analytically, and then progressing to more realistic but slightly more complicated theories for which we rely on numerical simulations.

From the point of view of economic theory our results are important because they reflect on basic questions about the most effective approach to understanding price formation. The standard neoclassic approach to understanding supply and demand is to assume that agents maximize utility subject to assumptions about the agent information processing model (e.g. rationality). Our results here are not inconsistent with that view, but they show that it is simply irrelevant in determining the shape of the market impact function. The central limit theorem is the dominant effect. Because market impact is closely related to excess demand, these results are not an esoteric side effect of financial economics, but rather bear on supply and demand, a central concept in economics. This is a good instance where structural considerations (in this case aggregation) dominate strategic considerations (utility maximization).

This paper is organized as follows: In the remainder of this section we discuss previous work. In Section V A we describe our data set and review the market structure in the LSE. Section II discusses the relationship between market impact and excess demand, first under asynchronous market clearing and then for the continuous double auction. Section III develops a theory for market impact under the idealized assumptions of IID order flow and permanent impact for individual transactions. Section IV extends this theory to apply to the more realistic situation in which order flow has long-memory and individual impacts are not permanent. In Section V we do a series of empirical tests to determine the factors that influence deviations from linearity. We then conclude and summarize in Section VI.

B. Different kinds of market impact

There are different kinds of market impact, depending on the market structure and the type of order flow, and it is important to distinguish them. The LSE has upstairs and downstairs markets. In the downstairs market trades are made by placing orders in a limit order book, and it is quite common to aggressively split large trading orders into many small pieces. The upstairs market trades are arranged bilaterally between individuals, and while there is some order splitting, it is a smaller effect. As a result of the different market structures the impacts can be quite different. In the upstairs market the trading volume is much more heavy tailed (see Lillo, Mike, and Farmer, **?**), which we will argue can make an important difference in determining the aggregate impact. For the LSE the data allows us to separate upstairs and downstairs trades, but because there are serious problems with the time stamps for upstairs trades, we only analyze downstairs trades.

A second factor that must be kept in mind is that large trading orders, which we will call *hidden orders*, are typically split into small pieces and executed incrementally. This is in contrast to *realized orders*, which are the actual orders that are traded, e.g. the pieces into which hidden orders are split. For realized orders the impacts may be part of a larger process of order splitting that is invisible with the data that we have here. The impacts of hidden orders may be quite different than those of realized orders. In this paper, with the data we are using, we can only study realized orders, but we can do this for both upstairs and downstairs trades.

C. Previous work

1. Empirical studies of market impact for single transactions

Many studies have examined the market impact for a single transaction, $N = 1$, and all have observed a concave function of $V = v_i$, i.e. one that increases rapidly for small v_i and more slowly for larger v_i. The detailed functional form, however, varies from market to market and even period to period. Early studies by Hasbrouck (**?**) and Hausman, Lo and MacKinlay (**?**) found strongly concave functions, but did not attempt to fit functional forms. Keim and Madhavan (**?**) also observed a concave impact function for block trades. Based on Trades and Quotes (TAQ) data for a set of 1000 NYSE stocks the concavity of the market impact was interpreted by Lillo, Farmer, and Mantegna (**?**) using the functional form

$$R(V,1) = \frac{sign(V)|V|^{\beta(V)}}{\lambda}. \tag{2}$$

The exponent $\beta(V)$ is approximately 0.5 for small volumes and 0.2 for large volumes. Even normalizing the volume V by daily volume, the liquidity parameter λ varies for different stocks; there is a clear dependence on market capitalization C that is well-approximated by the functional form $\lambda \sim C^{\delta}$, with $\delta \approx 0.4$. Potters and Bouchaud (2003) analyzed stocks traded at the Paris Bourse and NASDAQ and found that a logarithmic form gave the best fit to the data. For the London Stock Exchange, Lillo and Farmer (**?**) and Farmer, Patelli and Zovko (2005) found that for most stocks Equation 2 was a good approximation with $\beta = 0.3$, independent of V. Hopman (**?**) studied market impact on a thirty minute timescale in the Paris bourse for individual orders and found $\beta \approx 0.4$, depending on the urgency of the order. Thus all the studies find strongly concave functions, but report variations in functional form that depend on the market and possibly other factors as well.

2. Empirical studies of aggregate market impact

Studies of aggregated market impact have produced variable results, reaching different conclusions that we will argue depend substantially on the time scale for aggregation. The BARRA market impact model, an industry standard, uses the TAQ data aggregated on a half hour time scale (Torre, **?**). They compare fits using Equation 2 and find $\beta \approx 0.5$; they obtain similar results using individual block data. Kempf and Korn (**?**) studied data for futures on the DAX (the German stock index) on an five minute time scale and found a very concave functional form. Plerou et al. (**?**) studied data from the NYSE during 1994-95 ranging from 5 to 195 minute time scales and fit the market impact function with a hyperbolic tangent. They noted that at shorter time scales this functional form did not work well for small V; $\tanh(V)$ is linear for small V, but at short time scales (e.g. 5 or 15 minutes) they observed a nonlinear impact function, becoming more linear as they went toward longer time scales. Evans and Lyons (**?**) studied foreign exchange rate transactions data for DM and Yen against the dollar at the daily scale over a four month period. They used the number of buyer initiated transactions minus the number of seller initiated transactions as a proxy for the signed order flow volume V, and found a strong positive relationship to concurrent returns. They also developed a theory for interdealer and public trading. Under the assumption that the public's demand function is linear they (unsurprisingly) derive a linear market impact function. Evans and Lyons tested for nonlinearity in the data using

a quadratic term V^2, and found that it provided little benefit; unfortunately, they did not test using a cubic term, which in view of the expected approximate antisymmetry of $R(V)$ would have been much more informative. Chordia and Subrahmanyam **?** study impact for stocks in the S&P 500 at a daily time scale and perform linear regressions, but do not compare to other functional forms. For the Paris bourse Hopman (**?**) measures aggregate order flow as $\tilde{V} = \sum_i sign(v_i) v_i^\beta$, where the sum is taken over fixed time intervals. At a daily scale he finds he gets the best linear regression against contemporary daily returns $\tilde{R}$ with $\beta \approx 0.5$. He also documents that the slope of the regression decreases with increasing time scale. Finally, as discussed in more detail below, Gabaix et al. (**?** **?**) have made extensive studies of data from the New York, London and Paris stock markets on a fifteen minute time scale, and find exponents $\beta \approx 0.5$.

3. Empirical studies of hidden orders

Because data for hidden orders, which are sometimes also called tradiing packages, are difficult to obtain, there are only a few studies. Chan and Lakonishok (**?** , **?**). They find that order splitting can be spread over periods as long as a week. Another study is that of Gallagher and Looi (**?**). Reference Palermo groups paper when ready.

4. Why is market impact concave?

The standard reason given for the concavity of market impact is that it reflects the informativeness of trades. If small trades carry almost as much information as large trades, then the price changes caused by small trades should be nearly as big as those for large trades. For example, this could be due to "stealth trading", because informed traders keep their orders small to avoid revealing their superior knowledge [see Barclay and Warner, **?**)]. An alternative hypothesis due to Daniels et al. (**?**) is that it reflects the accumulation of liquidity in the limit order book. I.e., the depth in the order book as a function of the price will determine the market impact for a market order as a function of its size [see also Bouchaud, Mezard and Potters (**?**) and Smith et al. (**?**)]. Keim and Madhavan (**?**) have proposed a theory for block trades based on the hypothesis that there a cost for searching for counterparties, and for larger orders more searching is done and so more counterparties are found, thereby lowering the impact. Another hypothesis is that this is due to selective liquidity taking, i.e. that liquidity takers submit large orders when liquidity is high and small orders when it is low [see Farmer et al. (**?**), Weber and Rosenow (**?**), and Hopman (**?**)]. Finally, Gabaix et al. **?** **?**) have proposed that this is caused by a combination of first order risk aversion and the fact that larger trades take longer for liquidity providers to unwind. In Section II B we present evidence supporting the hypothesis that concavity is caused by selective liquidity taking.

5. Contrast to Gabaix et al.

Gabaix et al. (**?** **?**) address some of the same questions that we do and presents an alternative theory and data analysis, and so deserve special discussion. They hypothesize that the functional form of the market impact is driven by risk aversion. Under the assump-

tion that large orders are broken into pieces the time to fill will depend linearly on the size of the order. They assume first order risk aversion, i.e. that risk aversion increases with the standard deviation of price variations. Since the standard deviation of price fluctuations grows roughly proportional to the square root of time, the fair price for an order increases as the square root of its size. This was suggested earlier, though much less clearly, by Zhang (**?**).

To test whether the data supports their hypothesis Gabaix et al. regress R^2 vs. V using fifteen minute intervals. Their argument for testing their hypothesis this way goes as follows: Assume that

$$r_i = K\epsilon_i |v_i|^\beta + n_i, \tag{3}$$

where ϵ_i is the sign of v_i and n_i is an IID noise process. If we also assume that $E[\epsilon_i v_i] = 0$, aggregating over N transactions, squaring and taking expectations gives

$$E[R^2|\tilde{V}] = K^2(\tilde{V}^{2\beta} + 2\sum_{i\neq j}^{N} E[\epsilon_i\epsilon_j |v_i|^\beta |v_j|^\beta]), \tag{4}$$

$\tilde{V} = \sum_i |v_i|$ is the total absolute volume (and not the signed order flow imbalance). Gabaix et al. argue that the last term can be neglected if ϵ_i and ϵ_j are uncorrelated. They regress a sample estimate of $E[R_t^2|\tilde{V}]$ vs. , and show that that for large $\tilde{V}$ it has a slope close to one, implying $\beta \approx 1/2$.

There are serious problems with neglecting the second term. As shown by Bouchaud et al. (**?**) and Lillo and Farmer (**?**), real order flow has long memory, i.e. it has a positive autocorrelation $C(\tau)$ that decays as a power law $C(\tau) \sim \tau^\gamma$, where $0 < gamma < 1$. This means that v_i and v_j are strongly correlated, even when $|i-j|$ is large. As shown by Farmer and Lillo (**?**) this can dramatically alter the results. For example, using the v_i of real order flow but simulating returns with Eq. 3 one can get similar results to their, with a linear scaling for large V.

An even more serious problem with their test comes from using the total volume rather than the signed order flow imbalance. The total volume has the advantage of being easier to measure, but the disadvantage that the resulting test is much less powerful. Most disturbingly, their empirical results can be reproduced under trivial alternatives hypotheses. For example, suppose that r_i and v_i are independent, i.e. $r_i = n_i$, where n_i is a noise term that is independent of v_i. If the random processes $R = \sum_{i=1}^{N}$ and $V = \sum_{i=1}^{N}$ have the same number of increments, then if n_i is IID and its variance exists, R^2 will scale as N. Similarly because all the increments of $\tilde{v}$ are positive, $\tilde{V}$ scales as N, so $E[R^2|\tilde{V}] \sim \tilde{V}$. When seen in this light, it becomes surprising to get any result other than linear scaling. Although effects such as long-memory can cause nonlinear scaling of $E[R^2|\tilde{V}]$, this argument makes it clear that their are many alternatives to their hypothesis that will pass their test. This makes it critical to measure impact directly, e.g. by measuring $E[R|V]$ (which is identically zero if r_i and v_i are independent)[1].

[1] The theory of Gabaix et al. also predict that the heavy tails of returns are related to the functional form of $E[R|V]$. This is incompatible with observations that the heavy tails of returns are almost independent of V, and instead depend much more strongly on liquidity fluctuations, i.e. fluctuations of R_t around $E[R|V]$ [Farmer et al. (**?**), Farmer and Lillo and Gillemot (**?**), Weber and Rosenow (**?**)].

II. RELATION OF PRICE IMPACT TO SUPPLY AND DEMAND

Aggregate market impact is closely related to supply and demand. Because our results derive the form of aggregate market impact using methods that are quite different from the classic neoclassical approach, one of the interesting aspects of our work that extends beyond finance is what it implies about effective approaches to constructing economic theories. Because the relationship between market impact and supply and demand is not entirely obvious we develop this here. To simplify the discussion, throughout we will consider the demand minus the supply, $q(p) = D(p) - S(p)$, which is usually called the *excess demand*. This is justified because price formation only depends on the excess demand, and not on the supply and demand individually.

A. Market clearing with continuous excess demand

Consider a continuous double auction and assume market clearing. Let $q(p) = \sum_i q_i$ be the total excess demand at price p, where q_i is the excess demand of agent i, and assume $dq/dp < 0$ for all p. Suppose only one agent updates her excess demand at a time, to a new demand function q_i', while everyone else holds theirs' constant. It is useful to distinguish two types of updates:

- *Transaction causing.* The update occurs for the part of $q_i(p)$ whose domain includes the clearing price, and so causes a transaction.
- *Non-transaction causing.* The update occurs for the part of $q_i(p)$ whose domain does not include the clearing price, and so does not cause a transaction.

For a transaction causing update the market clearing condition is $q'(p') = q(p') + v_i = q(p) = 0$, where $v_i = \delta q_i = q_i' - q_i$. Since by assumption the agents update their individual demand functions one at a time, $v_i = \delta q$, i.e. agent i trades her full excess demand fluctuation, though the other side of the trade may be split among many counterparties. The sign of v_i indicates who *initiated* the trade – buyer initiated trades have positive signs, and seller initiated trades have negative signs. The magnitude indicates the total amount traded. Let $\delta p = p_{t_i} - p_{t_{i-1}}$ be the corresponding change in the clearing price. Under the assumption that q is differentiable and δp is sufficiently small, $q(p') \approx q(p) + (\partial q/\partial p)\delta p$, and the change in price due to a transaction at time t_j can be written

$$\delta p_j = -\frac{v_j}{\partial q/\partial p} = -\frac{\partial p(q, t_j)}{\partial q} v_j, \tag{5}$$

and the change in price for a series of N successive transactions $\{v_j\}$, where $j = k, \ldots, k+N$, is

$$\Delta p_k = \sum_{j=k}^{k+N} \delta p_j = -\sum_{j=k}^{k+N} \frac{\partial p(q, t_j)}{\partial q} v_j. \tag{6}$$

The transaction causing updates contribute directly to changes in price, and the non-transaction causing updates contribute indirectly by altering $\partial p(q, t_j)/\partial q$, which alters the response to each transaction.

Let r_j be the log-return $\log(p_j/p_{j-1})$ generated by a transaction v_j, and let

$$R_k(N) = \sum_{j=k}^{k+N} r_j = \log(p_{k+N}/p_k) \tag{7}$$

be the N step log-return starting at time k. The corresponding *market impact* is the pair (V_k, R_k), where $V_k = \sum_{j=k}^{k+N} v_j$. The return R_k is not a deterministic function of V_k, due to the dependence on the non-transaction excess demand updates and the sequence of trades $\{v_j\}$. Using Eq. 6 the return $R_k(N)$ can be written

$$R_k(N) = \sum_{j=k}^{k+N} \log(1 - \frac{v_j}{p_j}\frac{\partial p(q,t_j)}{\partial q}) \approx -\sum_{j=k}^{k+N} \frac{v_j}{p_j}\frac{\partial p(q,t_j)}{\partial q}, \tag{8}$$

where the approximation is valid as long as δp is sufficiently small. The quantity $\partial p(q,t_j)/p_j \partial q$ is the price elasticity. Thus, the return is (minus) the price elasticity weighted by the transaction volumes. The market impact is $R(V,N) = E_k[R_k(N)|V]$. In the special case where the transactions v_j are uncorrelated with the price elasticities this can be written in the simple form

$$R(V,N) = -E[\frac{1}{p}\frac{\partial p}{\partial q}]V. \tag{9}$$

I.e. it is just (minus) the average price elasticity times the total volume imbalance, and so is linear in V. In general, as we will demonstrate in a moment, because the transactions are correlated with the elasticities, the market impact is a nonlinear function of both V and N.

B. Revealed excess demand in the limit order book

Most modern financial markets use a continuous double auction for price formation. The market structure is similar to the market clearing framework described above, but with several important differences. Individual agents place trading orders to buy or sell in a queue called the limit order book. Each limit order for x shares at price π can be thought of as specifying an excess demand function $q_i(p)$ that is a step function. A buy order has $q_i(p) = x$ for $p \leq \pi$ and $q_i(p) = 0$ for $p > \pi$, and a sell order has $q_i(p) = 0$ for $p < \pi$ and $q_i(p) = -x$ for $p \geq \pi$. Since each individual excess demand function is discontinuous, the total excess demand function $q(p) = \sum_i q_i$ is also discontinuous. Orders that cross the best prices and general immediate transactions are called *effective market orders*, and orders that do not general immediate transactions are called *effective limit orders*. After each effective market order arrives there is always a *spread* $s = p_a - p_b$ between the best selling price p_a offered at any time (also called "the best ask"), and the best buying price p_b bid (also called "the best bid"). To avoid the alternation of transaction prices across spread, we compute market impact based on the midprice $p_m = 1/2(p_a + p_b)$.

The limit order book contains only the revealed excess demand, which is a small fraction of the total. As we discuss in more detail in Section IV B, for strategic reasons most agents do not like to reveal their true intentions, and indeed go to considerable effort to hide them. The majority of the excess demand $q(p)$ remains hidden and is only revealed incrementally. The revealed excess demand in he limit order book is typically only about XXX of the market capitalization. Thus, the trading orders visible in the limit order book are only the

tip of a very large iceberg, and the excess demand contained in the limit order book contains a highly incomplete picture of the true situation.

A striking feature of the revealed excess demand is its extreme variability. This is illustrated in Figure 2(a), which gives several snapshots of the excess demand $q(p-p_m, t_k)$ for the LSE stock Astrazeneca at different times t_k. We have plotted this relative to the midprice

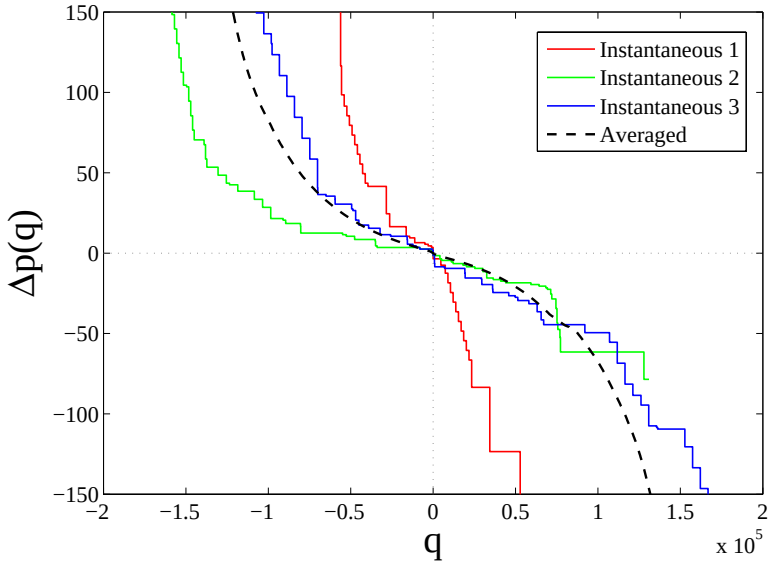

FIG. 2: Revealed excess demand $q(p-p_m, t_k)$ in the limit order book, for the stock Astrazeneca. (a) shows the excess demand function at randomly chosen times t_k, and (b) shows the time average $E_k[q(p, t_k)]$.

p_m. There is strong persistence from time t_k to t_{k+1}, but when sampled over sufficiently long time intervals, q looks like a random function. In panel (b) we take a time average, and show that while the average excess demand decreases rapidly near the best prices, as one moves away in either direction the excess demand curve tends to flatten [see Bouchaud, Mezard, and Potters (**?**) and Weber and Rosenow (**?**)]. This is because more demand is revealed near the best prices – far from the best prices the probability of a transaction is low, and so there is little incentive to pay the cost of revealing one's excess demand.

In contrast, the market impact is fully observable, and for this reason alone is a better target for economic theory than the excess demand. The market impact is based on what happens at the price levels where transaction prices are formed, and so is by definition an accurate reflection of the excess demand at these price levels. It is obviously much easier to test a theory for something that can be measured than to test a theory for something that can't. Because market impact involves only transactions and observed price changes, it is a better probe of the true excess demand that the revealed excess demand is.

III. AGGREGATION THEORY ASSUMING PERMANENT IID IMPACT

To develop intuition about this problem, we begin by developing a theory based on IID permanent impacts, which has the advantage that we can compute everything analytically.

In the next section, we argue that it is more realistic to consider impacts whose signs are strongly autocorrelated (so much so that they have long-memory), with impacts that either fluctuate in an appropriately correlated manner, or decay in time. In any case, the intuition developed from the IID case is very useful and gives a qualitative illustration of the main features of aggregate impact.

A. General theory

Consider a series of $N+1$ transactions with signed volumes v_i corresponding to total return $R_{N+1} = \sum_{i=1}^{N+1} r_i$ and total signed volume $V = \sum_{i=1}^{N+1} v_i$. (The reason for using $N+1$ will become clear later). Assuming that there exists a stationary probability distribution $P(R, V)$, the expected return given V can be written

$$R(V, N+1) \equiv E[R|V, N+1] = \int RP(R|V, N+1)\ dR = \frac{1}{P_{N+1}(V)} \int RP(R, V, N+1)\ dR, \tag{10}$$

where $P_{N+1}(V)$ is the probability density for V. We assume that the $N+1$ individual price impacts r_i due to the IID signed volumes v_i are given by a deterministic function $r_i = f(v_i)$. Let the distribution of individual v_i be $\pi(v_i)$. Then the joint distribution of v_i is

$$P(v_1, \ldots, v_{N+1}) = \pi(v_1)\pi(v_2)\ldots\pi(v_{N+1}).$$

The expected return given V is

$$\int RP(R, V, N+1)\ dR = \int dv_1 \ldots dv_{N+1}\pi(v_1)\ldots\pi(v_{N+1}) \sum_{i=1}^{N+1} f(v_i)\delta(V - \sum_{i=1}^{N+1} v_i), \tag{11}$$

where we introduced the Dirac delta function. The key idea is to use the integral representation of the Dirac delta, $\delta(x) = (2\pi)^{-1} \int \exp(-i\lambda x)\ d\lambda$, which allows us o rewrite the integral in (11) as

$$\int dv_1 \ldots dv_{N+1}\pi(v_1)\ldots\pi(v_{N+1}) \sum_{i=1}^{N+1} f(v_i) \frac{1}{2\pi} \int d\lambda e^{-i\lambda(V - \sum_{i=1}^{N+1} v_i)}. \tag{12}$$

By changing the order of integration in λ, in volume and of the summation we get

$$\frac{1}{2\pi} \int d\lambda e^{-i\lambda V} \sum_{i=1}^{N+1} \int dv_1 \ldots dv_{N+1}\pi(v_1)\ldots\pi(v_{N+1}) f(v_i) e^{i\lambda \sum_{i=1}^{N+1} v_i}. \tag{13}$$

Since all the terms in the summation are identical this can be rewritten as

$$\frac{N+1}{2\pi} \int d\lambda e^{-i\lambda V} \int dv_1 \ldots dv_{N+1}\pi(v_1)\ldots\pi(v_{N+1}) f(v_{N+1}) e^{i\lambda \sum_{i=1}^{N+1} v_i}. \tag{14}$$

and by decoupling the integration in v_{N+1} from the other integrations in volume we get

$$\frac{N+1}{2\pi} \int d\lambda e^{-i\lambda V} \int dv_{N+1} f(v_{N+1})\pi(v_{N+1}) e^{i\lambda v_{N+1}} \int dv_1 \ldots dv_N \pi(v_1)\ldots\pi(v_N)) e^{i\lambda \sum_{i=1}^{N} v_i}. \tag{15}$$

The last multiple integral can be written as the N^{th} power of a simple integral, so that

$$\int RP(R,V,N+1)\ dR = \frac{N+1}{2\pi}\int d\lambda e^{-i\lambda V}\int dv f(v)\pi(v)e^{i\lambda v}\left(\int dv\pi(v)e^{i\lambda v}\right)^N. \quad (16)$$

We now introduce the functions

$$h(\lambda)\equiv \ln(\int dv\pi(v)e^{i\lambda v}), \qquad g(\lambda)\equiv e^{-i\lambda V}\int dv f(v)\pi(v)e^{i\lambda v}, \quad (17)$$

and the integral can be rewritten as

$$\int RP(R,V,N+1)\ dR = \frac{N+1}{2\pi}\int d\lambda e^{Nh(\lambda)}g(\lambda). \quad (18)$$

By using Eq. 10 we finally write the aggregate price impact as

$$R(V,N+1) = \frac{N+1}{2\pi}\frac{1}{P_{N+1}(V)}\int d\lambda e^{Nh(\lambda)}g(\lambda) \quad (19)$$

This is an exact result that allows to calculate the aggregate impact when the probability distribution $\pi(v)$ of individual volumes and the impact function $f(v)$ are known. However in general it is not possible to perform the integral analytically. Since we are interested in the large N limit, we use saddle point approximation to obtain the asymptotic behavior of Eq. 19.

B. Theory for power-law impact and volume distribution

To get a concrete result we have to choose specific forms for π and f. We consider the important case that they are both power laws, of the form

$$\begin{aligned}\pi(v) &\sim \frac{1}{v^{\alpha+1}}\\ f(v) &\sim sign(v)|v|^{\beta}\end{aligned} \quad (20)$$

Note that in order to obtain the behavior of the aggregate impact we need to know the asymptotic behavior of the impact and of the volume distribution. We assume for simplicity that the volume distribution $\pi(v)$ is even, i.e. that it is symmetric for positive and negative returns. The distribution of individual return $r = f(v)$ behaves asymptotically as

$$\pi(r) \sim \frac{1}{r^{1+\frac{\alpha}{\beta}}} \quad (21)$$

and the returns have finite mean when $\beta < \alpha$. In Fig. 3 we integrate Eq. 19 for $\alpha = 1.5$ and $\beta = 0.3$ for several different values of N.

An asymptotic expression for the aggregate impact can be obtained using saddle point approximation. The detailed calculation is described in the Appendix. The behavior of $R(V, N+1)$ for large N has the general form

$$R(V,N+1) \sim \frac{V}{N^{\kappa}} \quad (22)$$

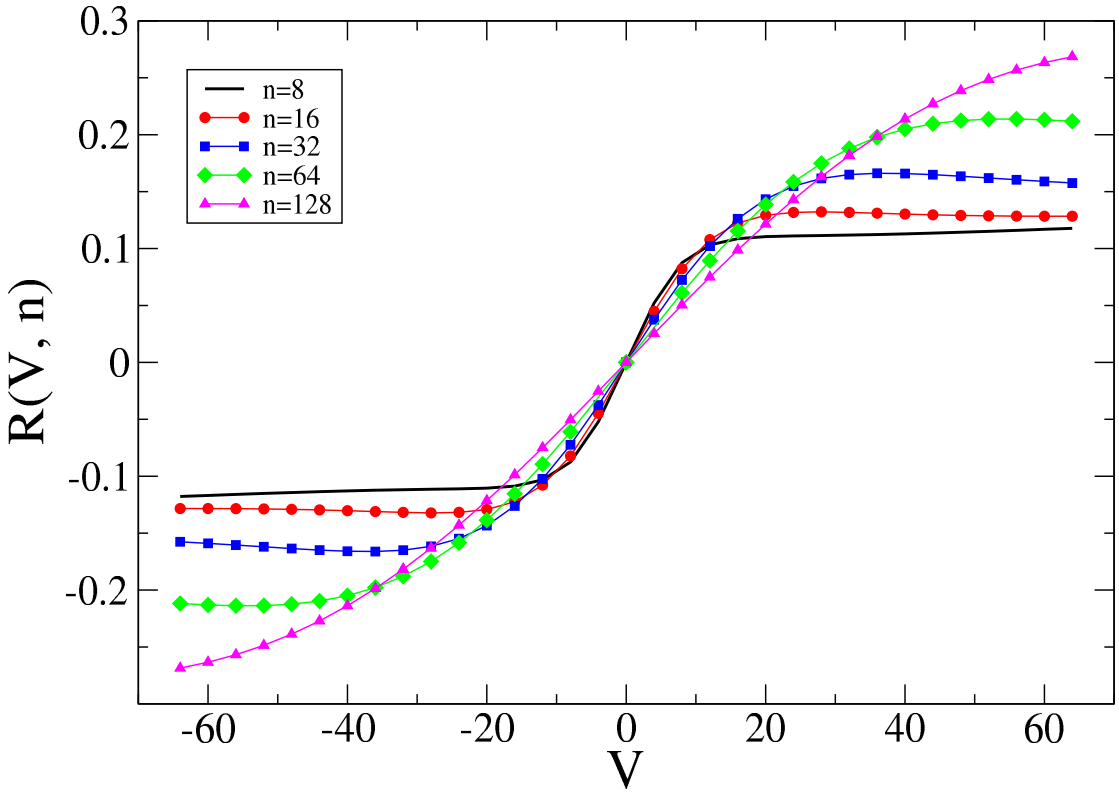

FIG. 3: Expected price impact vs volume imbalance for several different values of N for π and f power laws with $\alpha = 1.5$ and $\beta = 0.3$ (see Eq. 20).

TABLE I: Different regions for the scaling exponent κ.

region	conditions		κ
I	$\alpha > 2$	$\alpha > \beta + 1$	0
II	$1 < \alpha < 2$	$\alpha > \beta + 1$	$(2-\alpha)/\alpha$
III	$\alpha > 2$	$\alpha - 1 < \beta < \alpha$	$(\alpha - \beta - 1)/2$
IV	$\alpha < 2$	$\alpha - 1 < \beta < \alpha$	$(1-\beta)\alpha$
V	$\beta > \alpha$		not defined

where the aggregation exponent $\kappa(\alpha, \beta)$ depends on the exponent β and α describing the impact function and the volume distribution. Eq. 22 is valid in a region around the value $V = 0$ which becomes larger as N increases.

As illustrated in Fig. 4, the parameter space divides into five distinct regions, with different behavior for $\kappa(\alpha, \beta)$:

- **Region I**: $\alpha > 2$ and $\alpha > \beta + 1$. In this region $\kappa = 0$. Both r_i and v_i are sufficiently

thin tailed so that R and V are normally distributed for large N (actually in this region the convergence to linearity is very fast, e.g. it typically occurs to a good approximation with ten or so steps).

- **Region II**: $1 < \alpha < 2$ and $\alpha > \beta + 1$. In this region $\kappa = (2 - \alpha)/\alpha$. The variance of v_i does not exist, but the variance of r_i does exist, so V exhibits anomalous diffusion but R does not.

- **Region III**: $\alpha > 2$ and $\alpha - 1 < \beta < \alpha$. In this region $\kappa = (\alpha - \beta - 1)/2$. The variance of v_i exists, but the variance of r_i does not exist.

- **Region IV**: $\alpha < 2$ and $\alpha - 1 < \beta < \alpha$. In this region $\kappa = (1 - \beta)/\alpha$. Neither the variance of v_i nor the variance of r_i exists.

- **Region V**: $\beta > \alpha$. In this region the expected value of return diverges and therefore the aggregate impact $R(V, N)$ is not defined.

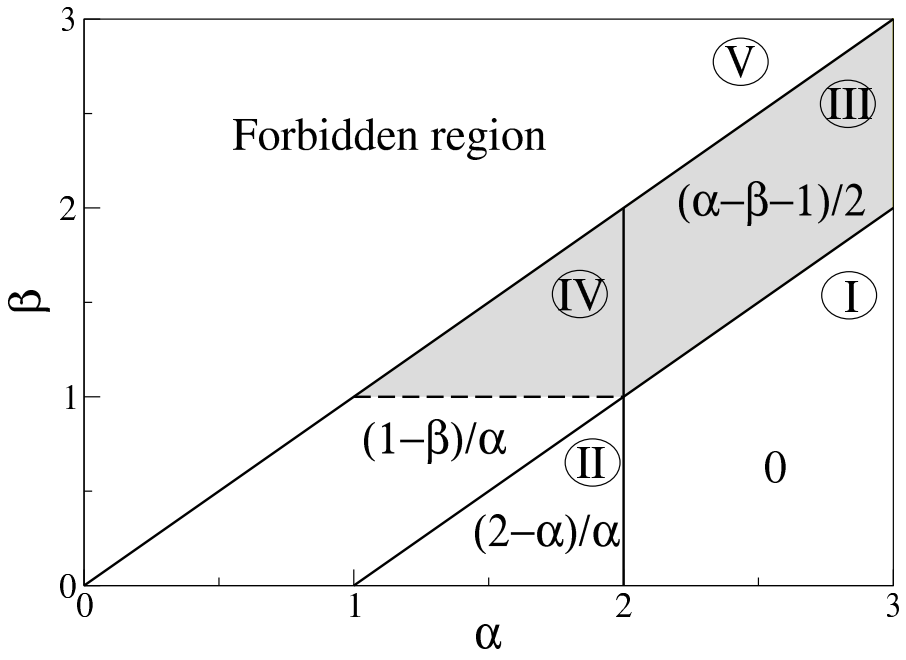

FIG. 4: The aggregation exponent κ as a function of the exponents α and β of Eq. 20. The circled numbers label the different regions described in the text. The grey area corresponds to the parameters giving a negative value of κ.

It is direct to see that κ is continuous at the boundaries between different regions. Moreover it is worth noting that there is a range of parameters (grey area in Figure 4) for which $\kappa < 0$. This implies that the slope of the aggregated impact *increases* with N. Although all the regimes corresponding to regions I-IV are in principle observable in real data, we want

here to focus our attention on regions I and II, also because these regions will be important in the following discussion. As we will see below, for real data it is $\kappa > 0$ and this rules out the possibility that region III is suitable for describing real data. In region I it is $\kappa = 0$, i.e. the slope of the linear part of the aggregate impact does not change with the aggregation N. We proved that this behavior can be observed also in cases not described by the form of Eq. 20, such as, for example, the case of power law impact and Gaussian or exponentially distributed volumes. In this case we prove that $\kappa = 0$ for any value of β. We showed that the behavior described by region I is observed more generally when the volume distribution has tails decaying faster than a power law (Gumbel regime in Extreme Value Theory). In region II it is $\kappa = (2 - \alpha)/\alpha$, i.e. the slope changes with an exponent which is independent of the exponent of the impact function.

C. Width of the region of linear impact

Beside the scaling of the slope of the linear region, the IID theory is also able to give quantitative prediction on other aspects of the aggregate impact. First of all, an important issue is to quantify the fraction of points that are explained by the linear behavior as a function of N. This is a delicate issue that requires to go to the next term in the asymptotic expansion where a non-linear term (in V) appears. Usually one can write such an expansion in the form

$$R(V, N+1) \sim \frac{V}{N^{\kappa}} - \frac{V^3}{N^{\kappa'}} \tag{23}$$

and by requiring that $V/N^{\kappa} \gg V^3/N^{\kappa'}$ we obtain an expression for the linear region $|V| << N^{\frac{\kappa'-\kappa}{2}}$. We proved that in region II the aggregate impact is linear is linear for $|V| \ll N^{1/\alpha}$. Together with the self-similarity property of the stable distribution this result implies that the fraction of points where the linear approximation holds tends to a fixed value smaller than 1 as N go to infinity. It is worth noting that for Gaussian distributed volumes (belonging to region I) the aggregate impact is linear for $|V| < N$. As a consequence the fraction of points where the impact is linear tends to 1 as N go to infinity.

D. Behavior of aggregate impact for large volumes

A second quantitative prediction of the IID model concerns the behavior of the aggregate impact $R(V, N+1)$ for a fixed value of N and for large V. An asymptotic theory shows that for large V it is always $R(V, N+1) \sim V^{\beta}$, i.e. for large volumes the aggregate impact behaves as the individual transaction impact.

E. Limitations of the model

1. Noisy impact

In the theory we have hypothesized that the impact function $f(v)$ is a deterministic function of the volume v. Empirical studies have shown that market impact of individual transaction is highly noisy. Typically fluctuations of impact are even larger than the mean

impact. One could therefore wonder whether the theory developed above still holds for noisy impact function.

Let us assume that the impact is not a deterministic function of the volume v_i, but it also an addittive random part

$$r = f(v_i) = f_d(v_i) + \xi_i \tag{24}$$

where ξ_i is an IID noise independent from v_i and $f_d(.)$ is a deterministic function. The expected return given V as

$$\begin{aligned} E[R|V] =& < \sum_{i=1}^{N+1} f(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} = < \sum_{i=1}^{N+1} (f_d(v_i) + \xi_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} \\ =& < \sum_{i=1}^{N+1} f_d(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v}} + < \sum_{i=1}^{N+1} \xi_i \delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} \end{aligned} \tag{25}$$

In the above derivation we used the symbol $< f >_g$ to indicate the mean value of the function f with respect to the random variable g. Since the noise is independent (even uncorrelated would probably works), the last term can be factorized as $< \sum_{i=1}^{N+1} \delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v}} < \xi_i > \xi$ which is zero because the the noise has zero mean.

Similarly, if the noise is multiplicative

$$r = f(v_i) = \xi_i f_d(v_i) \tag{26}$$

with $< \xi_i >= \bar{\xi} > 0$, one can write

$$\begin{aligned} E[R|V] =& < \sum_{i=1}^{N+1} f(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} = < \sum_{i=1}^{N+1} \xi_i f_d(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} \\ =& \sum_{i=1}^{N+1} < \xi_i f_d(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v},\xi} = \sum_{i=1}^{N+1} < \xi_i >_{\xi} < f_d(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v}} \\ =& \bar{\xi} < \sum_{i=1}^{N+1} f_d(v_i)\delta(V - \sum_{i=1}^{N+1} v_i) >_{\vec{v}} \end{aligned} \tag{27}$$

The aggregate impact with noisy impact is the same as the aggregate impact for deterministic impact except for an $N-$independent constant $\bar{\xi}$.

In conclusion, the *mean* impact is independent on the presence and intensity of the noise term. However the fluctuations of the aggregate impact around the mean will depend on the noise intensity.

2. Finite size effect

By using the saddle point approximation we have derived above the leading term of the asymptotic expansion of the aggregate impact. The derived behavior thus holds in the limit $n \to \infty$. An important question to ask is how well the leading term of the asymptotic expansion approximates the behavior of the aggregate impact for large but finite n. The theory of asymptotic expansion is quite complicated and in this paper we present a specific case which shows a general behavior. As a specific case we consider transaction volumes that are described by a Student distribution with α degrees of freedom, i.e.

$$\pi(v) = \frac{\Gamma(\frac{1+\alpha}{2})}{\sqrt{\pi}\Gamma(\alpha/2)} \frac{1}{(1+x^2)^{\alpha+1)/2}} \sim \frac{1}{v^{\alpha+1}} \tag{28}$$

with $\alpha > 0$ and a power law impact function

$$f(v) = sign(v)|v|^{\beta} \tag{29}$$

In order to calculate the aggregate impact we need to compute the functions $\mathcal{H}(\lambda)$, $\mathcal{G}(\lambda)$ and the distribution $P_n(V)$ of total volume needed to go from $< R >_V$ to $E[R|V]$.

Behavior of $\mathcal{H}(\lambda)$. The function $\mathcal{H}(\lambda)$ is the characteristic function and it is equal to

$$\mathcal{H}(\lambda) = \frac{2^{1-\alpha/2}}{\Gamma(\alpha/2)} |\lambda|^{\alpha/2} \; K_{\alpha/2}(|\lambda|) \tag{30}$$

where $K_\nu(z)$ is a Bessel function.

Behavior of $\mathcal{G}(\lambda)$. It is also possible to calculate also $\mathcal{G}(\lambda)$ which is

$$\mathcal{G}(\lambda) = \frac{i}{\sqrt{\pi}\Gamma(\alpha/2)} [\Gamma(1+\frac{\beta}{2})\Gamma(\frac{\alpha-\beta-1}{2})\lambda \;\; {}_1F_2(1+\frac{\beta}{2}; \frac{3}{2}, \frac{3+\beta-\alpha}{2}; \frac{\lambda^2}{4}) +$$
$$2\Gamma(\beta-\alpha)\Gamma(\frac{1+\alpha}{2}) \sin(\frac{\pi}{2}(\beta-\alpha)) sign(\lambda)|\lambda|^{\alpha-\beta} \;\; {}_1F_2(\frac{1+\alpha}{2}; \frac{1-\beta+\alpha}{2}, 1+\frac{\alpha-\beta}{2}; \frac{\lambda^2}{4})] \tag{31}$$

where ${}_1F_2()$ are hypergeometric functions.

The main problems due to finite size comes from the range $\gamma \leq 2$ therefore we focus on this range. According to the theory above the aggregate impact is the ratio between two functions. Let us consider them separately. The function $P_{N+1}(V)$ in the denominator cannot be calculated explicitly. However by using the series expansion of the Bessel function we obtain that for small values of λ the function $h(\lambda)$ behaves as

$$h(\lambda) = \log(\mathcal{H}(\lambda)) \simeq \frac{-\pi \csc(\alpha\pi/2)}{2\Gamma(\alpha/2)} \left(\frac{\lambda^\alpha}{2^{\alpha-1}\Gamma(1+\alpha/2)} - \frac{\lambda^2}{2\Gamma(2-\alpha/2)} + ... \right) \tag{32}$$

Moreover in this case, when $\alpha \to 2$ the two terms in brackets become equal and cancel out.

3. Real time vs. transaction time

IV. AGGREGATION THEORY WITH TEMPORARY, LONG-MEMORY IMPACT

A. Failure of IID model in describing aggregate impact of financial data

In order to compare the IID theory with real data we need to have an estimate of the shape of the impact function $f(v)$ and of the volume distribution $\pi(v)$. The impact of individual transactions has been studied in many different markets (Lillo, Farmer, and Mantegna 2003,, Potters and Bouchaud 2003,Lillo and Farmer2004). In all the investigated markets the impact is an highly concave function and for large volumes the impact can be approximated by a power law function with $\beta < 1$. Also the volume distribution for individual transaction has been widely studied. In electronic markets it has been found that the volume distribution is asymptotically distributed as a power law with an exponent α larger than 3 (Lillo and Farmer 2004). These empirical results, which are valid also for the data under investigation here, indicate that the IID theory should fit the real data with $\kappa = 0$, because the value of the parameters lie in region I. This is not the case because, as said above real data shows that $\kappa > 0$. In the next section we will discuss the origin of the discrepancy between IID theory and real data. As we will see below there is a non trivial way to use IID theory to describe aggregate impact for real data.

The main reason of the failure of the IID theory in explaining aggregate impact for real data is that theory assumes that the flow of orders can be approximated by an independent identically distributed random process, whereas empirical analysis shows the presence of strong time correlations. In two recent papers (Bouchaud *et al.* 2004 ,Lillo and Farmer 2004) it has been shown that the process defined by the signs of market order volumes, $\epsilon_i \equiv sign(v_i)$ is a long memory process. This means that the autocorrelation function of ϵ_i decays in time as $E[\epsilon_{i+\tau}\epsilon_i] \sim \tau^{-\gamma}$, where $0 < \gamma < 1$. Long memory processes are an important class of stochastic process that have found application in many different fields. The autocorrelation function of a long memory process is not integrable in τ between 0 and $+\infty$ and, as a consequence, the process does not have a typical time scale. Long memory processes can be characterized by the exponent γ describing the asymptotic behavior of the autocorrelation function or equivalently in terms of the Hurst exponent H that, for long memory processes, is $H = 1 - \gamma/2$. The strong autocorrelation of ϵ_i is in contrast with the hypotheses of the IID theory and explain why the theory does not work for real data.

The long memory property of signs has another important consequence that contrasts the hypothesis of the IID model. The predictability of ϵ_i leads to a very intriguing paradox. If buying tends to push the price up and selling tends to push the price down, and we know that buying and selling are highly autocorrelated (and therefore predictable) how is it that price returns remain uncorrelated and unpredictable and the market linearly efficient? In other words a strongly autocorrelated sign process and a fixed and permanent impact would give rise to an highly predictable price change process. As a consequence of the long memory of signs one must abandon the hypothesis of fixed and permanent impact which is postulated in the IID theory.

In the next two sections we review the hypotheses which have been recently proposed for the origin of long memory sign and to solve the efficiency paradox.

B. Theory for the origin of long memory of signed order flow

In a recent paper Lillo, Mike, and Farmer (2005) have suggested that the long memory of ϵ_i can be caused by delays in market clearing. Under the common practice of order splitting, large orders are broken up into pieces and executed incrementally. These large orders are termed *packages* or *hidden orders*. Large investors avoid to reveal their intentionality of buying or selling large quantities of shares and split the hidden order in smaller pieces which are traded incrementally in the market. Lillo, Mike, and Farmer have proposed a model in which power law distribution of hidden order size is the origin of power law correlated signs of executed orders. More precisely, under the assumption that the size of hidden orders U is asymptotically distributed as $P(U > x) \sim x^{-\alpha}$ and that hidden orders are split in a number of revealed orders proportional to U they showed that the resulting order sign process is asymptotically power law correlated with an exponent $\gamma = \alpha - 1$, i.e.

$$\alpha = 3 - 2H \tag{33}$$

The empirical testing of the hypothesis of the theory is difficult because data on hidden order size are not easily available. In the original paper (Lillo, Mike, and Farmer 2005) authors tested indirectly the theory by using the dual structure of the London Stock Exchange. More recently, a comprehensive empirical analysis of the Spanish Stock Exchange in which hidden orders are statistically inferred from data has shown that both the proportionality between U and the number of revealed orders and the power law distribution of U are observed in real data (Vaglica *et al.* 2007). In this paper we will take the theory of (Lillo, Mike, and Farmer 2005) as given and we will use Eq. 33 to infer the value of α from the observed exponent of the autocorrelation of signs.

It is important to point out that the model assumes that the long memory of executed orders is due to the persistence in buying or selling of traders individually and not to a kind of synchronization or herding between different traders. Here we show that this crucial assumption of the model is correct. Our database for the London Stock Exchange contains the brokerage code for the buyer and the seller of each transaction. We therefore can compute the autocorrelation function of market order sign by considering either orders placed by the same brokerage code or by different brokerage codes. Our analysis shows that when only transactions with the same brokerage code are considered the autocorrelation is still power law with a slightly smaller exponent than in the case irrespective of the brokerage code. Moreover for a fixed lag the autocorrelation function with the same brokerage code is one order of magnitude larger than the autocorrelation function irrespective of the brokerage code. Finally, when only transactions with different brokerage code are considered the autocorrelation function decays very rapidly to zero and it is clearly not consistent with a power law behavior. This indicates that the long memory of signs is due to the presence of investors that place many revealed orders of the same sign in order to execute large hidden orders and that there is no clear sign of herding behavior among different investors.

C. Reconciling efficiency and long-memory: Temporary price propagator vs. asymmetric liquidity

Explain how long-memory is resolved by fluctuating liquidity. Equivalence of our approach to that of Bouchaud et al. Need to replace permanent impact propagators by temporary propagators. As mentioned above the long memory of signs ϵ_i leads to the puzzle

that a fixed and permanent market impact would lead to an highly predictable price change at odds with what seen in markets. Two different solution to this puzzle has been proposed, one due to Lillo and Farmer (2004) and the other due to Bouchaud *et al.* (2004). Lillo and Farmer suggested that the efficiency puzzle is explained by permanent price impacts that fluctuate in size. These fluctuations are liquidity fluctuations and are dependent on the predictability of market order signs. To make a concrete example, if the recent past history of signs suggests that the next market order is going to be a buy, then price impact for buy market orders is decreased and for sell market orders is increased in such a way that the expected return is zero Bouchaud *et al.* suggested that efficiency is recovered by having a decaying price impact with fixed size. They state that impacts are on average fixed in size $\epsilon_i f(|v_i|)$, but vary i time with the propagator $G_0(\tau)$, where τ is the time since transaction i occurred and it is measured in transaction time. The total price impact measured at the time of transaction i is

$$r_i = G_0(1)\epsilon_i f(|v_i|) - \sum_{k>0}[G_0(k+1) + G_0(k)]\epsilon_{i-k} f(|v_{i-k}|) + \eta_i \tag{34}$$

They find that $G_0(\tau)$ decays asymptotically as a power law, and it is tuned in such a way that it cancels the effect of the autocorrelation of ϵ_i so that returns remain unpredictable. They show that the model gives uncorrelated returns if $G_0(\tau) \sim \tau^{-\lambda}$ with $\lambda = (1-\gamma)/2$.

The two theories for reconciling the long memory of order flow with market efficiency gives in general different predictions for the aggregate impact. Although it can be shown that also asymmetric liquidity theory leads to a decaying propagator (Farmer, Gerig, and Lillo, in preparation), this does not imply that the two theories are equivalent. In the following two sections we present two heuristic derivation of the scaling of the linear part of the aggregate impact for the two theories. A complete theory for the aggregate impact under the two theories is very difficult and it is outside the scope of this paper. We used numerical simulation to test the heuristic arguments and to investigate the overall shape of the aggregate impact. In Section V we present a comparison of heuristic arguments and numerical simulations with real data.

D. Aggregation theory for the propagator model

The propagator model makes the calculation of the aggregate impact much more difficult mainly due to the non-local character of the impact of individual transactions. Instead of develop a theory for aggregate impact for the propagator model, we developed an heuristic argument giving a prediction of the value of κ and then we tested this argument by extensive numerical simulations. The heuristic argument for computing κ is the following. Due to the propagator, a transaction at time u has an impact at a subsequent time $t > u$ given by $G_0(t-u) \sim (t-u)^{-\lambda}$, where, as said above, $\lambda = (1-\gamma)/2$ to ensure efficiency. By assuming incremental trading and by taking the continuous limit we obtain that the return due to transactions between 0 and t is

$$R \sim \int_0^t (t-u)^{-\lambda} du \sim t^{1-\lambda}$$

We now hypothesize that $R(V,N)$ scales in the same way as in the IID theory, i.e. $R \sim VN^{-\kappa}$. In order to determine κ for this model we assume proportional trading, i.e. that

$V \sim N$, which means $R \sim N^{1-\kappa}$. If we finally assume that the time interval is proportional to the number of transactions we obtain $1-\kappa = 1-\lambda$, i.e.

$$\kappa = \lambda = H - 1/2 \tag{35}$$

We perform extensive numerical simulations to support this heuristic argument for the relation between κ and H. Specifically, we simulated long series of market order flows by generating long memory correlated signs $\epsilon_i = sign(v_i)$ with different values of H. We generate uncorrelated volumes $|v_i|$ drawn from different distributions with finite variance. In the following we show the results for the case in which volumes $|v_i|$ are taken from a distribution asymptotically distributed as a power law with a tail exponent 3 (in the cumulative). We then assume a deterministic price impact function of the form $\epsilon_i |v_i|^\beta$, with a value beta smaller than 1. Finally we applied the Bouchaud propagator. We proved that the simulated returns are linearly efficient and we computed the aggregate volume for different values of transactions N. Also for the propagator model the aggregate impact is linear around $V = 0$ and the region where the linear approximation holds increases with N. For each value of N we performed a best linear fit of the slope of the aggregate impact around $V = 0$. Finally, we plotted the slope versus N and we estimate the exponent κ. The result of this analysis is summarized in Figure 5 where we considered 4 different values of the Hurst exponent H of the sign time series. In all cases the slope scales as a power law of N and the estimated value of the exponent κ is plotted in the inset of Figure 5 as a function of H. The relation $\kappa = H - 1/2$ fits very well the simulated data supporting the heuristic argument above.

E. Aggregation model for the asymmetric liquidity model

We have seen above that the IID theory is not suitable for describing the real data because the order flow is not and IID process. We have also discussed the fact that in order to maintain efficiency the impact of individual transaction cannot be a permanent and fixed function of the transaction volume. Thus the IID theory developed in Section III seems to be of little use for financial data. In this Section we show that one can reinterpret the IID theory in order to make quantitative predictions of the scaling properties of impact.

The main idea for this reinterpretation is to use the theory for the origin of the long-memory property of order flow developed in Lillo, Mike, and Farmer (2005) and described in Section IV B. Given a series of N revealed orders we can think to it as a series of $N_h < N$ hidden orders. The size of the hidden orders is drawn from a distribution with a power-law tail with exponent α which is related to the Hurst exponent of the transaction sign as in Eq. 33. As in the theory of Lillo, Mike, and Farmer (2005) we assume that the hidden order sign is an IID process. Therefore a correlated revealed order flow is interpreted as an IID hidden order flow. In order to use the IID theory we need (i) a relation between the number N of revealed orders and the number N_h of hidden orders in the same interval and (ii) a functional form $f(V)$ describing the impact of an hidden order of size V.

The number N_h of hidden orders active in an interval of N revealed orders is a random variable depending on the distribution of hidden order size. Here we make a "mean-field approximation" consisting in assuming that $N_h \simeq N/\bar{V}$, where $\bar{V}$ is the mean size of an hidden orders. Here size means number of revealed transactions. For example if $P(V)$ is Pareto distributed as $P(V) = \alpha/V^{\alpha+1}$ then $\bar{V} = \alpha/(\alpha-1)$ and $N_h \simeq (\alpha-1)N/\alpha$. In any case the important point of this assumption is that N_h and N are proportional.

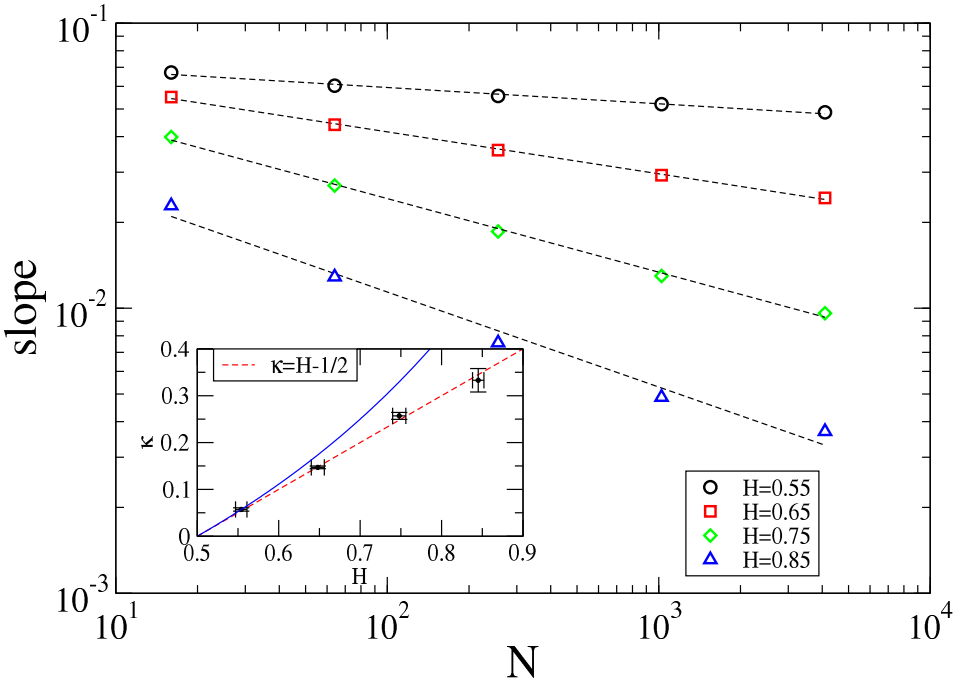

FIG. 5: Plot of simulated slope vs. N for several different H, and inset showing κ vs. H.

The functional form of the impact of hidden orders has been investigated theoretically and empirically in many papers. There is no consensus on the form and on the determinants of it. Many studies have shown that the impact of hidden orders is a concave function of the size, even if some theoretical studies conjecture that the impact should be a linear function **Farmer 2007, Farmer 2007b, Almgren-risk**. Some studies have suggested a power law function V^β with an exponent smaller than one whereas other studies conjectured a logarithmic functional form.

Given the value of α and β, it is possible to use the IID theory developed above and use the result of Fig. 4 to compute the values of the scaling exponent κ. In fact,

$$R(V,N) \sim R(V,N_h) \sim \frac{V}{N_h^\kappa} \sim \frac{V}{N^\kappa}, \tag{36}$$

where the last equality holds because of the proportionality between N and N_h.

The long memory of order flow suggest that $1 \leq \alpha \leq 2$ and if we assume that β is small enough, the scaling exponent κ is $(2-\alpha)/\alpha$ because the parameters are those of region II of Fig. 4. In this region the specific value of β does not affect the value of κ. Since α can be determined from Eq. 33 we get

$$\kappa = \frac{2-\alpha}{\alpha} = \frac{2H-1}{3-2H}. \tag{37}$$

This equation makes a quantitative prediction on the relation between two directly measurable quantities: the Hurst exponent H of the order flow and the scaling exponent κ of the

aggregate impact.

V. COMPARISON OF THEORY TO EMPIRICAL DATA

A. 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). AZN, LLOY, and VOD are among the most heavily traded names for the trading volume for each stock. The number of transactions in the investigated period is $569,321$ (AZN), $359,479$ (BSY), $599,739$ (LLOY), $392,020$ (PRU), $213,474$ (RTO), and $1,047,833$ (VOD). The data is from the on-book exchange (SETS) only, and is for the period from May 2000 to December 2002. This data contains a complete record of all order placements, so we are able to determine the signs of order unambiguously.

B. Testing the theory

In order to compare the empirical aggregate impact with the one predicted by the two theories we consider two aspects. Specifically, we first investigate how the slope of the linear part of the impact scales with N, i.e. we try to discriminate whether the empirical exponent κ scales as predicted by the Bouchaud model (Eq. 35) or as predicted by the model with permanent impact and IID order flow (Eq. 37). Secondly, we perform numerical simulations of the two models and we compare the aggregate impact of the simulations with that of the real data. The comparison is qualitative and visual.

We consider the scaling of the exponent κ with N. For the investigated stocks we computed the aggregate impact at different values of N and we fit the region close to $V = 0$ with a straight line. In Figure 6 we show in a double logarithmic plot the slope of the linear part as a function of N for the six stocks. For each stocks the points corresponding to different values of N stays on a straight line confirming that the slope of the linear part of aggregate impact scales as $N^{-\kappa}$ with $\kappa > 0$. Regressing the points in Figure 6 we obtain for each stock a value of κ. In the inset of Figure 6 we show the scatter plot of the estimated κ versus the Hurst exponent of the transaction order sign. The inset also shows the two lines expected under the Bouchaud theory (Eq. 35, dashed line) and under the permanent impact and IID order flow model (Eq. 37, solid line). For five of the six stocks the data are in agreement with the permanent impact model whereas data for Vodafone are in agreement with Bouchaud model. Even if this result indicates that permanent impact model fits better the data than Bouchaud model we believe that this data should be taken with caution. First our empirical analysis is based on a small set of stocks. Second it is not easy to assign proper error bars to the estimated parameters. Finally, our theory is asymptotic and it is not obvious how to take into account finite size effects, given also the small difference between the expected exponent κ under the two theories.

In order to have a more detailed view of how theories describe aggregate data we perform numerical simulations of the models. To test how well Bouchaud theory described aggregate

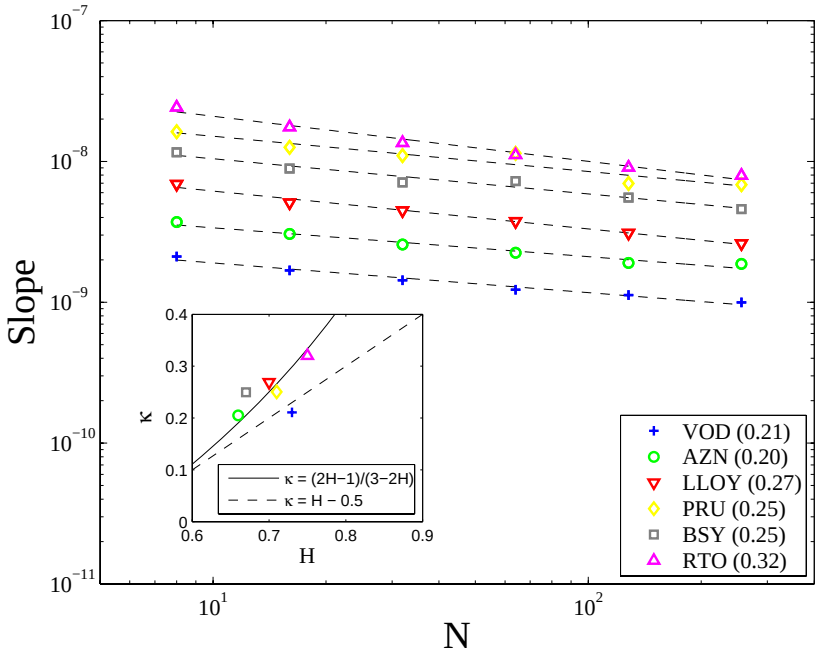

FIG. 6: Show some plots of slope vs. N.

impact we consider the actual order flow of real data and we apply the propagator for generating a surrogate time series of prices. For the impact of individual transactions we used the average impact computed unconditionally on all the transaction of the sample. Figure 7 shows the comparison between the aggregate impact of AstraZeneca and the impact obtained from these surrogate return time series. We see that the agreement between simulations and real data is reasonably good until $N = 512$. At this aggregation scale the surrogate impact is slightly steeper close to $V = 0$ and a little bit more concave than the real data.

C. Determinants of deviations from linearity in aggregate impact

Here we present the tests that Austin did to understand what influences the deviations from nonlinearity in the impact curves. This is described on page 73 of Austin's notes, 10.09.06 - 10.12.06. We identified seven different factors influencing the price impact curves, and did some empirical tests to determine the importance of each factor. Austin, give me the tex file for your notes and I will paste this in. Also, we need to discuss the right way to present this.

VI. CONCLUSIONS

Acknowledgments

We would also like to thank J-P. Bouchaud and Marc Potters for useful discussions. We would like to thank Barclay's Bank for supporting this research. Acknowledge NSF grant. FL acknowledges support from the research project MIUR 449/97 "High frequency dynamics

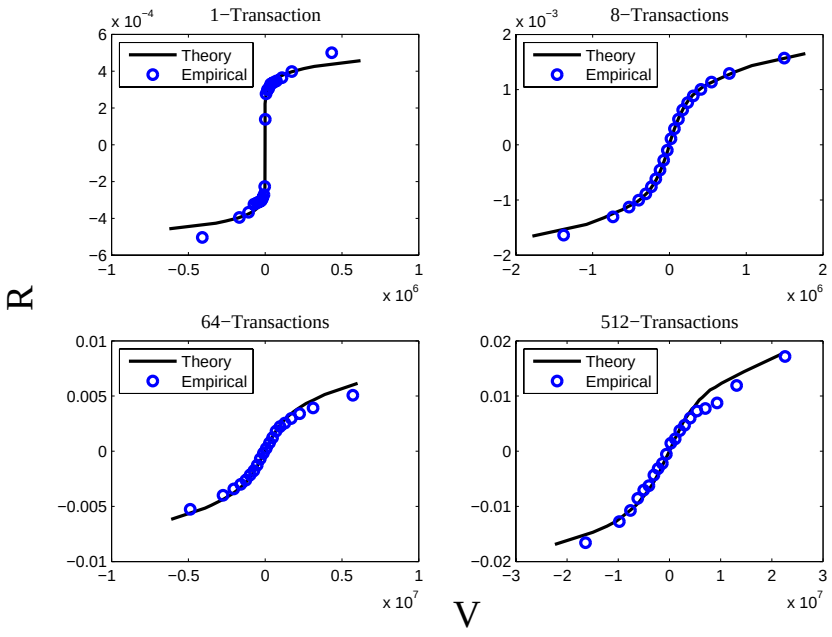

FIG. 7: Aggregate impact vs. volume for theory and real data. Can present theory in two different ways: First, we may want to use unconditional prediction from previous section (as Fabrizio has simulated), second use propagators based on actual v_i sequence (as Austin has computed).

in financial markets" and from the European Union STREP project n. 012911 "Human behavior through dynamics of complex social networks: an interdisciplinary approach.".

APPENDIX A: SADDLE POINT APPROXIMATION AND ASYMPTOTIC EXPANSION

The key result of this notes is Eq. 18 where we need to calculate the integral

$$I_n \equiv \int_{-\infty}^{+\infty} d\lambda e^{nh(\lambda)} g(\lambda)$$

because by applying saddle point approximation one can obtain the asymptotic behavior of this integral for large n. Here I review the saddle point approximation by following F.W.J. Olver, *Asymptotics and Special Functions*, page 80-81. First note that if $\pi(v)$ is even and the individual impact function $f(v)$ is an odd function, then $\mathcal{G}(\lambda)$ is purely imaginary and odd. Thus the only contribution to I_n comes from the real part of $g(\lambda)$ which is $\mathcal{G}(\lambda)$ $[-i\sin(\lambda V)]$ which is even. Finally, under the assumption that $\pi(v)$ is even, also $h(\lambda)$ is even and then we can write the integral as

$$I_n = -2i \int_0^{+\infty} d\lambda e^{nh(\lambda)} \mathcal{G}(\lambda)\ \sin(\lambda V) \equiv \int_0^{+\infty} d\lambda e^{nh(\lambda)} \tilde{g}(\lambda) \tag{A1}$$

For large values of n the asymptotic behavior of I_n is determined by the regions of λ where $h(\lambda)$ has a maximum. Since $h(\lambda)$ is the characteristic function of $\pi(v)$, it has a maximum

for $\lambda = 0$. Laplace's method in asymptotics prescribes that if for λ close to zero it is

$$h(\lambda) \simeq P\lambda^{\zeta_h} \qquad \tilde{g}(\lambda) \simeq Q\lambda^{\zeta_g} \tag{A2}$$

(and some other regularity conditions, see Olver) then

$$I_n \sim \frac{Q}{\zeta_h}\Gamma(\frac{\zeta_g+1}{\zeta_h})\frac{1}{(-P\ n)^{\frac{\zeta_g+1}{\zeta_h}}} \tag{A3}$$

Finally note that, since $\tilde{g}(\lambda) = -2i\mathcal{G}(\lambda)\ \sin(\lambda V)$, for small λ we have $\tilde{g}(\lambda) \propto V$, i.e. $I_n \propto Q \propto V$ independently on the details of the volume distribution and impact function.

Here I describe the asymptotic expansion of the integral

$$\int_a^b e^{-xp(t)}q(t)\ dt \tag{A4}$$

as described in F.W.J. Olver, *Asymptotics and Special Functions*, page 85-86 (Note that I use ν for the exponent of $q(t)$ whereas Olver uses λ, that in this notes has a different meaning).

Assume that $p(t)$ has a minimum at $t = a$ and that one can expand

$$p(t) = p(a) + \sum_{s=0}^{\infty} p_s(t-a)^{s+\mu} \tag{A5}$$

$$q(t) = \sum_{s=0}^{\infty} q_s(t-a)^{s+\nu-1} \tag{A6}$$

then under some regularity condition

$$\int_a^b e^{-xp(t)}q(t)\ dt \sim e^{-xp(a)}\sum_{s=0}^{\infty}\Gamma\left(\frac{s+\nu}{\mu}\right)\frac{a_s}{x^{(s+\nu)/\mu}} \tag{A7}$$

The coefficients a_s can be calculated and Olver gives the explicit expression for the first three,

$$a_0 = \frac{q_0}{\mu p_0^{\nu/\mu}} \qquad a_1 = \left(\frac{q_1}{\mu} - \frac{(\nu+1)p_1q_0}{\mu^2 p_0}\right)\frac{1}{p_0^{(\nu+1)/\mu}} \tag{A8}$$

$$a_2 = \left[\frac{q_2}{\mu} - \frac{(\nu+2)p_1q_1}{\mu^2 p_0} + [(\nu+\mu+2)p_1^2 - 2\mu p_0 p_2]\frac{(\nu+2)q_0}{2\mu^3 p_0^2}\right]\frac{1}{p_0^{(\nu+2)/\mu}} \tag{A9}$$