11

The Impact of Market Orders

To measure is to know. (Lord Kelvin)

In this chapter, we address the seemingly obvious notion of **price impact** (which we first discussed in Section 1.3.2): buy trades tend to push the price up and sell trades tend to push the price down. Expressed in the notation and language that we have subsequently developed, we might also express this notion by saying that price impact refers to the positive correlation between the sign (+1 for a buy order and −1 for a sell order) of an incoming market order and the subsequent price change that occurs upon or after its arrival. As we discuss in Section 11.3, these seemingly obvious statements are indeed verified by empirical data.

Price impact is an all-too-familiar reality for traders who need to buy or sell large quantities of an asset. To these traders, price impact is tantamount to a cost, because the impact of their earlier trades makes the price of their subsequent trades worse on average. Therefore, monitoring and controlling impact costs is one of the most active and rapidly expanding domains of research in both academic circles and trading firms.

Understanding and assessing price impact entails considering two different but related topics. The first is how *volume* creates impact: how much more impact does a larger trade cause? The second is the *temporal behaviour* of impact: how much of a trade's impact is permanent, how much decays over time, and how does this transient behaviour unfold? We will consider both of these topics in detail throughout this chapter. As we will discuss, both of these aspects of price impact are far from trivial. Is a transaction not a fair deal between a buyer and a seller? If so, which of the two is really impacting the price?

11.1 What Is Price Impact?

In much of the existing literature, there are two strands of interpretation for price impact, which reflect the great divide between efficient-market enthusiasts (who

believe that the price is always close to its fundamental value) and sceptics (who believe that the dynamics of financial markets is primarily governed by order flow). At the two extremes of this spectrum are the following stories:

(i) *Agents successfully forecast short-term price movements, and trade accordingly.* This is the **efficient-market point of view**, which asserts that a trader who believes that the price is likely to rise will buy in anticipation of this price move. This clearly results in a positive correlation between the sign of the trade and the subsequent price change(s), even if the trade by itself has no effect on prices. In this framework, a noise-induced trade that is based on no information at all should have no long-term impact on prices – otherwise, prices could end up straying very far from their fundamental values, which cannot be the case if markets are efficient. By this interpretation, if the price was meant to move due to information, it would do so even *without* any trades.

(ii) *Price impact is a reaction to order-flow imbalance.* This is the **efficient-market sceptic view**, which asserts that the fundamental value is irrelevant, at least on short time scales, and that even if a trade reflected no information in any reasonable sense, then price impact would still occur.
As an illustration of this viewpoint, recall the Santa Fe model (see Chapter 8), in which all order-flow events are described by independent, homogeneous Poisson processes. All else being held constant, then the mid-price will on average be higher (respectively, lower) conditional on the arrival of an extra buy (respectively, sell) market order than it would be conditional on that market order not arriving. This effect is readily apparent in Figure 11.2, where we plot the mean impact of a buy trade in the Santa Fe model. Clearly, there is a well-defined and measurable price impact, even though there is no notion of fundamental price or information in this zero-intelligence model.

Although both of the above explanations result in a positive correlation between trade signs and price movements, they are conceptually very different.[1] In the first story, as emphasised by J. Hasbrouck, *"orders do not impact prices. It is more accurate to say that orders forecast prices."*[2] Put another way, trades reveal private information about the fundamental value, creating a so-called **price discovery** process. In the second story, the act of trading itself impacts the price. In this case, one should remain agnostic about the information content of the trades, and should therefore speak of **price formation** rather than price discovery. If market

[1] On this point, see Lyons, R. (2001). *The microstructure approach to exchange rates*. MIT Press. Lyons writes: *Consider an example that clarifies how economist and practitioner worldviews differ. The example is the timeworn reasoning used by practitioners to account for price movements. In the case of a price increase, practitioners will assert "there were more buyers than sellers". Like other economists, I smile when I hear this. I smile because in my mind the expression is tantamount to the "price had to rise to balance demand and supply".*

[2] Hasbrouck, J. (2007). *Empirical market microstructure.* Oxford University Press.

participants believe that the newly established price is the "right" price and act accordingly, "information revelation" might simply be a self-fulfilling prophecy.

As mentioned above, the Santa Fe model (see Chapter 8) provides an illustration of the second story. In this model, the mechanism that generates impact can be traced back to the modelling assumption that at any given time, agents submitting orders always use the current mid-price as a reference. Any upwards (respectively downwards) change in mid-price therefore biases the subsequent order flow in an upwards (respectively downwards) direction. This causes the model to produce a diffusive mid-price in the long run, but only resulting from the permanent impact of a purely random order flow, in a purely random market.

Whether prices are formed or discovered remains a topic of much debate. At this stage, there is no definitive answer, but because trades in modern markets are anonymous, and because the line between real information and noise is so blurry, reality probably lies somewhere between these two extremes. Since some trades may contain real private information, and since other market participants do not know which trades do and do not contain such information, it follows that all trades must (on average) impact the price, at least temporarily. The question of how much real information is revealed by trades is obviously crucial in determining whether markets are closer to the first picture or the second picture. Several empirical results suggest that the impact of random trades is similar to that of putative informed trades (at least on the short run), and that the amount of information per trade is extremely small (see the discussions in Chapters 13, 16 and 20).

11.2 Observed Impact, Reaction Impact and Prediction Impact

From a scientific (but slightly ethereal) point of view, one would ideally like to assess the impact of a market order by somehow measuring the difference between the mid-price in a world where the order is executed and the mid-price in a world where all else is equal but where the given order is not executed. For a buy market order, for example,

$$\mathcal{I}^{\text{react.}}_{t+\ell}(\text{exec}_t \mid \mathcal{F}_t) := \mathbb{E}[m_{t+\ell} \mid \text{exec}_t, \mathcal{F}_t] - \mathbb{E}[m_{t+\ell} \mid \text{no exec}_t, \mathcal{F}_t], \tag{11.1}$$

where "exec$_t$" and "no exec$_t$" denote, respectively, the execution or non-execution of the market order at time t. We call this quantity the **reaction impact**, because it seeks to quantify how the market price reacts to the arrival of a given order. In this formulation, $\mathcal{F}_t$ represents the state of the world at time t. In particular, $\mathcal{F}_t$ contains all information that may have triggered the given buy order, but not whether the trade is executed or not. To aid readability, we will sometimes omit the conditioning on $\mathcal{F}_t$, but it is always implicitly present in our arguments.

The definition of reaction impact in Equation (11.1) is close in spirit to what natural scientists would like to consider: an experiment where the system is perturbed in a controlled manner, such that the result of that perturbation can be cleanly observed and quantified. Unfortunately, this definition cannot be implemented in a real financial system, because the two situations (i.e. the market order arriving or not arriving) are mutually exclusive, and history cannot be replayed to repeat the experiment in the very same conditions. Instead, what can be measured in a real financial market is the **observed impact**:

$$\mathcal{I}^{\text{obs.}}_{t+\ell}(\text{exec}_t) := \mathbb{E}\left[m_{t+\ell} \mid \text{exec}_t\right] - m_t, \tag{11.2}$$

where m_t is the observed mid-price just before the execution occurred.

Most studies of impact focus on measuring and studying observed impact, which is readily available ex-post in a given data set. If prices were martingales, then it would follow that $\mathbb{E}\left[m_{t+\ell} \mid \text{no exec}_t, \mathcal{F}_t\right]$ is equal to m_t, so observed impact would be precisely equal to reaction impact. In real markets, however, this equality does not hold (because $\mathcal{F}_t$ contains the information available to the trader, so one should expect the price to increase on average even in the absence of his or her trade, as the prediction motivating the trade is revealed). The amount of information contained in $\mathcal{F}_t$ can be written as:

$$\mathcal{I}^{\text{pred.}}_{t+\ell} := \mathbb{E}\left[m_{t+\ell} \mid \text{no exec}_t, \mathcal{F}_t\right] - m_t. \tag{11.3}$$

We call this quantity the **prediction impact**, in the spirit of Hasbrouck's view (recalled above).[3]

By Equations (11.1) and (11.2), the difference between observed impact and reaction impact is given by prediction impact:

$$\begin{aligned}\mathcal{I}^{\text{obs.}}_{t+\ell}(\text{exec}_t) - \mathcal{I}^{\text{react.}}_{t+\ell}(\text{exec}_t) &= \mathbb{E}\left[m_{t+\ell} \mid \text{exec}_t\right] - m_t - \mathbb{E}\left[m_{t+\ell} \mid \text{exec}_t\right] \\ &\quad + \mathbb{E}\left[m_{t+\ell} \mid \text{no exec}_t\right], \\ &= \mathbb{E}\left[m_{t+\ell} \mid \text{no exec}_t\right] - m_t.\end{aligned}$$

To recap: we have introduced three types of impact – namely, observed impact, reaction impact, and prediction impact – which are related by the equality:

$$\mathcal{I}^{\text{obs.}} = \mathcal{I}^{\text{react.}} + \mathcal{I}^{\text{pred.}}. \tag{11.4}$$

Prediction impact is very difficult to estimate empirically, because the full information set $\mathcal{F}_t$ used by market participants to predict future prices is extremely

[3] Note that we prefer here the term "prediction" to the term "information", to avoid any confusion with "fundamental information". The latter term suggests some knowledge of the fundamental price of the asset, whereas we prefer the agnostic view that some market participants successfully predict the future evolution of prices, whether or not this is justified by fundamentals. For an extended discussion on this point, see Chapter 20.

large and difficult to quantify. Reaction impact is somewhat easier to estimate, via one of the following methods:

(i) by performing experiments where trading decisions are drawn at random, such that $\mathcal{I}^{\text{pred.}}_{t+\ell} = 0$ by construction (up to statistical noise);
(ii) by choosing at random whether or not to execute an order with a given prediction signal (so as to measure both $\mathbb{E}[m_{t+\ell} \mid \text{exec}_t, \mathcal{F}_t]$ and $\mathbb{E}[m_{t+\ell} \mid \text{no exec}_t, \mathcal{F}_t]$ for the same strategy but at different times t), then subtracting the latter from the former;
(iii) ex-post, by conditioning on the order-sign imbalance of the rest of the market between t and $t + \ell$, and using this imbalance as a proxy for the presence of informed trading, i.e. for whether $\mathcal{I}^{\text{pred.}}$ is zero or not.[4]

None of these approaches are perfect. To measure anything meaningful, the first idea requires generating a large number of random trades, which is a costly and time-consuming experiment! A handful of studies have attempted to determine $\mathcal{I}^{\text{react.}}$ empirically, either by actually performing random trades or by carefully mining existing data sets to identify specific orders for which $\mathcal{I}^{\text{pred.}}$ can be regarded to be zero (such as trades initiated for cash-inventory purposes only). These studies all conclude that *on short time scales*, the mechanical impact estimated from random trades is to a good approximation identical to the mechanical impact estimated from proprietary (allegedly informed) trades, or from all trades in a given data set.[5] This shows that the prediction component $\mathcal{I}^{\text{pred.}}$ (if any) is only expected to show up at longer times, when the prediction signal that initiated the trade is realised (see Chapter 20).

11.3 The Lag-1 Impact of Market Orders

Although all types of order-flow events (market order arrivals, limit order arrivals, and cancellations) can impact prices, it is conceptually and operationally simpler to first study only the impact of market orders (see Chapter 14 for an extended discussion of the other events). One reason for doing so is that this analysis requires only trades-and-quotes data (i.e. the time series of bid-prices b_t, ask-prices a_t and trade prices p_t). Because we consider only market orders, in the following we count time t in market-order time, in which we increment t by 1 for each market

[4] This was suggested in Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the Bitcoin. *Market Microstructure and Liquidity*, 1(02), 1550008.
[5] See Section 14.5.2 and, e.g., Gomes, C., & Waelbroeck, H. (2015). Is market impact a measure of the information value of trades? Market response to liquidity vs. informed metaorders. *Quantitative Finance*, 15(5), 773–793, and Tóth, B., Eisler, Z., & Bouchaud, J.-P. (2017). The short-term price impact of trades is universal. https://ssrn.com/abstract=2924029.

order arrival, and in which b_t and a_t denote the values of the bid- and ask-prices immediately *before* the arrival of the t^{th} market order.

In high-quality trades-and-quotes data, there is an exact match between the transaction price p_t and either the bid- or the ask-price at the same time. If $p_t = a_t$, then the transaction is due to an incoming buy market order (which we label with the order sign $\varepsilon_t = +1$); if $p_t = b_t$, then the transaction is due to an incoming sell market order (which we label with the order sign $\varepsilon_t = -1$). Each entry in a trades-and-quotes data set also specifies the trade volume v_t. If an incoming buy (respectively, sell) market order's size does not exceed the volume at the best ask (respectively, bid) quote, then v_t is the full size of the incoming market order. If the volume of the incoming market order exceeds the volume at the ask (respectively, bid) quote, further transactions will occur at higher (respectively, lower) prices, within the limits of the market order volume and price. Any unmatched part of the market order will remain as a limit order in the LOB, at the price at which it was sent. Trades-and-quotes data sets typically report such activity as different, successive transactions with identical or very similar time stamps.

11.3.1 Unconditional Impact

The simplest measure of price impact is the mean difference between the mid-price just before the arrival of a given market order and the mid-price just before the arrival of the next market order.[6] To align activity for buy and sell market orders, our general definition of impact must also incorporate the order sign ε_t. Recalling that m_t denotes the mid-price immediately before the arrival of the t^{th} market order, we define the **lag-1 unconditional impact** as

$$\mathcal{R}(1) := \langle \varepsilon_t \cdot (m_{t+1} - m_t) \rangle_t, \tag{11.5}$$

where the empirical average $\langle \cdot \rangle_t$ is taken over all market orders regardless of their volume and regardless of the state of the world (including the LOB) just before the transaction. We could of course perform more precise measurements of price impact in specific situations by also conditioning on extra variables, but for now we consider the general definition in Equation (11.5).

Table 11.1 lists several statistics related to price impact. For all stocks, it is clear that $\mathcal{R}(1) > 0$ with strong statistical significance.[7] This demonstrates that

[6] The choice of the mid-price m_t as the relevant reference price is the simplest, but is not necessarily the most adequate. For large-tick stocks, in particular, we have seen in Section 7.2 that the volume imbalance I is a strong predictor of the sign of the future price change, so a better reference price could be defined as $\widetilde{m}_t = p_+(I)a_t + p_-(I)b_t$. Throughout the book, however, we stick with the mid-price m_t.

[7] Indeed, from the last column of Table 11.1, the total number of events used to compute $\mathcal{R}(1)$ is of the order of 10^6, which leads to a relative standard error of less than 1%.

Table 11.1. *The average spread just before a market order, $\langle s \rangle$; the lag-1 response functions, $\mathcal{R}(1)$ (all market orders), $\mathcal{R}^1(1)$ (price-changing market orders) and $\mathcal{R}^0(1)$ (non-price-changing market orders); the standard deviation of price fluctuations around the average price impact of a market order, $\Sigma_{\mathcal{R}} = \sqrt{\mathcal{V}(1) - \mathcal{R}(1)^2}$, all measured in dollar cents; the fraction of market orders that immediately change the price, $P[\mathrm{MO}^1]$; and the number of market orders observed between 10:30 and 15:00 during each trading day, N_{MO}, for 10 small- and large-tick stocks during 2015.*

	$\langle s \rangle$	$\mathcal{R}(1)$	$\mathcal{R}^1(1)$	$\mathcal{R}^0(1)$	$\Sigma_{\mathcal{R}}$	$P[\mathrm{MO}^1]$	N_{MO}
SIRI	1.06	0.058	0.516	0.006	0.213	0.112	623
INTC	1.08	0.246	0.769	0.029	0.422	0.293	4395
CSCO	1.09	0.206	0.735	0.022	0.386	0.256	3123
MSFT	1.09	0.276	0.769	0.039	0.441	0.322	7081
EBAY	1.10	0.348	0.745	0.059	0.502	0.419	3575
FB	1.21	0.481	0.818	0.124	0.674	0.514	10703
TSLA	12.99	2.59	3.79	0.403	4.49	0.649	3932
AMZN	21.05	3.63	5.57	0.597	6.38	0.618	4411
GOOG	26.37	3.97	6.65	0.644	7.62	0.557	3710
PCLN	94.68	15.30	24.97	2.17	28.77	0.579	1342

a buy (respectively, sell) market order is on average followed by an immediate increase (respectively, decrease) in m_t. We also point out three other interesting observations:

(i) For small-tick stocks, the value of $\mathcal{R}(1)$ is proportional to the mean spread $\langle s \rangle_t$ (see also Figure 11.1). In other words, the scale of mid-price changes induced by market order arrivals is of the same order as the bid–ask spread s_t. This turns out to capture a profound truth that we already alluded to in Section 1.3.2, and on which we will expand in Chapter 17 below. In a nutshell, this linear relation follows from the argument that market-making strategies must be roughly break-even: market-makers attempt to earn the bid–ask spread s, but face impact costs due to the adverse price move after a market order. The relation $\mathcal{R}(1) \propto \langle s \rangle_t$ means that, to a first approximation, adverse selection is compensated by the spread (see Chapters 16 and 17 for an extended discussion of this point).

(ii) For large-tick stocks, the average spread is bounded below by one tick, so the linear relationship saturates. However, $\mathcal{R}(1)$ itself is not bounded, because the proportion of trades that result in a one-tick price change may become arbitrarily small, resulting in a small average impact. In this situation, the market-making problem becomes more subtle and requires study of the full queuing systems (see Chapter 17).

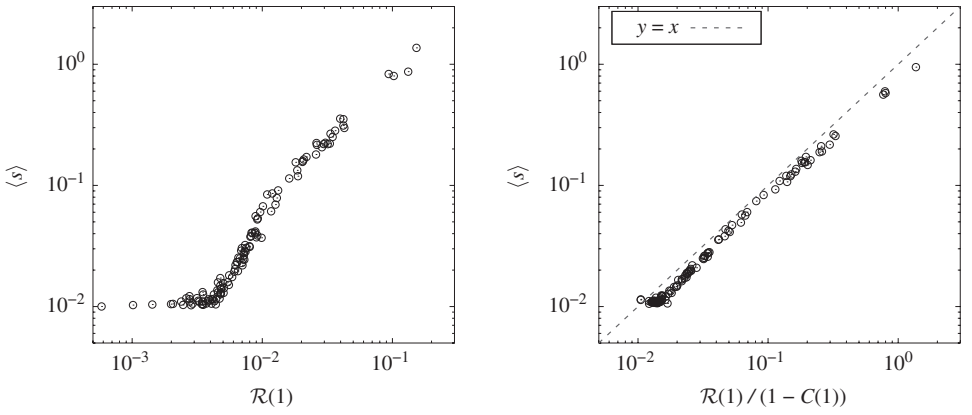

Figure 11.1. (Left panel) Average spread $\langle s \rangle$ versus the lag-1 impact $\mathcal{R}(1)$, for 120 stocks traded on NASDAQ, and (right panel) similar plot but with $\mathcal{R}(1)$ adjusted for the lag-1 correlation $C(1)$ in the market order flow (see Chapter 17 for a theoretical justification).

(iii) There is a substantial amount of noise around the mean impact $\mathcal{R}(1)$. One way to measure this dispersion is to calculate the lag-1 variogram of the mid-price:

$$\Sigma_{\mathcal{R}} := \mathcal{V}(1) - [\mathcal{R}(1)]^2; \qquad \mathcal{V}(1) = \langle (m_{t+1} - m_t)^2 \rangle.$$

As is clear from Table 11.1, the magnitude of the fluctuations is larger than the mean impact itself. As noted in (ii) above, some market orders do not change the price at all, others trigger large cancellations and hence have a very large impact, and some are even followed by a price change in the opposite direction! This highlights that $\mathcal{R}(1)$ does not simply measure the simple mechanical effect of the market order arrival, but instead incorporates the full sequence of other limit order arrivals and cancellations that occur between two successive market order arrivals.

There is no reason to limit our definition of price impact to the lag-1 case. For any $\ell > 0$, we can easily extend the definition from Equation (11.5) to the general case:

$$\mathcal{R}(\ell) := \langle \varepsilon_t \cdot (m_{t+\ell} - m_t) \rangle_t. \tag{11.6}$$

The function $\mathcal{R}(\cdot)$ is called the **response function**.

Figure 11.2 shows the shape of the response function for four stocks in our sample. In each case, $\mathcal{R}(\ell)$ rises from an initial value $\mathcal{R}(1)$ to a larger value $\mathcal{R}_\infty = \mathcal{R}(\ell \to \infty)$, which is 2–5 times larger than the initial response $\mathcal{R}(1)$. This occurs as a result of the autocorrelations in market order signs, which tend to push the price in the same direction for a while (see Section 13.2.1). This illustrates an important point, which we will return to in the next chapter: one should not confuse the response function with the mechanical impact of an isolated random trade. In the

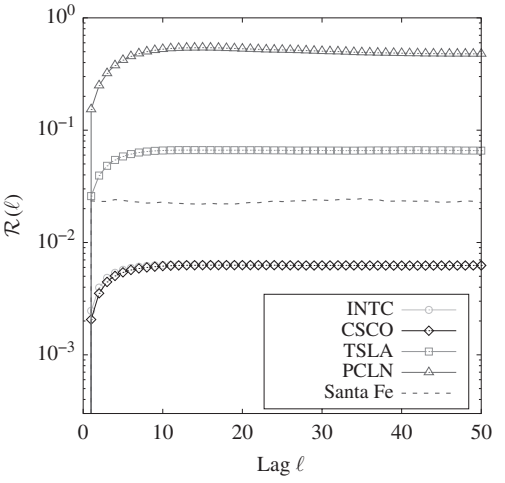

Figure 11.2. Response function $\mathcal{R}(\ell)$ in market order event-time, for (circles) INTC, (diamonds) CSCO, (squares) TSLA, (triangles) PCLN and (dashed curve) in the simulated Santa Fe model. The parameters of the Santa Fe model correspond to TSLA.

language of the previous section, there is an information contribution $\mathcal{I}^{\text{pred.}}$ that comes from the autocorrelation of the market order signs. Put another way, the response function $\mathcal{R}(\ell)$ also contains the reaction impact of future trades, which, as we saw in Chapter 10, are correlated with the present trade.

11.3.2 Conditioning on Trade Volume

So far, we have considered the impact of a market order irrespective of its volume. However, it seems natural that large market orders should somehow impact prices more than small market orders. As we discussed in Section 10.5.1, one possible reason that this should be the case is information leakage: if market orders reveal information, larger trades may indeed lead to larger subsequent price moves. Another possible reason is the purely statistical observation that a larger market order is more likely to consume all the available volume at the opposite-side best quote, and is therefore more likely to lead to a price change both instantaneously (by directly changing the state of the LOB) and subsequently (by causing other traders to modify their subsequent order flow).

Both of these intuitive arguments are indeed confirmed by empirical data. However, the effect of volume on impact is much weaker than might be naively anticipated. In fact, after normalising the volume of a market order v by the mean volume at the opposite-side best quote $\bar{V}_{\text{best}}$, one finds that the volume-dependence

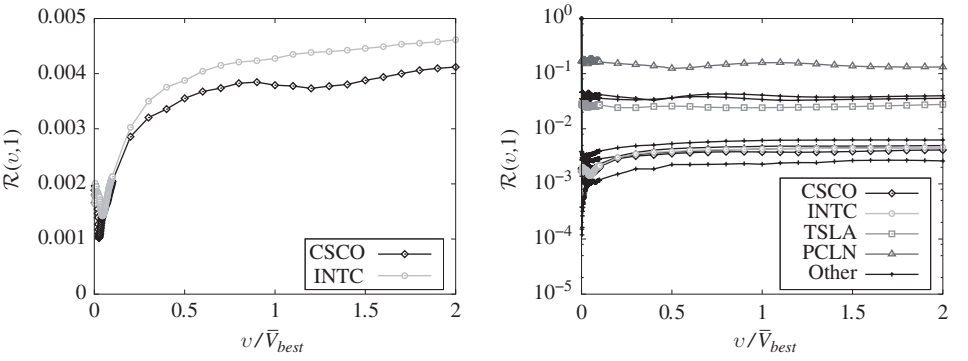

Figure 11.3. (Left panel) Lag-1 impact response function $\mathcal{R}(v,1)$ for INTC and CSCO, as a function of normalised market order volume (where we normalise by the average volume at the same-side queue). Observe the non-monotonic effect for small v. (Right panel) The same plot on a semi-logarithmic scale, for the ten stocks in Table 11.1. Small-tick stocks (with a large $\mathcal{R}(v,1)$) do not exhibit any significant dependence on v.

of the immediate impact is actually strongly sub-linear:

$$\mathcal{R}(v,1) := \langle \varepsilon_t \cdot (m_{t+1} - m_t) | v_t = v \rangle_t \cong A \left(\frac{v}{\bar{V}_{\text{best}}} \right)^{\zeta} \langle s \rangle_t, \tag{11.7}$$

where A is a constant of order 1 and ζ is an exponent that takes very small values. For the stocks in our sample, we find values $\zeta \cong 0 - 0.3$ (see Figure 11.3). In other words, the lag-1 impact of a single market order is a strongly concave function of its volume, and perhaps even a constant for small-tick stocks. As we discuss in the next section, this concavity mostly comes from a conditioning bias called selective liquidity taking.

Note a curious feature for large-tick stocks: the impact curve is non-monotonic for small volumes, such that small market orders seem to have an anomalously high impact. This is related to volume imbalance effects, as noted in Section 7.2: when the volume in one of the two queues is very small compared to the volume in the opposite queue, it is highly probable that a small market order will grab the small remaining quantity, resulting in a one-tick price jump.

11.3.3 Selective Liquidity Taking

Whether or not a given market order has instantaneous impact depends not only on the size of the market order, but also on the state of the LOB at its time of arrival. Therefore, attempting to analyse impact only as a function of market order size could be misleading. At the very least, one should distinguish between **aggressive market orders** and **non-aggressive market orders**.

Recall from Section 10.2 that we use the notation MO^1 to denote market orders that consume all volume available at the best opposite quote, which leads to an *immediate* price move, and MO^0 to denote market orders that do not. We introduce a similar notation for impact:

$$\mathcal{R}^1(1) := \langle \varepsilon_t \cdot (m_{t+1} - m_t) | \pi_t = \mathrm{MO}^1 \rangle_t,$$
$$\mathcal{R}^0(1) := \langle \varepsilon_t \cdot (m_{t+1} - m_t) | \pi_t = \mathrm{MO}^0 \rangle_t.$$

Table 11.1 lists the empirical values of $\mathcal{R}^1(1)$ and $\mathcal{R}^0(1)$ for the stocks in our sample. As might be expected, $\mathcal{R}^1(1) > \mathcal{R}^0(1)$, but note that even for MO^0 events, the response of the market is strictly positive: $\mathcal{R}^0(1) > 0$. This is due to the fact that there is a non-zero probability for the non-executed limit orders at the opposite-side best quote to be cancelled before the next market order arrival, and thereby to produce a price change in the direction of the initial trade. Clearly, one has

$$\mathcal{R}(1) = \mathbb{P}[\mathrm{MO}^0]\mathcal{R}^0(1) + \mathbb{P}[\mathrm{MO}^1]\mathcal{R}^1(1),$$

where $\mathbb{P}[\mathrm{MO}^1]$ is the probability that the market order is aggressive and $\mathbb{P}[\mathrm{MO}^0]$ is the probability that it is not (such that $\mathbb{P}[\mathrm{MO}^1] = 1 - \mathbb{P}[\mathrm{MO}^0]$).

It is interesting to study the distribution of market order volumes, conditioned to the volume at the opposite-side best quote V_{best} at their time of arrival.[8] Figure 11.4 shows the mean size of market order arrivals for given values of V_{best}. The plot suggests that these mean order sizes grow sub-linearly, and appear to be well described by the power-law

$$\langle \upsilon | V_{\mathrm{best}} \rangle \propto V_{\mathrm{best}}^{\chi}; \qquad \chi \cong 0.6.$$

Figure 11.5 shows the distribution $f(x)$ of the ratio $x = \upsilon / V_{\mathrm{best}}$. We observe that $f(x)$ mostly decreases with x and behaves qualitatively similarly for large- and small-tick stocks. Note also the important round-number effects: $f(x)$ has spikes when the market order size is equal to simple fractions of V_{best}. These round-number effects persist even when the market order is larger than the available volume (corresponding, for example, to $V_{\mathrm{best}} = 100$, $\upsilon = 200$). The largest spike occurs for $x = 1$, which means that traders submit a significant number of market orders with a size that exactly matches the available volume at the best. Finally, only few market orders are larger than the available volume at the best. In other words, when traders submit market orders, they often adapt their order volumes to the volume available at the opposite-side best quote. This phenomenon is known as **selective liquidity taking**: the larger the volume available at the opposite-side best quote, the larger the market orders that tend to arrive.

[8] As is clear from Section 4.2, V_{best} has a strong intra-day pattern. It can also fluctuate considerably from one trading day to the next.

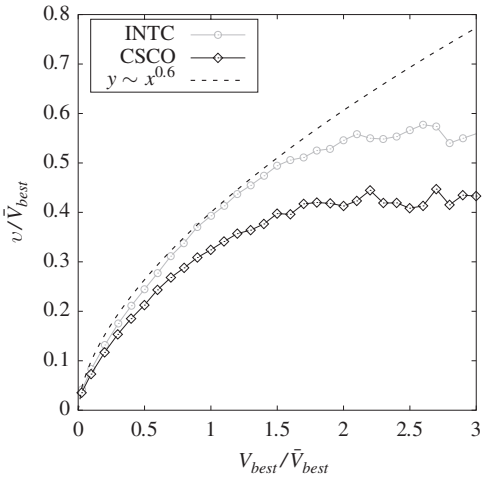

Figure 11.4. Average market order size (normalised by the average queue volume) versus the available volume at the same-side best quote (normalised by the average queue volume) for INTC and CSCO. For small queue volumes, the relationship roughly follows a power-law with exponent 0.6 (dashed curve), and saturates for large queue volumes.

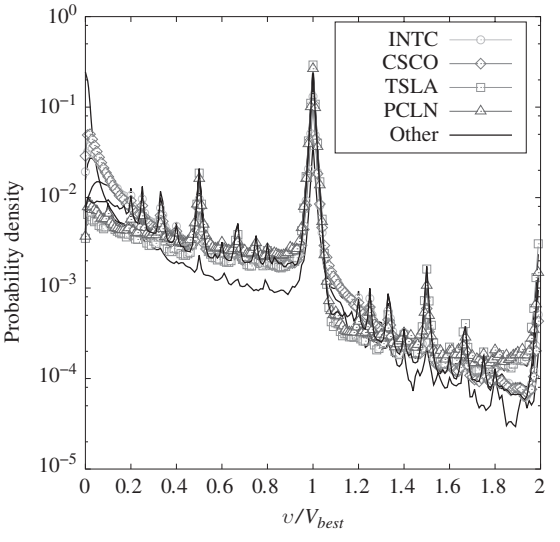

Figure 11.5. Empirical distribution of the fraction of executed queue volume q/V_{best} for the ten stocks in our sample.

One can make use of a simple caricature to understand the strong concavity of $\mathcal{R}(v,1)$ as a function of v, as we reported in the previous section (see Equation 11.7). Assume for simplicity that $\mathcal{R}^0(v,1) \approx 0$ and $\mathcal{R}^1(v,1) \approx \mathcal{I}$, such that market orders with a volume less than that of the opposite-side best quote have no impact at all, while market orders with a volume that matches that of

the opposite-side best quote impact the price by a fixed quantity $\mathcal{I}$ (which we also assume to be independent of volume and equal to some fixed fraction of the spread). In other words, we assume that market orders never eat more than one level, and that the impact of new limit orders and cancellations can be neglected. If this were the case in real markets, then it would follow that

$$\mathcal{R}(v, 1) \approx \mathbb{P}[\mathrm{MO}^1|v]\mathcal{I}.$$

In real markets, $\mathbb{P}[\mathrm{MO}^1|v]$ is an increasing function of v, is zero when $v = 0$ and converges to 1 when $v \to \infty$. The resulting $\mathcal{R}(v, 1)$ is therefore a concave function. In fact, using Bayes rule,

$$\mathbb{P}[\mathrm{MO}^1|v] = \int_0^v \mathrm{d}V_{\mathrm{best}}\, P(V_{\mathrm{best}}|v) = \frac{\int_0^v \mathrm{d}V_{\mathrm{best}}\, P(v|V_{\mathrm{best}})P(V_{\mathrm{best}})}{\int_0^\infty \mathrm{d}V_{\mathrm{best}}\, P(v|V_{\mathrm{best}})P(V_{\mathrm{best}})}.$$

Now suppose that the conditional distribution $P(v|V_{\mathrm{best}})$ is an arbitrary function of $x = v/V_{\mathrm{best}}$ and $P(V_{\mathrm{best}})$ is itself a power-law. In this case, one finds that $\mathbb{P}[\mathrm{MO}^1|v]$ is *independent* of v, leading to a volume-independent impact $\mathcal{R}(v, 1)$, as indeed observed for small-tick stocks (see Figure 11.3).

The conclusion of this toy calculation is that the strong concavity of lag-1 impact as a function of market order volume is (at least partially) a conditioning effect caused by the fact that most large market order arrivals happen when there is a large volume available at the opposite-side best quote.

11.3.4 Conditioning on the Sign of the Previous Trade

As we have shown throughout this chapter, market orders clearly impact the price in the direction of the trade. However, in Chapter 10 we showed that market order signs are autocorrelated. As illustrated by Table 11.2, the lag-1 autocorrelation coefficient $C(1) = \langle \varepsilon_t \varepsilon_{t-1} \rangle$ is quite large. Given that market orders impact prices, it might seem reasonable to expect that the autocorrelation in market order signs should also lead to some predictability in price moves.

If the lag-1 impact $\mathcal{R}(1)$ was independent of the past, then conditional on the last trade being a buy, the next price change would also be on average positive. Conditional on the last trade being a buy, the probability that the next trade is also a buy is $p_+ = (1 + C(1))/2$, and the probability that the next trade is a sell is $p_- = (1 - C(1))/2$. Therefore, it follows that:

$$\langle m_{t+1} - m_t | \varepsilon_{t-1} \rangle_{\mathrm{naive}} = p_+ \mathcal{R}(1) - p_- \mathcal{R}(1) = C(1)\mathcal{R}(1). \tag{11.8}$$

In this naive view, the presence of sign autocorrelations should thus lead to price predictability. However, price changes in financial markets are difficult to predict, even at high frequencies, so this naive picture is likely to be incorrect. To

Table 11.2. *The values of* $C(1)$, f, $\mathcal{R}_+(1)$ *and* $\mathcal{R}_-(1)$ *for the ten stocks in our sample. Impact is measured in dollar cents.*

	$C(1)$	$\mathcal{R}_+(1)$	$\mathcal{R}_-(1)$	f
SIRI	0.93	0.064	0.027	0.90
INTC	0.59	0.26	0.22	0.61
CSCO	0.66	0.22	0.17	0.71
MSFT	0.57	0.29	0.25	0.56
EBAY	0.48	0.36	0.33	0.47
FB	0.37	0.47	0.51	0.31
TSLA	0.69	2.53	2.84	0.52
AMZN	0.72	3.61	4.17	0.57
GOOG	0.75	3.94	4.87	0.60
PCLN	0.78	15.67	16.99	0.63

quantify this small level of predictability, let f denote the actual lag-1 mid-price predictability as a fraction of the above naive predictability. We define the fraction f as:

$$\langle \varepsilon_{t-1} \cdot (m_{t+1} - m_t) \rangle := fC(1)\mathcal{R}(1).$$

Table 11.2 shows the values of f for the stocks in our sample. In all cases, the value of f is smaller than 1, which is the value that the above naive picture would suggest.

How should we understand this empirical result? A first step is to note that a buy trade in fact impacts the mid-price less if it follows another buy trade than if it follows a sell trade. More formally, we define *two* impacts, as follows:

$$\begin{aligned} \mathcal{R}_+(1) &:= \mathbb{E}[\varepsilon_t \cdot (m_{t+1} - m_t) | \varepsilon_t \varepsilon_{t-1} = +1]; \\ \mathcal{R}_-(1) &:= \mathbb{E}[\varepsilon_t \cdot (m_{t+1} - m_t) | \varepsilon_t \varepsilon_{t-1} = -1]. \end{aligned} \tag{11.9}$$

By taking the product of successive trade signs, the conditioning selects successive trades in the same direction for $\mathcal{R}_+(1)$ and in the opposite direction for $\mathcal{R}_-(1)$.

As shown by Table 11.2, it holds that $\mathcal{R}_+(1) < \mathcal{R}_-(1)$ for small-tick stocks.[9] This illustrates an important empirical fact: the most likely outcome has the smallest impact. For example, if the previous trade is a buy, then due to the autocorrelation of market order signs, the next trade is more likely to also be a buy. When the next trade occurs, its impact will, on average, be smaller if it is indeed a buy than if it is

[9] For large-tick stocks, the situation is inverted as a consequence of the influence of volume imbalance on future price changes. The unpredictability argument is no longer about the mid-price m_t but instead about the modified mid-price $\widetilde{m}_t$ defined in Footnote 6.

a sell. This mechanism, which is a crucial condition for market stability, is called **asymmetric dynamical liquidity**.

Of course, considering the lag-1 autocorrelation is only the tip of the iceberg. As we saw in Chapter 10, market order signs actually have long-range autocorrelations that decay very slowly. Therefore, the sign of the next trade can be better predicted by looking at the whole history of trades, and not only the single most recent trade, as would be the case if trade signs were Markovian.[10] In Chapter 13, we extend the analysis from this section to incorporate autocorrelations at larger lags and to build a full theory of the delayed impact of market orders.

11.4 Order-Flow Imbalance and Aggregate Impact

To reduce the role of microstructural idiosyncrasies and conditioning biases (such as selective liquidity taking), price impact is often measured not at the trade-by-trade level (as we have done so far in this chapter), but instead over some coarse-grained time scale T. In many practical applications, choices of T range from about five minutes to a full trading day. In this framework, the goal is to characterise the positive correlations between aggregate signed order flow and contemporaneous price returns.

Consider all transactions that occur in a given time interval $[t, t+T)$, for some $T > 0$ and where t and T are expressed in either event-time or calendar-time. Throughout this section, we again choose to work in market order time, whereby we advance t by 1 for each market order arrival. For each $n \in [t, t+T)$, let ε_n denote the sign of the n^{th} market order and let v_n denote its volume. The **order-flow imbalance** is:

$$\Delta V = \sum_{n \in [t, t+T)} \varepsilon_n v_n.$$

If $\Delta V > 0$, then more buy volume arrives in the given interval than does sell volume, so one expects the price to rise.[11]

The **aggregate impact** is the price change over the interval $[t, t+T)$, conditioned to a certain volume imbalance:

$$\mathbb{R}(\Delta V, T) := \mathbb{E}\left[m_{t+T} - m_t \middle| \sum_{n \in [t, t+T)} \varepsilon_n v_n = \Delta V \right]. \tag{11.10}$$

This quantity can be studied empirically using only public trades-and-quotes data. If $T = 1$, then ΔV is simply the volume of the single market order at time t, and

[10] This Markovian assumption is precisely the starting point of the MRR model that we discuss in Section 16.2.1.

[11] This market scenario is often described as there being "more buyers than sellers"; in reality, of course, there is always an equal number of buy orders and sell orders for each transaction. What is usually meant by the phrase is that there are more buy market orders due to aggressive buyers, but even this is only a rough classification because buyers might also choose to use limit orders to execute their trades (see the discussion in Section 10.5.1 and Chapter 21).

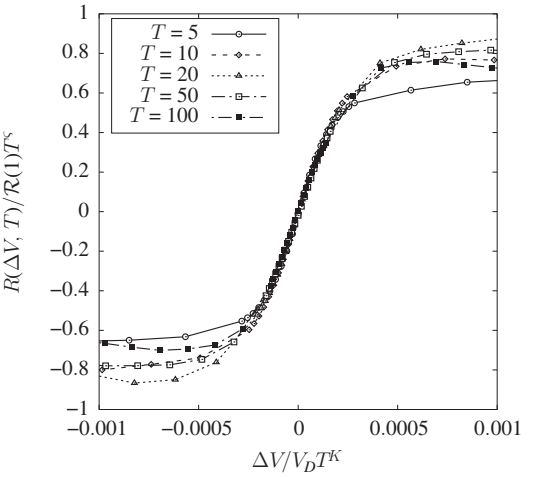

Figure 11.6. Aggregate impact scaling function $\mathscr{F}(x)$ for TSLA. By rescaling $\mathbb{R}(\Delta V, T)$ for $T = 5, 10, 20, 50$ and 100, with $\chi = 0.65$ and $\varkappa = 0.95$, the curves approximately collapse onto each other. Note that $\mathcal{R}$ and ΔV are rescaled each day by the corresponding values of $\mathcal{R}(1)$ and the daily volume V_{D}. A similar rescaling works similarly well for all stocks and futures.

one recovers the definition of lag-1 impact

$$\mathbb{R}(\Delta V, 1) \equiv \mathcal{R}(\upsilon = \Delta V, 1).$$

Therefore, by the same arguments as in Section 11.3.2, $\mathbb{R}(\Delta V, 1)$ is a strongly concave function of ΔV.

How does the behaviour of $\mathbb{R}(\Delta V, T)$ change for larger values of T? Empirically, when T increases, the dependence of $\mathbb{R}(\Delta V, T)$ on the volume imbalance becomes closer to a linear relationship for small ΔV, while retaining its concavity for large $|\Delta V|$. In fact, the following empirical **scaling law** appears to hold for a large variety of stocks and futures contracts:[12]

$$\mathbb{R}(\Delta V, T) \cong \mathcal{R}(1) T^{\chi} \times \mathscr{F}\left(\frac{\Delta V}{V_{\mathrm{D}} T^{\varkappa}}\right), \tag{11.11}$$

where V_{D} is the daily volume and $\mathscr{F}(u)$ is a scaling function (see Figure 11.6) that is linear for small arguments ($\mathscr{F}(u) \sim_{|u| \ll 1} u$) and concave for large arguments[13] and where empirically the exponents are given by $\chi \cong 0.5 - 0.7$ and $\varkappa \cong 0.75 - 1$. We will attempt to rationalise the values of these exponents in Section 13.4.3.

The slope of the linear region of $\mathbb{R}(\Delta V, T)$ is usually called **Kyle's lambda**, in reference to the Kyle model, which we discuss in Chapter 15. The value of this

[12] See: Patzelt, F., & Bouchaud, J. P. (2018). Universal scaling and nonlinearity of aggregate price impact in financial markets. *Physical Review E*, 97, 012304.

[13] A possible example is $\mathscr{F}(u) \sim_{|u| \gg 1} \mathrm{sign}(u)|u|^{\zeta}$ with a small exponent ζ, or other functions with a similar behaviour but that saturate for large u.

slope is often regarded as a measure of a market's (il-)liquidity. Using the scaling result in Equation (11.11), it follows that Kyle's lambda *decreases* with the time interval over which it is measured:[14]

$$\mathbb{R}(\Delta V, T) \sim_{|\Delta V| \to 0} \Lambda(T)\Delta V; \qquad \Lambda(T) \cong \Lambda(1) T^{-(\varkappa - \chi)}.$$

Interestingly, the empirical determination of $\varkappa - \chi \cong 0.25{-}0.3$ is more robust and stable across assets than χ and $\varkappa$ independently.

The saturation for larger values of ΔV is also interesting, but is likely to be related to selective liquidity taking (see Section 11.3.3) for single market orders, which is expected to persist on longer time scales $T \gg 1$. For example, imagine that there is an excess of sell limit orders in the LOB during the whole time interval $[t, t+T)$. This will likely attract buy market orders that can be executed without impacting the price too much. Therefore, one expects that these situations will bias $\mathbb{R}(\Delta V, T)$ downwards for large ΔV, following the very same conditioning argument that led to a concave shape for $\mathcal{R}(\Delta V, 1)$.

11.5 Conclusion

The primary conclusion of this chapter is that there exists a clear empirical correlation between the signs of trades and the directions of price moves. Trivial as it may sound, the mechanism underlying this behaviour is not immediately clear. Does this phenomenon occur because information is revealed? Or does the very occurrence of a trade itself impact prices? Generally speaking, one expects that both effects should contribute. However, empirical data suggests that the reaction part of impact is predominant on short to medium time scales, and is identical for all trades irrespective of whether they are informed or not. Since trades in modern financial markets are anonymous, it would be surprising if any distinction (between whether or not a given trade was informed) could be made before any genuine prediction contained in some of these trades reveals itself. By a similar argument, one expects that the impact of trades could also explain a large fraction of the volatility observable in real markets (see the discussion in Chapter 20).

From the point of view of this book, this is very good news! It suggests that there is hope for modelling the reaction part of the impact, which should follow reasonably simple principles, as we shall discuss in Chapters 13 and 19. The prediction part of impact is of course much harder to model, since it depends on a huge information set and on the specific strategies used to exploit mispricings. In any case, the information content of individual trades must be very small (see

[14] In the Kyle model (see Chapter 15), aggregate impact is linear and additive, so $\Lambda(T)$ must be constant and therefore independent of T. Since order flow is assumed to be completely random in this model, the value of the exponents corresponds to $\chi = \varkappa = \frac{1}{2}$ in that case, and the scaling function is linear: $\mathscr{F}_{\text{Kyle}}(x) = x$.

Chapters 13, 16 and 20). Furthermore, as we will see, the decay of the reaction impact is so slow that the distinction between transient impact (often associated with noise trades) and permanent impact (often associated with informed trades) is fuzzy.

Several empirical conclusions from this chapter will now trickle throughout much of the rest of the book. For example, we saw in Section 11.3.1 that the lag-1 impact of a market order is of the same order as the bid–ask spread. We will investigate this from a theoretical perspective in Chapter 17. We also saw in Section 11.3.1 that the response function $\mathcal{R}(\ell)$ is an increasing function of ℓ that saturates at large lags (see Figure 13.1). This can be traced back to the autocorrelation of the trade-sign series (see Sections 13.2 and 16.2.4, and Equation (16.22)), from which we will construct a theory for the reaction impact of trades (see Chapter 13). In Section 11.3.4, we saw that the impact of two trades in the same direction is less than twice the unconditional impact of a single trade. This is due to liquidity providers buffering the impact of successive trades in the same direction, and is a crucial condition for market stability, as we will see in Sections 13.3 and 14.4.

Finally, we have discussed the impact of trades at an aggregate level. We saw that small price returns are linearly related to small volume imbalances, with a slope (called "Kyle's lambda") that is found to decrease with the aggregation time scale T. We will return to this discussion in Section 13.4.3. As we discuss in the next chapter, however, this linear impact law, obtained by a blind aggregation of all trades, is of little use for understanding the price impact of a *metaorder* executed over a time horizon T. This necessary distinction is among the most striking results in market microstructure, and will require a detailed, in-depth discussion (see Section 13.4.4 and Part VIII).

Take-Home Messages

(i) Trade directions are positively correlated with subsequent price changes. There are two competing views that attempt to explain this phenomenon: that trades *forecast* prices or that trades *impact* prices.

(ii) The observed impact of a trade can be decomposed into two components: a reaction component, which describes how the market reacts to the trade itself, and a prediction component, which describes all other dynamics not directly related to the trade (e.g. exogenous information).

(iii) Measuring the reaction impact of a given order is difficult, because doing so would require replaying history to consider a world in

which all else was equal except that the given order was not submitted.

(iv) The response function describes the mean price trajectory at a given time lag after a trade. Empirically, the response function is found to be an increasing function of lag, starting from 0 and reaching a plateau with a scale that is of the same order as the bid–ask spread. Importantly, the response function must not be confused with the reaction impact of a market order, because it also contains the impact of future orders.

(vi) Because market order signs are strongly autocorrelated, if trades impacted prices linearly, then prices would be strongly predictable. This is not observed empirically. Therefore, the impact of trades must be history-dependent. For example, a buy trade that follows a buy trade impacts the price less than a buy trade that follows a sell trade.

(vii) It is non-trivial to assess the dependence of impact on volumes. At the scale of individual trades, conditioning effects lead to a strongly concave dependence.

(viii) The aggregate impact of volume imbalance is linear for small imbalances and saturates for large imbalances. The slope of the linear part decreases as a power-law of the aggregation time. However, this naive averaging only shows part of the story – as we will see in the next chapter.

11.6 Further Reading

Price Impact

Weber, P., & Rosenow, B. (2005). Order book approach to price impact. *Quantitative Finance*, 5(4), 357–364.

Gerig, A. (2007). *A theory for market impact: How order flow affects stock price*. PhD thesis, University of Illinois, available at: arXiv:0804.3818.

Farmer, J. D., & Zamani, N. (2007). Mechanical vs. informational components of price impact. *The European Physical Journal B-Condensed Matter and Complex Systems*, 55(2), 189–200.

Hasbrouck, J. (2007). *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press.

Bouchaud, J. P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In Hens, T. & Schenk-Hoppe, K. R. (Eds.), *Handbook of financial markets: Dynamics and evolution*. North-Holland, Elsevier.

Bouchaud, J. P. (2010). Price impact. In Cont, R. (Ed.), *Encyclopedia of quantitative finance*. Wiley.

Eisler, Z., Bouchaud, J. P., & Kockelkoren, J. (2012). The price impact of order book events: Market orders, limit orders and cancellations. *Quantitative Finance*, 12(9), 1395–1419.

Hautsch, N., & Huang, R. (2012). The market impact of a limit order. *Journal of Economic Dynamics and Control*, 36(4), 501–522.
Cont, R., Kukanov, A., & Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, 12(1), 47–88.
Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the Bitcoin. *Market Microstructure and Liquidity*, 1(02), 1550008.
Gomes, C., & Waelbroeck, H. (2015). Is market impact a measure of the information value of trades? Market response to liquidity vs. informed metaorders. *Quantitative Finance*, 15(5), 773–793.
Tóth, B., Eisler, Z. & Bouchaud, J.-P. (2017). The short-term price impact of trades is universal. https://ssrn.com/abstract=2924029.

(Weak) Volume Dependence of Impact

Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, 46(1), 179–207.
Jones, C. M., Kaul, G., & Lipson, M. L. (1994). Transactions, volume, and volatility. *Review of Financial Studies*, 7(4), 631–651.
Chen, Z., Stanzl, W., & Watanabe, M. (2002). *Price impact costs and the limit of arbitrage*. Yale ICF Working Paper No. 00–66.
Lillo, F., Farmer, J. D., & Mantegna, R. N. (2003). Econophysics: Master curve for price-impact function. *Nature*, 421(6919), 129–130.
Potters, M., & Bouchaud, J. P. (2003). More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 324(1), 133–140.
Zhou, W. X. (2012). Universal price impact functions of individual trades in an order-driven market. *Quantitative Finance*, 12(8), 1253–1263.
Taranto, D. E., Bormetti, G., & Lillo, F. (2014). The adaptive nature of liquidity taking in limit order books. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(6), P06002.
Gomber, P., Schweickert, U., & Theissen, E. (2015). Liquidity dynamics in an electronic open limit order book: An event study approach. *European Financial Management*, 21(1), 52–78.

History Dependence of Impact

Bouchaud, J. P., Gefen, Y., Potters, M., & Wyart, M. (2004). Fluctuations and response in financial markets: The subtle nature of random price changes. *Quantitative Finance*, 4(2), 176–190.
Bouchaud, J. P., Kockelkoren, J., & Potters, M. (2006). Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance*, 6(02), 115–123.
Farmer, J. D., Gerig, A., Lillo, F., & Mike, S. (2006). Market efficiency and the long-memory of supply and demand: Is price impact variable and permanent or fixed and temporary? *Quantitative Finance*, 6(02), 107–112.
Taranto, D. E., Bormetti, G., Bouchaud, J. P., Lillo, F., & Tóth, B. (2016). Linear models for the impact of order flow on prices I. Propagators: Transient vs. history dependent impact. Available at SSRN: https://ssrn.com/abstract=2770352.

Aggregate Impact

Kempf, A., & Korn, O. (1999). Market depth and order size. *Journal of Financial Markets*, 2(1), 29–48.
Plerou, V., Gopikrishnan, P., Gabaix, X., & Stanley, H. E. (2002). Quantifying stock-price response to demand fluctuations. *Physical Review E*, 66(2), 027104.

Chordia, T., & Subrahmanyam, A. (2004). Order imbalance and individual stock returns: Theory and evidence. *Journal of Financial Economics*, 72(3), 485–518.
Evans, M. D., & Lyons, R. K. (2002). Order flow and exchange rate dynamics. *Journal of Political Economy*, 110(1), 170–180.
Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2006). Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121(2), 461–504.
Lillo, F., Farmer J. D., & Gerig A. (2008). *A theory for aggregate market impact*. Technical report, Santa Fe Institute, unpublished research.
Hopman, C. (2007). Do supply and demand drive stock prices? *Quantitative Finance*, 7, 37–53.