2008

The Price of Immediacy

George C. Chacko

Jakub W. Jurek
University of Pennsylvania

Erik Stafford

Recommended Citation

Chacko, G. C., Jurek, J. W., & Stafford, E. (2008). The Price of Immediacy. *The Journal of Finance, 63* (3), 1253-1290. http://dx.doi.org/10.1111/j.1540-6261.2008.01357.x

At the time of publication, author Jakub W Jurek was affiliated with Harvard University. Currently, he is a faculty member at the Wharton School at the University of Pennsylvania.

The Price of Immediacy

Abstract

This paper models transaction costs as the rents that a monopolistic market maker extracts from impatient investors who trade via limit orders. We show that limit orders are American options. The limit prices inducing immediate execution of the order are functionally equivalent to bid and ask prices and can be solved for various transaction sizes to characterize the market maker's entire supply curve. We find considerable empirical support for the model's predictions in the cross-section of NYSE firms. The model produces unbiased, out-of-sample forecasts of abnormal returns for firms added to the S&P 500 index.

Disciplines

Finance | Finance and Financial Management

Comments

At the time of publication, author Jakub W Jurek was affiliated with Harvard University. Currently, he is a faculty member at the Wharton School at the University of Pennsylvania.

The Price of Immediacy

George C. Chacko, Jakub W. Jurek, and Erik Stafford*

Abstract

This paper develops a new model of transaction costs, arising as the rents that a monopolistic market maker is able to extract from impatient investors. The mechanism for trade is a limit order, and immediacy is supplied when the limit order is executed. We show that limit orders are American options and their value represents the cost of transacting. The limit prices inducing immediate execution of the order are functionally equivalent to bid and ask prices, and can be solved for various transaction sizes to characterize the market maker's entire supply curve. We find considerable empirical support for the model's predictions in the cross-section of NYSE firms. The model produces unbiased, out-of-sample forecasts of abnormal returns for firms being added to the S&P 500 index.

April 2007

JEL Classification: G12, G13

Keywords: Liquidity, limit order, American option, early exercise, transaction cost, monopoly

*Chacko: Kite Partners, LLC - e-mail: *gchacko@kitepartners.com*. Jurek: Harvard Department of Economics and Harvard Business School - e-mail: *jjurek@hbs.edu*. Stafford: Harvard Business School - e-mail: *estafford@hbs.edu*. This paper has previously been circulated under the title, "Pricing Liquidity: The Quantity Structure of Immediacy Prices." We thank John Campbell, Joshua Coval, Robin Greenwood, Will Goetzmann, Bob Merton, André Perold, David Scharfstein, Anna Scherbina, Halla Yang, and seminar participants at Harvard Business School, INSEAD, MIT, The Bank of Italy, and State Street Bank. We are especially grateful to an anonymous referee whose comments substantially improved the paper.

1 Introduction

Capital market transactions essentially bundle a primary transaction for the underlying security with a secondary transaction for immediacy. In this view, the price of immediacy explains the wedge between transaction prices and fundamental value, and therefore represents a cost of transacting. Despite widespread interest among investors and corporations alike, a useful characterization of transaction prices has been elusive. This paper addresses this challenge by developing a parsimonious model of the market for immediacy in capital market transactions, which yields an analytically tractable quantity structure of immediacy prices.

An inherent friction that limits liquidity in capital markets is the asynchronous arrival of buyers and sellers, each demanding relatively quick transactions. Grossman and Miller (1988) argue that the demand for immediacy in capital markets is both urgent and sustained, creating a role for an intermediary, or market maker, who supplies immediacy by standing ready to transact when order imbalances arise (Demsetz (1968)).[1] In this setting, the price of immediacy is determined by two factors: the costs of market making, and the amount of competition among market makers. Many models assume perfect competition in market making, allowing the price of immediacy to be determined as the marginal cost of supplying immediacy. There is a large literature exploring the nature of these costs, focusing on the market maker's cost of holding inventory (see for example, Garman (1976), Stoll (1978), Amihud and Mendelson (1980), and Ho and Stoll (1981), and the costs of adverse selection in market making, which arise when investors have access to information that is not yet reflected in the price.[2]

Abstracting from the costs of market making, we instead relax the assumption of perfect competition. Specifically, we study how the asynchronous arrivals of buyers and sellers grants the market maker transitory pricing power with respect to investors demanding immediate execution. In this sense, our framework is similar to the market structure in the search-based model of Duffie, Gârleanu, and Pedersen (2005), where all agents are symmetrically informed and market makers have no inventory risk because of perfect inter-dealer markets. This makes market making costless. However, because investors must search for viable trading counterparties, the market maker is able to extract some of the difference between investors' reservation values and fundamental value in exchange for providing immediacy, giving rise to a bid-ask spread. We specialize to the case where a single market maker is continually present and the investor is impatient. This setup effectively creates a market for immediacy operating around the determination of fundamental value, which is assumed to occur in a separate market.

Both the costly market making literature and the search literature focus on developing equilibrium models. In contrast, we develop a partial equilibrium model of a transaction in the market for immediacy, which results in explicit formulas for the price of immediacy. We study an

[1] Empirical evidence on order submission strategies generally supports this view (e.g. Bacidore, Battalio, and Jennings (2001); Werner (2003); He, Odders-White, and Ready (2006)).

[2] Bagehot (1971) was one of the first to consider the role of information in determining transaction costs in a capital market setting. Copeland and Galai (1983), Glosten and Milgrom (1985), and Kyle (1985) are important early models of the information component of transaction costs. See O'Hara (2004) for an overview of these models.

impatient investor seeking to transact Q units of a security. The investor trades via a stylized limit order, and in the spirit of individual portfolio choice problems, we assume that the fundamental value process and the arrival of other investors are unaffected by the individual's trading decisions. Similar to Duffie, Gârleanu, and Pedersen (2005), we assume imperfect competition in market making. In particular, we allow a single market maker to have exclusive rights to be continually present in the market for the security. The privileged position of the market maker, combined with the asynchronous arrival of immediacy demanding buyers and sellers, gives him some pricing power in setting transaction prices (or immediacy prices). The degree of pricing power is determined by the intensity of opposing order arrivals, and collapses to zero, as in perfect competition, when arrival rates are infinite.

To develop an analytical model of transaction prices, we exploit the fact that a request to transact via a limit order is essentially equivalent to writing an American option.[3] For example, consider a seller placing a limit order. The seller can be viewed as offering the right to buy at a specific price at some point prior to an expiration date. This is effectively an American call option, requiring delivery of the underlying block of shares upon execution. Similarly, a request to buy is like an American put option. To ensure immediate execution, the initiator of a transaction offer (the option writer) must offer a price at which it is currently optimal for the receiver of the transaction offer (the option owner) to exercise the option early. The strike prices, where immediate exercise is optimal, represent immediately transactable prices, and therefore are functionally equivalent to the prices bid and asked by the market maker.

The resulting formula for the price of immediacy is simple and intuitive, and can be simplified even further when the arrival rate of order flow is large relative to the riskfree rate. Figure 1 illustrates how the price of immediacy reflects the wedge between transaction prices and fundamental value for various transaction sizes. The approximate formula for the percentage transaction cost is simply the product of volatility and the square root of excess demand, $p(Q) \approx \sigma\sqrt{\frac{Q}{2\lambda}}$, where σ is the volatility of fundamental value, Q is the transaction size, and λ is the arrival rate of opposing order flow. The model predicts that bid-ask spreads are increasing in the volatility of fundamental value, and in the size of order imbalances, $\left(\frac{Q}{\lambda}\right)$. Larger transactions effectively require the immediacy demander to write longer maturity options, which translates into greater transaction costs. Additionally, when order flow arrives at an infinite rate, the monopolist market maker's pricing power collapses, and the price of immediacy is zero for all quantities. Finally, the model predicts that the price of immediacy is a concave function of the transaction size, which empirical evidence strongly supports.

An attractive feature of the model is that it delivers a formula for immediacy prices as a function of variables that can be estimated relatively easily, allowing us to test its performance in a variety of settings. In the first application, we use the model to predict the discount charged to the Amaranth Advisors hedge fund during the forced liquidation of its portfolio. We find that

[3] The notion that limit orders can be viewed as contingent claims is not new (see Copeland and Galai (1983) for a specific option-based model of prices bid and asked by a market maker; and Harris (2003) for general examples).

our model's estimate of a 35% charge for immediacy compares favorably with the $1.4 billion loss incurred by the fund, which represented a 30% discount relative to the previous day's closing NAV. In the second application, we use trade and quote (TAQ) data to fit our model to the quantity cross-section of transaction sizes for NYSE firms. This calibration exercise demonstrates how the model can be used to estimate the entire, generally unobserved, quantity structure of transaction costs for individual securities, including very large transactions like corporate issues and takeovers. To evaluate the performance of the calibration procedure, we then use the calibrated quantity structure of immediacy prices to predict the price reactions for a sample of firms when they are added to the S&P 500 index. The out-of-sample nature of this test is underscored by the fact that, on average, the volume of shares bought by indexers during the inclusion event is over 300 times bigger than the largest transaction used to calibrate the model. We find that the limit order model produces unbiased estimates of price impact in this situation, and is able to explain roughly three times more of the cross-sectional variation than other models previously reported in the literature.

The remainder of the paper is organized as follows. Section 2 describes the model. Section 3 discusses the properties of the quantity structure of immediacy prices. Section 4 explores the limit order placement of a patient trader. Section 5 proposes two methods for implementing the model and empirically evaluates the model's performance. Finally, Section 6 concludes the paper.

2 The Pricing of Limit Orders

A common feature of transaction offers across many markets is that they pre-specify price and quantity, and remain available for some potentially unknown amount of time. In financial markets, these offers are referred to as limit orders. So long as the value of the underlying asset can change over the life of the offer, viewing offers of this type as options is reasonable. The value of this option is naturally interpretable as a cost of transacting, since it represents the value foregone to obtain the desired execution terms. In particular, a limit order to sell (buy) Q shares at price K, gives arriving buyers (sellers) the right to purchase (sell) at a pre-specified limit price at some point prior to the expiration date of the limit order, and is therefore like an American call (put) option, with the Q-share block of the security acting as the underlying. By placing a limit order, the trade initiator can be viewed as surrendering an American call (put) option on the desired quantity of the underlying to the remaining market participants. Although the offer is potentially available to many counterparties, it is extinguished as soon as anyone exercises it or upon maturity. The option writer receives liquidity when the limit order is exercised. From the perspective of someone evaluating whether or not to exercise the option, the important considerations are their own liquidity demands and the potential for competition from other market participants.

The value of the limit order and its optimal exercise policy depend crucially on three factors: (1) the mechanism governing trading (market structure); (2) the arrival rate of shares eligible for execution against the order (market competition); and (3) the evolution of the fundamental value of the underlying security or basket of securities. Because these factors are likely to have complex

dynamics in reality, our model is best interpreted as a reduced-form characterization of transaction costs.

The challenge is to specify a suitable market structure that allows the demand and supply of immediacy to be isolated. Generally, each party to a trade is both demanding and supplying immediacy to some extent. To simplify, we assume that the limit order writer (the trade initiator) is impatient and demands immediate execution. In order to have the limit order filled instantaneously, he must write an option that is sufficiently deep in-the-money to make immediate exercise optimal. Although the option is available to both the market maker and opposing order flow, only the market maker can be relied upon to supply immediacy at any given time because order flow arrives stochastically. For the market maker, the threat of losing the order to opposing order flow acts like a stochastic dividend on the underlying block of shares, creating an incentive for the market maker to exercise the option early.

An attractive feature of this setup is that limit prices for which immediate exercise is optimal represent instantaneously transactable prices, and therefore are functionally equivalent to the prices bid and asked by a market maker. This allows us to characterize the generally unobserved bid and ask prices for large quantities (i.e. larger than the quantity posted at the best bid and ask). Moreover, the option-based model of transaction prices inherits the properties of ordinary options. The two drivers of transaction costs for any given quantity are the fundamental volatility and the effective option maturity, which is determined by the order flow arrival rate. A quantity structure of instantaneously transactable prices arises because larger trade sizes require the trade initiator to write options with longer effective maturities.

2.1 A Simple Model of Transaction Costs

Our model of transaction costs adopts a partial equilibrium framework similar in spirit to the one used for studying individual portfolio choice (Merton (1969, 1971)), in which the process for the asset's fundamental value is specified exogenously. We then focus on characterizing the determinants of the wedge between transaction prices and fundamental value, or equivalently, transaction costs. The separation of the determinants of fundamental value and liquidity costs present in our model is consistent with the conclusions of Cochrane's (2005) survey of the liquidity literature, in which he suggests that liquidity be interpreted "as an additional feature above and beyond the usual picture of returns driven by the macroeconomic state variables familiar from the frictionless view." By providing a theoretical model of the *level* of transaction prices we naturally complement the existing literature examining the effects of liquidity risk on the determination of expected *rates* of return (Pastor and Stambaugh (2003), Acharya and Pedersen (2005)).

The market for a security is composed of two symmetrically-informed agent types: investors and a market maker. The profit maximizing market maker acts as an intermediary, facilitating trades between asynchronously arriving investors, effectively creating a market for immediacy. However, unlike the individual investors, the market maker is assumed to additionally have continuous access to an inter-dealer market as in Duffie, Gârleanu, and Pedersen (2005), in which he

can instantaneously hedge his inventory risk. Trading in the inter-dealer market is frictionless and takes place at fundamental value, V_t, which is observable by all participants. The dynamics for fundamental value are described by a diffusion-type stochastic process:

$$\frac{dV_t}{V_t} = \mu dt + \sigma dZ_t \tag{1}$$

where μ and σ^2 are the instantaneous expected return and variance of the fundamental value, and dZ_t is a standard Gauss-Wiener process.[4] The price formation process giving rise to fundamental value, V_t, pins down the price of risk, γ_V, for exposure to the shocks dZ_t, and implies a pricing kernel of the form:

$$\frac{d\Lambda_t}{\Lambda_t} = -rdt - \gamma_V dZ_t \tag{2}$$

where r is the instantaneous riskless rate and $\gamma_V = \frac{\mu - r}{\sigma}$. If markets are incomplete, this pricing kernel will not be the unique kernel of the economy, but it will be the unique kernel in the span of dZ_t, allowing us to price any asset whose value is exposed only to innovations in dZ_t.

The inability of individual investors to participate in the market for fundamental value creates the scope for the market maker to provide liquidity services to the public and collect compensation in the form of a bid-ask spread. Although investors do not have access to the inter-dealer market, they can still trade with each other at fundamental value when opposing orders are present. Only in the absence of opposing order flow are they forced to submit limit orders to the market maker, who will buy (sell) the security at some discount (premium).[5] Providing a useful characterization of the wedge between fundamental value and the prices at which the market maker is willing to transact Q units of a security is the central goal of our investigation. To determine this wedge we first provide a more detailed specification of the mechanism by which limit orders are exercised.

Definition 1 (Trading Mechanism) *A limit order, $L^i(Q, K)$, specifies a quantity, price, and direction of trade (i.e. buy or sell, $i \in \{B, S\}$).*

1. *Limit orders can be exercised at any time by the market maker prior to the occurrence of an opposing Q-share order imbalance. Upon the occurrence of an opposing Q-share order imbalance, the limit order transacts with the order imbalance at the (then current) fundamental value, voiding the market maker's claim on the trade.*

2. *The instantaneous probability of observing a Q-share buy (sell) imbalance during the next instant is given by $\lambda^B(Q)dt$ ($\lambda^S(Q)dt$). Given this assumption, the expected time to the*

[4] Although the process for fundamental value is specified exogenously it can be naturally interpreted as the outcome of a rational expectations equilibrium arising in the inter-dealer market (Wang (1993), He and Wang (1995)).

[5] We require agents to submit limit orders, as opposed to market orders, to prevent the market maker from exploiting his instantaneous pricing power and filling sell (buy) market orders at a zero (infinite) price. In practice, this form of exploitation is precluded by legal restrictions and reputational considerations.

completion of a Q-share limit order to sell (buy) is distributed exponentially with mean $\frac{1}{\lambda^B(Q)}$ $\left(\frac{1}{\lambda^S(Q)}\right)$.[6]

To preserve tractability and abstract from modeling the evolution of the limit order book, we focus on the special case in which all limit order traders have zero patience and only place orders that are immediately exercisable by a profit maximizing market maker.[7] In order to obtain immediacy, an impatient limit order trader must set the limit price, K, such that the option embedded in the order is sufficiently in-the-money to make immediate exercise optimal. In general, the schedule of limit prices guaranteeing immediacy will depend on the factors determining the value of the option: the riskless rate, r; the volatility of the underlying, σ; and the arrival rate of opposing order flow, $\lambda^i(\cdot)$, which itself is a function of the order quantity, Q. We will denote the schedules of immediacy prices for Q-share sell and buy limit orders by, $K_B(Q, \alpha = 0)$ and $K_A(Q, \alpha = 0)$, respectively, with the spreads between fundamental value and these prices having the interpretation of the *price of immediacy*.[8] These schedules represent prices at which transactions can take place instantaneously and are functionally equivalent to bid and ask prices.

Proposition 1 *The strike price at which it is optimal to immediately exercise a sell (buy) limit order for Q shares determines the effective bid (ask) price for Q shares.*

In our baseline specification we assume that limit orders are not subject to cancellation by the limit order writer. This implies that the limit order option is *perpetual*, albeit subject to a stochastic liquidating dividend in the form of order execution by arriving order flow. The main virtues of the perpetual limit order feature are its analytical tractability and the fact that it provides an upper bound to immediacy costs. Since the value of the American option implicit in the limit order is monotonically increasing in time, a limit order writer forced to trade in perpetual limit orders is effectively surrendering options with the highest possible time value. Consequently, immediacy costs are maximized. In an appendix, we relax this assumption and consider limit orders subject to random cancelation by the limit order writer, as well as finite duration limit orders. We find that, as long as the expected lifetime of the limit order is non-zero, the qualitative predictions of the model are unaltered.[9]

The presence of the liquidating dividend is crucial in that it makes an early exercise strategy for the monopolist market maker optimal and facilitates the interpretation of option exercise

[6] The λ parameters can alternatively be interpreted as *search intensities* for eligible counterparties, in the spirit of Duffie, Gârleanu and Pedersen (2005) or Vayanos and Wang (2002).

[7] Grossman and Miller (1988) argue that there is high demand for immediacy in capital markets. Empirical evidence supports this view. Bacidore, Battalio, and Jennings (2001) and Werner (2003) report that between 37-47% of all orders submitted on the NYSE are liquidity demanding orders, comprised of market orders or marketable limit orders.

[8] The investor's patience level, $\alpha = 0$, is included to emphasize that immediacy is being demanded.

[9] In the degenerate case, when the limit order writer can credibly threaten to cancel the order instantaneously, all transactions take place at fundamental value. The credibility of such threats can be eliminated through the introduction of a small, fixed cost of order submission, which would render strategies with instantaneous cancelation infinitely costly.

as liquidity provision. The particular structure of the dividend process, controlled by a Poisson random variable with a quantity-dependent arrival intensity, is chosen for analytical tractability. In particular, the memoryless feature of the inter-arrival process preserves the time-stationary feature of the perpetual option valuation problem. This allows us to intuit that the optimal exercise boundary will be a barrier rule, which optimally trades off the preservation of the time-value of the option with the adverse consequences of the dividend.

2.2 Model Solution

Given the earlier assumptions, the value of the Q-share limit order with a strike price K, $L(V_t, Q, K, t)$, can be shown to satisfy the following ordinary differential equation (ODE):

$$L_F \cdot (rF_{Q,t}) + \frac{1}{2} L_{FF} \cdot (\sigma F_{Q,t})^2 - (r + \lambda^i(Q)) \cdot L = 0 \tag{3}$$

where subscripts are used to denote partial derivatives and $F_{Q,t} = Q \cdot V_t$ represents the fundamental value of the underlying block of shares. This ODE is solved subject to three boundary conditions. The first boundary condition is determined by the asymptotic behavior of the value of limit order as a function of $F_{Q,t}$, and the second pair of conditions arises from the value matching and smooth pasting at the optimal early exercise threshold. The equidimensional structure of the ODE suggests that the solution will be a linear combination of power functions in $F_{Q,t}$ with exponents given by:

$$\phi_{\pm}(\lambda^i) = \left(\frac{1}{2} - \frac{r}{\sigma^2}\right) \pm \sqrt{\left(\frac{1}{2} - \frac{r}{\sigma^2}\right)^2 + \frac{2(r + \lambda^i(Q))}{\sigma^2}} \tag{4}$$

Economic intuition allows us to exclude one of the two roots in both the case of a sell limit order and a buy limit order. In particular, since the value of a sell (buy) limit order is increasing (decreasing) in $F_{Q,t}$ we can exclude the negative (positive) root. Finally, to pin down the value of the constant of integration we make use of the fact that the optimal exercise rule for the option is a barrier rule. Consequently, the value of the limit order at optimal exercise is given by $Q \cdot (V_t^* - K)$ for a sell limit order and $Q \cdot (K - V_t^{**})$, where V_t^* and V_t^{**} are the optimal exercise thresholds for sell and buy limit orders, respectively. The expressions for the values of the limit orders and the associated optimal exercise thresholds are collected in the following proposition.

Proposition 2 *The value of a Q-share sell limit order is given by:*

$$L^S(V_t, Q, K, t) = \frac{QK}{\phi_+(\lambda^B) - 1} \cdot \left(\frac{\phi_+(\lambda^B) - 1}{\phi_+(\lambda^B)} \cdot \frac{V_t}{K}\right)^{\phi_+(\lambda^B)} \qquad V_t < V_t^* \tag{5}$$

and it is optimal for the market maker to exercise the implicit call option whenever fundamental value reaches the threshold $V_t^* = K \cdot \left(\frac{\phi_+(\lambda^B)}{\phi_+(\lambda^B)-1}\right)$ from below. The value of Q-share buy limit order

is given by:

$$L^B(V_t, Q, K, t) = \frac{QK}{1-\phi_-(\lambda^S)} \cdot \left(\frac{\phi_-(\lambda^S)-1}{\phi_-(\lambda^S)} \cdot \frac{V_t}{K} \right)^{\phi_-(\lambda^S)} \qquad V_t > V_t^{**} \tag{6}$$

and it is optimal for the market maker to exercise the implicit put option whenever fundamental value reaches the threshold $V_t^{**} = K \cdot \left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S)-1} \right)$ *from above.*

In order to induce immediate exercise of a sell (buy) limit order, the limit price (i.e. the option strike price) has to be set such that the prevailing fundamental value, V_t, is exactly equal to V_t^* (V_t^{**}), making it optimal for the market maker to exercise the order instantaneously. To do this, the limit order writer selects a limit price, K^*, which renders the time-value of the embedded option equal to zero at prevailing fundamental value, V_t. The distance, $V_t - K^*$, represents the value of the immediately exercisable option, and has the interpretation of the price of immediacy for a one-share transaction.

The strike prices for immediately executable buy (sell) transactions as a function of order quantity yield the *quantity structure of immediacy prices*. The analytical expressions for the immediacy prices depend on the order quantity, Q, through $\phi_+(\lambda^B)$ and $\phi_-(\lambda^S)$ and are summarized below.

Proposition 3 *The bid,* $K_B(Q, \alpha = 0)$*, and ask,* $K_A(Q, \alpha = 0)$*, prices are given by:*

$$K_B(Q, \alpha = 0) = V_t \cdot \left(\frac{\phi_+(\lambda^B)-1}{\phi_+(\lambda^B)} \right) \tag{7}$$

$$K_A(Q, \alpha = 0) = V_t \cdot \left(\frac{\phi_-(\lambda^S)-1}{\phi_-(\lambda^S)} \right) \tag{8}$$

and imply that the percentage immediacy costs for sales and purchases are given by:

$$\frac{K_B(Q, \alpha = 0) - V_t}{V_t} = -\frac{1}{\phi_+(\lambda^B)} \tag{9}$$

$$\frac{K_A(Q, \alpha = 0) - V_t}{V_t} = -\frac{1}{\phi_-(\lambda^S)} \tag{10}$$

The expressions for the proportional transaction costs can be further simplified by noting that under empirically plausible calibrations, the order arrival rates, $\lambda^i(Q)$, will be significantly larger than the riskless rate. This allows us to derive some simple approximations for $\phi_\pm(\cdot)$ and the percentage immediacy costs. In particular, whenever $\lambda^i(Q) \gg r$ we have:[10]

$$\phi_\pm(\lambda^i) \approx \pm \frac{\sqrt{2\lambda^i(Q)}}{\sigma} \tag{11}$$

[10] The proposed approximation underestimates (overestimates) the premia (discounts) at which assets can be bought (sold). The magnitude of this error is extremely small for plausible parameter values.

Consequently, the percentage immediacy costs predicted by our model are (approximately) proportional to $(\lambda^i(Q))^{-\frac{1}{2}}$ – the square root of the expected waiting time for the arrival of an opposing Q-share imbalance – and converge to zero as the arrival rates of opposing flow tend to infinity, as would be the case in a perfect capital market. The degree of non-linearity in the percentage immediacy costs is determined by the relationship governing the scaling of the order arrival intensity rate as a function of order quantity. For example, if the arrival rate of Q-share imbalances is Q^n times smaller than the arrival rate of single share imbalances, the percentage immediacy costs predicted by our model will be proportional to $Q^{\frac{n}{2}}$. In the remainder of the paper, we specialize to the case where the expected waiting time for the completion of a Q-share order is precisely Q times larger than the corresponding waiting time for a one-share order ($\lambda^i(Q) = \lambda^i(1) \cdot Q^{-1}$). This implies that the percentage immediacy premium implied by the ask prices will be concave in the order quantity and (approximately) proportional to $\sqrt{Q}$.

2.3 Discussion

Before turning to a characterization of the comparative statics of our model and its predictions under empirically calibrated parameter values, it is worthwhile to briefly re-iterate the two key modeling assumptions that allowed us to obtain a nonlinear quantity structure of transaction prices. First, the limit order must be interpretable as an option. This requires that the limit order have a fixed strike price and have the potential to remain outstanding for some non-zero length of time, allowing fundamental value to change. Second, the market must be structured such that the market maker has an *instantaneous* monopoly on the supply of immediacy, and is only forced to compete with order flow when it is present. Unlike a classical monopolist familiar from deterministic settings, in our stochastic setting, the market maker is perceived as a monopolist only by counterparties demanding immediacy. This can be seen more clearly by considering the (expected) number of trading counterparties, C, available to an agent interested in transacting Q shares in the next τ units of time. This patient agent can transact either with the market maker, who is always standing by, or the oncoming order flow, which appears randomly with a probability depending on the arrival rate of Q-share imbalances. Consequently, the number of trading parties perceived by the patient trader is given by:

$$E[C] = 1 + \left(1 - e^{-\lambda^i(Q)\cdot\tau}\right) = 2 - e^{-\lambda^i(Q)\cdot\tau} \tag{12}$$

As the agent becomes infinitely patient ($\tau \to \infty$), he perceives the market as being comprised of two trading counterparties, the market maker and oncoming order flow. As a result of the competition between these two counterparties, the agent is assured of transacting at fundamental value.[11] On the other hand, if the agent demands immediacy ($\tau = 0$), he perceives only one trading counterparty: a monopolistic market maker. More formally, the market maker can be thought of

[11] Notice that the same result would arise in a model in which two market makers were granted the right to be perpetually present in the market.

as having a probabilistic monopoly, since as $\tau \to 0$ the expected number of trading counterparties converges to one in probability.

Under these two assumptions, the market maker is effectively granted ownership of the option embedded in the limit order, and has to decide when and if to exercise the option, thereby delivering liquidity to the limit order writer. The incentive for the early exercise of this option by the market maker arises as a consequence of the presence of the opposing order flow, which acts like a stochastic liquidating dividend. To facilitate tractability and generate intuition, our baseline specification in Section 2.1 considered a perpetual American option with a Poisson liquidating dividend. This structure for the liquidating dividend implicitly assumes that limit orders are only subject to *one-shot execution* – there is no possibility for a limit order to be filled by a sequence of partial fills. Although this execution mechanism is simplified, it does have the added attraction that the mean inter-arrival time of opposing orders can be readily calibrated from empirical signed order flow data.

In an appendix, we show how to generalize our model to finite-lived limit orders, as well as, how to incorporate the possibility of order cancellation by the limit order writer. While these extensions can be accomplished in closed-form, similar modifications to the liquidating divided can only be accomplished at the expense of analytical tractability. Numerical simulations, using the Longstaff and Schwartz (2001) least squares methodology, show that the pricing of limit orders under a more sophisticated order flow process allowing for partial fills, yields results which are qualitatively indistinguishable from those obtained under the analytical model.[12]

3 The Quantity Structure of Immediacy Prices

Inelastic demand for immediacy is the limiting case, when patience goes to zero. The model imposes this condition to identify a quantity structure of instantaneously transactable prices– immediacy prices. In the model, the two primary drivers of the prices charged by the market maker are the volatility of fundamental value and the time rate of arrivals of opposing order flow. Matching intuition, the model predicts that bid-ask spreads are increasing in fundamental volatility and that there are economies of scale in transactions.

To illustrate the above results graphically, we exploit our auxiliary assumption that the expected waiting time for the completion of an order scales linearly in the order quantity, Q.[13] Using this assumption, Figure 2 graphs the schedule of percentage immediacy prices, (9) and (10), as a function of order quantity. In particular, we assume the annual volatility of fundamental value

[12] The numerical simulation modifies the definition of a limit order to allow partial execution by order flow and replaces the specification for the market order flow process. Under the augmented specification used for the numerical simulation the random maturity of the finite-lived limit order option is determined by the joint dynamics of order imbalance and fundamental value. These dynamics imply a time-varying instantaneous survival probability for the limit order and lead to a distribution of the times to completion that is not analytically tractable. In turn, it is not possible to obtain a closed-form expression for the value of the limit order option or its optimal early exercise rule, a feature which is shared by most American-type options.

[13] We verify the validity of this assumption empirically in the cross-section of NYSE firms in Section 5.

is 15% or 35%, the riskfree rate of interest is 5% per year, and that orders arrive at a rate of one share per second. Figure 2 shows that immediacy prices are nonlinear functions of the transaction size. Using the above definition of the cost of transacting, these costs are increasing and concave in transaction size. This is in contrast to most information-based models of liquidity, which typically produce constant marginal costs, or linear price functions of quantity (for example, Kyle (1985)). In Section 5, we evaluate models on the basis of these predictions.

3.1 Effect of Order Flow Arrival Rates

Demsetz (1968) argues that it is reasonable to expect scale economies in transactions. As order flow arrival rates for a security increase, the waiting times for transaction execution in that security decrease. In the limiting case of infinite arrival rates, waiting times go to zero. In the more typical case of finite arrivals, the waiting time of a transaction can make up a significant portion of the total transaction cost. When investors demand immediacy, the waiting time can be transferred to the market maker (or marginal supplier of liquidity) who specializes in providing this service, but the waiting time cannot be eliminated.

The key friction in the model is that order flow arrivals are finite, which gives rise to a positive waiting time for transaction execution. In the model, there is a direct mapping of waiting times to option maturity. The time rate of arrivals of opposing order flow determines the expected waiting time of any given order. This intuition is formally captured in expressions (9) and (10). First and foremost, as the arrival rate of order flow eligible for execution against the outstanding limit order, λ^i, increases, the market maker faces more competition from order flow and the percentage immediacy costs decline. In the perfectly liquid market, $\lambda^i \to \infty$, the market maker possesses no pricing power and the costs of immediacy collapse to zero. Conversely, as competition from exogenous order flow declines, $\lambda^i \to 0$, the market for immediacy becomes progressively less competitive (more illiquid), allowing the monopolist market maker to charge a wider bid-ask spread to counterparties seeking immediacy. When trading by other market participants ceases altogether, $\lambda^i = 0$, the market maker is the sole provider of immediacy through time, not just instantaneously, and the asset market breaks down completely. The value of the sell limit order converges to the value of the underlying, V_t, implying that, in order to obtain immediacy, the seller must part with the asset at a zero price. Intuitively, in this scenario, the market is a pure monopoly in which the market maker captures the entire surplus. On the other hand, buy transactions still remain possible, but only at significant premia to fundamental value. In the limiting case when $\lambda^i = 0$, the smallest percentage premium to fundamental value guaranteeing immediate execution is given by $\frac{\sigma^2}{2r}$.

Figure 3 displays the immediacy prices for fixed transaction sizes as a function of the order arrival rate. In general, immediacy prices do not equal fundamental value. As order flow arrival rates increase, expected waiting times shrink, and the bid and ask prices converge towards fundamental value. The increase in efficiency is largest when arrival rates begin low and increase. The figure shows a changing rate of convergence in immediacy prices towards fundamental value–

initially very fast at low arrival rates, then becoming more gradual as arrival rates increase.

3.2 Effect of Fundamental Volatility

In our model, immediacy prices offered by the market maker deviate farther from fundamental value as the volatility of fundamental value increases, for any given quantity (an illustration is presented in Figure 2). This is a direct consequence of the option-based approach. Option values are increasing in volatility, and this property flows through to the strike price at which immediate exercise is optimal. The more valuable the option, the larger the distance must be between the strike price and fundamental value for the market maker to exercise immediately. In particular, in the limit as $\sigma \to \infty$, the value of a Q-share sell limit order with a limit price of K approaches Q times the fundamental value. A similar buy limit order approaches Q times the limit price. Because immediate exercise requires that the limit order writer give the market maker an option that is in-the-money, the percentage immediacy cost for sell orders goes to 100%. Buy limit orders, on the other hand, are never executed. Conversely, in the absence of any price risk, i.e. when the volatility of fundamental value is zero ($\sigma = 0$), the options implicit in the order flow have no value, so no premium is required to induce the market maker to exercise immediately.

3.3 Liquidity Events

The analytical model presented in Section 2 allows us to examine how shocks to the arrival rate of buy/sell orders and the fundamental value of the underlying may compound during a liquidity crisis to affect immediacy prices. The arrival rate of buy (sell) orders will determine the expected maturity of the options written by a seller (buyer) demanding immediacy. Therefore, from the seller's (buyer's) perspective, a liquidity crisis is likely to involve a significant decrease in the current rate of buy (sell) order arrivals, relative to the equilibrium rate. This asymmetry in arrival rates may become more severe if the current rate of sell order arrivals also increases. This captures the notion that a liquidity crisis involves some sort of order imbalance. As a consequence of a temporary order imbalance a significant asymmetry in buy and sell immediacy prices may emerge at all quantities, causing the midpoint of the bid-ask spread to become a biased estimator of the fundamental value.

Figure 4 displays the effects of an order imbalance on the quantity structure of immediacy prices. In particular, the figure assumes that the current rate of sell order arrivals increases fivefold, while the current rate of buy order arrivals falls by this factor. This represents a major "running for the exit" in the security. Immediacy prices for buyers become much more elastic, such that an investor wishing to buy can now immediately transact very large quantities at a price much closer to fundamental value. However, investors wishing to sell immediately must pay a large premium, even for relatively small quantities. In other words, the immediacy prices facing sellers are now less elastic at all quantities.

Figure 4 also displays immediacy prices in the case when an order imbalance coincides with

an increase in fundamental volatility. The increased volatility offsets the reduced waiting time for buy orders, attenuating the increased elasticity of immediacy ask prices slightly. On the other hand, the higher volatility further increases the premium for immediacy for sellers, making prices even less elastic at all quantities.

4 Robustness and Extensions

The market structure considered in this paper is highly stylized and a number of restrictive assumptions were required to arrive at our analytical predictions for transaction prices. First, the market maker is given monopoly in the right to "hang around," while other market participants must take an action and move on. The only competition the market marker faces with respect to current demand is from offsetting future orders, which arrive stochastically and play the role of a liquidating dividend. Consequently, while there is competition in the supply of immediacy through time, instantaneously the market maker is a monopolist. Second, we restrict our attention to the case of traders demanding immediate execution, which allowed us to skirt the difficult task of modeling the evolution of the limit order book. Although the assumption of inelastic demand is crucial in allowing us to trace out the market marker's supply function for immediacy-demanding transactions, it conceals the importance of patience in determining transaction prices. Finally, our model abstracts away from issues regarding the costs of market making and asymmetric information, which have been at the center of the microstructure literature.

Relaxing these assumptions is likely to bring the model closer in line with the true richness of the problem faced by market makers and traders in the real world. In this section, we examine the robustness of our model's predictions with respect to such extensions, and suggest directions for future research.

4.1 Search and Pricing Power in Market Making

The assumption of a monopolistic market maker, who enjoys the privileged position of being a continuously available trading counterparty, plays a central role in our model. It grants ownership of the option implicit in a limit order to the market maker, and allows us to solve for its value under the optimal exercise rule. The introduction of a competitive market making function would alter the pricing of a limit order through its early exercise rule. In particular, an individual placing a limit order in this market structure could expect their limit order to be exercised either by opposing order flow, as before, or by the market maker any time the intrinsic value of the option exceeds the marginal cost of the market maker's adjustment to inventory. The introduction of a competitive market making function would therefore modify the early exercise boundary to read $V_t - K^*(Q) = mc(Q)$, necessitating an explicit characterization of the market maker's cost function, as is commonly required in traditional models of market microstructure. Conversely, if the market maker is a monopolist, we can determine the price of immediacy through the optimal exercise policy of the limit order, with no knowledge of the market maker's cost function.

Sidestepping the difficult problem of characterizing the cost of market making in terms of unobservable variables like information asymmetries and individual preferences, requires an alternative friction to generate transaction costs. We assume imperfect competition in market making, consistent with the notion that supplying immediacy is sometimes profitable. This brings our model much closer in spirit to the search literature. In search models, transaction prices are determined through bilateral bargaining, which makes the markets they describe inherently uncompetitive. Generally, each party to a trade is both demanding and supplying immediacy to some degree. The relative market power of each party is specified exogenously through bargaining parameters, which determine the division of surplus between two willing trading counterparties. We specialize to the situation where a single market maker continuously supplies immediacy to investors with inelastic immediacy demands.[14]

Our decision to examine the price of immediacy in partial equilibrium yields two advantages over the more general frameworks employed in search models. First, we are able to consider the pricing of an asset with a stochastic fundamental value, whereas search models examine transaction prices around a deterministic fundamental value. The time-varying fundamental value gives the offer to transact an option-like property. Second, our specification can be readily calibrated using empirical data and is the first to deliver a usable quantity structure of immediacy prices. Of course, it is important to keep in mind that our model only studies price determination in a single, stylized transaction, with no regard for patience or the potential for interactions between the determination of fundamental value and transaction prices (O'Hara (2003)). Consequently, we view our model as describing the *nanostructure* of a market transaction, which may be an important component of extensions of search models to settings with stochastic variation in fundamental value.

4.2 Patience

In this section, we relax the assumption that each trader demands immediate execution, and offer a reduced form examination of the effect of patience on limit price selection. Specifically, we propose an intuitive parametrization for the agent's patience level, which nests the special case of zero patience considered earlier. Of course, in equilibrium, the magnitude of the patience parameter depends on myriad factors including the trader's utility function, the opportunity cost of delaying order execution, and actions of other market participants. Rather than explicitly model each of these factors, we continue in the partial equilibrium spirit of our earlier analysis, and specify the patience parameter exogenously. We show that the limit buy (sell) prices selected by traders are monotonically decreasing (increasing) functions of their patience, and depend on properties of the underlying (order arrival rates, drift, volatility, etc.), as well as the trader's decision horizon (i.e. frequency with which limit prices are reset). Formally, in a model with a limit order book, these buy (sell) orders would be below the prevailing ask (bid) prices. However, because there is no

[14] Since our model features a single market maker who is continuously present in the market, it is most similar to the case of the Duffie, Garleanu and Pedersen (2005) search model with a "fast monopolistic market maker" discussed in Theorem 3.3.

limit order book in our model, the limit prices selected by patient traders are better thought of as reservation values at which they would be willing to place an immediacy demanding limit order.[15]

To examine the impact of patience on limit price (reservation value) selection, we parameterize traders by the probability, α, that their order fails to be executed within τ units of time. Traders with zero patience, who demand immediate execution, are characterized by $\alpha = 0$, and traders with infinite patience, who do not mind seeing their order go unexecuted, have an $\alpha = 1$. Consequently, we refer to the value of α as the patience level. The value of τ has the interpretation of a decision horizon, and represents the horizon at which it becomes optimal for a trader to recompute their reservation value (Merton (1987)).

The reservation buy price, $K_A^*(Q, \alpha, \tau)$, of a trader with a decision horizon, τ, and patience level, α, is set such that the probability of observing the market ask price reaching $K^*(Q, \alpha, \tau)$ or lower within τ units of time is exactly $1 - \alpha$. Intuitively, the more patient the trader (i.e. the larger the value of α), the further the reservation buy price will be below the prevailing ask value, $K_A(Q, \alpha = 0)$, guaranteeing immediate execution.[16] Using the notation introduced earlier in the paper, we know that the market maker will exercise a buy limit order with limit price, $K_A^*(Q, \alpha, \tau)$, at time h, if and only if,

$$V_h = K_A^*(Q, \alpha, \tau) \cdot \left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S) - 1} \right) \tag{13}$$

To determine the reservation buy price we make use of the probability distribution function of the running minimum of a geometric Brownian motion. Specifically, the above condition requires that the minimum of the security's fundamental value, V_t, over the time interval $[t, t + \tau]$ be less than or equal to $K_A^*(Q, \alpha, \tau) \cdot \left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S) - 1} \right)$ with probability $1 - \alpha$,

$$\text{Prob} \left[\min V_h \leq K_A^*(Q, \alpha, \tau) \cdot \left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S) - 1} \right), h \in [t, t + \tau] \right] = 1 - \alpha \tag{14}$$

Using the distribution of the running minimum of a GBM this condition can be rewritten as,

$$\alpha = \Phi(d_1) - \exp \left\{ \frac{2}{\sigma^2} \cdot \left(\mu - \frac{\sigma^2}{2} \right) \cdot \ln \left(\frac{K_A^*}{V_t} \cdot \left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S) - 1} \right) \right) \right\} \cdot \Phi(d_2) \tag{15}$$

[15] Given our model assumptions, if we allowed the patient trader's order to be the sole outstanding order in the limit order book, it would also be subject to execution by oncoming order flow. In reality, however, because orders submitted by patient traders are likely to be away from the prevailing market prices, they are unlikely to be the first to be executed by oncoming market orders. Consequently, we view it as a better approximation to interpret the limit prices selected by patient traders as their reservation values, rather than the prices of actual submitted orders.

[16] The analysis of the reservation sell price is symmetric.

where μ is the drift in fundamental value and:

$$d_1 = \frac{-\ln\left(\frac{K_A^*}{V_t}\cdot\left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S)-1}\right)\right)+\left(\mu-\frac{\sigma^2}{2}\right)\cdot\tau}{\sigma\sqrt{\tau}} \tag{16}$$

$$d_2 = \frac{\ln\left(\frac{K_A^*}{V_t}\cdot\left(\frac{\phi_-(\lambda^S)}{\phi_-(\lambda^S)-1}\right)\right)+\left(\mu-\frac{\sigma^2}{2}\right)\cdot\tau}{\sigma\sqrt{\tau}} \tag{17}$$

This implicit characterization of buy reservation values, $K_A^*(Q,\alpha,\tau)$, nests our earlier solution for immediacy demanding trades. To see this, note that traders demanding no uncertainty about execution ($\alpha = 0$) are characterized by a reservation value, $K_A^*(Q,\alpha,\tau)$, equal to the price guaranteeing immediate execution, $K_A(Q,\alpha=0)$. In other words, the only way to be certain of executing a buy transaction is to transact immediately at the prevailing ask price, regardless of the decision horizon.

In Figure 5, we illustrate the impact of patience on the trader's reservation value for the case of two decision horizons. In both cases, we fix the riskless rate and the drift of the underlying asset at 5% and 12% *per annum*, respectively. Consistent with intuition, the figure indicates that investors who are more patient, and thus more willing to absorb uncertainty about execution (larger value of α) or having longer decision horizons (larger τ), can obtain meaningful savings relative to impatient traders. For example, consider a trader demanding a 50% probability of having a ten thousand share buy order executed within 100 (1000) seconds, given an arrival rate of 100 shares per second, when the underlying has a annualized volatility of 25%. While the reservation value of the trader with a 100 second horizon is roughly equal to the prevailing fundamental value, the reservation value of the trader with a 1000 second horizon is 15 bps *below* it. By comparison, a trader demanding immediate execution would have to submit an order that is 7 bps above fundamental value.

5 Applications

This section considers two types of applications for the model introduced in Section 2, and examines the resulting model-based predictions for the price of immediacy in capital market transactions. In the first application, we test our model using a real-world scenario in which the impatient demand for a large transaction by a non-information motivated trader was met with very little competition in the supply of liquidity. Under these circumstances, the model's extreme assumptions are likely to be valid, allowing us to apply it very literally. We therefore estimate the parameters of the fundamental value and order flow processes from observable data and then plug these estimates into the model to determine the cost for this rare, but important type of transaction. This application also highlights the potential for our model to be used as a stress-testing platform for deriving liquidity-adjusted estimates of portfolio losses in cases of market stress. In the second application, we consider the possibility that there may be more competition – perhaps in the form

of latent liquidity supply – than the model assumes. In this situation, it is more appropriate to calibrate, or imply out, the order flow arrival rate parameter using the transaction-level data generated under ordinary conditions, and then evaluate how the model forecasts out-of-sample transactions.

5.1 The Forced Liquidation of Amaranth's Energy Position

As a first example, we examine the collapse of the Amaranth Advisors hedge fund in September 2006. After sustaining massive losses on its positions in natural gas contracts, the fund was forced to liquidate its energy book to two financial institutions at a discount of roughly 30%. Because this asset sale was both rapid and non-informational, it represents an ideal scenario in which to test our model's prediction regarding the price of immediacy.

The Amaranth crisis stemmed from a series of calendar trades on natural gas contracts put on by the firm. In the US, there is insufficient storage capacity for natural gas to meet peak winter heating demand. As a result, the natural gas futures market for summer/fall gas contracts and winter gas contracts is typically in contango, where prices of summer and fall natural gas contracts typically trade at a discount relative to the winter contracts. The market therefore provides a return for purchasing and storing natural gas in the summer and fall and delivering it in the winter. This, in turn, incents storage operators to store more natural gas and sell it in the winter. However, the spread between the summer/fall futures prices and winter futures prices is extremely volatile, so the storage operator takes a substantial risk.[17] Hedge funds, such as Amaranth, typically sold summer/fall contracts and bought winter contracts, thus allowing the storage operators to hedge their risk. Essentially, what Amaranth and other energy funds did was to provide liquidity for longer dated contracts, allowing storage operators to manage longer-dated risks better.

During the weeks of September 11, 2006 and September 18, 2006 the spread between summer/fall contracts and winter contracts for delivery in 2007 through 2011 narrowed considerably. On some of these days the decrease in the spread represented a multiple standard deviation event relative to how these spreads had moved in the past. Because Amaranth was essentially long these spreads (selling fall/summer contracts and buying winter contracts), they suffered substantial losses. Moreover, Amaranth was also long winter contracts, which were hedged with short spring contracts – again for delivery in 2007 through 2011 – and these spreads decreased substantially too (Burton and Strasburg (2006), Davis (2006)). The fund lost approximately \$560 million on September 14^{th} alone, and it lost about 35% during the week of September 11^{th} (Burton and Strasburg (2006), White (2006)).[18] Using this information and the decrease in the calendar spreads during these two

[17] In fact, the business of storage can be viewed in real option terms. The value of the storage facilities is essentially equal to the value of an option on the calendar spread on natural gas. As the near-term contracts cheapen and the longer-term contracts become more expensive, the value of storage operators' facilities become more valuable as these operators can buy the near-term contracts and sell the longer-delivery contracts and realize the value difference via storing natural gas.

[18] Trincal (2006) estimates that the fund was worth \$9.2 billion at the end of August, so a 35% loss would be approximately \$3.2 billion.

time periods,[19] we can follow the simple procedure laid out in Till (2006) to infer that Amaranth held approximately 100,000 total contracts (these contracts were accumulated through trading on the NYMEX and the ICE). On September 20th, all of the energy positions of Amaranth were transferred to JP Morgan Chase and Citadel Investment Group in an overnight transaction forced by the fund's brokers.

To determine the model predicted discount for this liquidating transaction, we need to estimate three parameters reflecting market conditions on September 19, 2006: the riskless rate, r; the volatility of fundamental value, σ; and the buy order flow arrival rate, λ^B. We estimate the riskless rate to be 4.72% using the yield on 30-day Treasury bills. The volatility of fundamental value to is estimated to be 95%, which is the implied volatility of one-month, at-the-money natural gas options. Finally, we estimate the daily buy order arrival rate to be 1,000 based on daily trading volume on NYMEX and ICE. With these parameter estimates, the model predicted transaction cost for selling 100,000 contracts is a staggering 34.8%. In reality, the liquidating transaction resulted in a loss of $1.4 billion relative to the market value of these positions at the end of day September 19th (see Till (2006)), representing an actual discount of 30%.

5.2 Calibrating the Quantity Structure of Immediacy Prices

An alternative approach to using the model is to calibrate the parameters to fit an observed relation between transaction costs and transaction sizes. In this method, the model is used as a device for inferring one of the underlying parameters, in the same spirit that the Black-Scholes formula is used to "back out" implied volatility. Given the ease of obtaining accurate proxies for the riskless rate of interest and the volatility of fundamental value, typically the parameter of interest will be the order arrival rate, λ^i. By using empirical transaction data to infer the order arrival rate, we implicitly relax the model's assumption of the existence of a single, privileged market marker, which is unlikely to hold in "normal" times. To imply out the order arrival rate one simply matches the model-predicted immediacy cost to a particular data point, or cross-section of points, for which one has an ample number of observations, e.g. the cross-section of the most frequently observed order quantities. Then, fixing the fundamental volatility from the daily return series, one can solve for the implied order arrival rate and use the analytical structure of the model to generate the entire, unobserved quantity structure of immediacy prices.

5.2.1 Transaction Costs for NYSE Firms

In this section, we illustrate how our model can be used to generate estimates of transaction costs – as a function of order quantity – for publicly traded securities. To calibrate the model we use transaction-level trade and quote data (TAQ) for NYSE firms in 2004, as well as daily data on

[19] During the week of September 11th, winter-spring spread decreased by $31,000 per contract (the contract multiplier for natural gas contracts on the NYMEX is 10,000 mmBtu), while on September 14th the spread decreased by $6,000 per contract. The summer/fall-winter spread decreased by $49,000 per contract during the week of September 11th and by $4,000 per contract on September 14th.

stock returns and US Treasury bond yields. We proxy the volatility of a firm's fundamental value using the standard deviation of its daily stock returns over the year, and use the yield on one-month Treasury bonds as a measure of the prevailing risk free rate of interest. Using the quantity cross-section of the realized percentage transaction costs, we then imply out the order arrival rate, effectively producing a transaction cost function for use at the end of the year.

In order to measure the percentage realized transaction costs, we first need to specify a measure of fundamental value. Here, we take the standard approach in the microstructure literature, and use the midpoint of the prevailing, best bid and ask quotes as our proxy for fundamental value, $\hat{V}_t$. We then define the proportional transaction cost as $p(Q) = \frac{P_t - \hat{V}_t}{\hat{V}_t}$, where P_t is the observed transaction price. This procedure is similar in spirit to the Lee and Ready (1991) tick-signing algorithm, which is used to classify data into buyer- or seller-initiated transactions. In this classification scheme, trades occurring at prices above (below) the prevailing midquote are considered to be buyer-initiated (seller-initiated). At the end of this procedure, we have two datasets for each firm containing transaction quantity-cost pairs, one for buyer-initiated transactions and one for seller-initiated transactions. Finally, to attenuate the effect of noise, we require each unique transaction quantity bucket to contain at least 20 observations, and calibrate our model to the sample mean of the proportional transaction cost within each quantity bucket. By maximizing a model fit criterion, we can then imply out the buy and sell order flow arrival rates that are most consistent with the structural model.

To ensure that we choose a plausible specification for the scaling order of the waiting times in quantity, we first examine this relation empirically in the cross-section of NYSE listed firms. Specifically, we use signed transaction data for 2004 to estimate the mean waiting times for cumulative flows of Q shares, $E[\tau^i(Q)]$, and examine their scaling order with respect to quantity by estimating the following non-linear least squares regression,

$$E[\tau^i(Q)] = \alpha_0 + \alpha_1 \cdot Q^n. \tag{18}$$

Using our sample of $1,488$ firms, we find that the cross-sectional mean estimate of n is 0.9805 (0.9785) for buys (sells). When we restrict our attention to firms with concave arrival time scaling, the mean estimate of n is 0.9788 (0.9766) for buys (sells), suggesting very minute deviations from linearity. The main concern, however, are firms with convex arrival time scaling which is sufficiently extreme ($n > 2$) to offset the concavity of the square root function appearing in our approximation to the price of immediacy, (11). Overall, we find that no firm delivers a point estimate for n in excess of two, and even when firms do exhibit convex scaling, departures from linearity are similarly small, with a mean n estimate of 1.0047 for buys and 1.0099 for sells. In fact, less than 7% of the firms in the sample exhibit statistically significant and convex scaling in waiting times. Moreover, because the entire trading record is used repeatedly to construct the mean interarrival times at various quantity sizes, the strength and frequency of the rejection of linearity is already likely to be overstated.

To calibrate the order flow arrival rates we use a simple linear model based on the approximate formula (11), combined with the auxiliary assumption of a constant arrival rate. Under the constant arrival rate assumption, the waiting times scale linearly in the transaction quantity, $\frac{1}{\lambda^i(Q)} = \frac{1}{\lambda^i(1)} \cdot Q$, which enables us to re-write the percentage transaction cost as,

$$p(Q) \approx \sigma \cdot \sqrt{\frac{Q}{2\lambda^i(1)}}. \tag{19}$$

Within each quantity category, we calculate the average transaction cost, $\overline{p(Q)}$, and the average dollar transaction value, $\overline{Q}$, based on the median stock price over the period. Because we measure transaction quantities in terms of dollars, the value of the order arrival rate implied from our model will be in terms of dollars per unit time. This convention allows us to interpret the measured values as a fraction of market capitalization, giving them the flavor of a scale-free turnover metric.

For each firm with at least 10 quantity categories, we estimate the following specification:

$$\overline{p(Q)} = \beta_0 + \beta_1 \cdot \sqrt{\overline{Q}}. \tag{20}$$

The regressions are estimated using ordinary least squares (OLS) as well as a weighted-least squares procedure (GLS) that weights observations by the total value of transactions within each quantity category. The order flow arrival rates are recovered through the equation: $\hat{\lambda}^i = \frac{\sigma^2}{2\hat{\beta}_1^2}$. Formally, the model does not suggest the intercept, but any fixed cost warrants its inclusion.

Figure 6 displays the average quantity structure of immediacy prices for NYSE firms by size quintile. In addition, the table below the figure summarizes a variety of characteristics of the underlying firms. Consistent with intuition, small firms have higher volatility and lower implied order flow arrival rates than large firms. This translates into considerably larger immediacy prices for small firms. Our estimates of the fixed cost of transactions are meaningfully different across the size quintiles, and range from 18bps for the smallest firms to 2bps for the largest firms. Figure 6a illustrates that this relation holds across dollar transaction sizes. On average, the price of immediacy is about 10 times larger for a firm in the smallest quintile of NYSE firms than for a firm in the largest quintile. Overall, our calibrated order arrival rates imply annual turnovers of about 20 times market capitalization. We also display the quantity structure of immediacy prices as functions of the fraction of shares outstanding (Figure 6b). The calibrated model generates interesting predictions for more extreme capital market transactions. For example, a cash tender offer can be thought of as a demand for the instantaneous acquisition of all the shares outstanding of the target company. The liquidity component of such a transaction is predicted to be 7.5% on average for a small firm and 4.3% for a large firm. While actual takeover premia tend to be larger than these estimates, it is intriguing to consider that a substantial portion of takeover premia may represent a premium for immediacy.

Finally, our model predicts that immediacy prices are concave in quantity. We evaluate this

prediction by comparing the explanatory power of our limit order model (square root model) to a linear model, suggested by a traditional microstructure framework. A summary of this analysis is reported in Table 1. Specifically, we report the average R^2 by size decile for both the square root and linear models, as well as the fraction of times that the square root model produces a larger R^2. Overall, the square root model "beats" the linear model 86% of the time using the order flow arrival rates estimated via OLS and 82% of the time using the GLS estimates. This suggests that the relation between immediacy price and transaction size is indeed concave, even for moderate transaction sizes observed on a daily basis. The improvement of the square root model over the linear model is largest for the biggest firms.

5.2.2 Estimating the Liquidity Component of S&P 500 Index Inclusions

The real test of any model requires analyzing how it performs out-of-sample. Consequently, to evaluate the joint effectiveness of the previously described calibration procedure and our model, we apply the procedure to a sample of firms that are being included in the S&P 500 index and examine how the model's predicted transaction costs compare to the actual realized abnormal returns around the inclusion. To get a sense a how extreme this test is, it is worth noting that – on average – the inclusion represents a transaction for 10% of shares outstanding, whereas the maximum transaction size in the TAQ data used to calibrate the model averages only 0.0033% of shares outstanding.

Index inclusions are widely recognized as large liquidity events. Harris and Gurel (1986) and Shleifer (1986) estimate abnormal returns for firms added to the S&P 500 index to be three percent on the inclusion day.[20] Both papers argue that inclusions to the S&P 500 index convey little new information about future return distributions, but cause outward shifts in excess demand by investment strategies that track the S&P 500. Harris and Gurel interpret their findings as supportive of price pressure (Scholes (1972)) because they find nearly complete price reversal over a two-week interval. On the other hand, Shleifer views his results as evidence of downward sloping long-run demand curves for securities because he finds little price reversal. Recently, Wurgler and Zhuravskaya (2002) test the downward sloping demand curve hypothesis by classifying firms added to the S&P index on the basis of whether they have close substitutes. Consistent with the hypothesis that excess demand curves slope downward, the inclusion effect is greater for firms that lack close substitutes, where it is riskier for arbitrageurs to keep demand curves elastic.

Given the extreme size of the transactions associated with index inclusions, explaining the cross-section of abnormal returns around the event poses a significant challenge, particularly for a structural model. For example, Wurgler and Zhuravskaya (2002) carry out a cross-sectional regression of abnormal returns around index inclusions on the level of arbitrage risk, proxied by residual variance from a market model regression. Although the level of arbitrage risk is highly statistically significant, it delivers an R^2 of only about 0.04. Moreover, their regression model lacks

[20] More recent inclusions are associated with larger abnormal returns. For our sample covering 1994 through 2004, the average cumulative abnormal return from announcement to inclusion is 7%.

the structure to predict the *ex ante* liquidity cost of the index inclusion for individual firms.

We collect a sample of firms added to the S&P 500 index between 1994 and 2004, requiring that TAQ data are available. This results in a sample of 255 firms. For each event firm, we calculate abnormal returns as the residuals from a one-factor market model. The market model parameters are estimated over a 150-day window ending 21 days prior to the announcement date using the value-weighted CRSP index as a proxy for market returns. The individual firm abnormal stock returns are cumulated (CARs) from the announcement date to the inclusion date. To estimate the quantity of shares that need to be purchased by funds indexed to the S&P 500, we use information obtained from the 2005 Annual Survey of S&P Indexed Assets issued by Standard & Poor's, which reports annual estimates of the total value of capital indexed to the S&P 500. Over our sample period, the total value of indexed capital corresponds to roughly 10% of the market capitalization of the firms in the index. In addition, from CRSP we collect the market value of the stocks comprising the S&P 500 index and the shares outstanding for the newly added firms at the end of the month prior to the announcement. Finally, we calibrate the model parameters for each sample firm using the procedure described in the previous section with data from the year prior to the announcement.

The analysis involves running cross-sectional regressions of CARs on the variables predicted by our model to explain the cross-section of price impacts. In particular, the model predicts that CARs should be positively related to both volatility and the square root of the ratio of transaction size and the calibrated order flow arrival rate, $\sqrt{\frac{Q}{\lambda}}$. More precisely, the model predicts that CARs should be proportional to the interaction of these two terms.

Table 2 reports the results from our cross-sectional regressions. As the model predicts, both volatility and the square root term are individually statistically significant (specifications 1, 2, and 6). Both variables remain statistically significant in multiple regressions (specifications 3 and 7). When the interaction term is used as the single explanatory variable the adjusted R^2 increases to 0.11, which represents a significant improvement relative to the R^2 of 0.04 reported by Wurgler and Zhuravskaya (2002).[21] Furthermore, in regressions that include both the interaction term and the individual terms, only the interaction term is significant, suggesting that the specific form recommended by the model is better than an *ad hoc* specification. Finally, our results hold independent of whether the order flow arrival rates are calibrated using the OLS or GLS procedure.

The final analysis involves regressing the CARs on the model-predicted price impact, $p(Q)$,

$$CAR_i = \gamma_0 + \gamma_1 \cdot E[p(Q)] \tag{21}$$

Although qualitatively similar to the previous analysis, here the expected price impact is properly scaled according to (11) and includes the intercept, $E[p(Q)] = \hat{\beta}_0 + \sigma \cdot \sqrt{\frac{Q}{2\hat{\lambda}}}$. Consequently, if the model is an unbiased predictor of the liquidity component of the event abnormal returns, the

[21] In unreported analysis (available upon request), we find that total volatility has more explanatory power than residual volatility (R^2 of 0.06 and 0.04, respectively). Residual volatility is not statistically significant when total volatility or our estimate of price impact are included in the regression, while the variables suggested by the model retain statistical significance.

slope coefficient, γ_1, should be precisely equal to one. Under this specification, the intercept, γ_0, can be interpreted as a measure of the average information (or other) effect associated with the event. The results from this regression for a few model specifications are displayed in Table 3. The estimates of the slope coefficients for the limit order model are roughly equal to 1.2, and are not statistically distinguishable from one at conventional significance levels. For comparison, we include the results for a linear model of transaction costs. Here, the slope coefficients are roughly zero and the R^2s are minuscule relative to those from the limit order model.

It is particularly encouraging that the improvement of the limit order model over the linear model is so extreme in this setting, in contrast to the modest improvement it delivered in the earlier calibration exercise using all NYSE firms. Because transactions associated with index inclusions are over 300 times larger than the maximum transaction size included in the calibration, we take this as evidence that the limit order model performs well out-of-sample. This presents strong evidence in favor of the concave price impact specifications, and confirms that the model is able to deliver unbiased predictions in situations in which data are limited, and therefore a model is most needed.

6 Conclusion

This paper views the wedge between fundamental value and capital market transaction prices as emerging from an imperfect market for immediacy. In a setting with stochastic arrivals of buyers and sellers, we grant the market maker the privilege of being the sole trading counterparty for investors with inelastic demands for immediacy, enabling him to extract rents from impatient order flow. The magnitude of these rents depends on the competition implicit in opposing order flow and defines the price of immediacy.

The mechanism for trade in our model is a limit order, and immediacy is supplied when the limit order is executed by either the market maker or opposing order flow. We view limit orders as options, and their value as a measure of the cost of transacting. Because of his unique position, the market maker is the effective owner of the option embedded in a limit order, and must decide when and if to exercise this option. The incentive for exercising a limit order option early arises through competition for the order with the opposing order flow, which from the market maker's perspective acts like a stochastic liquidating dividend. In this setting, limit prices that induce immediate exercise of the American-type limit orders, determine the price of immediacy at various quantities and are functionally equivalent to bid and ask prices.

The option-based model of immediacy proposes that immediacy prices are determined by the product of the fundamental volatility of the security and the square root of a scaled measure of instantaneous excess demand, i.e. the ratio of order quantity to the share arrival rate. Larger transactions effectively require writing options with longer maturity, and option values increase with the square root of time to maturity. This simple formula can be readily calibrated using empirical data, and used to generate the entire unobserved quantity structure of transaction costs. Empirical analysis of stock market transactions for NYSE firms supports the predictions of the

model, confirming that our model with full-information, but imperfect market making, is able to describe a range of properties of real world transaction costs.

Two simple implications of our basic setup are that market makers are long volatility and that they earn profits in the presence of order imbalances. This seems to fit with common intuition and empirical evidence on supplying liquidity. For example, in the price pressure hypothesis proposed by Scholes (1972), uninformed shifts in excess demand can cause prices to temporarily diverge from their information-efficient values in order to compensate those that provide liquidity. This should not occur with perfectly competitive market making in the absence of imperfect information. Our model captures this notion of price pressure through imperfect competition. A larger order flow imbalance represents a weakening of competition for the monopoly market maker, allowing him to extract larger rents. In other words, investors with common liquidity demands are forced to write options with longer effective maturities (i.e. more valuable options) when order imbalances grow and/or become somewhat persistent. A consequence of this type of price pressure is that supplying immediacy in these situations is profitable.

An attractive feature of the limit order framework is that the model can be estimated as a function of relatively observable variables. We propose a method for implementing the model using forecasts of volatility and order flow data. We also jointly test the model and the calibration procedure by predicting the price reaction for firms being added to the S&P 500 index. As the model predicts, we find that volatility and the square root of the ratio of transaction size to order flow are significant variables in explaining price reactions. Moreover, the model produces unbiased forecasts of the price reaction in a setting where the average transaction size is over 300 times bigger than the largest transaction used in the calibration. This compares very favorably to alternative models, which produce highly biased estimates when used this far out-of-sample.

The practicality of the option-based framework suggests that it may be an interesting platform for future theoretical and empirical research. In particular, the model could be used to estimate the immediacy component of corporate transactions like security issuance, repurchases, and takeovers. Finally, the model may be a useful step towards a new measure of liquidity risk. The uncertainty over transactable prices, relative to fundamental value, produces a liquidity risk. As such, the time series variation of the price of immediacy is a natural measure of this risk. This suggests extending the baseline model to incorporate time-varying arrivals and investigating commonality in the time dynamics of the resulting quantity structure of immediacy prices.

References

[1] Acharya, V. and L. H. Pedersen, 2005, Asset Pricing with Liquidity Risk, *Journal of Financial Economics* 77, pp. 375-410.

[2] Amihud, Y. and H. Mendelson, 1980, Dealership market: Market making with inventory, *Journal of Financial Economics* 8, 31-53.

[3] Bacidore, J., R. Battalio, R. Jennings, 2003, Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange, *Journal of Financial Markets* 6, 337-362.

[4] Bagehot, W., 1971, The only game in town, *Financial Analysts Journal* 22, 12-14.

[5] Burton K. and J. Strasburg (2006): "Amaranth Plans to Stay in Business, Maounis Says," *Bloomberg News*, Sept. 22.

[6] Carr, Peter, 1998, Randomization and the American Put, *Review of Financial Studies* 11(3), 597-626.

[7] Cochrane, John, 2005, Asset pricing program review: Liquidity, trading, and asset prices, *NBER Reporter*, Winter.

[8] Copeland, Thomas E., and Dan Galai, 1983, Information effects and the bid-ask spread, *Journal of Finance* 38, 1457-1469.

[9] Davis, A. (2006): "How Giant Bets on Natural Gas Sank Brash Hedge-Fund Trader," *Wall Street Journal*, Sept. 19.

[10] Demsetz, Harold, 1968, The cost of transacting, *Quarterly Journal of Economics* (82), 33-53.

[11] Duffie, Darrell, Nicolae Gârleanu, and Lasse Pedersen, 2005, Over-the-counter markets, *Econometrica* 73, 1815-1847.

[12] Garman, M., 1976, Market microstructure, *Journal of Financial Economics* 3, 257-275.

[13] Glosten, L. and P. Milgrom, 1985, Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 13, 71-100.

[14] Grossman, Sanford J., and Merton H. Miller, 1988, Liquidity and market structure, *Journal of Finance* 43, 617-633.

[15] Guo, X., and Zhang, Q., 2004, Closed-form solutions for perpetual American put options with regime switching, *SIAM Journal of Applied Mathematics* 64(6), 2034-2049.

[16] Harris, Larry, 2003, *Trading and Exchanges: Market Microstructure for Practitioners*, Oxford University Press, New York.

[17] Harris, Larry, and Eitan Gurel, 1986, Price and volume effects associated with changes in the S&P 500: New evidence for the existence of price pressures, *Journal of Finance* 41, 815-829.

[18] He, Chen, Elizabeth R. Odders-White, and Mark J. Ready, 2006, The impact of preferencing on execution quality, *Journal of Financial Markets* 9, 246-273.

[19] He, Hua, Jiang Wang, 1995, Differential information and dynamic behavior of stock trading volume, *Review of Financial Studies* 8, 919-972.

[20] Ho, Thomas and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 47-73.

[21] Lee, Charles M. C. and Mark J. Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733-746.

[22] Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315-1335.

[23] Longstaff, Francis A., Eduardo S. Schwartz, 2001, Valuing American options by simulation: a simple least squares approach, *Review of Financial Studies* 14, 113-147.

[24] Merton, Robert C., 1969, Lifetime portfolio selection under uncertainty: The continuous time case, *Review of Economics and Statistics* 51, 247-257.

[25] Merton, Robert C., 1971, Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* 3, 373-413.

[26] Merton, Robert C., 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483–511.

[27] O'Hara, Maureen, 2003, Presidential Address: Liquidity and Price Discovery, *Journal of Finance* 58, 1335-1354.

[28] O'Hara, Maureen, 2004, *Market Microstructure Theory*, Blackwell Publishing, Oxford, UK.

[29] Pastor, L. and R. Stambaugh, 2003, Liquidity Risk and Expected Stock Returns, *Journal of Political Economy* 111, 642–685.

[30] Scholes, Myron, 1972, The market for corporate securities: Substitution versus price pressure and the effects of information on stock prices, *Journal of Business* 45, 179-211.

[31] Shleifer, Andrei, 1986, Do demand curves for stocks slope down? *Journal of Finance* 41, 579-590.

[32] Stoll, Hans R., 1978, The supply of dealer services in securities markets, *Journal of Finance* 33, 1133-1151.

[33] Till, H. (2006): "EDHEC Comments on the Amaranth Case: Early Lessons from the Debacle," Working Paper, EDHEC Risk and Asset Management Research Center, EDHEC Business School.

[34] Trincal, E. (2006): "Amaranth is Bleeding Assets, But Keeps Talking to Investors," *Hedge-World*, Sept. 21.

[35] Vayanos, D, and T. Wang, 2007, Search and endogenous concentration of liquidity in asset markets, *Journal of Economic Theory*, forthcoming.

[36] Wang, Jiang, 1993, A model of intertemporal asset prices under asymmetric information, *Review of Economic Studies* 60, 249-282.

[37] Werner, Ingrid, 2003, NYSE order flow, spreads, and information, *Journal of Financial Markets* 6, 309-335.

[38] White, Ben, 2006, Amaranth Chief Defends Policies, *Financial Times*, 9/23-24/2006.

[39] Wurgler, Jeffrey and Katia Zhuravskaya, 2002, Does arbitrage flatten demand curves for stocks? *Journal of Business* 75, 583-608.

A Finite Maturity Limit Orders

The baseline model of Section 2 is solved under the assumption that limit order writers never cancel a submitted order. This feature allows us to treat the limit order option as a perpetual option, subject to a stochastic liquidating dividend in the form of execution by the opposing order flow. However, since this assumption is clearly open to challenge, we devote this appendix to showing that it can be relaxed considerably without any effect on the qualitative features of the quantity structure of immediacy prices.

We begin our analysis by considering limit orders that are subject to random cancelation by the limit order writer, and then turn our attention to limit orders with a finite maturity date. Although limit orders are *de facto* unlikely to be canceled at randomly selected times, random cancelation will be observationally indistinguishable from a deterministic cancelation rule, so long as the market maker cannot infer this rule. We therefore assume that a limit order is canceled by the order writer at the N-th arrival time of a Poisson process with intensity η. Under this auxiliary assumption, the order cancelation time, τ, will have an Erlang distribution with:

$$Pr\{\tau \in dt\} = \frac{\eta^N}{(N-1)!} t^{N-1} e^{-\eta t} dt \tag{1}$$

and the expectation and variance of the cancelation time, τ, will be given by:

$$E[\tau] = \frac{N}{\eta} \qquad Var[\tau] = \frac{N}{\eta^2} \tag{2}$$

In the base case, when $N = 1$, the Erlang distribution collapses to an exponential distribution. In this case it is easy to show that the value of the buy and sell limit orders continues to be given by the expressions provided in Section 2, but with slightly modified cancelation intensities.

Proposition A.1 *The value of a sell (buy) limit order that is subject to cancelation by the limit order writer at the first jump time of a Poisson process with intensity η, is equivalent to the value of a limit order that is not subject to cancelation, but is subject to a stochastic liquidating dividend arriving at rate $\tilde{\lambda}^B(Q)$ ($\tilde{\lambda}^S(Q)$), is given by:*

$$\tilde{\lambda}^S(Q) = \lambda^S(Q) + \eta \tag{3}$$
$$\tilde{\lambda}^B(Q) = \lambda^B(Q) + \eta \tag{4}$$

A formal proof of this result can be found in Section C of the technical appendix.[1]

The simple isomorphism between limit orders that are not subject to cancelation by the limit order writer (i.e. perpetual limit orders) and those that are, shows that the qualitative features of the quantity structure of transaction prices will be unaffected by the introduction of the cancelation feature.

To establish that our results continue to hold in the case of finite duration limit orders (i.e. orders that will be canceled at a future date T), we exploit the randomization device of Carr (1998). This mathematical device takes advantage of the scaling of the moments of an Erlang distributed random variable in the Poisson arrival intensity, η, to synthesize a random variable with a pre-specified mean and zero variance. To see this, suppose we let $\eta = \frac{N}{T}$, and allow $N \to \infty$. Asymptotically, the moments of the limit order cancelation time, τ, collapse to $E[\tau] \to T$ and $Var[\tau] \to 0$. In other words, the limit order is canceled at time T with unit probability.

To determine the value of a limit order subject to cancelation at time T, it is therefore sufficient to determine the value of the limit order subject to Erlang cancelation when $\eta = \frac{N}{T}$, and $N \to \infty$. Under the Erlang cancelation scheme, the value of a limit order will depend on the fundamental value of the underlying, $F_{Q,t}$; the arrival intensity of the opposing order flow, $\lambda^i(Q)$; and the number of periods left to the termination of the option, n. Given these assumptions, the value of the limit order will be given by the solution to the following system of N ($n = N \ldots 1$) ordinary differential equations:

$$L_F^{(n)} \cdot (rF_{Q,t}) + \frac{1}{2} L_{FF}^{(n)} \cdot (\sigma F_{Q,t})^2 - r \cdot L^{(n)} = \lambda^i(Q) \cdot (L^{(n)} - 0) + \eta \cdot (L^{(n)} - L^{(n-1)}) \tag{5}$$

with $L^{(0)} = 0$ (i.e. the limit order becomes worthless upon cancelation). The terms on the left hand side of the equality represent the evolution of the limit order value in the absence of jumps, while the terms on the right hand side represent the probability weighted losses from order exercise by oncoming order flow and the passage of time, as measured by the jumps in the $Poisson(\eta)$ variable. To solve this system of ODEs we proceed by backwards recursion, starting with state $n = 1$. The solution is comprised of a sequence of state-dependent value functions and

[1] The technical appendix can be downloaded from the authors' websites.

the associated optimal early exercise thresholds. A characterization of the complete, recursive solution is given in the following proposition.

Proposition A.2 *The value of a sell limit order in state n is given by:*

$$L^{(n,S)} = \begin{cases} \alpha_{0,n} \cdot F_{Q,t}^{\phi_+(\lambda^B)} + \alpha_{1,n} \cdot F_{Q,t}^{\phi_-(\lambda^B)} + \left(\frac{\eta}{\eta+\lambda^B(Q)}\right) \cdot F_{Q,t} - \left(\frac{\eta}{r+\eta+\lambda^B(Q)}\right) \cdot QK & V_t \geq V_{n-1}^* \\ \beta_{0,n} \cdot F_{Q,t}^{\phi_+(\lambda^B)} + L_p^{(n,S)}(V_t < V_{n-1}^*) & V_t < V_{n-1}^* \end{cases} \tag{6}$$

where V_n^* denotes the optimal early exercise threshold for state n and $L_p^{(n,S)}(V_t < V_{n-1}^*)$ is an analytical expression related to the value function, $L^{(n-1,S)}$, in the continuation region for state $n-1$. The values for $(\alpha_{0,n}, \alpha_{1,n}, \beta_{0,n}, V_n^*)$ can be determined by solving a system of equations described in the technical appendix. The corresponding solution for a buy limit order can be found in Section C.

It is possible to show that the sequence of optimal exercise thresholds for a sell limit order, V_n^*, is increasing in n, reflecting the increasing time-value of the limit order option. Despite the complexity of the full solution, the form of the value function characterizing the limit order option in an arbitrary state, n, is closely related to the solution from Section 2. The recursive analytical solution, combined with numerical solution of the system of equations parameterizing the coefficients of the value function and optimal exercise threshold, confirms that the quantity structure of transaction prices (now indexed by state n) retains all the qualitative features examined in Section 3.

Table 1: Summary of Transaction Cost Calibration for NYSE Stocks by Size Decile (2004).

This table reports the average R^2 from firm-level regressions of percentage transaction costs on dollar transaction sizes for NYSE firms in 2004, grouped by market capitalization decile. The dependent variable, p, is the average percentage transaction cost within a quantity category. The independent variable is either the average dollar transaction size with in a quantity category, Q, (linear) or the square root of Q (square root). The quantity categories are defined separately for each firm based on the unique transaction quantities in the 2004 TAQ data. The regressions are estimated via ordinary least squares (OLS) and a weighted least squares procedure (GLS) that weights observations by total dollar value rather than the number of transactions in each quantity category. The fraction of times where the square root model produces a larger R^2 than the linear model is also reported. The regressions are estimated separately for buy and sell transactions. Buy transactions are identified as those with a transaction price above the midpoint of prevailing bid and ask prices. The number of observations is denoted by N.

Buy Transactions							
	OLS			**GLS**			
Size Decile	Linear	Limit Order Model (Square Root)	Fraction with Square Root > Linear	Linear	Limit Order Model (Square Root)	Fraction with Square Root > Linear	N
1	0.17	0.18	0.68	0.41	0.43	0.70	136
2	0.21	0.24	0.80	0.65	0.67	0.73	138
3	0.18	0.20	0.76	0.64	0.66	0.69	137
4	0.21	0.23	0.80	0.67	0.69	0.62	137
5	0.18	0.22	0.91	0.69	0.73	0.84	138
6	0.19	0.23	0.92	0.70	0.73	0.84	137
7	0.18	0.23	0.93	0.69	0.74	0.89	137
8	0.15	0.20	0.97	0.67	0.73	0.96	137
9	0.20	0.24	0.93	0.71	0.76	0.93	138
10	0.19	0.24	0.96	0.69	0.76	0.98	137
All NYSE	0.18	0.22	0.86	0.65	0.69	0.82	1385

Sell Transactions							
	OLS			**GLS**			
Size Decile	Linear	Limit Order Model (Square Root)	Fraction with Square Root > Linear	Linear	Limit Order Model (Square Root)	Fraction with Square Root > Linear	N
1	0.26	0.27	0.68	0.50	0.51	0.69	137
2	0.32	0.34	0.79	0.71	0.72	0.72	137
3	0.30	0.34	0.76	0.75	0.77	0.70	138
4	0.31	0.35	0.81	0.76	0.78	0.63	137
5	0.31	0.36	0.90	0.79	0.81	0.83	138
6	0.31	0.36	0.92	0.78	0.80	0.84	138
7	0.28	0.34	0.93	0.77	0.81	0.89	137
8	0.22	0.29	0.97	0.72	0.78	0.96	138
9	0.27	0.33	0.93	0.77	0.82	0.93	137
10	0.23	0.30	0.96	0.74	0.80	0.98	138
All NYSE	0.28	0.32	0.86	0.73	0.76	0.82	1388

Table 2: Regressions Explaining Abnormal Returns Around S&P 500 Additions.

This table reports estimated coefficients from cross-sectional regressions of abnormal returns around S&P 500 index inclusions. The dependent variable is the abnormal stock return cumulated from the announcement date to the inclusion date. Abnormal returns are the residuals from a one-factor market model. Market model parameters are estimated over a 150 day window ending 21 days prior to the announcement date using the value-weighted CRSP index as a proxy for market returns. Volatility is the annualized standard deviation of stock returns. Order Flow is the square root of the ratio of transaction value to calibrated buy order arrival rate, $\sqrt{Q/\lambda}$. The calibrated buy order arrival rate, λ, is estimated via ordinary least squares (OLS) and a weighted least squares procedure (GLS) that weights observations by total dollar value. The estimate of transaction size, Q, is the dollar value of shares expected to be purchased by funds mimicking the S&P 500 index. The adjusted R-square is denoted as R^2, t-statistics are in parentheses, and the number of observations is denoted as N.

Variable	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Intercept	0.0325	0.0533	0.0173	0.0459	0.0409	0.0565	0.0188	0.0463	0.0414
	(3.24)	(7.63)	(1.59)	(7.38)	(2.83)	(7.78)	(1.66)	(7.02)	(2.65)
Volatility	0.0799		0.0798		0.0204		0.0817		0.023
	(4.13)		(4.20)		(0.66)		(4.27)		(0.67)
Order Flow (OLS)		0.0001	0.0001		0.0204				
		(3.11)	(3.20)		(0.66)				
Volatility x Order Flow (OLS)				0.0003	0.0004				
				(5.69)	(2.45)				
Order Flow (GLS)						0.0001	0.0001		-0.0000
						(2.31)	(2.54)		(-0.66)
Volatility x Order Flow (GLS)								0.0003	0.0004
								(4.99)	(2.08)
R^2	0.060	0.033	0.093	0.110	0.111	0.017	0.079	0.086	0.091
N	255	255	255	255	255	255	255	255	255

Table 3: Regressions of Actual Event Reactions on Expected Event Reactions.

This table reports estimated coefficients from cross-sectional regressions of abnormal returns around S&P 500 index inclusions. The dependent variable is the abnormal stock return cumulated from the announcement date to the inclusion date. Abnormal returns are the residuals from a one-factor market model. Market model parameters are estimated over a 150 day window ending 21 days prior to the announcement date using the value-weighted CRSP index as a proxy for market returns. For the limit order model, $E[\text{Price Impact}] = \beta_0 + \sigma \cdot \sqrt{Q/\lambda}$, is calculated for each firm being included in the S&P 500 index.. The estimate of transaction size, Q, is the dollar value of shares expected to be purchased by funds mimicking the S&P 500 index. The per second buy order arrival rate, λ, and the intercept, β_0, are calibrated for each firm from TAQ data over the year prior to the announcement date using either an ordinary least squares (OLS) or weighted least squares (GLS) regression as in Table 1. The per second volatility, σ, is the standard deviation of daily returns over the year prior to the announcement date, scaled by the square root of trading seconds per day. For the linear model, $E[\text{Price Impact}] = b_0 + b_1 \cdot Q$, where b_0 and b_1 are calibrated for each firm from TAQ data over the year prior to the announcement date using either an ordinary least squares (OLS) or weighted least squares (GLS) regression as in Table 1. The adjusted R-square is denoted as R^2, t-statistics are in parentheses, and the number of observations is denoted as N.

Model	Intercept	E[Price Impact]	R^2 [N]
Limit Order Model calibrated via OLS	0.04	1.22	0.11
	(6.77)	(5.79)	[255]
Limit Order Model calibrated via GLS	0.04	1.21	0.09
	(6.40)	(5.10)	[255]
Linear Model calibrated via OLS	0.07	0.01	0.00
	(2.31)	(1.36)	[255]
Linear Model calibrated via GLS	0.07	0.00	0.00
	(2.61)	(0.93)	[255]

Figure 1: **The price of immediacy.** This figure illustrates the relationship between transaction prices and the fundamental value in two capital markets. In a perfect capital market all transactions – independent of quantity demanded – take place at the fundamental value, V_t. In an imperfect capital market, where the market maker has pricing power, transaction prices, $K^*(Q, \lambda)$, diverge from fundamental value and depend on the size of the transaction, Q, relative to the share arrival rate, λ. The wedge between fundamental value and the transaction price represents the price of immediacy.

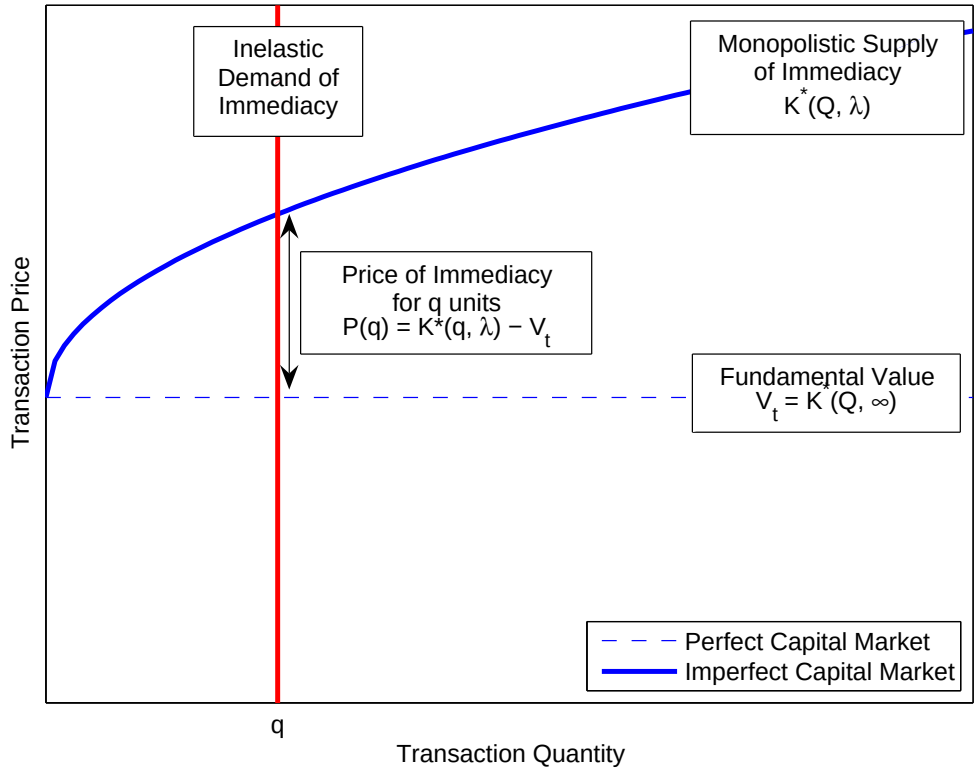


Figure 2: **Quantity structure of immediacy prices.** This figure illustrates the price of immediacy as a function of order quantity. The price of immediacy is computed as the fraction of fundamental value which has to be forgone to induce the market maker to execute a limit order instantaneously. It is plotted against the limit order quantity assuming an arrival rate of one share per second ($\lambda^i = 1$), and an annualized riskless rate of 5%.

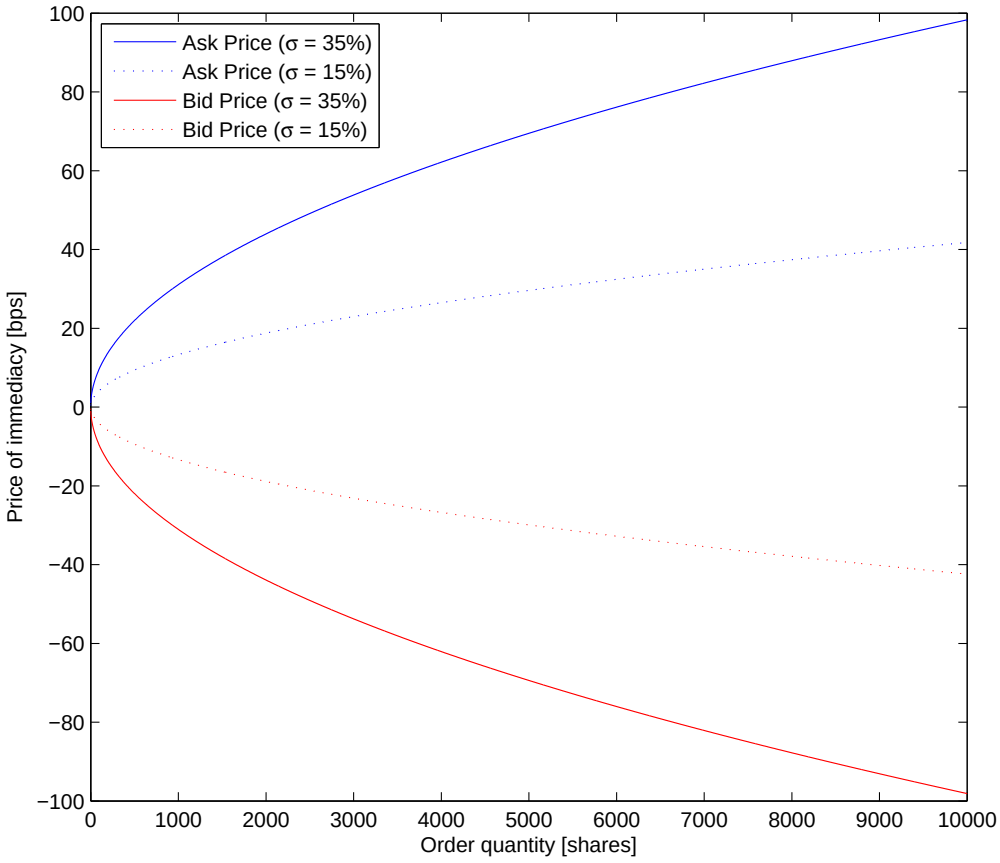


Figure 3: **Immediacy prices as a function of the order arrival rate.** The figure depicts the price of immediacy for a buy (sell) transaction for Q = 1,000 (Q = 10,000) shares as a function of the order arrival rate. The x-axis plots the base 10 logarithm of the share arrival rate λ^S (λ^B) for sells (buys). The fundamental volatility equals 35% *per annum*, and the riskless rate is fixed at 5% *per annum*

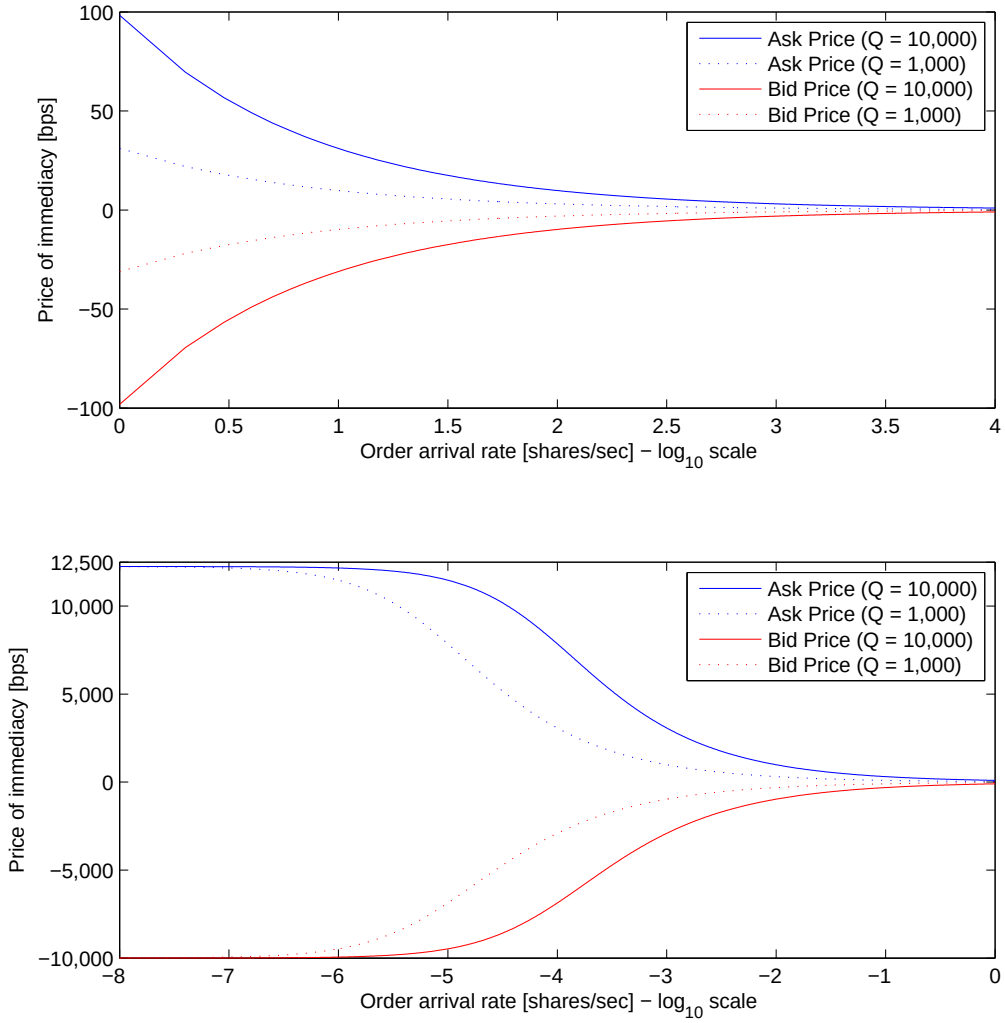


Figure 4: **Immediacy prices during liquidity events.** The figure depicts the percentage cost of obtaining immediacy for a buy (sell) transaction - as a function of the demanded quantity - during a liquidity event. In the base case, buy and sell orders are assumed to arrive at a rate of one share per second; the riskless rate is fixed at 5% and the volatility of fundamental value is 15%. In the order imbalance scenario, the intensity of sell (buy) arrivals, λ^S (λ^B), increases (decreases) fivefold, but the volatility of fundamental value remains unchanged. In the liquidity event, the change in the order arrival intensities is accompanied by an increase in the fundamental volatility from 15% to 35%.

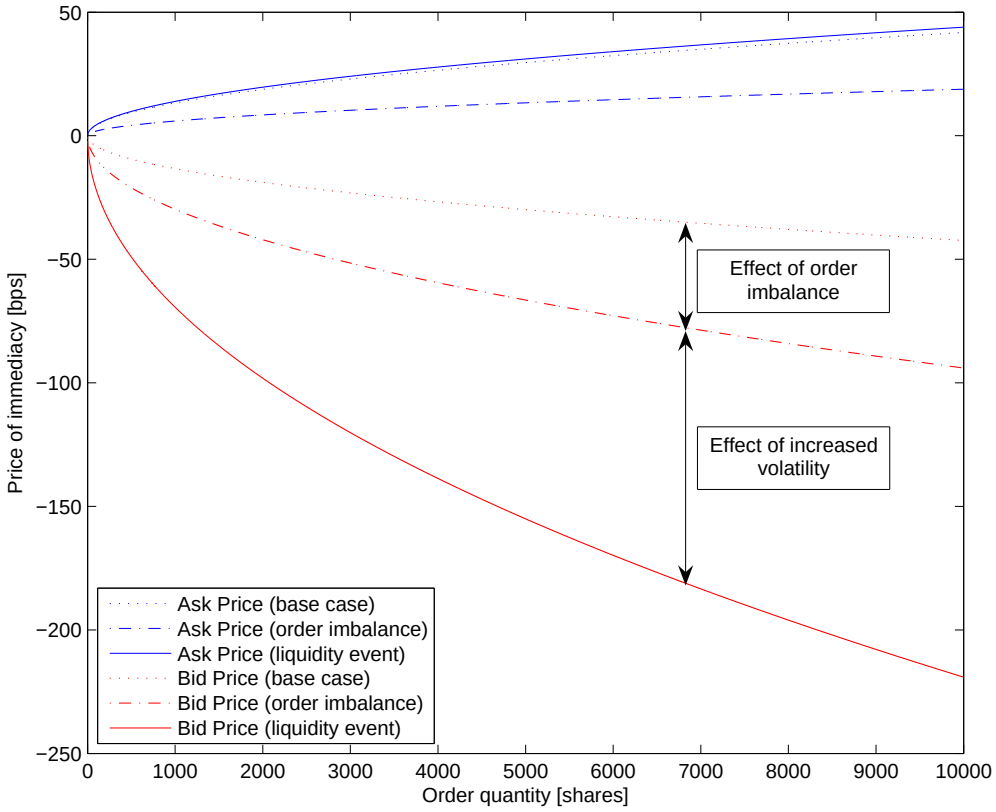


Figure 5: **Reservation values for patient traders**. This figure graphs the reservation value of a patient trader seeking to acquire 10,000 shares of a security with an order arrival rate of 100 shares per second, as a function of the trader's patience. The reservation value is expressed as a premium/discount relative to the prevailing fundamental value, V_t. Patience, α, is parameterized by the probability of not being executed within the decision horizon τ. The decision horizon is fixed at 100 (1000) seconds in the left (right) panel. Each plot considers two values of the underlying's volatility. The riskless rate and drift of the security are fixed at 5% and 12% *per annum*, respectively.

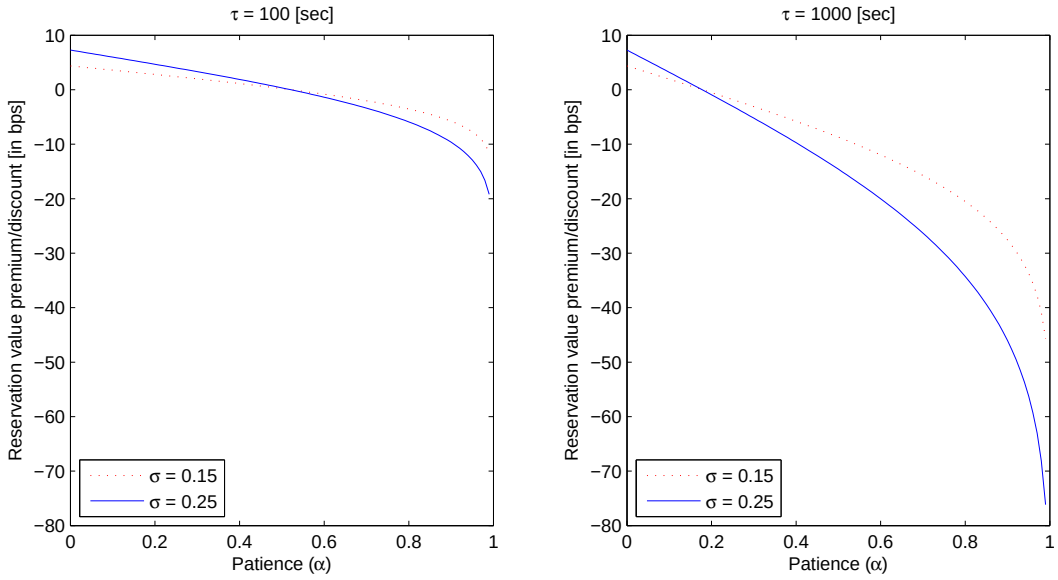
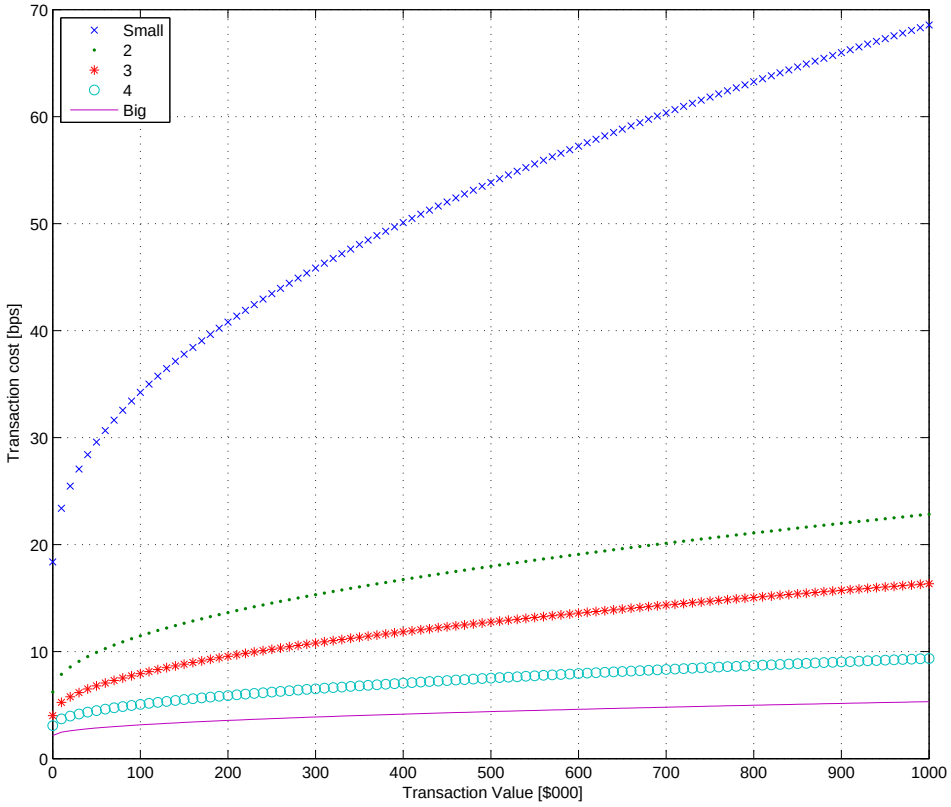


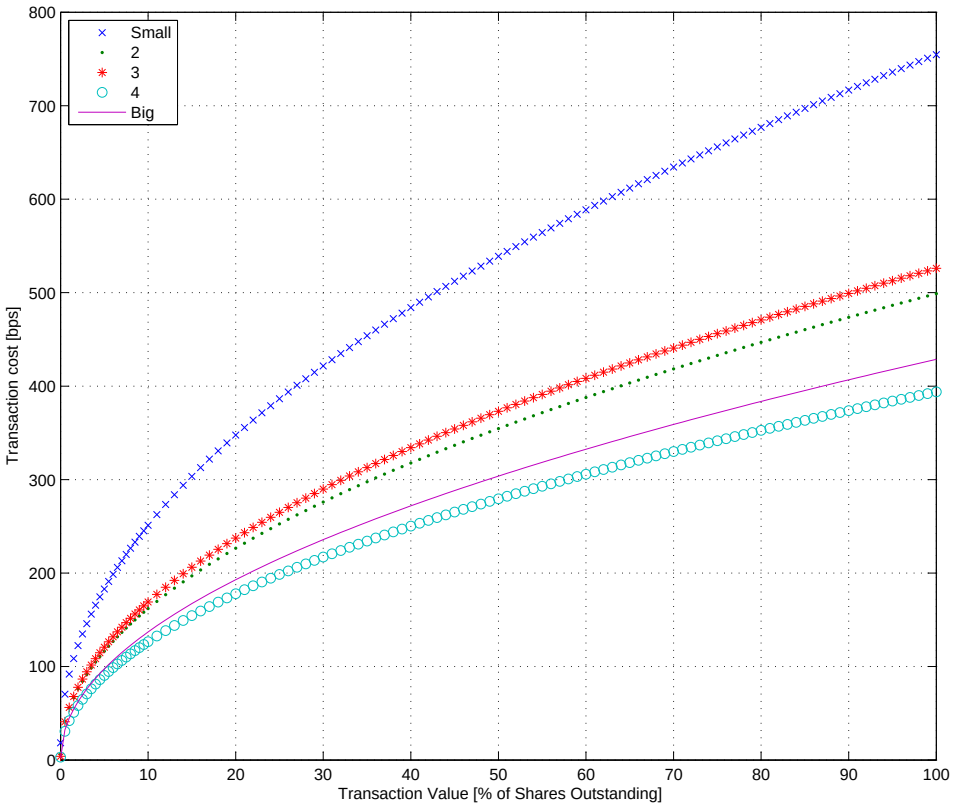
Figure 6: **Calibrated Quantity Structure of Immediacy Prices for NYSE Stocks by Size Quintile.** This figure displays the price of immediacy as a function of transaction size for NYSE firms in 2004, grouped by market capitalization quintile. Panel A presents immediacy prices as a function of dollar transaction size. Panel B presents immediacy prices as a function of fraction of shares outstanding. For each NYSE firm, the average percentage transaction cost within a quantity category is regressed on the square root of the average dollar transaction size in that quantity category. The quantity categories are defined separately for each firm based on the unique transaction quantities in the 2004 TAQ data. The regressions are estimated via a weighted least squares procedure that weights observations by total dollar value. Market capitalization is equal to the end-of-year total value of shares outstanding. Volatility is the standard deviation of daily stock returns. Lambda represents the calibrated order flow arrival rate in $10,000 blocks per second.

Panel A: Immediacy prices as a function of dollar transaction size.



Size Quintile	Median Market Cap [MM$]	Median Stock Price [$]	Median Volatility	Median Lambda [$10k/s]	Intercept [bps]	Avg. Price for $10K [bps]	Avg. Price for $100K [bps]	Avg. Price for $1M [bps]
1	303	15.10	0.37	0.10	18	23	34	69
2	871	26.80	0.30	0.44	6	8	11	23
3	1782	32.50	0.28	0.79	4	5	8	16
4	4020	36.71	0.25	2.24	3	4	5	9
5	16582	45.54	0.21	5.62	2	2	3	5

Panel B: Immediacy prices as a function of fraction of shares outstanding.



Size Quintile	Median Market Cap [MM$]	Median Stock Price [$]	Median Volatility	Median Lambda [$10k/s]	Intercept [bps]	Avg. Price for 10% [bps]	Avg. Price for 50% [bps]	Avg. Price for 100% [bps]
1	303	15.10	0.37	0.10	18	251	539	755
2	871	26.80	0.30	0.44	6	162	355	499
3	1782	32.50	0.28	0.79	4	169	373	526
4	4020	36.71	0.25	2.24	3	127	279	394
5	16582	45.54	0.21	5.62	2	137	304	429