Equity market impact

The impact of large trades on prices is very important and widely discussed, but rarely measured. Using a large data set from a major bank and a simple but realistic theoretical model, Robert Almgren, Chee Thum, Emmanuel Hauptmann and Hong Li propose that impact is a 3/5 power law of block size, with specific dependence on trade duration, daily volume, volatility and shares outstanding. The results can be directly incorporated into an optimal trade scheduling algorithm and pre- and post-trade cost estimation

Transaction costs are widely recognised as an important determinant of investment performance (see, for example, Freyre-Sanders, Guobuzaite & Byrne, 2004). Not only do they affect the realised results of an active investment strategy, but they also control how rapidly assets can be converted into cash should the need arise. Such costs generally fall into two categories:

■ Direct costs are commissions and fees that are explicitly stated and easily measured. These are important and should be minimised, but are not the focus of this article.

■ Indirect costs are not explicitly stated. For large trades, the most important component of these is the impact of the trader's own actions on the market. These costs are notoriously difficult to measure, but they are the most amenable to improvement by careful trade management and execution.

This article presents a quantitative analysis of market impact costs based on a large sample of Citigroup US equity brokerage executions. We use a simple theoretical model that lets us bring in the very important role of the rate of execution.

The model and its calibration are constructed to satisfy two criteria:

■ Predicted costs are quantitatively accurate, as determined by direct fit and by out-of-sample back-testing, as well as extensive consultation with traders and other market participants.

■ The results may be used directly as input to an optimal portfolio trade scheduling algorithm. (The scheduling algorithm itself is non-trivial and will be published elsewhere.)

The results of this study are currently being implemented in Citigroup's Best Execution Consulting Services (BECS) software, for use internally by all desks as well as clients of the equity division. The current work is focused on the US market but work is under way to extend it to global equities. BECS is the delivery platform for the next generation of Citigroup's trading analytic tools, both pre- and post-execution.

The pre-trade model presented here is an extension of the market standard existing model that has been delivered through the StockFacts Pro software for the past 14 years (Sorensen *et al*, 1998). The new pre-trade model is based on better developed empirical foundations: it is based on real trading data taking time into consideration while verifying the results through post trade analysis. Table A summarises the advantages and some disadvantages of our approach.

Much work in both the academic and the industrial communities has been devoted to understanding and quantifying market impact costs. Many academic studies have worked only with publicly available data, such as the trade and quote (TAQ) tick record from the New York Stock Exchange (NYSE). Breen, Hodrick & Korajczyk (2002) regress net market movement over five-minute and half-hour time periods against the net buy-sell imbalance during the same period, using a linear impact model. A similar model is developed in Kissell & Glantz (2003). Rydberg & Shephard (2003) develop a rich econometric framework for describing price motions; Dufour & Engle (2000) investigate the key role of waiting time between successive trades. Using techniques from statistical physics, Lillo, Farmer & Mantegna (2003) look for a power-law scaling in the impact cost function, and find significant dependence on total market capitalisation as well as daily volume, and Bouchaud *et al* (2004) discover non-trivial serial correlation in volume and price data.

The publicly available data sets lack reliable classification of individual trades as buyer- or seller-initiated. Even more significantly, each transaction exists in isolation; there is no information on sequences of trades that form part of a large transaction. Some academic studies have used limited data sets made available by asset managers that do have this information, where the date but not the time duration of the trade is known (Chan & Lakonishok, 1995, Holthausen, Leftwich & Mayers, 1990, and Keim & Madhavan, 1996).

The transaction cost model embedded in our analysis is based on the model presented by Almgren & Chriss (2000) with non-linear extensions from Almgren (2003). The essential features of this model, as described in below, are that it explicitly divides market impact costs into a permanent component associated with information, and a temporary component arising from the liquidity demands made by execution in a short time.

Data

The data set on which we base our analysis contains, before filtering, almost 700,000 US stock trade orders executed by Citigroup equity trading desks for the 19-month period from December 2001 to June 2003. (The model actually used within the BECS software is estimated on an ongoing basis, to reflect changes in the trading environment.) We now briefly describe and characterise the raw data, and then the particular quantities of interest that we extract from it.

☐ **Description and filters.** Each order is broken into one or more transactions, each of which may generate one or more executions. For each order, we have the following information:

■ The stock symbol, requested order size (number of shares) and sign (buy or sell) of the entire order. Client identification is removed.

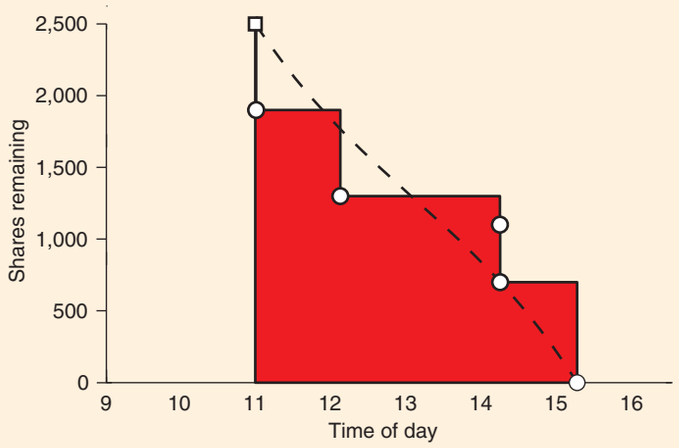
■ The times and methods by which transactions were submitted by the Citigroup trader to the market. We take the time t_0 of the first transaction to be the start of the order. Some of these transactions are sent as market orders, some are sent as limit orders, and some are submitted to Citigroup's automated VWAP server. Except for the starting time t_0, and except to exclude VWAP orders, we make no use of this transaction information.

■ The times, sizes and prices of execution corresponding to each transaction. Some transactions are cancelled or only partially executed; we use only the completed price and size. We denote execution times by $t_1, \ldots, t_n$, sizes by $x_1, \ldots, x_n$, and prices by $S_1, \ldots, S_n$.

A. Distinguishing features of our model

Advantages	Disadvantages
■ Calibrated from real data	■ Based only on Citigroup data
■ Includes time component	■ Little data for small-cap stocks
■ Incorporates intra-day profiles	■ Little data for very large trades
■ Uses non-linear impact functions	
■ Confidence levels for coefficients	

1. A typical trading trajectory



The vertical axis is shares remaining and each step downwards is one execution. The trajectory starts at the first transaction recorded in the system; the program ends when the last execution has been completed. The dashed line is our continuous-time approximation

B. Summary statistics of orders in our sample

	Mean	Min	Q1	Median	Q3	Max
Total cost (%)	0.04	−3.74	−0.11	0.03	0.19	3.55
Permanent cost (*I*, %)	0.01	−3.95	−0.17	0.01	0.19	2.66
Temporary cost (*J*, %)	0.03	−3.57	−0.11	0.02	0.17	2.33
Shares/ADV (|*X*|, %)	1.51	0.25	0.38	0.62	1.36	88.62
Time (days)	0.39	0.00	0.10	0.32	0.65	1.01
Daily volatility (%)	2.68	0.70	1.70	2.20	3.00	12.50
Mean spread (%)	0.14	0.03	0.08	0.11	0.16	2.37

Note: mean and quartile levels for each of several descriptive variables. The three cost variables are very nearly symmetrically distributed about zero (*I* and *J* are defined in the 'Trajectory cost model' section)

All orders are completed within one day (though not necessarily completely filled).

Figure 1 shows a typical example. A sell order for 2,500 shares of DRI was entered into the system at t_0 = 10:59am. The transactions submitted by the trader generated $n = 5$ executions, of which the last one completed at t_n = 15:15. The dashed line in the figure shows the continuous-time approximation that we use in the data fitting: execution follows the average day's volume profile over the duration of the trade execution.

In addition, we have various additional pieces of information, such as the instructions given by the client to the trader for the order, such as 'over the day', 'market on close', 'market on open', 'VWAP' or blank.

The total sample contains 682,562 orders, but for our data analysis we use only a subset.

■ To exclude small and thinly traded stocks, we consider only orders on stocks in the Standard & Poor's 500 index, which represent about half of the total number of orders but a large majority of the total dollar value. Even within this universe, we have enough diversity to explore dependence on market capitalisation, and we have both New York Stock Exchange and over-the-counter stocks.

■ We exclude approximately 400 orders for which the stock exhibits more than 12.5% daily volatility (200% annual).

Furthermore, we want only orders that are reasonably representative of the active scheduling strategies that are our ultimate goal.

■ We exclude orders for which the client requested 'market on close' or 'market on open'. These orders are likely to be executed with strongly nonlinear profiles, which do not satisfy our modelling assumption. (There are only a few hundred of these orders.)

■ We exclude orders for which the client requested VWAP execution. These orders have consistently long execution times and represent very small rates of trading relative to market volume. (These are about 16% of the total number of orders.)

■ Also, we exclude orders for which any executions are recorded after 4:10pm, approximately 10% of the total. In many cases, these orders use Citigroup's block desk for some or all of the transactions, and the fills are reported some time after the order is completed. Therefore, we do not have reliable time information.

This exclusion, together with our use of filled size in place of originally requested size, could be a source of significant bias. For example, if clients and traders consistently used limit orders, orders might be filled only if the price moved in a favourable direction. Analysis of our data set suggests that this effect is not significant – for example, we obtain almost exactly the same coefficients with or without partially filled orders – and informal discussions with traders confirm the belief that partial fills are not the result of limit order strategy.

Most significantly, we exclude small orders since our goal is to estimate transaction costs in the range where they are significant. Specifically, we include only orders that:

■ have at least two completed transactions;

■ are at least 1,000 shares; and

■ are at least 0.25% of average daily volume in that stock.

The results of our model are reasonably stable under changes in these criteria. After this filtering, we have 29,509 orders in our data set. The largest number of executions for any order is $n = 548$; the median is around five. The median time is around 30 minutes.

Table B shows some descriptive statistics of our sample. Most of our orders constitute only a few per cent of typical market volume, and our model is designed to work within this range of values. Orders larger than a few per cent of daily volume have substantial sources of uncertainty that are not modelled here, and we cannot claim that our model accurately represents them.

□ **Variables.** The goal of our study is to describe market impact in terms of a small number of input variables. Here we define precisely what market impacts we are measuring, and what primary and auxiliary variables we will use to model them.

■ Observables. Let $S(t)$ be the price of the asset being traded. For each order, we define the following price points of interest: S_0 is the market price before this order begins executing; S_{post} is the market price after this order is completed; and $\bar{S}$ is the average realised price on the order

The realised price $\bar{S} = \Sigma x_j S_j / \Sigma x_j$ is calculated from the transaction data set. The market prices S_0 and S_{post} are bid-ask mid-points from TAQ.

The pre-trade price S_0 is the price before the impact of the trade begins to be felt (this is an approximation, since some information may leak before any record enters the system). We compute S_0 from the latest quote just preceding the first transaction.

The post-trade price S_{post} should capture the 'permanent' effects of the trade program. That is, it should be taken long enough after the last execution that any effects of temporary liquidity have dissipated. In repeatedly performing the fits described below ('Cross-sectional description'), we have found that 30 minutes after the last execution is adequate to achieve this. For shorter time intervals, the regressed values depend on the time lag, and at about this level the variation stops. That is, we define:

$$t_{post} = t_n + 30 \text{ minutes}$$

The price S_{post} is taken from the first quote following t_{post}. If t_{post} is after market close, then we carry over to the next morning. This risks distorting the results by including excessive overnight volatility, but we have found this to give more consistent results than the alternative of truncating at the close.

Based on these prices, we define the following impact variables:

$$\text{Permanent impact}: \quad I = \frac{S_{post} - S_0}{S_0} \qquad \text{Realised impact}: \quad J = \frac{\bar{S} - S_0}{S_0}$$

The 'effective' impact J is the quantity of most interest, since it determines the actual cash received or spent on the trade. In the model below, we will define temporary impact to be J minus a suitable fraction of I, and this temporary impact will be the quantity described by our theory. We cannot sensibly define temporary impact until we write this model.

On any individual order, the impacts I, J may be either positive or negative. In fact, since volatility is a very large contributor to their values, they are almost equally likely to have either sign. They are defined so that positive cost is experienced if I, J have the same sign as the total order X: for a buy order with $X > 0$, positive cost means that the price $S(t)$ moves upward. We expect the average values of I, J, taken across many orders, to have the same sign as X.

■ Volume time. The level of market activity is known to vary substantially and consistently between different periods of the trading day; this intra-day variation affects both the volume profile and the variance of prices. To capture this effect, we perform all our computations in volume time τ, which represents the fraction of an average day's volume that has executed up to clock time t. Thus a constant-rate trajectory in the τ variable corresponds to a VWAP execution in real time, as shown in figure 1. The relationship between t and τ is independent of the total daily volume; we scale it so that $\tau = 0$ at market open and $\tau = 1$ at market close.

We map each of the clock times $t_0, \ldots, t_n$ in the data set to a corresponding volume time $\tau_0, \ldots, \tau_n$. Since the stocks in our sample are heavily traded, in this article we use a non-parametric estimator that directly measures differences in τ: the shares traded during the period corresponding to the execution of each order. Figure 2 illustrates the empirical profiles. The fluctuations in these graphs are the approximate size of statistical error in the volume calculation for a 15-minute trade; these errors are typically 5% or less, and are smaller for longer trades.

■ Explanatory variables. We want to describe the impacts I and J in terms of the input quantities:

$$X = \sum_{j=1}^{n} x_j \quad = \text{Total executed size in shares}$$

$$T = \tau_n - \tau_0 \quad = \text{Volume duration of active trading}$$

$$T_{post} = \tau_{post} - \tau_0 \quad = \text{Volume duration of impact}$$

As noted above, X is positive for a buy order, negative for sell. We have explored defining T using a size-weighted average of execution times but the results are not substantially different. We make no use of the intermediate execution times $\tau_1, \ldots, \tau_{n-1}$, and make no use of the execution sizes except in calculating the order size and the mean realised price.

In eventual application for trajectory optimisation, the size X will be assumed given, and the execution schedule, here represented by T, will be optimised. In general, the solution will be a complicated time-dependent trajectory that will be parameterised by a time scale T. For the purposes of data modelling, we ignore the trajectory optimisation and take the schedules to be determined only by the single number T.

■ Auxiliary variables. Although our goal is to explain the dependence of the impact costs I, J on order size X and trade time T, other market variables will influence the solution. The most important of these are: V, which is the average daily volume in shares, and σ, which is the daily volatility.

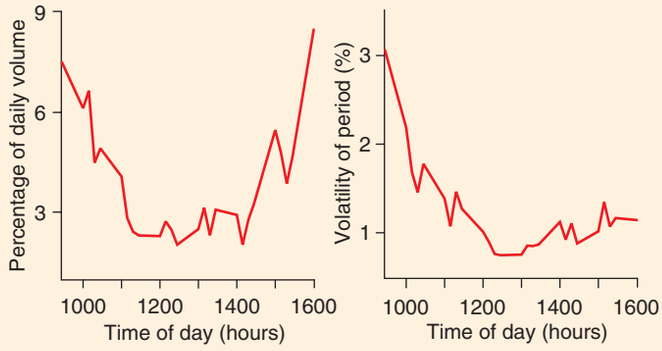
V is a 10-day moving average. For volatility, we use an intra-day estimator that makes use of every transaction in the day. We find that it is important to track changes in these variables not only between different stocks but also across time for the same stock.

These values serve primarily to 'normalise' the active variable across stocks with widely varying properties. It seems natural that order size X should be measured as a fraction of average daily volume V: X/V is a more natural variable than X itself.

In our model presented below, order size as a fraction of average volume traded during the time of execution will also be seen to be important. We estimate VT directly by taking the average volume that executed between clock times t_0 and t_n over the previous 10 days. In fact, since in our model trade duration T appears only in the combination VT, this avoids the need to measure T directly.

We use volatility to scale the impacts: a certain level of participation in the daily volume should cause a certain level of participation in the 'normal' motion of the stock. Our empirical results show that volatility is the most important scale factor for cost impact.

2. Ten-day average intra-day volume and volatility profiles, on 15-minute intervals



Our approach defines to a new time scale determined empirically by the cumulative volume profile; implicitly this takes the volatility profile to be the same, which is approximately valid. Our estimation introduces statistical error, which is roughly the same size as the fluctuations in these graphs

Trajectory cost model

The model we use is based on the framework developed by Almgren & Chriss (2000) and Almgren (2003), with simplifications made to facilitate the data fitting. The main simplification is the assumption that the rate of trading is constant (in volume time). In addition, we neglect cross-impact, since our data set has no information about the effect of trading one stock on the price of another.

We decompose the price impact into two components:

■ A permanent component that reflects the information transmitted to the market by the buy/sell imbalance. This component is believed to be roughly independent of trade scheduling; 'stealth' trading is not admitted by this construction. In our data fit, this component will be independent of the execution time T.

■ A temporary component reflects the price concession needed to attract counterparties within a specified short time interval. This component is highly sensitive to trade scheduling; here, it will depend strongly on T.

More detailed conceptual frameworks have been developed (Bouchaud *et al*, 2004), but this easily understood model has become traditional in the industry and in academic literature (Madhavan, 2000).

The realised price impact is a combination of these two effects. In terms of the realised and permanent impact defined above and observed from the data, our model may be summarised as:

$$\text{Realised} = \text{Permanent} + \text{Temporary} + \text{Noise}$$

with suitable coefficients and scaling depending on T. Thus the temporary impact is obtained as a difference between the realised impact and the permanent impact; it is not directly observable although we have a theoretical model for it.

We start with an initial desired order of X shares. We assume this order is completed by uniform rate of trading over a volume time interval T. That is, the trade rate in volume units is $v = X/T$, and is held constant until the program is completed. Constant rate in these units is equivalent to VWAP execution during the time of execution. Note that v has the same sign as X; thus $v > 0$ for an buy order and $v < 0$ for a sell order. Market impact will move the price in the same direction as v.

□ **Permanent impact.** Our model postulates that the asset price $S(\tau)$ follows an arithmetic Brownian motion, with a drift term that depends on our trade rate v. That is:

$$dS = S_0 g(v)\, d\tau + S_0 \sigma\, dB$$

where $B(\tau)$ is a standard Brownian motion (or Bachelier process), and $g(v)$ is the permanent impact function; the only assumption we make so far is that $g(v)$ is increasing and has $g(0) = 0$. As noted above, τ is volume time, representing the fraction average of an average day's volume that has executed so far. We integrate this expression in time, taking v to equal X/T for $0 \le \tau \le T$, and get the permanent impact:

$$I = Tg\left(\frac{X}{T}\right) + \sigma\sqrt{T_{post}}\,\xi \tag{1}$$

where $\xi \sim \mathcal{N}(0, 1)$ is a standard Gaussian variate.

Note that if $g(v)$ is a linear function, then the accumulated drift at time τ is proportional to $X\tau/T$, the number of shares we have executed to time τ, and the total permanent impact I is proportional to the total order size X, independently of the time scale T.

□ **Temporary impact.** The price actually received from our trades is:

$$\tilde{S}(\tau) = S(\tau) + S_0 h\left(\frac{X}{T}\right)$$

where $h(v)$ is the temporary impact function. For convenience, we have scaled it by the market price at the start of trading, since the time intervals involved are all less than one day.

This expression is a continuous-time approximation to a discrete process. A more accurate description would be to imagine that time is broken into intervals such as, say, one hour or 30 minutes. Within each interval, the average price we realise on our trades during that interval will be slightly less favourable than the average price that an unbiased observer would measure during that time interval. The unbiased price is affected on previous trades that we have executed before this interval (as well as volatility), but not on their timing. The additional concession during this time interval is strongly dependent on the number of shares that we execute in this interval.

At a constant liquidation rate, calculating the time average of the execution price gives the temporary impact expression:

$$J - \frac{I}{2} = h\left(\frac{X}{T}\right) + \sigma\left(\sqrt{\frac{T}{12}\left(4 - 3\frac{T}{T_{post}}\right)}\,\chi - \frac{T_{post} - T}{2\sqrt{T_{post}}}\,\xi\right) \tag{2}$$

where $\chi \sim N(0, 1)$ is independent of ξ. The term $I/2$ reflects the effect on the later execution prices of permanent impact caused by the earlier parts of the program.

The rather complicated error expression reflects the fluctuation of the middle part of the Brownian motion on $[0, T]$ relative to its end point at T_{post}. It is used only for heteroscedasticity corrections in the regression fits below.

Equations (1) and (2) give us explicit expressions for the permanent and temporary cost components I, J in terms of the values of the functions g, h at known trade rates, together with estimates of the magnitude of the error coming from volatility. The data-fitting procedure is in principle straightforward: we compute the costs I, J from the transaction data, and regress those values against the order size and time as indicated, to extract directly the functions $g(v)$, $h(v)$.

□ **Choice of functional form.** We now address the question of what should be the structure of the permanent impact function $g(v)$ and the temporary impact function $h(v)$. Even with our large sample, it is not possible to extract these functions purely from the data; we must make a hypothesis about their structure.

We postulate that these functions are power laws, that is, that:

$$g(v) = \pm\gamma|v|^{\alpha} \quad \text{and} \quad h(v) = \pm\eta|v|^{\beta}$$

where the numerical values of the (dimensionless) coefficients γ, η and the exponents α, β are to be determined by linear and non-linear regression on the data. The sign is to be chosen so $g(v)$ and $h(v)$ have the same sign as v.

The class of power law functions is extremely broad. It includes concave functions (exponent < 1), convex functions (exponent > 1) and linear functions (exponent = 1). It is the functional form that is implicitly assumed by fitting straight lines on a log-log plot, as is very common in physics and has been done in this context, for example, by Lillo, Farmer & Mantegna (2003).

We take the same coefficients for buy orders ($v > 0$) and sell orders ($v < 0$). It would be a trivial modification to introduce different coefficients $\gamma_{\pm}$ and $\eta_{\pm}$ for the two sides, but our exploratory data analysis has not indicated a strong need for this. Similarly, it would be possible to use different coefficients for stocks traded on different exchanges but this does not appear to be necessary.

We are far from neutral in the choice of the exponents. For the permanent impact function, there is strong reason to prefer the linear model with $\alpha = 1$. This is the only value for which the model is free from 'quasi-arbitrage' (Huberman & Stanzl, 2004). Furthermore, the linear function is the only one for which the permanent price impact is independent of trading time; this is a substantial conceptual simplification, though, of course, it must be supported by the data.

For temporary impact, there is ample empirical evidence indicating that the function should be concave, that is, $0 < \beta < 1$. This evidence dates back to Loeb (1983) and is strongly demonstrated by the fits in Lillo, Farmer & Mantegna (2003). In particular, theoretical arguments (Barra, 1997) suggest that the particular value $\beta = 1/2$ is especially plausible, giving a square-root impact function.

Our approach is as follows. We shall make unprejudiced fits of the power law functions to the entire data set, and determine our best estimates for the exponents α, β. We will then test the validity of the values $\alpha = 1$ and $\beta = 1/2$, to validate the linear and square-root candidate functional forms.

Once the exponents have been selected, simple linear regression is adequate to determine the coefficients. In this regression, we use heteroscedastic weighting with the error magnitudes from (1) and (2). The result of this regression is not only values for the coefficients, but also a collection of error residuals ξ and χ, which must be evaluated for normality as the theory supposes.

Cross-sectional description

Above we have assumed an 'ideal' asset, all of whose properties are constant in time. For any real asset, the parameters that determine market impact will vary with time. For example, one would expect that execution of a given number of shares would incur higher impact costs on a day with unusually low volume or with unusually high volatility.

We therefore assume that the natural variable for characterising the size of an overall order or of a rate of trading is not shares *per se* but the number of shares relative to a best estimate of the background flow for that stock in the time period when trading will occur. That is, the impact cost functions should be expressed in terms of the dimensionless quantity X/VT rather than X itself, where V is the average number of shares per day defined above (see 'Variables').

Furthermore, we suppose the motion of the price should not be given as a raw percentage figure, but should be expressed as a fraction of the 'normal' daily motion of the price, as expressed by the volatility σ.

With these assumptions, we modify the expressions (1) and (2) to:

$$I = \sigma Tg\left(\frac{X}{VT}\right) + \langle\text{noise}\rangle \tag{3}$$

$$J - \frac{I}{2} = \sigma h\left(\frac{X}{VT}\right) + \langle\text{noise}\rangle \tag{4}$$

where the 'noise' is the error expressions depending on volatility. Now $g(\cdot)$ and $h(\cdot)$ are dimensionless functions of a dimensionless variable. They are assumed to be constant in time for a single stock across days when σ and V vary. We now investigate the dependence of these functions on cross-stock variables.

☐ **Model determination.** To bring the full size of our data set into play, we must address the more complicated and less precise question of how the impact functions vary across stocks, that is, how they might depend on variables such as total market capitalisation, shares outstanding, bid-ask spread or other quantities. We consider permanent and temporary impact separately.

■ Permanent. We insert a 'liquidity factor' $\mathcal{L}$ into the permanent cost function $g(v)$, where $\mathcal{L}$ depends on the market parameters characterising each individual stock (in addition to daily volume and volatility). There are several candidates for the inputs to $\mathcal{L}$:

☐ Shares outstanding. We constrain the form of $\mathcal{L}$ to be

$$\mathcal{L} = \left(\frac{\Theta}{V}\right)^{\delta}$$

where Θ is the total number of shares outstanding, and the exponent δ is to be determined. The dimensionless ratio Θ/V is the inverse of 'turnover', the fraction of the company's value traded each day. This is a natural explanatory variable, and has been used in empirical studies such as Breen, Hodrick & Korajczyk (2002).

☐ Bid-ask spread. We have found no consistent dependence on the bid-ask spread across our sample, so we do not include it in $\mathcal{L}$.

☐ Market capitalisation. This differs from shares outstanding by the price per share, so including this factor is equivalent to including a 'price effect'. Our empirical studies suggest there is a persistent price effect, as also found by Lillo, Farmer & Mantegna (2003), but that the dependence is weak enough that we neglect it in favour of the conceptually simpler quantity Θ/V.

■ Temporary. In extensive preliminary exploration, we have found that the temporary cost function $h(v)$ does not require any stock-specific modification: liquidity cost as a fraction of volatility depends only on shares traded as a fraction of average daily volume.

■ Determination of exponent. After assuming the functional form explained above, we confirm the model and determine the exponent δ by performing a non-linear regression of the form:

$$\frac{I}{\sigma} = \gamma T \operatorname{sgn}(X)\left|\frac{X}{VT}\right|^{\alpha}\left(\frac{\Theta}{V}\right)^{\delta} + \langle\text{noise}\rangle \tag{5}$$

$$\frac{1}{\sigma}\left(J - \frac{I}{2}\right) = \eta \operatorname{sgn}(X)\left|\frac{X}{VT}\right|^{\beta} + \langle\text{noise}\rangle \tag{6}$$

where 'noise' is again the heteroscedastic error term from (1), and sgn is the sign function. We use a modified Gauss-Newton optimisation algorithm to determine the values of alpha, delta and beta that minimise the normalised residuals. The results are:

$$\alpha = 0.891 \pm 0.10 \qquad \delta = 0.267 \pm 0.22 \qquad \beta = 0.600 \pm 0.038$$

Here, as throughout this article, the error bars expressed with ± are one standard deviation, assuming a Gaussian error model. Thus the 'true' value can be expected to be within this range with 67% probability, and within a range twice as large with 95% probability.

From these values, we draw the following conclusions:

☐ The value $\alpha = 1$, for linear permanent impact, cannot reliably be rejected. In view of the enormous practical simplification of linear permanent impact, we choose to use $\alpha = 1$.

☐ The liquidity factor is very approximately $\delta = 1/4$.

☐ For temporary impact, our analysis confirms the concavity of the function with β strictly inferior to one. This confirms the fact that the bigger the trades made by fund managers on the market, the less additional cost they experience per share traded. At the 95% confidence level, the square-root model $\beta = 1/2$ is rejected. We will therefore fix on the temporary cost exponent $\beta = 3/5$. In comparison with the square-root model, this gives slightly smaller costs for small trades, and slightly larger costs for large trades.

Note that because $\delta > 0$, for fixed values of the number X of shares in the order and the average daily volume V, the cost increases with Θ, the total number of shares outstanding. In effect, a larger number of outstanding shares means that a smaller fraction of the company is traded each day, so a given fraction of that flow has greater impact.

Therefore, these results confirm empirically the theoretical arguments of Huberman & Stanzl (2004) for permanent impact that is linear in block size, and the concavity of temporary impact as has been widely described in the literature for both theoretical and empirical reasons.

☐ **Determination of coefficients.** After fixing the exponent values, we determine the values of γ and η by linear regression of the models (5) and (6), using the heteroscedastic error estimates given in (1) and (2). We find:

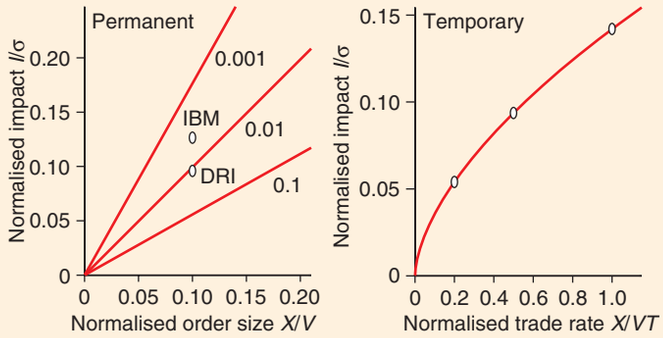
$$\gamma = 0.314 \pm 0.041 \quad (t = 7.7) \qquad \eta = 0.142 \pm 0.0062 \quad (t = 23)$$

The t statistic is calculated assuming that the Gaussian model expressed in (1) and (2) is valid; the error estimates are the value divided by the t statistic. Although the actual residuals are fat-tailed as we discuss below, these estimates indicate that the coefficient values are highly significant.

The R^2 values are typically less than 1%, indicating that only a small part of the value of the dependent variables I and J is explained by the model in terms of the independent variables. This is precisely what is expected, given the small size of the permanent impact term relative to the random motion of the price due to volatility during the trade execution.

This persistent cost, though small, is of major importance since it is on average the cost incurred while trading by fund managers. Furthermore,

3. Permanent and temporary price impact



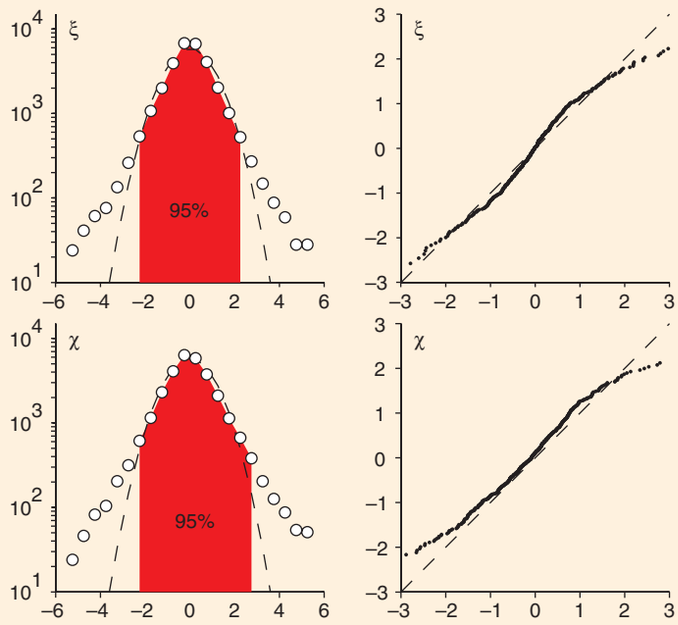
The left graph shows permanent price impact, giving normalised price motion in terms of normalised order size for three values of daily turnover $V/\Theta = 0.001, 0.01, 0.1$. The right graph is temporary impact cost function, in terms of normalised order rate. The examples from table C are also shown. For permanent cost, the location on the graph depends on asset properties but not on time of trade execution; for temporary cost, the location depends on time but not on asset

C. Example of impact costs

		IBM			DRI		
Average daily volume	V	6.561m			1.929m		
Shares outstanding	Θ	1,728m			168m		
Inverse turnover	Θ/V	263			87		
Daily volatility (%)	σ	1.57			2.26		
Normalised trade size	X/V	0.1			0.1		
Normalised permanent	I/σ	0.126			0.096		
Perm. price impact (bp)	I	20			22		
Trade duration (days)	T	0.1	0.2	0.5	0.1	0.2	0.5
Normalised temporary	K/σ	0.142	0.094	0.054	0.142	0.094	0.054
Temp. impact cost (bp)	K	22	15	8	32	21	12
Realised cost (bp)	J	32	25	18	43	32	23

Note: examples of permanent and temporary impact costs are shown, for a purchase of 10% of the day's average volume, in two different large-cap stocks. The permanent cost is independent of time of execution. The temporary cost depends on the time, but across different assets it is the same fraction of daily volatility. We write $K = J - I/2$.

4. Permanent and temporary error residuals



Permanent error residuals ξ and temporary residual χ (29,509 points). In the left column, the vertical axis is the number of sample values in the bin, on a log scale; the dashed line shows the value that would be expected in a standard Gaussian distribution of zero mean and unit variance (not adjusted to sample mean and variance). In the right column, the horizontal axis is the residual, and the vertical axis is values of the cumulative normal; if the distribution were normal, all points would lie on the dashed line. The distribution is clearly fat-tailed, but the standard Gaussian is a reasonable fit to the central part

since most orders are part of large portfolio trades, the volatility cost actually experienced on the portfolio level is considerably lower than exhibited in the stock-level analysis, increasing the significance of the fraction of impact cost estimated. As previously mentioned, the non-linear optimisation of the volatility versus impact cost trade-off at the portfolio level is a subject of current work.

The dimensionless numbers γ and η are the 'universal coefficients of market impact'. According to our model, they apply to every order and every asset in the entire data set. To summarise, they are to be inserted into the equations:

$$I = \gamma\sigma\frac{X}{V}\left(\frac{\Theta}{V}\right)^{1/4} + \langle\text{noise}\rangle \qquad J = \frac{I}{2} + \text{sgn}(X)\eta\sigma\left|\frac{X}{VT}\right|^{3/5} + \langle\text{noise}\rangle$$

giving the expectation of impact costs; in any particular order the realised values will vary greatly due to volatility. Recall that I is not a cost, but is simply the net price motion from pre-trade to post-trade. The actual cost experienced on the order is J.

We have chosen these simple forms in order to have a single model that applies reasonably well across the entire data set, which consists entirely of large-cap stocks in the US markets. More detailed models could be constructed to capture more limited sets of dates or assets, or to account for variations across global markets. In practice, we expect that the coefficients, perhaps the exponents, and maybe even the functional forms, will be continually updated to reflect the most recent data.

■ Examples. In figure 3, we show the impact cost functions, and in table C we show specific numerical examples for two large-cap stocks, when the customer buys 10% of the average daily volume. Because DRI turns over 1/87 of its total float each day, whereas IBM turns over only 1/263, trading one-tenth of one day's volume causes a permanent price move of only 0.1 times volatility for DRI, but 0.13 times for IBM; half of this is experienced as cost. Because the permanent impact function is linear, the permanent cost numbers are independent of the time scale of execution.

□ **Residual analysis.** The result of our analysis is not simply the values of the coefficients presented above. In addition, our error formulation provides specific predictions for the nature of the residuals ξ and χ for the permanent and temporary impact as in equations (1) and (2). Under the assumption that the asset price process is a Brownian motion with drift caused by our impact, these two variables should be independent standard Gaussians. We have already used this assumption in the heteroscedastic regression, and now we want to verify it.

Figure 4 shows histograms and Q-Q plots of ξ and χ. The means are quite close to zero. The variances are reasonably close to one, and the correlation is reasonably small. But the distribution is extremely fat-tailed, as is normal for returns distributions on short time intervals (Rydberg, 2000, has a nice illustration), and hence does not indicate that the model is poorly specified. Nonetheless, the structure of these residuals confirms that our model is close to the best that can be done within the Brownian motion framework.

Summary

We have used a large data sample of US institutional orders, and a simple but realistic theoretical model, to estimate price impact functions for equity trades on large-cap stocks. Within the range of order sizes considered (up to about 10% of daily volume), this model can be used to give quantitatively accurate pre-trade cost estimates, and is in a form that can be directly incorporated into optimal scheduling algorithms. Work is under way to refine the calibration to handle global markets, and the model is currently being incorporated into Citigroup's Best Execution Consulting Services software. ■

Robert Almgren is associate professor in the departments of mathematics and computer science at the University of Toronto, and Chee Thum, Emmanuel Hauptmann and Hong Li are senior analysts at Citigroup Global Quantitative Research in New York and London. **They are grateful to Stavros Siokos of Citigroup Equity Trading Strategy and Neil Chriss of SAC Capital Management for helpful feedback and perspective.** **Email: Robert.Almgren@utoronto.ca**

REFERENCES

Almgren R, 2003
Optimal execution with nonlinear impact functions and trading-enhanced risk
Applied Mathematical Finance 10, pages 1–18

Almgren R and N Chriss, 2000
Optimal execution of portfolio transactions
Journal of Risk 3(2), pages 5–39

Barra, 1997
Market impact model handbook

Bouchaud J-P, Y Gefen, M Potters and M Wyart, 2004
Fluctuations and response in financial markets: the subtle nature of 'random' price changes
Quantitative Finance 4(2), pages 176–190

Breen W, L Hodrick and R Korajczyk, 2002
Predicting equity liquidity
Management Science 48(4), pages 470–483

Chan L and J Lakonishok, 1995
The behavior of stock prices around institutional trades
Journal of Finance 50, pages 1,147–1,174

Dufour A and R Engle, 2000
Time and the price impact of a trade
Journal of Finance 55(6), pages 2,467–2,498

Freyre-Sanders A, R Guobuzaite and K Byrne, 2004
A review of trading cost models: reducing trading costs
Journal of Investing 13, fall, pages 93–115

Holthausen R, R Leftwich and D Mayers, 1990
Large-block transactions, the speed of response, and temporary and permanent stock-price effects
Journal of Financial Economics 26, pages 71–95

Huberman G and W Stanzl, 2004
Price manipulation and quasi-arbitrage
Econometrica 72(4), pages 1,247–1,275

Keim D and A Madhavan, 1996
The upstairs market for large-block transactions: analysis and measurement of price effects
Review of Financial Studies 9, pages 1–36

Kissell R and M Glantz, 2003
Optimal trading strategies
Amacom

Lillo F, J Farmer and R Mantegna, 2003
Master curve for price-impact function
Nature 421, pages 129–130

Loeb T, 1983
Trading cost: the critical link between investment information and results
Financial Analysts Journal 39(3), pages 39–44

Madhavan A, 2000
Market microstructure: a survey
Journal of Financial Markets 3, pages 205–258

Rydberg T, 2000
Realistic statistical modelling of financial data
International Statistical Review 68(3), pages 233–258

Rydberg T and N Shephard, 2003
Dynamics of trade-by-trade price movements: decomposition and models
Journal of Financial Economics 1(1), pages 2–25

Sorensen E, L Price, K Miller, D Cox and S Birnbaum, 1998
The Salomon Smith Barney global equity impact cost model
Technical report, Salomon Smith Barney