Algorithmic Trade Execution and Market Impact

Richard Coggins†, Marcus Lim‡, Kevin Lo†
† Discipline of Finance,
School of Business, H69
University of Sydney, New South Wales 2006,
Australia

‡ Computer Engineering Laboratory,
School of Electrical and Information Engineering,
University of Sydney, New South Wales, 2006,
Australia

Email: r.coggins@econ.usyd.edu.au

Algorithmic trade execution has potential to reduce the costs of implementing investment decisions. A difficult aspect of trade execution is estimating, forecasting and minimising market impact costs. Techniques for implementing algorithmic trade execution addressing these problems are being developed. Results for these techniques will be presented in the context of the full order book data available from the ASX and provided by the Capital Markets CRC.

1 Introduction

When a large equity fund manager makes a decision to purchase or liquidate a holding in a particular security the realisation of the transaction or its execution is not straight forward, as is the case for the small investor. The quality of execution is measured by transaction costs. There is no formal definition of the best execution [31], but in practice it means execution at the most favourable prices. Common execution benchmarks include implementation shortfall, VWAP and closing price. Implementation shortfall can be measured by considering the difference between a real portfolio and a paper portfolio [37]. Perold called this difference an implementation shortfall [34]. The shortfall can be divided into two components. The explicit component represents broker fees and taxes. The implicit component captures market impact cost, opportunity cost, and the bid-ask spread. For sake of argument if we consider a large trade which is to be worked over an entire day then an implementation shortfall approach could use the opening price as a bench mark for market impact, and consider the opportunity cost of not completing the trade by days end. Similarly, a VWAP bench mark compares the volume weighted prices achieved by the trader compared to all other trades in the market on that day. By this means, VWAP attempts to measure the traders performance relative to the average execution price for that day in the market. Finally, the closing price is an often desired price since it is often used to benchmark fund performance and is used for derivative margining and settlement.

In this paper we will discuss approaches to achieving such benchmarks algorithmically and how we might assess the performance of such algorithms prior to their use in an actual market. In section 2 we describe optimal trade execution in the context of an implementation shortfall framework. Next in section 3 we consider the problem of evaluating a trade execution algorithm's performance and encounter the problem of how can we forecast market impact when large volumes (more than 5% of average daily market volume) are involved. In section 4 we consider some simple execution rules for achieving the VWAP benchmark. In particular, we compare and contrast limit order versus market order submission strategies and the difficulties of forecasting their performance. Having considered these execution problems, we establish that modelling market impact is an important aspect of forecasting and optimising execution performance for large volume targets. Sections 5, 6 and 7 discuss a cross section of market impact models ranging from considering the impact of single trades, market wide impact relationships and the temporally aggregated dynamics of market impact. Finally we conclude in section 8.

2 Optimal Execution

The material in this section is based on [9]. Best execution is difficult to achieve in practice due to limited liquidity and price volatility in the market over the time period that the transaction is to be completed. Limited liquidity implies that as a large trade is attempted, the trader becomes a large demander of liquidity and as a result will incur a premium to complete the trade. This premium represents the cost of immediacy, a transaction cost which we will refer to as the trade execution shortfall per share (the difference in the price of the security when the trade commences and the average price achieved for the completed trade for a sell and visa versa for a buy) [15]. The goal of a trade execution strategy is to minimise the trade execution shortfall subject to time and specified risk level constraints. Intuitively, one observes that by distributing trading volume across time and choosing peaks in liquidity, to the extent that they are known, trading shortfall can be reduced at the risk that the price of the security (as measured by the midpoint of the bid-ask spread) moves adversely in relation to the type of transaction to be completed. Hence, we are seeking to balance different sources of risk, namely a risk characterising our uncertain knowledge of future liquidity versus a risk corresponding to our uncertain knowldege of future prices. The realisation of a particular strategy is then in turn, determined by our propensity to take risks to minimise the trade execution shortfall.

In this paper, we consider the problem of trade execution in the context of a fully electronic limit order market, namely, the Australian Stock Exchange (ASX). A brief description of the microstructure of the ASX is provided in appendix A. The method we propose is an extension of a framework derived by Almgren and Chriss [2, 3] which in turn builds on work by [6]. The original framework and our extensions of it are described in section 2.1.

Obtaining high performance trade execution plans also depends on our ability to predict any patterns in the security's behaviour. Section 2.2 describes the methods we use to characterise a security's liquidity and price behaviour. In section 2.3 we will compare our results with three other trade execution benchmarks which we denote as one-interval, uniform and VWAP execution. One interval execution is a risk averse strategy, where we take the currently available liquidity and avoid the risk of future volatility and liquidity risk. A uniform trade execution strategy recognises a need to reduce the peak demand for liquidity, but is agnostic with respect to the intra-day levels and patterns in the security's liquidity and price. A VWAP strategy seeks to minimise the difference in volume weighted average price of the transaction and the volume weighted average price of the entire market for the security and typically uses a historical trading volume as a proxy for liquidity [22], and may incorporate technical rules to take into account any knowledge of price behaviour.

Hence, a VWAP strategy makes use of intra-day liquidity and price patterns, however, it differs from our optimal approach in the sense that it has an implicit approach to the transaction risk and may not necessarily involve the optimisation of a trade execution objective. We consider this approach in section 4. We present results comparing the four trade execution techniques by measuring the trade shortfalls on an out of sample interval of the limit order book for the security National Australia Bank (NAB). The next section describes our general approach to trade execution shortfall minimisation.

2.1 Mimising Execution Shortfall

The formulation of the optimal trade execution problem described in this section builds on the work of Almgren and Chriss [2, 3]. The formulation can be summarised as follows. Firstly, we will define a price process for the security. Here we adopt a discrete time arithmetic random walk parameterised by the volatility over a time interval. This is a reasonable approximation to the more plausible geometric random walk over the intra-day time scales we are considering. Next we will introduce a temporary market impact function which measures the average price concession when we sell a given quantity of shares (from this point onwards we will consider the liquidation of a security for the purpose of explanation, however the buy case follows directly). Associated with the market impact function is a measure of uncertainty of the market impact, which we will call impact risk. Given normally distributed random shocks for price and market impact we can determine the expected execution shortfall and the variance of the shortfall. A utility function is then constructed from the mean and variance which is a risk constrained shortfall to be optimised. For certain forms of the volatilty and market impact functions trading trajectories can be derived analytically [2, 3]. Here, we extend the formulation to include non-stationary market impact, impact risk and volatility which require numerical solution, for which we employ deterministic discrete time dynamic programming [32]. Finally, we will show the connection between the mean/variance utility function for shortfall and the value at risk objective.

Let X denote the quantity of shares which we wish to sell in a time T. We wish to derive trade execution plans which specify the number of shares to sell at N discrete time intervals. Let $\tau = T/N$ denote the length of the time intervals. Let k be an index over the N time intervals. A trading trajectory is then defined by a list $x_0, ..., x_N$ where x_k is the number of shares we still hold after time interval k, $x_0 = X$ and we require $X_N = 0$. Equivalently we can specify a trade by a list of shares to be traded at each time interval specified by $n_k = x_k - x_{k-1}$. The price process we will assume for the security is,

$$S_k = S_{k-1} + \sigma \tau^{1/2} \xi_k, \tag{1}$$

where S_k is the security price at time k, σ is the volatility and ξ_k is a set of IID zero mean unit variance normal random variables. In contrast to [2], for the purposes of our empirical work presented in this paper, we do not attempt to characterise permanent market impact. However, we will consider temporary market impact. By temporary market impact, we mean that as we trade we may exhaust the liquidity available at a series of price levels leading to a lower price as we sell during a given time interval k. We assume that the effect of our liquidity consumption is limited to each individual trading interval, with a new equilibrium price being established at the start of the next interval as if we hadn't traded. The price that we achieve as a consequence of this temporary market impact is given by,

$$\tilde{S_k} = S_{k-1} - h(\frac{n_k}{\tau}) + \tau^{-1/2} f(\frac{n_k}{\tau})\tilde{\xi_k}, \tag{2}$$

where $\tilde{\xi_k}$ are zero mean, unit variance normal random variables independent of the ξ_k. In section 2.2 we will indicate how $h()$ and $f()$ can be determined empirically. Hence, the total trading revenue on completion of all trades is $\sum_{k=1}^{N} \tilde{S}_k n_k$. The trade execution shortfall is therefore given by,

$$XS_0 - \sum_{k=1}^{N} n_k \tilde{S}_k \tag{3}$$

From this we can determine the expectation and variance of the shortfall with respect to the random innovations of the price process ξ,

$$E(x) = \tau \sum_{k=1}^{N} v_k h(v_k) \tag{4}$$

$$V(x) = \sigma^2 \tau \sum_{k=1}^{N} x_k^2 + \tau \sum_{k=1}^{N} v_k^2 f(v_k)^2 \tag{5}$$

where we have used $v_k = n_k/\tau$. We now construct a utility function,

$$U(x) = E(x) + \lambda V(x) \tag{6}$$

For $h(v)$ of the form $h(v) = \eta v^i$ where i is an integer, $U(x)$ can be minimised analytically [2]. In this paper we consider two practical extensions of this model. To account for the bid-ask spread and the fact that market impact is fixed below a critcal volume h_{v0} we consider $h(v)$ of the form,

$$h(v) = \begin{cases} \eta(v - h_{v0}) + \epsilon & v > h_{v0} \\ \epsilon & v \leq h_{v0} \end{cases} \tag{7}$$

Further, we condider the case where the market impact function $h(v)$ has a volume dependent uncertainty associated with it. Following Almgren's approach and extending it to take take into account the bid-ask spread and critical volume we define the impact risk by a function,

$$f(v) = \begin{cases} \beta(v - f_{v0}) + \alpha & v > f_{v0} \\ \alpha & v \leq f_{v0} \end{cases} \tag{8}$$

This non-linearity in $h(v)$ requires us to minimise $U(x)$ numerically. Secondly we consider the case where $\eta(k)$, $\sigma(k)$ are functions of time as measured by index k. When $h(v)$ has a linear form $h(v) = \eta(k)v$ and $f(v)$ is constant with respect to volume traded $f(v) = \alpha$, we can obtain an analytical solution to the trading trajectory as follows. By writing $v_k = x_{k-1} - x_k$ and differentiating $U(x)$ with respect to each x_k we obtain,

$$\frac{\partial U(x_k)}{\partial x_k} = 2\{\lambda\sigma_k^2 x_k - (\eta_k + \lambda\alpha_k^2)(x_{k-1} - 2x_k + x_{k+1})\} \tag{9}$$

where we have shown the subscript k to indicate the parameters as a function of time. The set of equations in x_k are second order difference equations. To solve the equations we impose the boundary conditions, $x_0 = X$ and $x_N = 0$, implying an initial holding of X shares and complete liquidation of the holding by time step N. Given these constraints the ratio $r(k)$ defined by $x_k = r(k)x_{k-1}$ can be derived,

$$r(k) = \frac{A_{N-k}}{A_{N-k} + A_{N-k+1} + \lambda\sigma_{N-k}^2 - A_{N-k+1}r(k-1)} \tag{10}$$

where A_k is given by,

$$A(k) = \eta_k + \lambda\alpha_k^2 \tag{11}$$

Although (10) does not take into account non-linearities in the impact function, it does none the less take into account the non-stationarities in the parameters, and hence is useful for rapidly calculating approximate strategies for a large number of securities.

In order to compute more accurate optimal trading trajectories we employ the following reward function and two dimensional state space model (the model is defined in the context of a sell program, a buy formulation follows straightforwardly). Denote the state variable for the problem to be $S = (s_1, s_2)$ where s_1 represents the stock being held and s_2 measures time intervals (half hour bins). Let a_k denote the action variable, specifying the number of shares to be sold at each time step with the constraint $0 \leq a_k \leq s_1(k)$ implying that we never buy shares and we can never sell more shares than we currently hold. Let $R(S, a)$ denote the reward function which is a function of the current state S and the action taken by the agent a_k ($a_k = n_k$). Then,

$$R = E(S, a) + \lambda V(S, a) \tag{12}$$

where $E()$ and $V()$ have the general form of (4) and (5) respectively and,

$$\begin{aligned} s_1(k) &= s_1(k-1) - a(k) \\ s_2(k) &= s_2(k-1) + 1 \end{aligned} \tag{13}$$

For a sell program, the states are initialised as $s_1(1) = X$ and $s_2(1) = 1$. Note, the time interval $s_2(k)$ is modelled explicitly as part of the state space due to the reward function R depending on the non-stationary impact coefficient $\eta(k)$ and the non-stationary volatility $\sigma(k)$ and the other parameters used to model market impact. Numerical solutions were computed using a Matlab discrete dynamic programming toolbox [32] incorporating the non-linear impact function (7) and impact risk (8).

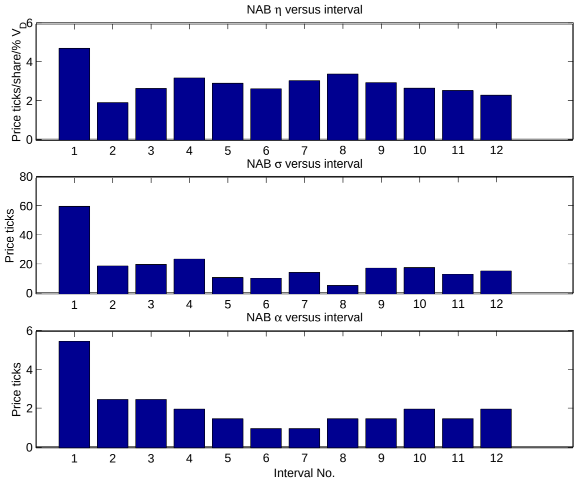

Figure 1: Intraday impact coefficient η, volatility σ and impact risk α for NAB stock averaged over a 3 week period.

2.2 Measuring Security Characteristics

In our experiments we consider a three week trading period for each security in order to parameterise our models. This provides us with fifteen days of trading data which we treat identically (we have chosen to ignore possible weekly seasonal effects). We divide each trading day into half hour intervals, giving us 12 intervals, since the ASX opens at 10am and closes at 4pm. For the data we consider we ignore off-market trades, crossings and undisclosed orders. For each security we then collect the following data:

- total daily trading volume for each of the 15 days.
- the difference in midpoint quote at the start and end of each half hour interval.
- the average, normalised, time weighted cumulative order volume available at each price tick in the order book relative to the midpoint price, for each half hour interval. Normalisation is with respect to the daily traded volume.

A plot of the transformed normalised cumulative order volume as a function of price ticks from the midquote for a particular half hour bin provides a volume normalised, price invariant measure of market impact as a function of traded volume from which η is derived. Figure 1 shows average intraday impact coefficient $\eta(k)$, volatility $\sigma(k)$ and impact risk $\alpha(k)$.

Given these measurements the expected shortfall and variance of the shortfall may be written:

$$E(S,a) = \tau \sum_{k=1}^{N} a_k h(a_k, \eta_k, \epsilon_k, h_{kv0}) \tag{14}$$

$$V(S,a) = \tau \sum_{k=1}^{N} \sigma_k^2 (s_1(k) - a_k)^2 + \tau \sum_{k=1}^{N} a_k^2 f^2(a_k, \beta_k, \alpha_k, f_{kv0}) \tag{15}$$

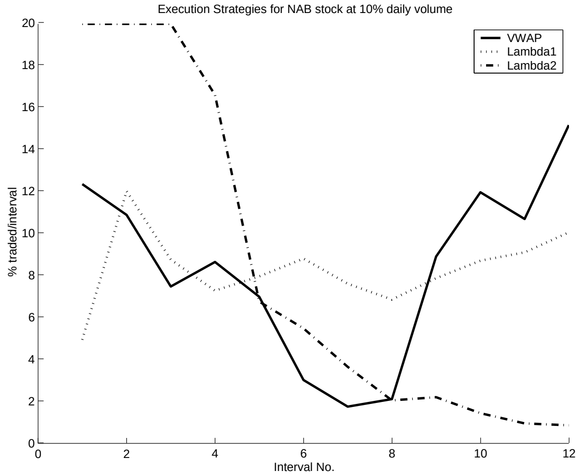

Figure 2: Two trading trajectories computed for the NAB stock at a level of 10% of daily volume with the risk parameter $\lambda = 10^{-6}$ and $\lambda = 10^{-1}$ respectively. $\lambda = 10^{-6}$ corresponds to the case where the risk of adverse price movements and our uncertainty about future liquidity later in the day are negligible and is therefore a risk neutral strategy. $\lambda = 0.1$ corresponds to a shortfall value at risk $Var(p) = 66$c/share, with $p = 0.99$ implying that this shortfall is predicted to be exceeded 1% of the time. Note that for this second trajectory, the first two intervals trade 20%, which is due to a rule we introduced that limits trading in a half hour period to a maximum of 2% of forecast daily volume. We did this to ensure our impact was temporary. The VWAP trajectory is based on the average fraction of trading volume, traded in the stock over the three week training period during the respective intra-day interval.

2.3 Simulated Trading Performance

By parameterising the reward function R as described in the previous section we are then able to compute optimal trading trajectories for a given stock. Three typical trading trajectories are shown in Figure 2.

Given a trading trajectory x it then remains to test its performance. We do this by considering 63 days of trading immediately after the three weeks training period which was used to parameterise our models. Each of the trading trajectories we calculate sets half hour trading targets during the day. We then simulate trading over the half hour by trading a fixed fraction of the trading target against the orders available in the schedule every five minutes. This entails the following difficulties. It is possible that we match against the same orders a number of times. It also assumes that orders we consume, don't ultimately influence future order flows, that is, we only take account of instantaneous market impact. In section 3 we consider more sophisticated approaches to assessing trading performance.

To assess the performance of the optimal approach that we have presented in this paper we also evaluated a number of simpler strategies:

- One interval (ONEINT): we liquidate all the stock during the first half hour of trading.
- Uniform (UNIFORM): we liquidate our holding uniformly throughout the day.
- (VWAP) trader: we use historical fractional trading volumes averaged over the 3 week training period as targets for the half hour intervals.

- Optimal (Lambda1, Lambda2): the algorithm presented in this paper, constrained to trade all the holding in one day, at two different risk settings λ (Lambda1=10^{-6}, Lambda2=10^{-1}).

In order to quantify trading performance we consider two metrics. The first is execution shortfall as given in (3). The second is a volume weighted average price metric,

$$\Delta VWAP = \{\frac{VWAP_{trade}}{VWAP_{market}} - 1\} * 10^4 \tag{16}$$

where $VWAP_{trade} = \sum_k \tilde{S}_k n_k / X$ and the units are basis points ($\frac{1}{100}$th of a percent). Execution shortfall directly measures the transaction costs resulting from trading. $\Delta VWAP$ on the other hand measures the trading performance against the rest of the market. For the execution shortfall measure we are also able to calculate a value at risk for any given trade execution strategy. We do this as follows. Firstly, we determine,

$$\lambda_v = -\frac{\partial E}{\partial V^{\frac{1}{2}}} \tag{17}$$

for any trading strategy x. We then calculate,

$$p = \int_{\infty}^{\lambda_v} N(z)dz \tag{18}$$

where $N(z)$ is the standard normal density function and the value at risk is given by $Var(p) = E(x) + \lambda_v V(x)^{\frac{1}{2}}$.

The results are shown in table 1. We test our strategies for both the buy and sell cases so as to reveal any bias in our results due to the specific out of sample test period used. We note that the one interval strategies for each stock have low or zero values of p, with low or negative transaction costs, indicating that although considered risk averse strategies, transaction costs are difficult to predict due large variations in price and liquidity during the opening half hour. We note that in a small number of instances our value at risk calculation appears spurious, which we believe is due to the non-differentiable nature of our market impact function shown in (7). These are indicated in the tables by a '-'. We see that the UNIFORM, VWAP and Lambda1 strategies produce similar shortfall forecasts and measured shortfalls and the value at risk calculation shows that they are often close to risk neutral strategies ($p = 0.5$). The Lambda2 strategies correspond to more risk averse strategies as indicated by the value at risk forecasts and standard deviations of shortfall forecasts. We note that the measured transaction costs do show reduced standard deviations, but not to the extent that our simple model predicts. We note that volatility effects are dominant in our results. This suggests, that in order to improve our trading strategies, we would need an improved price process model which exploits intraday regularities in prices. Alternatively, we could also consider limit order submission strategies to as a mechanism to avoid the cost of the spread and reduce market impact costs. This however, introduces additional difficulties for trade evaluation to be discussed in the following section.

Table 1: Performance of trade execution strategies for NAB stock. Each of the strategies are tested at nominally 1%, 5% and 10% of average daily traded volume ($\hat{V_D}$). V_D is actual average daily traded volume for the test period. VaR(p) is the shortfall upper bound that can be achieved with probability p. $\hat{sf}$ and sf are the forecast shortfall and measured shortfall in cents per share traded respectively. $\Delta VWAP$ indicates the fractional difference in volume weighted price achieved for the whole trade against the entire market for the day expressed in percentage basis points. Standard deviations over the 63 trading day test period appear in brackets.

Trade Type (Units)	$\hat{V_D}$ (%)	V_D (%)	p	VaR(p) (c/share)	$\hat{sf}$ (c/share)	sf (c/share)	$\Delta VWAP$ (basis pts.)
ONEINT							
Buy	1	1.1	0.30	-3.9	5.1 (17.4)	4.7 (26.0)	1.2 (48.2)
Sell	1	1.1	0.28	-3.5	6.0 (16.2)	-0.4 (22.7)	-11.1 (55.1)
Buy	5	5.4	0.34	-3.6	24.1 (66.3)	10.8 (26.4)	18.7 (50.9)
Sell	5	5.4	0.32	-3.3	28.2 (68.3)	7.4 (27.4)	-33.6 (59.4)
Buy	10	10.9	0.34	-3.6	47.9 (127.5)	21.0 (27.8)	48.2 (57.1)
Sell	10	10.9	0.33	-3.2	56.1 (133.5)	24.9 (38.3)	-83.9 (84.1)
UNIFORM							
Buy	1	1.1	0.54	4.0	0.5 (31.5)	5.5 (32.2)	3.3 (9.1)
Sell	1	1.1	0.55	4.5	0.6 (31.5)	-3.5 (32.0)	-2.3 (9.0)
Buy	5	5.4	0.67	15.3	1.5 (31.5)	5.7 (32.2)	3.9 (9.2)
Sell	5	5.4	0.68	16.2	1.6 (31.5)	-3.3 (32.1)	-3.0 (9.2)
Buy	10	10.9	0.80	29.3	2.7 (31.5)	5.9 (32.2)	4.6 (9.2)
Sell	10	10.9	0.81	30.9	2.9 (31.5)	-3.1 (32.0)	-3.6 (9.1)
VWAP							
Buy	1	1.1	0.54	3.8	0.6 (30.2)	5.2 (32.0)	2.5 (6.5)
Sell	1	1.1	0.55	4.3	0.6 (30.2)	-3.2 (31.6)	-3.1 (6.9)
Buy	5	5.4	0.65	13.7	1.8 (30.2)	5.5 (32.0)	3.3 (6.6)
Sell	5	5.4	0.67	15.0	1.9 (30.2)	-3.0 (31.6)	-3.9 (7.0)
Buy	10	10.9	0.78	26.8	3.2 (30.3)	5.8 (32.0)	4.1 (6.7)
Sell	10	10.9	0.80	29.1	3.6 (30.3)	-2.7 (31.7)	-4.7 (7.0)
Lambda1							
Buy	1	1.1	0.54	3.9	0.5 (32.4)	5.5 (33.0)	3.4 (9.4)
Sell	1	1.1	0.55	4.4	0.6 (33.3)	-3.6 (33.0)	-2.3 (9.2)
Buy	5	5.4	0.66	14.9	1.4 (32.4)	5.8 (32.4)	4.0 (8.9)
Sell	5	5.4	0.67	16.2	1.6 (32.4)	-3.4 (33.0)	-2.8 (9.8)
Buy	10	10.9	0.79	28.4	2.6 (32.4)	6.0 (32.4)	4.7 (8.9)
Sell	10	10.9	0.79	29.5	2.7 (33.3)	-3.2 (33.1)	-3.3 (10.0)
Lambda2							
Buy	1	1.1	-	-	4.4 (15.3)	4.9 (25.9)	1.8 (41.4)
Sell	1	1.1	-	-	5.2 (14.4)	-1.5 (23.4)	-8.0 (46.3)
Buy	5	5.4	-	-	5.8 (21.9)	6.4 (27.0)	6.0 (30.8)
Sell	5	5.4	-	-	7.9 (21.8)	-1.9 (25.9)	-6.9 (30.9)
Buy	10	10.9	0.97	55.0	5.2 (26.1)	6.6 (28.9)	6.7 (20.5)
Sell	10	10.9	0.99	66.6	6.6 (26.1)	-2.4 (28.3)	-5.6 (20.8)

3 Trade Evaluation Environments

Traditionally technical traders have backtested their trading strategies over historical price series in order to assess their profitability. These historical price series are typically the daily open, high, low and close prices. Depending on the trade being considered is it often debatable whether these prices can be achieved in the market. A further step is often taken to model trade slippage by using a penalty function of these prices, providing less favourable prices as trading volumes increase. Such a slippage function attempts to capture an average of cost of implementation of the technical trading strategy. In the context of execution strategies considered in this paper, we are considering a whole extra level of detail, whereby on one hand we have access to the complete transaction by transaction and order event by order event data for the market. We then are faced with the challenge of how we can best make use of all this additional data to accurately assess implementation costs. (The same issues arise if one is to consider the problem of assessing the performance of intraday technical trading rules).

In the last section we considered an optimal execution strategy using market orders only. A pure market order strategy is more straightforward to backtest on data, because we don't need to consider the problem of what will happen to an order if it is submitted as a limit order at a particular point in the schedule. Rather, we could just make the assumption, that a given market order will consume available liquidity on the other side of the book. The question then arises however, how would the market react to the liquidity consumption arising from our market orders? Two simplistic approaches could be considered:

- We could assume that the liquidity we consume is instantaneously replenished. That is, we could assume that the market is perfectly elastic and that the market impact costs we face are just the instantantaneous costs associated with paying the spread and any levels of the order book we consume.
- We could assume that none of the liquidity we consume is replenished. In this case we have an on-going effect on the market.

In section 2 we took the first of these approaches to evaluate the performance of our algorithm. This was straight forward to implement, however we note that, for large market orders the perfect elasticity assumption behind it will not hold. The second approach is less straightforward to implement. Essentially it entails rebuilding the historical order flows for the market and merging our back tested order flow with the historical flow. This has the benefit that we are now considering the on-going effect of our liquidity consumption. By building such a system, it also allows us to backtest limit order submission strategies. However, it none the less does not model the reactions of market participants to our order flow. Thus we consider this in section 7. In the following section we now present an approach to evaluating the performance of execution strategies which supply liquidity, that is use limit order submission. We will demonstrate this in the context of using some rules to achieve the VWAP execution benchmark.

4 VWAP execution strategies

VWAP is a popular execution bench mark in the industry for trades that need to be worked over a few hours or days. The essence of a VWAP trading strategy is to split a

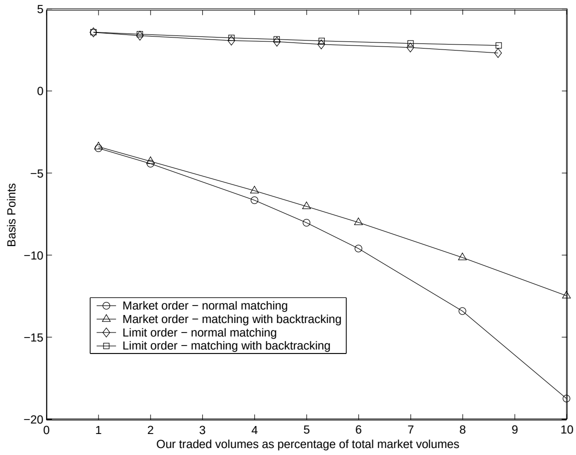

Figure 3: Comparison of a pure market order sell strategy to a pure limit order sell based strategy averaged over the year 2002 for the security WBC. The vertical axis measures the VWAP of the trades realised relative to the VWAP of the market for a day. The horizontal axis is the percentage daily volume being executed.

large order up into a series of smaller orders. The sizes of the smaller orders are scaled to correspond to the forecast intra-day volume pattern for a stock. The rationale for this approach is to minimise the market impact cost of trading those orders, by trading more when liquidity is high and less when liquidity is low. These strategies can be further refined by considering the types of orders that are used. A straight forward approach could use a pure market order strategy, where a series of maret orders are distributed across the day. An improved approach uses a combination of limit and market orders. In general, optimisation techniques as described in section 2 could be used to decide an optimal strategy. This is beyond the scope of this paper. Instead we will consider the following straight forward rule base approach. For the sake of clarity consider a VWAP buy execution.

1. Divide the day up into time slots, allocating a given percentage of trade volume to each time interval.
2. At each time slot submit a limit order of the specified size at the best bid.
3. If within x minutes the best bid has gone up and our order has not executed, amend the order to the best bid.
4. If by the end of the time slot the order has not fully completed amend it to become a market order to force completion.

Thus, we observe that with a limit order based submission approach we are now faced with some additional problems of order management to obtain the best algorithm and as mentioned in section 3 we also need a more sophisticated approach to backtesting in order to assess the performance of the algorithm. Figure 3 shows a comparison of a pure limit order VWAP strategy versus a pure market order based approach. The figure compares two approaches to backtesting the strategies as mentioned previously, that is where we assume liquidity is consumed and replenished, versus when liquidity is consumed and not replenished. We note that, as expected on average a limit order based approach to execution improves the performance relative to VWAP. In all of what has been discussed so far we have treated the problem of market impact using some simplistic extreme assumptions, being that we either only have a shortlived temporary impact, we have an on-going impact due to the effect of our liquidity supply or consumption and that ultimately, market participants make no other responses to our order flow. In the following sections we now address the important problem of modelling market impact in more detail.

5 Immediate price impact

This section is concerned with the issue of price impact, and its response to the size of trades. The price impacts of block trades have been shown to be large in small cap stocks and are systematically related to trade size and market capitalization (Loeb [30], Kraus & Stoll [23], Holthausen et. al. [18], Keim & Madhaven [19]). A recent study [5] analyzed the interplay between the price impact of each trade and the balance of liquidity supply and demand.

Here we investigate the relationship between the immediate price impact and traded volume for trades on the Australian Stock Exchange (ASX). The ideas and methodology in this paper are directly based on the work of Lillo et. al. [26] and Lillo et. al. [27]. They have examined the relationship between price impact and traded volume empirically, but their results were based on the stocks in the NYSE which has a specialist managing the order book. In this paper we reproduce some results from a study described in [28]. By investigating the relationship between price impact and traded volume, we are indirectly examining the relationship between price formation and trading activity. Liquid stocks will generally incur less price impact than an illiquid stock. Liquid stocks tend to be more actively traded, and this is inherently linked to the stock's market capitalization.

The study uses a fairly large amount of data (four years of transaction-level data), looks at the short term response to a *single* trade, and measures market activity in units of transactions. The 300 largest stocks of the Australian Stock Exchange from 2001 to 2004, are studied with the goal of understanding how much price changes on average in response to a buy or sell order.

Section 5.1 describes the data used in the study. In order to illustrate the overall approach, we present price impact curves for single stocks in section 5.2. In section 5.3, we investigate in more general terms how price impact varies with market capitalization, by grouping stocks with similar market cap together, and comparing the price impact across groups of stocks. Section 5.4 brings liquidity into the picture, and considers its relationship with market capitalization. Section 5.5 addresses the issue of obtaining a single market curve that describes price impact with respect to trade size.

5.1 Data

The sample selected spans a period of 4 years (2001 to 2004), encompassing the top 300 stocks in the S&P asx300 index, listed on the Australian Stock Exchange (ASX). The data is transaction-level data, which means that it covers every single event that happens in the market, and each event has a time-stamp attached to it. Many kinds of events are possible, including order submission, amendment and deletion. For the purposes of this paper, we are mainly interested in the *trade* events that occur during normal trading hours. Consequently, the data needs to be filtered down to the trade level and processed further. Details regarding the data and preprocessing steps can be found in A.

5.2 Single Stock Analysis

We start off by considering the simplest case of analyzing the price impact in response to traded volume, on a stock by stock basis. To illustrate, we compare the impact behavior of a high-cap stock to that of a mid-cap one.

5.2.1 Method

The methodology used to determine price impact is similar to that of Lillo et. al. [26]. For the stock under investigation, the algorithm goes over every trade in turn for a period of one year. Trades which are buyer-initiated (Bi) were treated separately from seller-initiated (Si) trades.

The price impact Δp of a single trade of size ω was obtained by taking the first difference of the logarithm between the price of the orderbook before the trade and the price immediately after the trade occurs. This difference is the immediate price change or shift caused by the trade. More specifically, we have the price shift

$$\Delta p(t_{i+1}) = \log_{10} p(t_{i+1}) - \log_{10} p(t_i) \tag{19}$$

where t_i is the time of the pre-trade price, and t_{i+1} is the time of the price following the trade. We take the price p to be the best *ask* price when we are considering Bi trades, while in the case of Si trades, we take p to be the best *bid* price. In other words, we are investigating the ask price shift caused by a Bi trade, and the bid price shift caused by a Si trade. The intuition is that a Bi trade will cause an increase in the best ask price, while a Si trade will cause a decrease in the best bid price.

Δp represents the *percentage change* in p before and after each Bi or Si trade. The use of relative as opposed to absolute differences, lets us compare the impact across stocks more meaningfully — after all, a 'small' absolute price shift can still be quite substantial in the case of small stocks. The transaction size ω for a single trade is defined as the *normalized daily-normalized volume*. This means the original trade size was first normalized by the total trade volume for the day, which helps to filter out any inter-day liquidity variation effects. Then the result was further divided by the average of all the normalized volumes for the entire one-year period for that stock:

$$\omega_{ij} = \frac{v_{ij}}{\sum_{m=1}^{T_i} v_{im}} \left(\frac{N}{\sum_{d=1}^{N} T_d} \right)^{-1} \tag{20}$$

where ω_{ij} is the normalized daily-normalized volume for trade j on day i, T_i is the total number of trades on day i, and N is the total number of days ($N \approx 250$ for one year). The numerator in equation 20 is the daily-normalized volume for v_{ij} i.e., fraction of daily volume for day i, while the denominator is its average for the entire N day period. This puts all stocks on a roughly equal footing, and thus makes any comparison between stocks more meaningful.

Ultimately, for both Bi and Si cases, we are interested in investigating the average Δp as a function of ω. For this study, we examine Δp at 20 distinct values of ω, covering the entire ω range. This involves dividing the data points into twenty bins, with each bin having approximately the same number of points, and hence the same number of trades. This was accomplished by first dividing the total number of trades for the stock by 20, and taking the result as the target number of trades per bin, i.e., the width of each bin

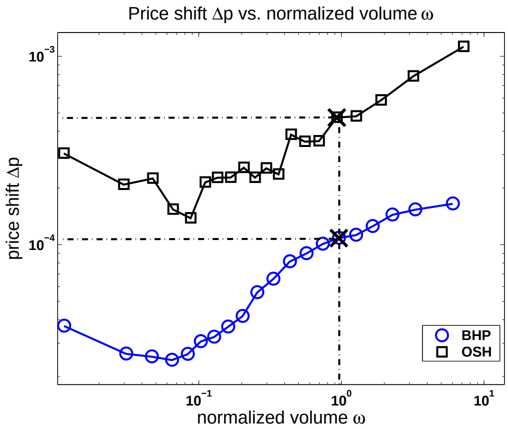

Figure 4: Price shift versus normalized transaction size for buyer initiated (Bi) trades for two representative stocks, Oil Search Limited (squares) and BHP Billiton (circles). The year being investigated is 2004.

was chosen such that it contains close to the target number. Because we are interested in the *average* price impact, the mean of the price shifts was then calculated for each of the twenty bins, and plotted against the corresponding median volume for that bin.

5.2.2 Results and Discussion

The results of doing this for two representative stocks BHP Billiton and Oil Search Limited are shown in figure 4. The plots are on a log scale for both axes. The 'X' marks on the BHP and OSH curves are visual aids that mark the point on the curves where the normalized volume is at the mean (10^0 on the x-axis). Keeping in mind that each of the bins contains roughly the same number of trades, it is immediately apparent from the curves that about 75% of trade sizes are below the average normalized volume, for both stocks.

The 'X' marks show that for BHP, a trade with an average normalized volume incurs an immediate average percentage price increase of around $10^{-4} = 0.01\%$ or one basis point; a similar trade for OSH with an average normalized volume incurs a slightly higher price increase, from the figure, this is around $10^{-3.5} = 0.03\%$ or three basis points. In other words, the price impact for BHP (a high-cap stock) is lower (by a factor of three) than that for OSH (a mid-cap stock) for the same normalized trade size. In fact, the figure also shows that the highest percentage price increase for BHP only approaches the lowest percentage price increase for OSH.

In both cases, the impact starts to increase monotonically beyond a certain normalized trade volume, specifically around 10^{-1} or 10% of the average normalized volume for both stocks. Notice that the impact for BHP increases more rapidly than that of OSH. The other observation is that the impact for BHP tends to level off for trades with average normalized volume and above, but the OSH curve continues to increase. These observations suggest, at least for this simple case of two stocks, that high-cap stocks experience lower impact than lower cap stocks, for the same normalized trade volume. In addition, the rate of increase of the impact with respect to volume, is greater for high-cap stocks than for lower-cap stocks. Also, the impact for high-cap stocks has a tendency to level off for higher trade volumes.

5.3 Grouped Stock Analysis

The results from the previous section indicate three key points:

1. Higher cap stocks yield lower price impact for the same normalized transaction size.
2. The rate of increase of impact with respect to traded volume is greater for higher cap stocks than for lower cap stocks.
3. The impact of high-cap stocks levels off at high volume.

The next step is to try and understand more systematically whether this behavior applies to the market as a whole. With this in mind, it is necessary to consider more than two stocks. We have considered the top 300 stocks in the S&P asx300 [28]. This index represents approximately 90% of the Australian market. The stocks in the index were grouped based on their free float market cap, so that the groups can then be compared to one another. Note that we are now making comparisons between *groups* of stocks, whereas in the previous section, *individual* stocks were compared.

5.3.1 Method

As with the single stock case, the experiments were performed on a year by year basis. The sequence of steps for each of the four years under investigation follows the pattern of ranking, grouping and data processing, in that order.

1. The top 300 stocks were determined, and ranked according to each company's *free float market capitalization* for a particular year. This was calculated based on the following formula:

   $$free\,float\,market\,cap = price\ *\ shares\,on\,issue\ *\ IWF$$

   where IWF is the Investable Weight Factor, which represents the proportion of a company that is freely available for trading in the market[6].
2. The 300 companies were formed into ten groups, with each group chosen to have approximately the same number of transactions. This means that the total number of trades for all the different stocks in each group is roughly the same across all ten groups. The total number of trades was obtained by adding up all the 'on market' trades for all stocks in the group over the whole year. Because higher-cap stocks like NAB are more actively traded, group size i.e. the number of stocks in each group, varies from one group to the next. For instance, in 2004, the highest cap group for both Bi and Si cases consisted of three stocks (NAB, BHP and CBA), while the last group had more than 100 small-cap stocks.
3. For each group of stocks, the price impact of trades with respect to the trade size was determined. This was done for each stock in the group. The procedure for accomplishing this is exactly identical to the single stock case described in section 5.2; the additional step was to aggregate the trades for all stocks in the group together. The subsequent step of binning the data into 20 bins is also identical to the single stock case — bins were chosen for each group such that each bin has approximately an equal number of transactions. Table 2 shows the average number of elements per bin for all four years.

[6]For the remainder of the paper we abbreviate 'free float market capitalization' to market capitalization.

Table 2: Average number of trade transactions per bin for all four years.

	2001	2002	2003	2004†
Bi	14900	17000	19700	16000
Si	14600	16700	18500	15800

†Jan - Oct 2004

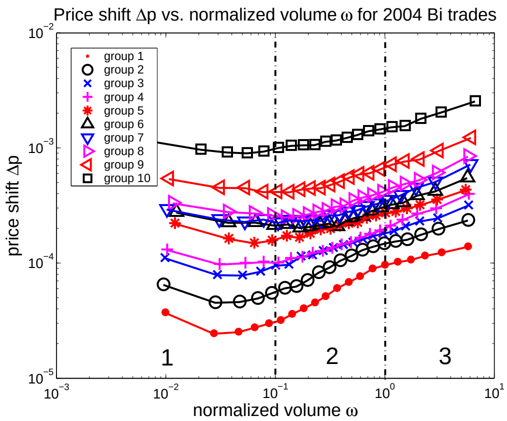

Figure 5: Percentage price shift versus normalized transaction size for buyer initiated (Bi) trades for ten groups of stocks sorted by market capitalization. The investigated year is 2004. The mean market capitalization decreases from group 1 to group 10. The impact curves all tend to increase beyond 10% of the average normalized volume.

5.3.2 Results and Discussion

The results obtained for Bi trades for 2004 are shown in figure 5. As in the single stock case, the plots are on a log scale for both axes. Also, for the Si case, the absolute value of the price shift i.e., $|\Delta p|$ is used, which has the effect of flipping the negative Si impact curves about the x-axis. This puts both Bi and Si cases on a common point of reference, making comparisons easier.

The initial period decreases in impact in most of the impact curves in figure 5 might be exlained as follows. For tiny trades ranging from one to ten percent of the average normalized trade volume, there is less incentive on the part of market participants to pay much attention to the liquidity conditions, hence on average the impact over this volume range exhibits no increasing pattern with increasing volume. The next section examines liquidity and market capitalization in more detail, using the results of this section as the basis for analysis.

5.4 Liquidity and Market Capitalization

Figure 5 shows the price impact functions ordered in increasing market cap from top to bottom. As mentioned, this clear pattern suggests that there is a direct relationship between liquidity and the size of the stock i.e. its market cap. We now investigate this relationship empirically.

5.4.1 Method

Linear regression was performed on each of the price impact functions (each group of stocks). In mathematical terms, for each curve, the goal is to estimate β (the gradient) and $\log_{10}\lambda$ (intercept) in the following linear equation:

$$\log_{10}|\Delta p| = \beta \log_{10}\omega - \log_{10}\lambda \tag{21}$$

β is the gradient of each of the curves in the figures. As indicated in the previous section, higher-cap groups seem to have a higher rate of increase, and so we would expect higher-cap groups to have larger β values. For the specific Bi case of 2004 (figure 5), we perform the regression separately for each of the three regions (in addition to an overall regression for all three regions together), in order to quantify the difference in behavior both between the regions, and across the groups.

λ in equation 21 is the liquidity parameter for a particular group. To see why this is so, solve for λ, which yields $\lambda = \frac{\omega^{\beta}}{|\Delta p|}$. This is a measure of the volume that can be traded to incur a certain impact; a highly liquid stock would therefore have a bigger value of λ than a less liquid stock i.e. we can trade more volume in high liquid stocks than in illiquid stocks for the same impact incurred. Again note that we have taken the absolute value of Δp, which defines impact as positive for both Bi and Si cases. The regression was done separately for both cases.

The mean market cap C for each group was also determined, and then for each of the groups, its liquidity parameter λ was plotted against its corresponding mean market cap C. Thus, we ended up with ten points, one per group, and the last step was to do another linear regression on these points, in order to determine the relationship between liquidity and market cap. This process was carried out separately for Bi and Si trades. We are interested in how much liquidity increases with an increase in the market cap for both Bi and Si trades. In short, we wish to estimate γ in the following equation:

$$\log_{10}\lambda = \gamma \log_{10} C + D \tag{22}$$

Knowing the value of γ for a particular year tells us the rate of increase of λ with respect to C. To see this, solve for λ in equation 22, which gives $\lambda = C^{\gamma} \cdot 10^{D}$ — we can then describe the relationship between λ and C in the following way: "λ increases as C^{γ}" or "liquidity increases as the mean market cap scaled to a power". The aim is therefore to estimate the power γ.

The regression (22) was repeated for all the four years under investigation. An overall regression was also performed over all 4 years at the same time. Bi and Si cases were always treated separately.

5.4.2 Results and Discussion

The results of performing the linear regression on equation 21 resulted in an average β ranging from 0.1 to 0.4 for all groups and all 4 years. Higher cap groups tend towards the higher β values, while the lower-cap groups have lower values. This makes sense, because higher-cap impact functions tend to have higher rate of change than lower-cap groups, as was described before.

Table 3 shows the estimated β parameters for the specific case of Bi trades for 2004. The β values for region 1 are negative for all groups, suggesting that the impact is actually

Table 3: β parameter for each group of stocks for each region, for Bi trades of 2004.

	Region 1	Region 2	Region 3	All†
Group 1	0.10	0.52	0.21	0.32
Group 2	0.09	0.47	0.29	0.29
Group 3	0.08	0.30	0.31	0.22
Group 4	0.10	0.30	0.37	0.22
Group 5	0.13	0.26	0.30	0.16
Group 6	0.10	0.23	0.34	0.13
Group 7	0.10	0.20	0.41	0.15
Group 8	0.11	0.28	0.39	0.17
Group 9	0.11	0.25	0.33	0.15
Group 10	0.09	0.18	0.31	0.14

† β parameter estimated across all three regions

Table 4: Estimated γ parameter for all 4 years.

	2001	2002	2003	2004	All†
Bi	0.4317***	0.4017***	0.4098***	0.4249***	0.4169***
Si	0.4289***	0.3849***	0.4120***	0.4094***	0.4097***

† γ estimate for regression over all four years

*** γ statistically significant at the < 0.001 level

decreasing for small trade volumes — an interesting and somewhat surprising characteristic that is consistent across all the groups. The β values for region 2 indicate that for moderate trade volumes, the higher-cap groups 1 and 2 have a higher rate of increase (0.52 and 0.47) in the impact compared to lower-cap groups. However, this rate of increase decreases again for higher volumes, as is evident from the β values in region 3. In contrast, the lower-cap groups have a relatively uniform increase rate from region 2 to region 3. The values of γ obtained as a result of performing the linear regression (equation 22), are shown in table 4. Two regressions were performed each year, one for Bi trades, another for Si trades. An overall regression was also performed for all four years at once, for both Bi and Si cases. For all regressions, the estimated parameter γ is statistically significant, in each case rejecting the null hypothesis $H_0 : \gamma = 0$ at the < 0.001 level. Figure 6 illustrates a clear relationship between λ and market capitalization C. The plots are on a log scale for both axes. The fact that the points lie on a straight line in a loglog plot immediately suggests the existence of a power law relationship between λ and C. The straight line on the figures is the line of best fit through all the 40 points (10 points per year, one per group), and has a gradient of 0.42 (the last column of table 4). As mentioned, this value was estimated by performing the regression (equation 22) over all the 40 points. Since $\gamma \approx 0.42$ for both Bi and Si cases, we can then conclude that liquidity increases roughly as $C^{0.42}$. Note that this value is close to that obtained by [26] and Lillo et. al. [27], who obtained a value of 0.39. This seems to suggest that despite the differences that exist among different markets, some of the fundamental economic principles that govern them remain the same.

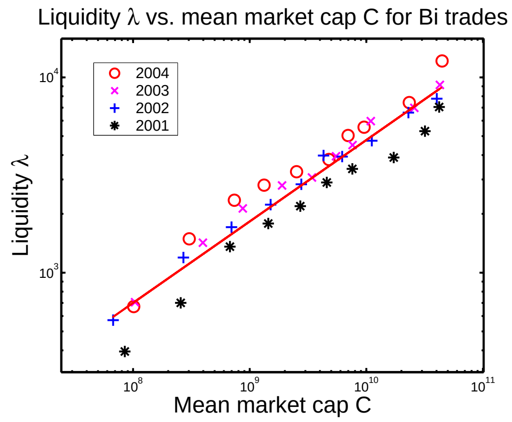

Figure 6: Liquidity λ of each group of stocks as a function of the group's mean market capitalization for 2001 (*), 2002 (+), 2003 (x) and 2004 (o). The thick line is the power law best fit line on all points, with gradient of 0.42 i.e., exponent $\gamma = 0.42$. This plot applies to the Bi case only.

5.5 Scaling and Data Collapse

This final section brings together all the concepts discussed in the previous sections with the aim of consolidating them into a unified representation — to explore the extent to which the price impact curves can be explained by a single "market" curve. The relationship between market capitalization and liquidity is the key element that we exploit.

5.5.1 Method

The previous section indicated that liquidity is closely related to market capitalization, and this relationship is a power law. By extension, this means that impact Δp and market cap C also share a power law relationship, which we can derive by equating equation 21 with 22, and solving for Δp:

$$\Delta p(\omega, C) = K \cdot \frac{\omega^\beta}{C^\gamma} \tag{23}$$

where K is a constant. This has the functional form of the scaling relationship $m(t, L) = L^d f(t/L^c)$, with m, t, and L replaced with Δp, ω and C respectively. We can relate this to the curves in figures 5. Each set of impact curves from these figures comes from the same *family* of curves, with the functional form $\Delta p(\omega, C)$ of equation 23. The curves represent this 2-variable function graphically as Δp against ω for a sequence of different values of C, where C is the average market capitalization for each group. A quantitative way of showing scaling is a data collapse (scaling plot) based on the observation that the coexistence curves for many systems could be made to fall on a single curve. According to [35] and [4], the scaling hypothesis predicts that all curves of the same such family can be "collapsed" onto a single curve provided one plots not Δp against ω, but rather a *scaled* Δp via the transformation $\Delta p \to \Delta p/C^d$ against a *scaled* ω via the transformation $\omega \to \omega/C^c$. In this case, we plot $\Delta p/C^d$ against ω/C^c for all groups in one particular year, where d and c are two different values of the exponents. The methodology we employ is based on the minimization of a measure proposed by [4]; minimization of this measure can be achieved numerically, allowing us to extract the exponents d and c. Also, before applying the transformation, we have renormalized the market cap C for each group by dividing it by its mean for all ten groups. This was done in order to make the R^2 statistic (described next) scale-invariant.

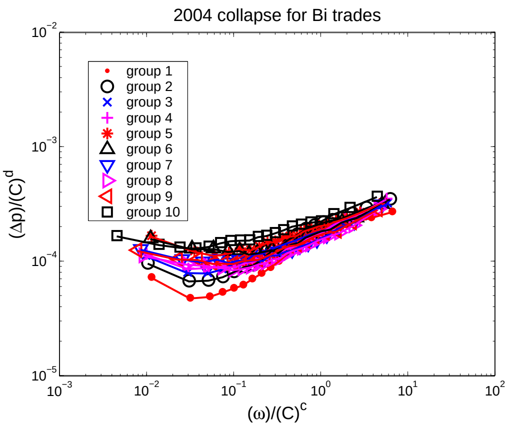

Figure 7: The price shift vs. transaction size, for Bi trades in 2004 with the x-axis and y axis renormalized as described in the text to make the curves collapse onto a single curve.

Table 5: R^2 for the single curve collapse for 2003 and 2004.

	2004 Bi	2004 Si	2003 Bi	2003 Si
R^2	0.989	0.987	0.984	0.991

The extent of the collapse for each year was measured by computing the R^2 statistic. It is computed by looking at two sources of variation, $SS_{collapse}$ and SS_{total}:

$$R^2 = 1 - \frac{SS_{collapse}}{SS_{total}}$$

For a particular year, SS_{total} is the variability (sum of squared errors) of the uncollapsed curves about their mean. $SS_{collapse}$ is the variability of the collapsed curves about their mean. SS_{total} can be thought of as the error if we did not use any information about the scaling factors C^d and C^c. Then $SS_{collapse}$ is the error if we do use these factors to collapse the curves. If $SS_{collapse}$ is much smaller than SS_{total}, then we can conclude that the collapse is satisfactory.

5.5.2 Results and Discussion

The value of exponent d obtained ranges from -0.36 to -0.45, while c ranges from -0.1 to 0.2, for Bi and Si cases in the years 2003 and 2004. Using these values to do the scaling $\Delta p \rightarrow \Delta p/C^d$ and $\omega \rightarrow \omega/C^c$ yields the curves in figure 7. Table 5 gives the R^2 values for the collapsed curves for two years, treating Bi and Si cases separately. The R^2 values in table 5 show that a lot of the variance is accounted for by the collapse. For instance, we see from the table that for both years and for both Bi and Si trades, the collapse has an R^2 that is close to 1. Figure 7 confirms these values — we can see that most of the groups fall neatly onto one curve. The exception seems to be group 2 (the group with the black circles), which appears to have a distinctly different characteristic from the other groups, as indicated back in section 5.4. It does not conform to the shape of the overall market impact curve, and interestingly, the common stock in group 2 in all these cases is Telstra Corporation Limited (TLS), a very heavily traded stock on the ASX. Its percentage volume for the group is 52% in 2004 and 58% in 2003, making TLS the dominant stock of group 2. We surmise that TLS is unusual in that it is partially privatized (50%), and has been subject to two floats, with ongoing speculation making market capitalization a lesser explanatory factor for its liquidity.

6 A Time Aggregated Market Impact Model

The results and disussions in this section summarize work on market impact models described in [29].

6.1 Price formation and Trading activity

The essence of market microstructure is the analysis of how the process of liquidity supply and demand influences price formation. A group of market microstructure models, commonly known as asymmetric information models, assumes there are informed traders in the market who have superior information, and uninformed traders. Informed traders would trade in order to transform their informational advantage into profits. In contrast, uninformed traders trade for liquidity reasons (Easley and O'Hara 1992 [11]).

Much of market microstructure research has been concerned with how trade-related variables can convey information. Theoretical models (e.g. Easley and O'Hara 1987 [12]) and empirical studies (e.g. Hasbrouck [14]) suggest that trade size influences price. The basic intuition is that since informed traders would want to trade larger quantities when they have valuable information, uninformed traders would interpret large orders as an indication of private information and will adjust prices accordingly. Hasbrouck [14] also found evidence of a non-linear and concave relationship between prices and traded volumes. Large trades seem to convey more information than smaller trades, but the marginal information content appears to decrease.

Studies such as Chordia [8] and Brown [7] have used an alternative measure of trading activity: order imbalance, which is buy orders less sell orders. These studies show prices are affected by order imbalances. Chordia [8] further suggests that order imbalances can provide additional power beyond trading volumes in explaining price formation. More extreme order imbalances, regardless of volume, can signal private information or excessive investor interest. Informed traders, who want to conceal their informational advantage, find it optimal to split their orders over time to minimize the price impact of trades, thus causing persistent order imbalances.

6.2 Database

The database includes a complete record of all the new orders, trades, amendments and cancellations of orders that occurred on SEATS from 2000 to 2002. Depending on the type of analysis carried out, our samples are selected either from the 10 most actively traded stocks or the top 100 stocks on the ASX.

We partition the trades into buyer initiated (Bi) and seller initiated (Si) trades. In contrast to the NYSE, for the ASX data we are able to infer initiation directly from the submitted orders, thus ensuring 100% accuracy in determining the direction of trades.

6.3 Price Impact Function

For this model, we investigate how trading and liquidity conditions affect prices, which is a key mechanism of the market. For informational and liquidity reasons, trading and liquidity conditions cause price changes, which are described by the price impact function. We choose two different variables to represent trading activity: transaction volume and order imbalance. We also investigate the price impact of a trade in response to both liquidity and trading conditions by introducing a measure of liquidity fluctuations.

We select order imbalance as a measure of trading activity as in other studies done by Chordia [8], Brown [7], Gabaix et al. [13] and Kissell and Glantz [21]. As mentioned earlier, order imbalance can signal private information. This argument is based on two fundamental concepts:

1. Stock prices change due to the arrival of new information in an efficient market.
2. Based on the fundamental concepts of supply and demand, stock prices would change if there is an imbalance between buy and sell orders for a stock, i.e. prices go up if demand exceeds supply and go down if supply exceeds demand.

These two ideas can be connected by assuming an informed trader would trade a large number of shares causing a supply and demand imbalance, and that an imbalance between supply and demand communicates the private information to the market. We define Q_t as the net order imbalance over a time interval $[t, t + \Delta t]$. We select Δt to be 60 minutes. Q_t measures the imbalance between buy and sell orders, i.e. $Q_t \equiv Q_{B_t} - Q_{S_t}$, where Q_{B_t} is the volume exchanged in buyer initiated trades and Q_{S_t} is the volume exchanged in seller initiated trades. Values of $Q_t > 0$ ($Q_t < 0$) indicates a buy (sell) imbalance during that interval. Following Kissell and Glantz[21], we use normalized order imbalance (Z) rather than absolute imbalance (Q) to relate the imbalance to daily trading activity. The normalized imbalance (Z) is expressed as a percentage of average daily trading volume, where ADV is the average daily volume traded in the stock over the previous 30 intervals.

$$Z = \frac{Q}{ADV} * 100 \tag{24}$$

Since order imbalance is a time-aggregated measure, a time-aggregated measure should also be applied to price impact. Our interpretation of price impact $\Delta P_{Q_t}$is how far the average execution price is moved by all the trades over a time interval. We choose volume-weighted average price (VWAP) as the measure of average execution price.

In general, volume-weighted average price is calculated as:

$$vwap_t = \frac{\sum_{j=1}^{N} P_j * V_j}{\sum_{j=1}^{N} V_j} \tag{25}$$

where P_j and V_j is the price and volume of j_{th} trade respectively, N is the total number of trades over the interval Δt.

In this analysis, trades which are buyer-initiated (Bi) are treated separately from the seller-initiated (Si) trades. Thus, two separate VWAP terms are obtained: $vwap_{B_t}$ represents the VWAP for all buyer-initiated trades and $vwap_{S_t}$ is the VWAP for all seller-initiated trades. The VWAP for t_{th} interval depends on the sign of order imbalance and is defined as:

$$vwap_{Q_t} = \begin{cases} vwap_{B_t} & Q_t > 0 \\ vwap_{S_t} & Q_t < 0 \end{cases} \tag{26}$$

Therefore, price impact is the change in execution price:

$$\Delta P_{Q_t} = vwap_{Q_t} - vwap_{Q_{t-1}} \tag{27}$$

We analyzed all the trades on the top 100 stocks from the 01/08/2000 to 31/07/2001. We then aggregated the trade data for each fixed time interval $\Delta t = 60$ mins. There were a total of 137390 observations. To place all stocks on an equal basis, price impact ΔP_{Q_t} and normalized order imbalance Z_t were normalized by dividing them by their mean values over the sample respectively.

We then investigated the empirical relationship between price impact and order imbalance by estimating the coefficients of equations (28), (29) and (30). The three relationships considered were: linear, power and tanh functions. Both power and tanh functions assume a non-linear relationship between price impact and order imbalance. Buy and sell order imbalances were estimated separately to account for the asymmetry between buy and sell.

Since we are interested in deriving an *average* relationship between price impact and order imbalance, we can group the data into different bins and thus eliminate other factors apart from the price impact of trades. This was accomplished by first sorting the data according to order imbalance, then by grouping the data into fifty bins with each bin having approximately the same number of data points. The mean of the price shifts $\Delta \bar{P}_{Q_t}$ and the median of the order imbalances $\bar{Z}_t$ were then calculated for each of the fifty bins.[1]

Linear function:

$$\Delta \bar{P}_{Q_t} = a_1 \bar{Z}_t \tag{28}$$

Power function:

$$\Delta \bar{P}_{Q_t} = a_1 \bar{Z}_t^{a_2} \tag{29}$$

Tanh function:

$$\Delta \bar{P}_{Q_t} = a_1 \tanh(a_2 \bar{Z}_t) \tag{30}$$

6.3.1 Results and Discussion

Figure 8 illustrates linear, power and tanh functions. Power and tanh functions are the more appropriate fits of the relationship. The coefficients estimated for the functions (equations 28, 29 and 30) are also shown in table 6. The tanh and power functions both have high $R^2(bin)$ values, supporting the power and tanh price impact plots in figure 8, that these two functions are appropriate representation of the empirical relationship between price impact and order imbalance. These results are consistent with the findings by Gabaix et al. [13] that price impact increases at a decreasing rate as order imbalance increases. $R^2(bin)$ is calculated using the binned data which is aggregating many trades for a given imbalance, in contrast to $R^2(raw)$ which is calculated using individual trades. Therefore, $R^2(raw)$ are relatively small because the raw data has a much higher variation than the binned data. On top of that, the average margin between the asymptotic means of the bins and their 95% confidence levels was found to be 2.7342. These two pieces of evidence suggest that there are large error margins on raw data estimations, and price impact costs of single trades are only a small determinant of prices. Price impact of trade packages, if they could be identified, might have more explanatory power. Furthermore, as our analysis is focused only on the liquidity demand imbalance, should the liquidity supply imbalance and the temporal relationship between liquidity demand and supply also be included, then the model might also have more explanatory power.

[1] The median instead of the mean of the order imbalance is chosen due to the asymmetrical distribution of volumes in a bin.

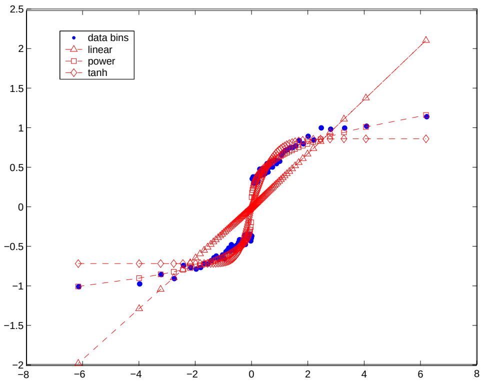

Figure 8: The price impact functions for order imbalance. The binned data (○) are fitted to linear (△), power(□) and tanh(◇) functions. The y-axis is price impact ΔP_{Q_t}, the x-axis is normalized order imbalance Z_t, with positive (negative) values representing buy (sell) order imbalance.

	Buy	(t-stat.)	Sell	(t-stat.)
Linear				
a_1	0.3418	(5.1679)*	0.3203	(4.7983)*
$R^2(bin)$	0.0408		−0.6030	
$R^2(raw)$	−0.0093		−0.0176	
Power				
a_1	0.6464	(5.8947)*	0.6245	(5.8947)*
a_2	0.3277	(1.9188)*	0.2574	(1.4754)
$R^2(bin)$	0.9140		0.8746	
$R^2(raw)$	0.0222		0.0138	
Tanh				
a_1	0.8742	(2.4845)*	0.7073	(2.5069)*
a_2	1.2320	(0.8898)	2.3025	(0.5807)
$R^2(bin)$	0.6024		0.2503	
$R^2(raw)$	0.0151		0.0041	

Table 6: Estimated Coefficients for linear, power and tanh functions which measure the average relationship between price impact and and order imbalance. (*Significant at the 0.05 level)

7 A Time Series Price Impact Model

The average price impact functions discussed in section 6.3 are *time-averaged* measures of the empirical relationship between price impact and trading activity, and the information content of trades. In general, any average price impact function relies on the assumption that price impact is stationary in time (time-independent). However, this assumption is far from realistic. The average price impact function provides a starting point for more sophisticated price impact models which consider the dynamics of price impact of trades.

Hasbrouck [14] was the first to study the stock price and trade dynamics using a vector autoregressive model (VAR). Hasbrouck's analysis revealed that price changes depend on the characteristics of trades, in addition to the current and past levels of prices. Most importantly, Hasbrouck has shown not only trading volume affects prices but also it has a persistent impact on prices. Dufour and Engle [10] extended Hasbrouck's model by using an Autoregressive Conditional Duration (ACD) model, and has shown that the time duration between consecutive trades also affects price behavior. In this section, a similar methodology to that of Hasbrouck [14] is adopted to analyze the dynamics between price changes and order imbalances by applying a vector autoregressive model (VAR) and summarises results presented in [29].

7.1 Intra-day patterns

We selected net order imbalance Q_t as the trade-related variable. Volume-weighted average price (VWAP) was used to represent execution price, and the execution returns are defined as: $r_t = vwap_t - vwap_{t-1}$. Order volume and price volatility have typically been found to exhibit *U-shaped* patterns, with activities concentrated at the beginning and closing of the day. These findings are consistent with the theoretical predictions of Admati and Pfleiderer [1] which show trading concentration is due to the strategic behavior of informed and uninformed traders.

As our model focuses on the impact of net supply and demand imbalance on prices, it is desirable to remove the intra-day patterns to ensure our model is free from any effect of trading concentration. For example, as trades are concentrated at the beginning and end of day, the magnitudes of order imbalance calculated in these intervals are higher not because of greater net supply and demand imbalance, but simply as a result of larger average trading volumes during these intervals.

The effects of intra-day patterns can be removed from the data by the *variable time intervals* method. A daily interval was divided into N intra-day intervals, with each period containing an appropriately equal amount of trading volume, however the window size Δt for each period in real time was different. Specifically, Δt was shorter at the beginning or end of day due to the high concentration of trading activity.

7.2 Optimal time scales and Optimal lags

The number of periods per day (N) and the number of lags (L) varied according to the particular specification of the model that was fitted. To determine the best fitted model with the optimal time scales (i.e. optimal periods per day) and optimal lags, simulation on multiple models over a range of different time scales and lag terms were performed. The data were divided into $N = 1, 2, ..., 6$ different time intervals per trading day. The number of lags L varied from $L = 1, ..., N$, i.e. we restricted $L <= N$ to investigate any relationship within but not beyond a single trading day. Therefore, there were a total of 21 different models to choose from for each stock. The best model was chosen based on the *Bayesian Information Criterion* (BIC).[2]

7.3 Order Imbalances - Returns VAR Model

Formally, the VAR model is:

$$r_t = \sum_{i=1}^{L} a_i r_{t-i} + \sum_{i=1}^{L} b_i Q_{t-i} + \nu_{1,t} \tag{31}$$

$$Q_t = \sum_{i=1}^{L} c_i r_{t-i} + \sum_{i=1}^{L} d_i Q_{t-i} + \nu_{2,t} \tag{32}$$

[2]Models on different time scales were first converted to a daily basis for direct comparisons of BIC values.

ANZ		Periods	3	Lags	1
R_t	Coeff	T-stat.	Q_t	Coeff	T-stat.
a_1	0.2348	(2.5268)*	c_1	-0.2626	$(-2.8616)^*$
b_1	0.2134	(2.3057)*	d_1	0.3853	(4.2152)*
ν_1	-0.0204	(-0.2227)	ν_2	0.1973	(2.1803)*
R^2	0.2799		R^2	0.1272	

Table 7: Estimates of the VAR model for Australia and New Zealand Bank(ANZ). (*Significant at the 0.05 level)

where a, b, c and d the parameters, $\nu_{1,t}$ and $\nu_{2,t}$ are the disturbance terms, t is the variable time period and L is the number of lags.

We analyzed trade data for each of the 10 actively traded stocks on the ASX from the 03/01/2002 to 30/04/2002. The models were estimated via the ordinary least squares (OLS) method.[3]

7.4 Results and Discussion

The optimal model for the stock ANZ is $N = 3$ periods per day with $L = 1$ lagged periods. Estimated coefficients are shown in table 7. Similar results were obtained for the other nine stocks. The optimal models for most of the other stocks have $L = 1$ lagged periods, and N are in the range of 2 to 5.

The dynamic properties of the system can be investigated by examining the impulse response of the returns process. Hasbrouck[14] has suggested that there are both temporary and permanent components of price impact, and the permanent price impact of an unexpected trade is naturally interpreted as the information content of the trade. However, the role played by the slow decaying autocorrelation in trades (suggested by Bouchaud et al. [5]) is an important consideration in the specification and estimation of the VAR model. Slow decaying autocorrelation in trades implies that the trade reversal is weak, this supports our finding that the VAR model has very short lags, as it is unlikely that the addition of trades at longer lags would add significant explanatory power to the estimation. However, the omission of longer lags introduces a different problem. When long-run decaying reversion is present, the model will take into account the initial positive impact on the returns, but ignores the subsequent long run reversion. Therefore, the permanent impact estimated by the VAR model is overestimated.

The main benefit of the VAR system is its ability to demonstrate the dynamic properties of price changes and order imbalances. However, the weakness of this VAR model is that it assumes the relationship between order imbalance and price impact is linear. As we have demonstrated in section 6.3, the relationship between price shifts and order imbalances is unlikely to be linear.

Another limitation of the VAR model is that it relies on the assumption that the price process is a unit root process. Bouchaud et al. [5] found that the correlation function of the sign of trades reveals slowly decaying correlations. This suggests that a fractionally integrated model may be more appropriate.

[3] Since there are identical regressors in each equation, OLS estimates are consistent and asymptotically efficient.

The VAR model has shown that price impact has two components: temporary and permanent effects. Kissell and Glantz [21] have explicitly separated price impact into temporary and permanent components. They have found that temporary impact contributes to 95% of the overall impact.

8 Conclusions

We have reviewed some techniques for algorithmic trade execution. In particular we considered optimised market order strategies for minimising implementation shortfall. The models presented were static models. In the future, we will investigate state space models which also include state variables for price and liquidity. Investigation of how often the models should be recomputed is also important. We envisage extending our models to include improved daily trading volume forecasts, improved price process models encompassing non-gaussian innovations, price jumps and serial correlation, medium term and permanent trading impacts and effects of undisclosed orders.

We then considered trade execution rules for achieving/beating the VWAP execution benchmark. We saw that a market order VWAP strategy, while simple to test, does not on it's own allow us to get all that close to VWAP. We then considered limit order submission strategies. Limit order strategies are more difficult to backtest because we have to track the position of the order in the queue. However, a hybrid market and limit order strategy does show benefits in improving execution performance.

Having considered these approaches to trade execution, we note that as our approach becomes more sophisticated the need to forecast the likely performance of the approaches becomes critical. A backtesting approach using a full order matching system takes us part of the way, however, fails to accurately take into account the reaction of other market participants to our trading. This leads us into a study of market impact. We have summarised some work on market impact from a number of perspectives. In particular, we looked at how market impact varies across the market, the dependence of market impact on liquidity demand and supply, liquidity demand imbalance and the ratio of liquidity supply to liquidity demand. Finally, we looked at how to estimate the dynamics of market impact, and indeed presented a modelling approach based on vector autoregressions as one approach to measuring the reponse of markets to trading and order submission.

Acknowledgements

The authors wish to acknowledge the support of the Capital Markets Co-operative Research Centre (CMCRC) and its industrial partners, for making available the transaction-level data on the Australian Stock Exchange (ASX).

A Microstructure of the Australian Stock Exchange

The ASX[4] is an order-driven market, but in contrast with the NYSE, there are no appointed market makers in the market. Liquidity is instead supplied by traders who submit

[4]Australian Stock Exchange

limit orders. Limit orders specify the quantity to be traded at a specified price. Traders who demand liquidity submit orders at prices on the opposite side of the order book, thus these orders are executed immediately. On the other hand, liquidity suppliers submit orders at prices on the same side of the order book. These orders would remain on the order book, and thus providing liquidity for potential liquidity demanders.

The ASX is a purely electronic order-driven market. All events that occur throughout a trading day are automatically recorded in precise detail. Four of the most important events are ENTER, AMEND, TRADE and DELETE. Together, they provide adequate information to recreate the orderbook.

1. ENTER event refers to an order submission event. It contains all order details - orderid, price, volume, buy or sell etc.
2. AMEND event refers to the amendment of an existing order.
3. TRADE event occurs when a buy order and a sell order meet to result in a trade. This is the event of interest to us in this paper. However, a trade event by itself does not provide enough information to infer details like trade initiation (Bi or Si); the ENTER and AMEND events are needed in the inference process.
4. DELETE event refers to the deletion of an existing order.

The ASX goes through several market phases over the course of each trading day. For this paper, we concentrate only on events that happen during normal trading (10am to 4pm). In other words, trades that result from opening or closing auctions were all discarded. In addition, any trades immediately after a trading halt (suspension of trading during normal trading hours) were also ignored. The inference of trade direction procedure revolves around the fact that every trade is immediately preceded by either an ENTER or an AMEND event. It is therefore possible to determine precisely the attributes of a trade by going back and examining the most recent ENTER or AMEND event, whichever comes first.

As described in section 5.1, trade initiation direction is about which side — buyer or seller — initiated the trade. If the most recent ENTER or AMEND event was a buy order, then the trade is buyer initiated (Bi). If the most recent ENTER or AMEND event was a sell order, then the trade is seller initiated (Si). It is also possible to have a crossing initiated trade, caused by the submission of a crossing order[7]. However, in this paper, crossing initiated trades were discarded in the analysis.

It is instructive to compare our inference procedure to that of [25]'s method of inferring trade direction by comparing quote midpoint and trade price which inherently entails a small degree of uncertainty. Historically the algorithm reportedly classified correctly 85% of all trades. In contrast, automated order-driven trading systems like the ASX report accurately in strict chronological order of the events, and the detailed information about every single order submission event (instead of quote changes) allows the inference of trade direction with complete precision.

References

[1] Anat R. Admati and Paul C. Pfleiderer. A theory of intraday trading patterns: Volume and price variability. *Review of Financial Studies*, 1(1),1988,3-40.

[7] Crossings are a type of order where the buying and selling broker are the same.

[2] Robert F. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk, Working paper, October 2001.

[3] Robert F. Almgren and Neil Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–39, 2000.

[4] Bhattacharjee S M and Seno F, 2001 A measure of data collapse for scaling *Journal of Physics A: Mathematics and General* **34** 6375–6380

[5] Bouchaud J P, Gefen Y, Potters M and Wyart M, 2004 Fluctuations and response in financial markets: the subtle nature of 'random' price changes *Quantitative Finance* **4** 176–190

[6] Dimitris Bertsimas and Andrew W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50, 1998.

[7] Philip Brown, David Walsh, and Andrea Yuen. The interaction between order imbalance and stock price. *Pacific-Basin Finance Journal*, 5(5),1997,539-557.

[8] Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. Order imbalance, liquidity, and market returns. *Journal of Financial Economics*, 65(1),2002,111-130.

[9] Richard Coggins, Adam Blazejewski, and Michael Aitken. Optimal execution of trades in a limit order market. *International Conference on Computational Intelligence for Financial Engineering*, March, 2003,371-378.

[10] Dufour A and Engle R F 2000 Time and the Price Impact of a Trade *Journal of Finance* **46** 2467–2498

[11] David Easley and Maureen O'Hara. Time and the process of security price adjustment, journal of finance. *Journal of Finance*, 47(2),1992,576-606.

[12] David Easley and Maureen O'Hara. Price, trade size, and information in securities markets. *, Journal of Financial Economics*, 19(1),1987,69-90.

[13] Xavier Gabaix, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423,2003,267-270.

[14] Joel Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46(1),1991,179-207.

[15] Lawrence E. Harris. Optimal dynamic order submission strategies in some stylized trading problems. *Financial Markets, Institutions and Instruments*, 7(2):1–76, 1998.

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

[17] Hasbrouck J 1991 Measuring the information content of stock trades *Journal of Finance* **46** 178–208.

[18] Holthausen R, Leftwich R and Mayers D 1987 The effect of large block transactions on security prices: a cross-sectional analysis *Journal of Financial Economics* **19** 237–267.

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

[20] Donald B. Keim and Ananth Madhavan. The cost of institutional equity trades. *Financial Analysts Journal*, 54(4):50–69, 1998.

[21] Robert Kissell and Morton Glantz. *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk.* (American Management Association, 2003).

[22] Hizuru Konishi. Optimal slice of a vwap trade. *Journal of Financial Markets*, 5(2):197–221, 2002.

[23] Kraus A and Stoll H 1972 Price impacts of block trading on the New York stock exchange *Journal of Finance* **27** 569–588

[24] Kyle A 1985 Continuous auctions and insider trading *Econometrica* **53** 1315–1335

[25] Lee C M C and Ready M 1991 Inferring trade direction from intraday data *Journal of Finance* **46** 733–746

[26] Lillo F, Farmer J D and Mantegna R N 2002 Single Curve Collapse of the Price Impact Function for the New York Stock Exchange *Preprint* cond-mat/0207428

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

[28] Lim, M and Coggins R 2004, The immediate price impact of trades on the ASX, *Journal of Quantitative Finance*, to appear.

[29] Lo, K and Coggins R 2004, The price impact of trades on the ASX, *Financial Engineering and Applications.*

[30] Loeb T F 1983 Trading cost: the critical link between investment information and results *Financial Analysts Journal* **39** 39–43

[31] Jonathan R. Macey and Maureen O'Hara. The law and economics of best execution. *Journal of Financial Intermediation*, 6:188–223, 1997.

[32] Mario J. Miranda and Paul L. Fackler. *Applied Computational Economics and Finance.* MIT Press, 1st edition, 2002.

[33] O' Hara M 1995 *Market Microstructure Theory* (Blackwell Publishers)

[34] Andre F. Perold. The implementation shortfall: paper versus reality. *Journal of Portfolio Management*, 14:4–9, 1988.

[35] Stanley H E 1999 Scaling, Universality and Renormalization: Three Pillars of Modern Critical Phenomena *Reviews of Modern Physics* **71 (1999)** 358–366

[36] Torre N 1997 *BARRA Market Impact Model Handbook* (BARRA Inc, Berkeley CA)

[37] Jack L. Treynor. What does it take to win the trading game? *Financial Analysts Journal*, 37(1):55–60, 1981.