Order Protection through Delayed Messaging

Eric M. Aldrich*
Department of Economics
University of California, Santa Cruz

Daniel Friedman†
Department of Economics
University of California, Santa Cruz

March 3, 2019

Abstract

Several financial exchanges recently introduced messaging delays (e.g. IEX and NYSE American) to protect ordinary investors from high-frequency traders who exploit stale orders. To capture the impact of such delays, we propose a simple parametric model of the continuous double auction market format. The model shows how messaging delays can lower transactions costs but typically increase queuing costs. Recently available field data support two key predictions: the distribution of queued pegged orders is highly leptokurtotic (discrete Laplace), and opportunities to profit from dynamic choice of pegged vs market orders are small and fleeting.

Keywords: Market design, high-frequency trading, continuous double auction, IEX.

JEL Classification: C91, D44, D47, D53, G12, G14.

Acknowledgments:
We are indebted to the Center for Analytical Finance (CAFIN) at UCSC for seed funding of the Experimental High-Frequency Trading project of which this paper is a part, and to Dan Aisen and Ben Connault for access to data and for insight into the IEX system. We benefited from helpful comments from Peter Bossaerts, Peter Cramton, Pete Kyle, Natalia Lazzati, Kristian Lopez Vargas, John Musacchio, Axel Ockenfels, Ioanid Rosu, and Nirvikar Singh, and from participants at the 2018 Experimental Finance Workshop at the Max Plank Institute at Bonn, the 2018 Expbot Workshop in Luxembourg and Economics Department Seminars at Monash University and the University of Maryland. None of those individuals or organizations should be held responsible for the paper's remaining idiosyncrasies and shortcomings.

*Email: ealdrich@ucsc.edu.
†Email: dan@ucsc.edu.

1 Introduction

Financial firms have invested billions of dollars to speed up order placement and execution. For such high-frequency trading (HFT) firms, communication lags in major financial markets have shrunk from seconds to milliseconds in recent decades, and to tens of microseconds in recent years. With HFT now constituting a majority of transactions in major exchanges worldwide (SEC, 2014), those exchanges face competing incentives: they profit by accommodating HFT firms, yet still must retain traditional slower clients (O'Hara, 2015), many of whom feel that HFT puts them at a disadvantage. Some reform proposals intended to protect ordinary traders (e.g., by Budish et al. (2015), Du and Zhu (2017), and Kyle and Lee (2017)) would fundamentally change the market format by batching orders or by making allocations continuous functions of time.

As a practical matter, several exchanges have already responded with incremental changes to allocation rules, notably Investors Exchange (IEX), NYSE American, Thomson Reuters, Electronic Broking Services (EBS)[1], and TSX Alpha, all of which have imposed a deterministic or randomized delay applied to all inbound and outbound messages processed by the exchange. Such rule changes have provoked heated policy debates regarding acceptable exchange design, and even regarding the definition of time itself.[2]

Does HFT indeed harm ordinary traders under the traditional continuous market format? Does a messaging delay help ordinary traders, and does it have unintended consequences? The present paper contributes to the growing theoretical literature that addresses such questions. We develop an equilibrium model that spotlights the consequences of imposing a uni-

[1]The "latency floor" mechanisms of EBS and Thomas Reuters are somewhat unique among randomized delay interventions. In brief, a new order triggers a batching period of random length (e.g. one, two, or three milliseconds) during which all subsequent orders are collected and randomized by trader ID. The orders are then processed in a round-robin procedure, beginning with the first order of each trader, and subsequently the second order of each trader, and so on.

[2]For example, the September 2015 application by IEX to the SEC to become a national securities exchange provoked polarized comments regarding the appropriateness of a public exchange deliberately delaying orders. Prior to approving IEX's application in June 2016, the SEC changed a rule to define "immediacy" as 1 millisecond. This change has enormous impact on Regulation National Market System (Reg NMS) Rule 611, known as the "Order Protection Rule", which requires exchanges to immediately pass orders to markets in the national system with better prices.

form delay on new orders while a class of previously submitted hidden liquidity-providing orders ("pegged" orders) are automatically repriced without delay. Our model thus captures the essential elements of HFT-inspired reforms at IEX and NYSE American, and is closely related to those of TSX Alpha, as well as a reform proposal at the Chicago Stock Exchange (see Appendix C.1).

We build on earlier theoretical models. Easley et al. (2012) note that snipers — traders who use tiny advantages in messaging speed to pick off stale limit orders posted by market makers — force market makers to widen their bid-ask spread in equilibrium (Glosten and Milgrom (1985), Copeland and Galai (1983)), and in that sense are toxic. Budish et al. (2015) compare the ability of alternative market formats to deal with such toxicity. To sharpen comparisons, their model drastically simplifies the market by representing ordinary investors as exogenous order flows that balance in expectation around an exogenous fundamental value V. Their active participants — market makers and snipers — observe V, which jumps at random times. These active participants can choose, at a given flow cost, to subscribe to a service ("speed") that enables them to respond more quickly to jumps in V. Each market maker sets a bid and ask price symmetrically around V, choosing from a continuous range.

Our model retains many of the simplifications of Budish et al. (2015), but also requires substantial modifications. In particular, we need a discrete price grid in order to properly model the importance of timing and position in the queue of orders tied at a given price. We also need to distinguish between lit (publicly displayed) and hidden orders. A few articles have considered these modifications, albeit in a very different context. For example, Buti et al. (2017) find that opening a "dark" trading venue, with trading at the midpoint between best bid and offer, increases fill rates at a pre-existing lit market, but reduces liquidity, increases spreads and reduces welfare for all traders. Werner et al. (2015) investigate the impact of the price grid increment size in a single, lit venue.

Menkveld and Zoican (2017) analyze the impact of an exogenous increase in execution speed in a single venue, and show that it has offsetting effects on the equilibrium spread. Brolley and Cimon (2018) consider two venues, one of which delays incoming market orders but allows snipers to bypass the delay if they pay a fee. All orders are lit in their three period model, and they assume directly that the probability that a pre-existing order at the delayed venue is protected against sniping is proportional to the length of the delay that the venue imposes. Our paper is complementary in that we more explicitly model the protection technology and its consequences within a single venue, but do not consider interactions across multiple trading venues.

Our paper is also informed by the empirical literature on HFT. Such papers often distinguish between aggressive (liquidity removing) and passive (liquidity adding) HFT strategies. Passive HFT is generally associated with improved market performance; see e.g. Jovanovic and Menkveld (2015), Hagströmer et al. (2014), Menkveld and Zoican (2017), Malinova et al. (2014), and Brogaard et al. (2017). Although aggressive HFT is generally associated with informed price impact, especially over short horizons, it can increase adverse selection costs for other traders, increase short-term volatility, and raise trading costs for institutional and retail traders, as shown by Brogaard et al. (2014), Zhang and Riordan (2011), and Menkveld and Zoican (2017). The estimated net benefits of aggressive and passive HFT are often positive overall, but usually with the acknowledgment of non-negligible costs, e.g., Brogaard and Garriott (2017), Hasbrouck and Saar (2013), Bershova and Rakhlin (2012), and Breckenfelder (2013). The findings in Hirschey (2017) suggest that HFT behavior provides a net improvement to liquidity, but increases costs to non-HFT traders.

Popular accounts of financial market reforms involving a messaging delay (e.g., Lewis, 2015; Pisani, 2016) have focused on the 350 microsecond "speed bump" caused by routing communications through a 38-mile cable coiled in a "shoe box". On its own, a speed bump of this form offers no protection to slow traders, as it does not change the order in which fast and slow messages are received at an exchange. However, such a delay allows the exchange to have a timely view of the National Best Bid and Offer (NBBO) — an aggregation of competitive price quotes across all public equities exchanges — and to automatically reprice pegged orders before predatory orders arrive at the matching engine. As a result, slow traders using pegged orders are protected from fast traders who attempt to "snipe" stale orders when new information enters the market.

Pegged orders are available on all national securities exchanges in the United States. They are commonly "hidden," i.e., not shown in the publicly available limit order book. Exchanges typically charge a fee for placing and/or executing such orders and encourage the submission of visible ("lit") orders by giving priority to lit orders at any given price, even those that arrive after hidden orders. Pegged orders thus face the implicit cost of always being queued behind visible orders. This cost is non-trivial due to the fixed price grid (mandated by the Securities and Exchange Commission) used at all equities exchanges — it is not possible to "just barely" beat another trader on price, so position in the queue at a given price typically matters. Indeed, we shall see that microsecond speed advantages are valuable only because ties on price are so common on a discrete price grid.

Despite their essential role in protecting investors at exchanges featuring messaging delay, pegged orders have not been analyzed in the existing literature. Our model fills this gap by focusing on how pegged orders trade off and interact with traditional order types: market orders and limit orders. In emphasizing interactions among order types, our paper follows in the tradition of Foucault (1999) and Hoffmann (2014), who study the trade-off between market and limit orders in the presence of fast and slow traders. Our model also complements the work of Buti et al. (2015) and Buti et al. (2017) in that it studies the consequences of dark trading, but *within* an exchange that also maintains a lit order book.

Our model allows for the simultaneous expression of both passive and aggressive proprietary trading strategies. Proprietary liquidity providers (referred to as makers) and fast liquidity consumers (referred to as snipers) are in some respects similar to agents in Budish et al. (2015), Baldauf and Mollner (2016) and Menkveld and Zoican (2017). In addition to proprietary agents, we also model a population of investors who (via brokers) endogenously choose whether to enter the marketplace with pegged orders (which transact at better prices but may incur queuing delay) or with market orders (which obtain immediacy but may transact at less favorable prices). The model nests both the traditional continuous double auction market format and the uniform-delay market format as special cases.

The key to equilibrium in our model is the endogenous steady-state distribution of the pegged order queue, which we derive in closed form. The expected delay associated with pegged orders creates a queuing cost, which is a function of the pegged order queue distribution. This queuing cost in turn determines the endogenous fraction of investors who place pegged orders when messaging delays offer protection, and also the (different) endogenous fraction when protection is not offered. In either case, the equilibrium fraction of pegged orders is shifted by exogenous parameters of the model, notably by investor impatience and by the frequency of price movements (sniping opportunities) relative to the frequency of investor arrivals.

The model thus yields a wealth of predictions that can be tested against laboratory or field data. For example, in equilibrium a messaging delay that protects pegged orders will, under a wide range of exogenous parameter values, result in (a) a substantially higher proportion of pegged orders, (b) a lower sniper/maker ratio, (c) transactions prices that deviate less from fundamental value, (d) lower transactions costs, but (e) higher queuing cost. The model also identifies parametric conditions under which some of these effects are diminished or even reversed.

Section 2 sets the stage by describing order types, the traditional continuous double auction (CDA) format, and a CDA variant that delays messages to and from an exchange. It also presents summary data from IEX, the first exchange to provide pegged order protection through a fixed delay. Section 3 lists our simplifying assumptions and obtains closed-form equilibrium expressions for the usage rates of market orders vs. pegged orders and for the prevalence of trader types. It also discusses dynamic strategies for selecting order types away from the stead-state equilibrium. Section 4 obtains parallel results for CDA markets while also checking robustness by relaxing some of the more restrictive assumptions.

Section 5 develops the practical implications of our model. Using IEX data for hidden orders, it shows that the empirical distribution conforms remarkably closely to the discrete Laplace distribution predicted by our model. It calibrates other parameters to the data and introduces appropriate performance metrics. It then systematically analyzes the welfare impact of order protection. Section 6 provides a concluding discussion, including some empirically testable implications. Proofs, additional technical details, and additional institutional details can be found, respectively, in the Appendices.

2 Institutional Background

A financial market format specifies how orders are processed into transactions. In this section we provide a general description of continuous double auctions (CDA) and a more specific description of a format that protects pegged orders with a uniform messaging delay, first implemented by the Investors Exchange (IEX). We then present a summary of data from IEX and use that data to motivate elements of the model introduced in Section 3.

2.1 Continuous Double Auction format

Most modern financial markets use variants of the *continuous double auction* (CDA) format, also known as the continuous limit order book. A *limit order* is a message to the exchange comprised of four basic elements: (a) direction: buy (sometimes called a bid) or sell (sometimes called an ask or offer), (b) limit quantity (maximum number of units to buy or sell), (c) limit price (highest acceptable price for a bid, lowest acceptable price for an offer), and (d) time in force (indicating when the order should be canceled). The CDA limit order book collects and sorts bids by (1) price and (2) time received (at each price), and likewise collects and sorts offers. The highest bid price and the lowest offer price are referred to, respectively, as the *best bid* and *best offer*, and the difference between them is called the *spread*.

The CDA processes each limit order as it arrives. If the limit price locks (equals) or crosses (is beyond) the best contra-side price — e.g., if a new bid arrives with limit price equal to or higher than the current best offer — then the limit order immediately transacts ("executes" or "fills") at that best contra-side price, and the transacted quantity is removed from the order book. On the other hand, if the price is no better than the current best same-side price, then the new order is added to the order book, behind other orders at the same price.

The SEC mandates that prices displayed in equities markets order books are discrete. Specifically, Regulation National Market System (Reg NMS) Rule 612 requires the minimum price increment for nearly all equity instruments to be a penny, prohibiting displayed quotations in fractions of a penny. In contrast, time remains essentially continuous. We will see that the disjunction between discrete price and continuous time creates interesting complications for the CDA format.

At present, there are 12 SEC-approved "national securities exchanges" in the United States that trade U.S. equities instruments. Under Reg NMS, these exchanges are required to report transactions and quotations to a centralized processor, known as the Securities Information Processor (SIP). The SIP monitors all bids and offers at all 12 exchanges, and constantly updates the official National Best Bid and Offer (NBBO), consisting of the National Best Bid (NBB) and National Best Offer (NBO). However, since the speed of light is finite and the 12 exchanges have different physical locations, there is no "true" NBBO —

at best there is an NBBO from the perspective of the SIP. For this reason, unlike the order books internal to exchanges, which never lock or cross, it is possible for the NBB or NBO to temporarily lock or cross with the best bid or offer at a specific exchange. These instances are fleeting, as Reg NMS requires other exchanges with less aggressive quotations to pass orders on to those with better bids or offers.

Most CDA exchanges recognize a variety of order types beyond simple limit orders. *Market orders* are the most common variation, specifying a very high bid or very low offer price and essentially zero time in force. Most exchanges also recognize "hidden" orders which are not publicly displayed in the order book and which are given lower priority than ordinary "lit" (displayed) orders. The lexicographic priority system is: price, display, time. For example, all hidden bid orders are prioritized after the lit bids at the same price; among themselves they are prioritized on a first-come, first-served basis, even if there are different types of hidden orders.

An important type of hidden order is a *pegged* limit order. An NBB peg is a limit order that enters the book at the current NBB and is automatically re-priced by the exchange whenever the NBB changes. Similarly, an NBO peg is automatically re-priced to track the NBO. The SEC also permits exchanges to offer hidden (but not lit) *midpoint pegs*: bids or offers that track (often at half-penny prices) the midpoint of NBB and NBO.

2.2 The IEX format

The IEX market format (also implemented by NYSE American) is a CDA variant that delays all inbound and outbound messages to its messaging server by 350 microseconds. This delay is long enough to allow the system a fresh view of the NBBO and to reprice pegged orders ahead of new messages that are coincident with changes in the NBBO. As a result, pegged orders are protected from fast traders who would profit from transacting at stale prices when the NBBO changes.

Besides traditional (lit) limit and market orders, IEX (and NYSE American) offers the following types of (hidden) pegged orders:

- Midpoint peg. Limit orders pegged to the NBBO midpoint. By virtue of their more aggressive price, they have priority over traditional limit orders.

- Primary peg. Limit orders that are booked one price increment (typically $0.01) away from NBB or NBO, but which are promoted to transact at NBB or NBO if sufficient trading interest arrives at those prices.

- Discretionary peg. Limit orders which first enter the (non-displayed) order book at the midpoint but, if not executed immediately, rest at either the NBB or NBO; see Appendix C for more details.

Unlike other US exchanges, IEX charges fees only for midpoint transactions and for nothing else.

2.3 Some Data

	Other Nonroutable				Primary Peg			
	Hidden		Lit		Hidden		Lit	
	BBO	Mid	BBO	Mid	BBO	Mid	BBO	Mid
Agency Remover	3.49	0.688	3.95	0	0	0	0	0
Prop Remover	4.74	0.717	2.55	0	0	0	0	0
Agency Adder	0.781	0.998	6.91	0	2.64	0	0	0
Prop Adder	0.798	0.207	4.73	0	2.19	0	0	0
	Midpoint Peg				**Discretionary Peg**			
	Hidden		Lit		Hidden		Lit	
	BBO	Mid	BBO	Mid	BBO	Mid	BBO	Mid
Agency Remover	0	10.1	0	0	0	10.0	0	0
Prop Remover	0	7.19	0	0	0	2.33	0	0
Agency Adder	0.631	7.16	0	0	3.91	19.7	0	0
Prop Adder	0.0210	2.12	0	0	0.0951	1.36	0	0

Table 1: IEX percentage volume shares for December 2016 by order type and transaction price. Excludes routable orders and transactions in locked or crossed market conditions.

Table 1 reports transaction volume statistics at IEX during the month of December 2016[3]. The data exclude periods when markets were locked or crossed with the NBBO (3.4% of volume) and exclude transactions involving orders routable to other exchanges (12.3% of volume; see Appendix C for a discussion of routable orders). The table entries are normalized to sum to 100%, and so they are shares of the remaining 84.3% of all transactions.

IEX classifies traders into two broad types: (1) agencies (brokers), who provide services to and receive fees from external clients and who compete to offer rapid order execution at favorable prices, and (2) proprietary firms, who trade on their own account, maintaining net positions close to zero, and who earn revenue by buying at prices a bit lower on average than selling prices (either by adding liquidity at a spread or removing liquidity when stale quotes persist in the order book). Firms that do both are classified as agencies.

Table 1 shows that Agency firms represent over 70% of volume at IEX; volume at other exchanges is typically more evenly split between agencies and proprietary traders. Agency volume has three main components:

1. Adding orders at BBO: 7.7% of transaction volume. Our model in the next section will attribute this to the proprietary arm of integrated agency firms.
2. Removing orders at BBO: 7.4% of volume. Our model will attribute this to impatient investor clients.
3. Midpoint and discretionary peg orders transacting at midpoint: 47.0% of volume. Our model will attribute this to less impatient clients.

Following is a similar breakdown for proprietary firms; again see Appendix C for more details.

1. Adding orders at BBO: 5.5% of volume. Our model attributes this to market making by proprietary firms.
2. Removing orders at BBO: 7.3% of volume. The model attributes this to proprietary "snipers," who exploit unprotected stale limit orders when the NBBO changes.
3. Midpoint and discretionary peg orders transacting at midpoint: 13.0% of volume. For simplicity, and since they comprise only 25% of all midpoint and discretionary orders,

[3]The IEX data is proprietary and was made available by direct request.

our model will group this order flow with midpoint orders transmitted by agencies on behalf of their clients.

3 Baseline Model

Our baseline model is a continuous double auction that protects midpoint pegs. The model highlights tradeoffs between order types under simplifying assumptions on the grid of asset prices and on exogenous variables specifying investor arrival and changes in the asset's fundamental value. This baseline model also makes stark assumptions regarding who buys speed and which orders are protected from fast traders; the extended model in Section 4 will eliminate protection and will examine speed purchase decisions.

3.1 Assumptions

A1. The market consists of a single asset trading at a single exchange, one indivisible unit at a time.

A2. Prices lie on a discrete, uniform grid $\mathcal{P} = 1, 2, \ldots, \hat{P}$. A price unit, i.e., the grid step size, represents half of the minimum price increment (e.g. a half penny per share).

A3. The fundamental value of the asset, V, follows a marked Poisson process on $\mathcal{P}$. The fundamental value changes to $V' \in \{V-2, V+2\}$ with equal innovation rate $\nu > 0$. That is, the total innovation rate is 2ν, with one-sided rate ν of a two-increment (e.g., one penny) upward jump and one-sided rate ν of a two-increment downward jump.

A4. An exogenous flow of impatient investors with unit demands arrive independently at Poisson rate $\rho > 0$ on each side of the market.

a. Investors have gross surplus $\varphi > 1$ per unit of the asset.

b. An investor may have the broker transmit a market order. If there is a contra-side midpoint order resting in the (hidden) order book, then the market order executes immediately at midpoint and incurs execution fee $d \in (0, 1)$. Otherwise, the market order executes immediately at the BBO and incurs trading cost of 1 (e.g. a half penny per share).

c. Alternatively, an investor may have the broker transmit a midpoint peg order. If there is a contra-side midpoint order, then the transmitted order executes immediately at midpoint and incurs execution fee $d \in (0, 1)$. Otherwise the transmitted order goes to the end of the same-side (hidden) order queue. If the transmitted order is not executed immediately, its net surplus is discounted at rate $\delta > 0$.

A5. The cost of speed, $c > 0$, and time lags in responding to innovations in V are such that

a. Traders placing orders at BBO do not purchase speed and, when V jumps, are susceptible to sniping by other proprietary traders who do purchase speed.

b. Snipers can reverse transactions immediately at V'.

c. Pegged orders track V with so short a lag that they are protected from sniping.

Assumption A1 is a straightforward simplification to sharpen the analysis, and is used in many of the theoretical models mentioned in the introduction. As noted, a few of those models also assume a discrete price grid as in our A2, and all major exchanges currently impose such a grid. A2 is important for our purposes because (i) it is not clear how to define pegs when (possibly tiny) orders can be priced arbitrarily close together, and (ii) messaging speed is far less important when ties on price are rare, as might be the case with a continuous price space.

Assumptions A3 and A4 are unrealistic[4] but together they say that the fundamental value V equilibrates supply and demand even while experiencing exogenous shifts. We think of V as representing the NBBO midpoint, which is observed on other exchanges, but taking it as exogenous sharpens the comparison of market formats on a single exchange. As noted earlier, Budish et al. (2015) and other previous papers assume that V is publicly observable as in A3, and several previous papers (e.g. Foucault, 1999; Hoffmann, 2014) also assume that V is equally likely to jump up or down a fixed amount.

Assumption A4a is a standard way to capture the idiosyncratic component of investor valuation and the potential gains from trade in a financial market; e.g., $\varphi = L$ in the notation

[4]Indeed, at extreme prices ($V = 1, 2$ and $\hat{P} - 1, \hat{P}$), some jumps are infeasible so A3 must be modified. As a practical matter, the SEC permits the grid to be redefined in such extreme cases. Here, to keep the focus on matters of greater interest, we assume that such modifications are negligible because we are operating far away from the extremes.

of Hoffmann (2014, p.158). Assumption A4b implicitly assumes a deep order book at BBO, and explicitly assumes that $d < 1$, as seems reasonable; Appendix C.2 notes that $d = 0.18$ at IEX. Assumption A4c is intended to capture investors' impatience in a simple way. Note that it conflates discretionary pegs with midpoint pegs and ignores primary pegs; see Appendix C for a justification of this simplification.

Speed in contemporary financial markets is typically acquired via a subscription fee; Assumption A5 represents this cost as an amortized flow to facilitate comparisons to expected flow benefits and costs of transacting. To streamline the analysis of trading profits we impose A5b, which follows contemporary inventory valuation practice as well as models such as that of Budish et al. (2015). Likewise A5a and A5c succinctly summarize the impact of timing parameters; Section 4 below will analyze timing issues in more detail.

3.2 Action Space and State Space

In order to focus on the tradeoffs between pegged orders and the orders featured in previous financial market models, we use a streamlined set of just three order types:

r: single unit $r$egular lit limit orders added at the best bid $(V-1)$ and best offer $(V+1)$, which we attribute to proprietary traders.

p: single unit midpoint $p$eg limit orders added at price V, attributed to brokers (agencies). These orders are hidden and are subject to fee $d \in [0, 1)$ upon execution.

m: single unit $m$arket orders, also attributed to brokers, that remove liquidity at the midpoint if it is occupied by contra-side orders, in which case they also incur the fee d, and otherwise remove r orders at the best bid or best offer.

The attribution to different sorts of market participants (r to proprietary traders, and p and m to brokers) is a convention aligned with exchange accounting practice, but not a restriction on participants' behavior, since any participant can assume either or both roles.

As noted, we assume that the equilibrium lit order book is deep and so we need not track its transitory changes. We do, however, need to model transitions in the hidden order book at midpoint. Since buy and sell p orders execute against each other (by Assumptions A4b-c, and in reality), midpoint orders can rest on only one side of the market at any given time.

Therefore the state of the market is described by the level of the fundamental value, $V \in \mathcal{P}$, together with the *order imbalance* $k \in \mathbb{Z}$ at the midpoint price. By convention, $k < 0$ means that there are precisely $-k > 0$ midpoint peg buy orders resting in the hidden order book and no sell orders, $k > 0$ indicates k midpoint peg sell orders in the hidden order book and no buy orders, and $k = 0$ indicates an empty queue at the midpoint price V. See Figure 1.

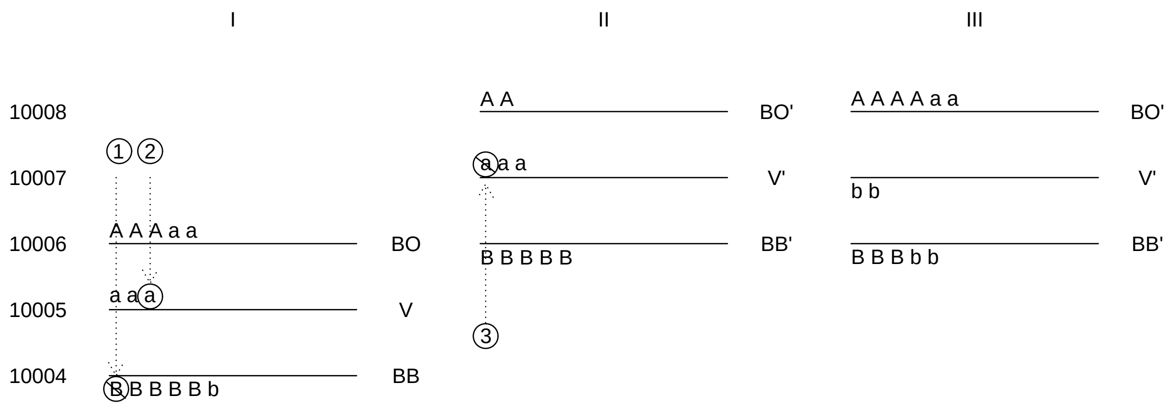

Figure 1: Example states in IEX market. Uppercase (resp. lowercase) denotes lit (resp. hidden) orders, B/b for buy and A/a for sell; those to the left have higher priority at that price. Panel I: initial state is $k = 2$ and $V = 10005$ half-pennies (i.e., $V = \$50.025$ per share); event (1) is a market sell order which 'crosses the spread' to transact at BB $10004 = \$50.02$; event (2) is a midpoint offer which rests at V implying a transition to $k = 3$. Panel II: V has jumped to $10007 = \$50.035$; event (3) is a market buy order or midpoint bid which transacts at V and triggers transition $k = 3 \to 2$. Panel III: V remains at 10007 but an excess of bids relative to offers has driven k to -2.

New investor arrivals can trigger transitions in k. Let ω denote the fraction of arrivals that brokers transmit as midpoint peg orders, with the remaining fraction, $1-\omega$, transmitted as market orders. Given the symmetry in Assumption A4, a new arrival generates a midpoint peg buy or sell order with probability $\omega/2$ each, or a market buy or sell order with probability $(1-\omega)/2$ each. A new pegged sell (resp. buy) order always generates a transition $k \to k+1$ (resp. $k \to k-1$). A new market sell (resp. buy) order generates a transition $k \to k+1$ when $k < 0$ (resp. $k \to k-1$ when $k > 0$) and otherwise executes at BBO and generates no transition. The following proposition, proved in Appendix A, characterizes the stationary distribution of k that emerges from those transitions.

Proposition 3.1. *Let $\omega \in (0,1)$ be the probability that an investor arrival on either side of the market results in a midpoint peg order. Given Assumptions A1-A5, there is a unique steady state distribution $\mathbf{q} = (q_k)_{k \in Z}$ of the order imbalance, with*

$$q_k = \left(\frac{1-\omega}{1+\omega}\right) \omega^{|k|}, \quad k \in \mathbb{Z}. \tag{3.1}$$

Equation (3.1) tells us that the steady state distribution is discrete Laplace, a distribution discussed in Inusah and Kozubowski (2006): symmetric and unimodal at 0 and decreasing at exponential rate as $|k|$ increases. The distribution has a single parameter, the fraction ω of investor orders that are transmitted as midpoint peg orders rather than as market orders. That fraction is determined endogenously in equilibrium, as we now will see.

3.3 Investor Surplus

Following assumption A4, investors choose between midpoint peg orders and market orders.

Peg. Upon execution, a midpoint peg order generates surplus φ less the execution fee d. These orders transact immediately with contra-side midpoint peg orders if any are present, and otherwise are placed at the back of the midpoint queue (e.g., at position $k+1$ when the state is $k \geq 0$), and thus incur queuing costs expressed in terms of the given discount rate δ. The relevant discount factor is $\beta = \exp\left(-\frac{\delta}{\rho}\right) \leq 1$ when contra-side orders arrive at rate $\rho > 0$. Thus, by A4, the conditional expected net surplus is $(\varphi - d)\beta^{k+1}$ for a pegged sell order when $k \geq 0$.

Using steady state probabilities (3.1) for order imbalance k, the unconditional expected net surplus for a midpoint peg sell order is

$$\begin{aligned}
\pi_p &= (\varphi - d)\left[\sum_{k=-\infty}^{-1} q_k + \sum_{k=0}^{\infty} q_k \beta^{k+1}\right] \\
&= (\varphi - d)\left[\frac{1-\omega}{1+\omega}\right]\left[\frac{\omega}{1-\omega} + \frac{\beta}{1-\beta\omega}\right] \\
&= \left(\frac{\varphi - d}{1+\omega}\right)\left[\omega + \frac{\beta(1-\omega)}{1-\beta\omega}\right]
\end{aligned} \tag{3.2}$$

The model's symmetry ensures that (3.2) also applies to pegged buy orders.

Market order. Like a pegged order, with probability $\sum_{k=-\infty}^{-1} q_k$ a market order will execute immediately against a contra-side midpoint peg and earn $\varphi - d$. With probability $\sum_{k=0}^{\infty} q_k$, there will be no contra-side midpoint orders and, unlike a pegged order, a market order will then execute immediately against an r order at BBO. In that case, since the price is 1 half-spread away from V, it earns $\varphi - 1$. Thus the expected net surplus for a market order (either buy or sell) is

$$\pi_m = (\varphi - d) \sum_{k=-\infty}^{-1} q_k + (\varphi - 1) \sum_{k=0}^{\infty} q_k = \left(\frac{\varphi - d}{1 + \omega} \right) \omega + \frac{\varphi - 1}{1 + \omega}. \tag{3.3}$$

3.4 Proprietary Trader Profits

We model makers and snipers as disjoint subsets of proprietary traders. Since our model does not fully pin down market scale, and since integer constraints in this context seem unhelpful, we treat the numbers N_r and N_s as real numbers (population masses) rather than as integers (population counts). Our focus is on their relative, not absolute, magnitudes.

Snipers trade off the flow cost, c, of buying speed against profits from sniping stale r orders following a jump in the fundamental value V. When an opportunity arises, each sniper uses fast market orders to obtain on average $\frac{N_r}{N_s}$ successful snipes. By assumptions A3 and A5, each successful snipe of a resting r order involves buying (or selling) a single share at $V + 1$ (or $V - 1$) and reversing the transaction at $V' = V + 2$ (or $V' = V - 2$), yielding a profit of 1 half-spread. Since opportunities arrive at both sides of the market at rate ν, the expected flow profit for a sniper is

$$\pi_s = 2\nu \frac{N_r}{N_s} - c. \tag{3.4}$$

Market makers place r orders at the BBO, trading off the single half-spread gain of transacting with a market order against a possible half-spread loss to a sniper. Investor market orders arrive at each side of the market at rate $(1 - \omega)\rho$, and with probability $\sum_{k=0}^{\infty} q_k = \frac{1}{1+\omega}$ encounter no contra-side liquidity at the midpoint. With jumps occurring on each side of the market at rate ν, the expected flow profit to a market maker who maintains both a bid and an offer at BBO therefore is

$$\pi_r = 2 \frac{(1 - \omega)\rho}{(1 + \omega) N_r} - 2\nu, \tag{3.5}$$

given mass N_r of makers.

3.5 Equilibrium

Definition 3.1. *Given an exogenous flow cost of speed $c > 0$, midpoint execution fee $d \geq 0$, investor gross surplus $\varphi \geq 1$, discount factor $\beta = \exp\left(-\frac{\delta}{\rho}\right) \in (0,1)$, and arrival rates $\rho, \nu > 0$ for investors and fundamental value innovations, the vector (ω^*, N_r^*, N_s^*) constitutes a* market equilibrium *if*

1. *at $\omega^* \in (0,1)$ (resp. $\omega^* = 0$), a midpoint peg order has the same (resp. no more than the) expected net surplus as a market order, and*

2. *with $N_s^* \geq 0$ snipers and $N_r^* \geq 0$ market makers, proprietary traders earn zero expected profit from either activity.*

The idea behind the first equilibrium condition is that investors and brokers will increase the fraction $\omega \in (0,1)$ of pegged orders whenever the expected surplus differential $\pi_p - \pi_m$ is positive, and decrease ω when the differential is negative. Hence expected (net discounted) surplus should be equal at an interior steady state, while at $\omega^* = 0$ we should have $\pi_p \leq \pi_m$. Later we will see that $\omega^* = 1$ is not consistent with impatient investors. The second equilibrium condition arises from the reasonable assumption that there are no substantial barriers to entry or exit for either of the two proprietary activities.

Under current assumptions, equilibrium takes a simple form.

Proposition 3.2. *Under assumptions A1 - A5 and parameter restrictions $\varphi \geq 1 > d \geq 0$, $\beta = \exp\left(-\frac{\delta}{\rho}\right) \in (0,1)$, and $c, \rho, \nu > 0$, there is a unique market equilibrium (ω^*, N_r^*, N_s^*). The equilibrium fraction of brokers/investors choosing midpoint pegs vs. market orders is*

$$\omega^* = \max\left\{0, \frac{\varphi - d - \beta^{-1}(\varphi - 1)}{1 - d}\right\}, \tag{3.6}$$

and the equilibrium masses of proprietary traders choosing to act as market makers and snipers are, respectively,

$$N_r^* = \frac{\rho}{\nu}\left(\frac{1-\omega^*}{1+\omega^*}\right) \tag{3.7}$$

$$N_s^* = \frac{2\rho}{c}\left(\frac{1-\omega^*}{1+\omega^*}\right). \tag{3.8}$$

Proof. Applying the first market equilibrium condition we obtain equation (3.6) as follows:

$$\begin{aligned}
\pi_p = \pi_m \quad &\iff \quad \left(\frac{\varphi - d}{1+\omega}\right)\left[\frac{\beta(1-\omega)}{1-\beta\omega}\right] = \frac{\varphi - 1}{1+\omega} \\
&\iff \quad (\varphi - d)\beta(1-\omega) = (\varphi - 1)(1-\beta\omega) \\
&\iff \quad \omega = \frac{\varphi - d - \beta^{-1}(\varphi - 1)}{1-d}.
\end{aligned} \tag{3.9}$$

If the last expression in (3.9) is negative, then it is straightforward to show that $\pi_p(0) \leq \pi_m(0)$ and so $\omega^* = 0$. Note that $\beta < 1$ and the other parameter restrictions ensure that $\omega^* < 1$ in (3.9). To obtain Equations (3.7) and (3.8), apply the second market equilibrium condition $\pi_r = \pi_s = 0$ to Equations (3.5) and (3.4) and solve for N_r and N_s. □

Equation (3.6) shows that the equilibrium fraction of midpoint peg orders, ω^*, is not sensitive to sniping risk, as captured by the innovation rate ν. This is a direct result of order protection offered through the message delay. However, ω^* is negatively related to gross surplus φ, and is positively related to the discount factor, since greater patience naturally lowers the effective queuing cost of a midpoint peg.

Not surprisingly, Equation (3.7), shows that the equilibrium mass of market makers is positively related to the arrival rate of investor orders, ρ, but negatively related to the fraction of those orders that are placed as pegs, ω^*, and to the risk of sniping, ν. The equilibrium mass of snipers exhibits a similar relationship to the investor arrival rate and the fraction of midpoint pegs, but, perhaps surprisingly, it is unrelated to the number of sniping opportunities. This is due to the offsetting equilibrium decrease in N_r^* as ν increases. Finally, as expected, the mass of snipers is inversely related to the cost of fast communication technology.

3.6 Dynamic strategies

The foregoing equilibrium analysis assumes that investors/brokers respond optimally to the steady-state distribution of the midpoint queue, and not to particular realizations of the queue state k. Since these queued orders are hidden, that assumption seems reasonable at first. But might investors use active or passive order submission strategies to glean useful information about the current realization of k? Consider the following active strategy: an investor always initiates a bid (offer) submission with a peg, followed by a cancellation if the order does not immediately fill at the midpoint, thus ascertaining that $k \leq 0$ ($k \geq 0$) and that expected queuing costs of pegged orders are higher than in steady state. The investor would then follow the peg cancellation with a market order. However, closer inspection of this "ping first" strategy reveals that it is dominated by initiating the submission with a market order. Like the ping (the initial pegged order), the direct market bid (offer) will immediately transact at midpoint to earn $\varphi - d$ if $k > 0$ ($k < 0$). Otherwise, it will earn $\varphi - 1$, whereas the "ping first" strategy will earn $\beta^{\epsilon}(\varphi - 1) < \varphi - 1$, where $\epsilon > 0$ reflects the messaging delays in cancelling and replacing the initial ping.

Alternative strategies arise from observing the publicly displayed transaction stream. Transactions at midpoint might reflect removal of either a pegged bid or a pegged ask, and therefore are not informative regarding the current state k. However, observing a transaction at BBO is informative. A transaction at BB, for instance, reveals that there is no midpoint bid to remove, and thus that $k \geq 0$. Intuition suggests that in this case (a) investors who arrive on the buy side will find it advantageous to place a pegged bid due to reduced expected queuing cost, and (b) investors who arrive on the sell side will find it advantageous to place a market offer due to increased expected queuing cost on the sell side.

Appendix A.4 develops analytical machinery to investigate both intuitive conjectures. Taking the discrete Laplace distribution in Proposition 3.1 as the Bayesian prior, it identifies a class of distributions that closely approximate the Bayesian posterior distributions following an observation of a BBO transaction. It also derives an explicit formula for how the state distribution $\mathbf{p}(t)$ evolves from an arbitrary initial distribution $\mathbf{p^o}$. One consequence is that strategy (b) doesn't work. Placing a market order when the midpoint queue is known to be heavier than usual does not increase (or decrease) profit in market equilibrium. The mathematically informed intuition is that the posterior expected queuing cost is the same in case (b) as it is when the investor's own order fails to execute at midpoint, and that posterior expectation is compatible with market equilibrium as in Proposition 3.2.

Appendix A.4 does, however, partially vindicate the intuition behind strategy (a). Indeed, if the investor were able to react immediately to the empty queue, strategy (a) would increase profit above that earned in equilibrium (as given in Equation (3.2) or (3.3)) by the proportion $\frac{\beta\omega(1-\beta)(1-\omega)(1+\omega)}{\omega(1-\beta\omega)+\beta(1-\omega)}$. However, for reasonable parameters, this maximal proportional advantage is small. Our best numerical estimate is that the expected advantage (taking into account the lag until investors arrive, and relaxation towards the steady state distribution) is less than 1%. Moreover, available data shows no evidence of such dynamic strategies; see Appendix B. The next two sections, therefore, will focus on steady-state equilibrium.

4 Extension: Unprotected Midpoint Pegs: $\xi = 1$

The equilibrium in Section 3 was derived under the assumptions that pegged orders are protected from sniping, that there are always r orders resting at the BBO, and that market makers do not purchase speed technology. We now explore what happens when some of those assumptions are relaxed.

Timing notation. Jumps in the fundamental value V are registered at the Securities Information Processor (SIP) and, with latency $\tau_{SIP} > 0$, resting pegged orders automatically adjust in parallel fashion. Traders' messages to the exchange have default round-trip latency τ_{slow}, but at flow cost $c > 0$, traders can reduce their latency to $\tau_{fast} \in (0, \tau_{slow})$. The exchange imposes an additional uniform delay $\eta \geq 0$ so that traders experience overall latencies $\tilde{\tau}_{fast} = \tau_{fast} + \eta$ and $\tilde{\tau}_{slow} = \tau_{slow} + \eta$. To avoid trivialities, we assume that $\tilde{\tau}_{slow} = \tau_{slow} + \eta > \tau_{SIP}$. These parameter values are known by all participants.

In the traditional CDA format $\eta = 0$ while, at exchanges like IEX, $\eta > 0$ is chosen so that $\tilde{\tau}_{fast} = \tau_{fast} + \eta > \tau_{SIP}$. To compress notation, define the binary composite parameter

$$\xi = \begin{cases} 1 & \text{if } \tilde{\tau}_{fast} = \tau_{fast} + \eta \leq \tau_{SIP} \\ 0 & \text{if } \tilde{\tau}_{fast} = \tau_{fast} + \eta > \tau_{SIP}. \end{cases} \tag{4.1}$$

When $\xi = 0$, pegged orders are fully protected from sniping and jump in tandem with V before any other messages reach the exchange, as in assumption A5c. When $\xi = 1$, pegged orders are not protected. We assume that snipers have the capacity to snipe them all; see Appendix A.5 for further discussion.

The steady state distribution of order imbalance is different than in the protected case, because sniping now induces direct transitions $k \to 0$. The following generalizes Proposition 3.1 to cover the unprotected case.

Proposition 4.1. *Let $\omega \in (0,1)$ be the probability that an investor arrival on either side of the market results in a midpoint peg order. Given assumptions A1-A5b, there is a unique steady state distribution $\mathbf{q} = (q_k)_{k\in Z}$ of the order imbalance, with*

$$q_k = \left(\frac{1-\lambda}{1+\lambda}\right)\lambda^{|k|}, \quad k \in \mathbb{Z}, \tag{4.2}$$

where

$$\lambda = \frac{1}{2}\left(1+\frac{\xi\nu}{\rho}+\omega\right) - \frac{1}{2}\sqrt{\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2 - 4\omega} \qquad \in (0,\omega], \tag{4.3}$$

and the variable $\xi = 0$ (resp. $\xi = 1$) indicates that pegged orders are protected from sniping (resp. are not protected).

See A.6 for a proof, which uses queuing theory and difference equations to derive (not just verify) the formulas. Note that in the protected case $\xi = 0$, the expression (4.3) collapses to $\lambda = \omega$ and (4.2) reduces to (3.1).

We proceed as before to obtain the equilibrium values of λ and ω. To facilitate comparisons, we continue to assume that liquidity adders do not use speed technology; see Appendix A.4 for justification. Thus, resting p orders are vulnerable when V jumps and $\xi = 1$. A successful sniper gains (and the liquidity adder loses) $|V'-V| = 2$ half-spreads on a midpoint peg, rather than the usual 1 half-spread on an order resting at BBO.

When a midpoint peg offer is queued behind k other pegged offers, it will be sniped if and only if a positive jump in V occurs before $k+1$ buy orders arrive from brokers. Thus, the conditional probability of not being sniped is $\left(\frac{\rho}{\rho+\xi\nu}\right)^{k+1}$, with expected profit $(\varphi-d)\beta^{k+1}$, where the discount factor is still? $\beta = \exp\left(-\frac{\delta}{\rho}\right)$. With complementary probability, the offer is sniped, resulting in a 2 half-spread loss discounted in the same manner. By symmetry, the same probabilities apply to midpoint peg bids.

Thus the expected net surplus for a midpoint peg order generalizes to

$$\begin{aligned}\pi_p &= (\varphi-d)\left[\sum_{k=-\infty}^{-1} q_k + \sum_{k=0}^{\infty} q_k\left(\frac{\beta\rho}{\rho+\xi\nu}\right)^{k+1}\right] - 2\sum_{k=0}^{\infty} q_k\beta^{k+1}\left[1-\left(\frac{\rho}{\rho+\xi\nu}\right)^{k+1}\right] \\ &= (\varphi-d)\left[\frac{\lambda}{1+\lambda} + \frac{1-\lambda}{1+\lambda}\frac{\beta\rho}{(\rho+\xi\nu-\beta\rho\lambda)}\right] \\ &\qquad - \left[\frac{1-\lambda}{1+\lambda}\right]\frac{2\beta}{(1-\beta\lambda)} + \left[\frac{1-\lambda}{1+\lambda}\right]\frac{2\beta\rho}{(\rho+\xi\nu-\beta\rho\lambda)}.\end{aligned} \tag{4.4}$$

An investor choosing a market order will earn the same expected net surplus as in Equation (3.3), generalizing from ω to λ:

$$\pi_m = (\varphi - d)\sum_{k=-\infty}^{-1} q_k + (\varphi - 1)\sum_{k=0}^{\infty} q_k = (\varphi - d)\frac{\lambda}{1+\lambda} + (\varphi - 1)\frac{1}{1+\lambda}. \tag{4.5}$$

As in the protected case, the first terms in (4.4) and (4.5) represent execution against a contraside midpoint peg, and these identical terms cancel in the equal surplus condition.

Snipers now have $N_r + \xi N_p$ potential targets: the regular orders plus unprotected pegged orders. Since the profit is 2 half-spreads on the latter, Equation (3.4) becomes

$$\pi_s = 2\nu\frac{N_r + 2\xi N_p}{N_s} - c. \tag{4.6}$$

Conditions for market makers are unchanged, so (3.5) still characterizes their profitability.

Proposition 4.2. *Under assumptions A1 - A5b and parameter restrictions* $\varphi \geq 1 > d \geq 0$, $\beta = \exp\left(-\frac{\delta}{\rho}\right) \in (0,1)$, *and* $c, \rho, \nu > 0$, *there is a unique market equilibrium, with*

$$\tilde{\omega}^* = \tilde{\lambda} + \xi\left(\frac{\tilde{\lambda}}{1-\tilde{\lambda}}\right)\frac{\nu}{\rho} \tag{4.7a}$$

$$N_r^* = \frac{\rho}{\nu}\left(\frac{1-\tilde{\omega}^*}{1+\tilde{\lambda}}\right) \quad \textit{and} \tag{4.7b}$$

$$N_s^* = 2\nu\frac{N_r^* + 2\xi N_p^*}{c} = \frac{2\rho}{c}\left(\frac{1-\tilde{\omega}^*}{1+\tilde{\lambda}}\right) + \frac{4\xi\nu}{c}\frac{\tilde{\lambda}}{1-\tilde{\lambda}^2} \tag{4.7c}$$

where

$$\tilde{\lambda} = \frac{1}{2\beta^2\rho(1-d)}\Bigg[\beta^2\rho(\varphi - d) - \beta\xi\nu(\varphi+1) - \beta\rho(\varphi+d-2) - \Big(\big(\beta\xi\nu(\varphi+1) + \beta\rho(\varphi+d-2) - \beta^2\rho(\varphi-d)\big)^2 - 4\beta^2\rho(1-d)\left(\beta\rho(\varphi-d) - \rho(\varphi-1) - \xi\nu(\varphi-1+2\beta)\right)\Big)^{1/2}\Bigg]_+ \tag{4.8}$$

$$\textit{and } N_p^* = \frac{\tilde{\lambda}}{1-\tilde{\lambda}^2}. \tag{4.9}$$

Proposition 4.2 is proved in a manner similar to its special case, Proposition 3.2; see Appendix A.6. The expressions are more complicated mainly because the equal profit condition here yields a quadratic expression in λ rather than a linear expression in ω.

Corollary 4.1. *In the limiting case $\delta/\rho \to 0$ (or $\beta \to 1$), the steady-state value of λ is*

$$\hat{\lambda} = 1 - \xi \frac{(\varphi + 1)}{(1 - d)} \frac{\nu}{\rho}, \tag{4.10}$$

which is valid for $\rho \geq \frac{\varphi+1}{(1-d)}\nu$.

Proof. When $\beta = 1$, the equal profit condition $\pi_p = \pi_m$ simplifies to

$$(\varphi - d)(1 - \lambda)\rho + 2(1 - \lambda)\rho = (\varphi + 1)(\rho + \xi\nu - \rho\lambda) \tag{4.11}$$

from which (4.10) follows. □

5 Practical implications

What do our analytical results tell us about the advantages and disadvantages of order protection via delayed messaging? To answer that question, we validate and calibrate the model with available data. We then derive relevant performance metrics and use them to compare institutions as we vary parameters, especially the protection parameter, ξ.

5.1 Validation and calibration

Using proprietary data for the distribution of pegged orders provided by IEX, we estimate the equilibrium parameter, ω. The data consist of total time (measured in nanoseconds) during December 2016 market hours that the midpoint peg and discretionary peg order queues spent in each possible state, k. For the purpose of fitting the discrete Laplace distribution in Equation (3.1), we combine the discretionary and midpoint peg state times, since 89% of discretionary peg transactions occur at midpoint (see Table 1 and Appendix C.3), and confine attention to state values $|k| \leq 6$, which account for 99.97% of the total trading time (4.912751×10^{14} nanoseconds) during December 2016. The (log) empirical frequencies, w_k, are represented by the connected black points in Figure 2; for $|k| \geq 7$ the fractions are noisy and minuscule.

To estimate the discrete Laplace parameter, ω, we use the maximum likelihood estimator of Inusah and Kozubowski (2006),

$$\hat{\omega} = \frac{\bar{K}}{1 + \sqrt{1 + \bar{K}^2}}, \tag{5.1}$$

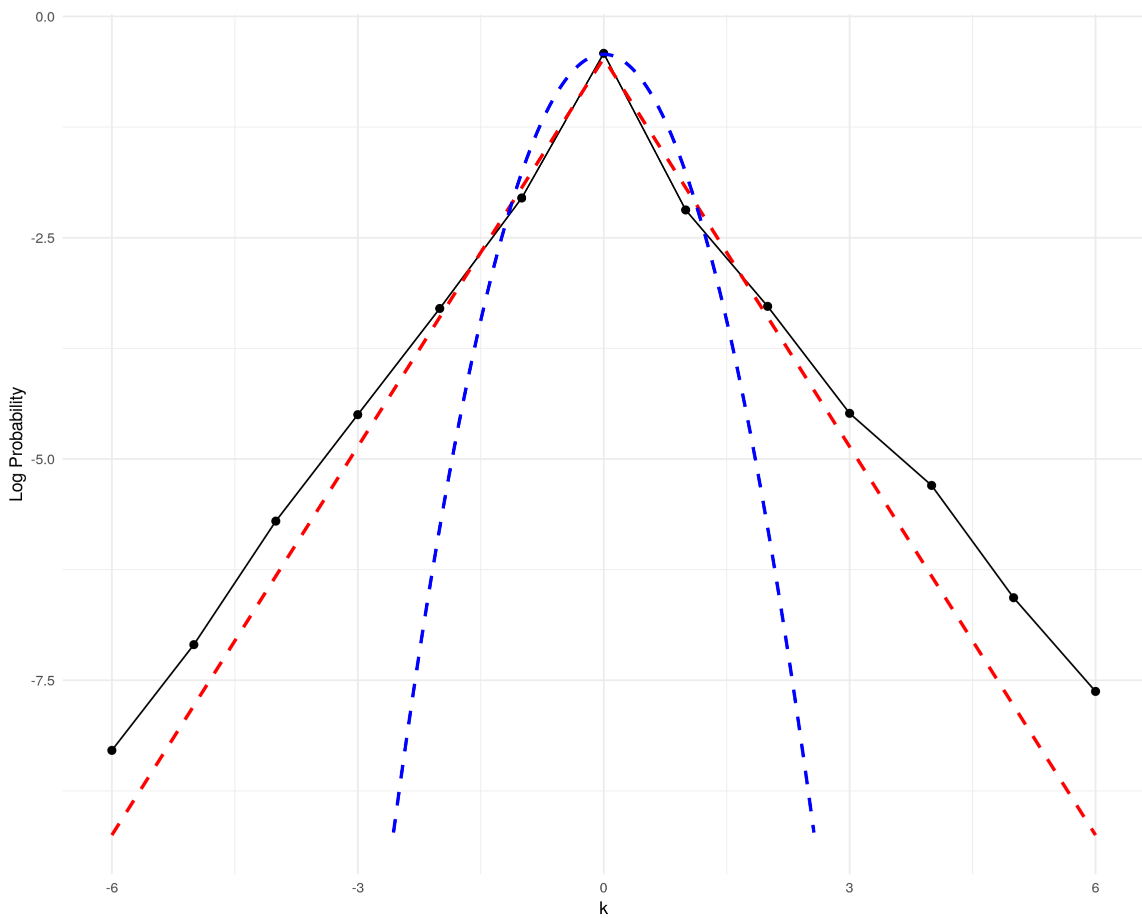

Figure 2: Pegged order queue (log) distributions: empirical (connected black dots), estimated discrete Laplace (red dashed lines), and estimated Normal (blue dashed curve).

where $\bar{K} = \sum_{k \in \mathcal{K}} w_k |k|$ and where $\mathcal{K} = \{-6, \ldots, 0, \ldots, 6\}$. We find $\hat{\omega} = 0.23$ and depict the corresponding discrete Laplace distribution in Figure 2 as a dashed red line. The figure also shows the Normal distribution obtained by minimizing w_k-weighted squared errors (dashed blue curve). It is apparent that, consistent with our theoretical model, the discrete Laplace distribution is a far better description of the data than a Normal distribution. We emphasize that our model was not designed to fit the empirical distribution of pegged orders; rather, the distribution of Equation (3.1) is a byproduct of our model primitives and equilibrium. However, it appears to be an excellent characterization of unconditional pegged order submission behavior in the IEX market.

The estimated value of ω has implications for the underlying parameters of the model. An informal analysis of available financial market data, summarized in Appendix D, leads us to baseline values $(c, d, \varphi, \beta, \rho, \nu) = (10, 0.18, 1.8, 0.56, 50, 1.0)$.

5.2 Performance metrics

We take the investor's perspective in assessing market performance since the other participants in our model earn zero profit in equilibrium and, more fundamentally, one can argue that the social value of financial markets lies in serving investors, not in extracting revenue from them. We therefore focus on the following two performance metrics.

Transaction cost. Investors pay brokerage fee b, which is typically 0.6 to 1.0 half spreads (\$0.003 – \$0.005); thus, our baseline value is 0.8. With probability $P = \sum_{k=1}^{\infty} \tilde{q}_k = \frac{\lambda}{1+\lambda}$, a market order executes immediately at midpoint and is charged an additional explicit fee of d, while with probability $1 - P$ it executes at BBO and pays an additional implicit fee of one half spread via worse execution price. Thus, for a market order, the per-share mean transaction cost is

$$TC = b + d \cdot P + 1 \cdot (1 - P) = b + \frac{1 + d\lambda}{1 + \lambda}. \tag{5.2}$$

In market equilibrium, TC will be the same for either type of order when $\omega > 0$, so equation (5.2) also applies to pegged orders.

Queuing cost. The expected fractional loss of surplus due to discounting is zero except for orders transmitted as midpoint pegs that go to the back of the queue. By the logic of the previous section, conditional on same-side imbalance $k \geq 0$, the expected discount factor is $\left(\frac{\beta\rho}{\rho+\xi\nu}\right)^{k+1}$, implying a proportional loss $\left[1 - \left(\frac{\beta\rho}{\rho+\xi\nu}\right)^{k+1}\right]$ of net surplus. We define QC as the unconditional expected proportional loss,

$$\begin{aligned} QC &= \omega \sum_{k=0}^{\infty} \tilde{q}_k \left[1 - \left(\frac{\beta\rho}{\rho + \xi\nu}\right)^{k+1}\right] \\ &= \frac{\omega}{1+\lambda} - \frac{\omega\beta\rho}{\rho+\xi\nu}\left(\frac{1-\lambda}{1+\lambda}\right)\sum_{k=0}^{\infty}\left(\frac{\beta\rho\lambda}{\rho+\xi\nu}\right)^k \\ &= \frac{\omega}{1+\lambda}\left(\frac{\rho+\xi\nu-\beta\rho}{\rho+\xi\nu-\beta\rho\lambda}\right). \end{aligned} \tag{5.3}$$

We see no natural way to combine the two metrics into an overall welfare measure; their relative importance would differ across market participants and among policy analysts.

Therefore, we retain both metrics as performance measures, noting that QC represents a deadweight loss while TC is a transfer of surplus from investors to proprietary traders (who ultimately fully dissipate it via snipers' speed purchases).

5.3 Impact of order protection

What impact do model parameters have on equilibrium and performance? We focus here on the parameters ν (controlling the frequency of jumps in the fundamental value) and β (patience of investors), and track their effects on the equilibrium peg fraction ω^*, the sniper ratio $\frac{N_s^*}{N_r^*}$, and the two performance metrics.

Figure 3 depicts those equilibrium ratios and performance metrics as we vary the fundamental jump arrival rate $\nu \in (0, 20)$, with investor arrival rate ρ held constant at its baseline value of 50. Panel (a) shows that with protection, the equilibrium share of pegged orders, ω^*, is independent of the jump rate ν; the horizontal dashed red line remains constant at its baseline value of 0.23 and the dashed blue line is constant at a substantially higher value ($\omega^* = 0.892$) when investors are more patient. The solid lines show that when protection is removed, $\xi = 1$, midpoint pegs disappear for $\nu \gtrsim 2.5$ in the baseline, and for $\nu \gtrsim 12.5$ when investors are more patient; in those regions, the high probability of sniping renders midpoint pegs unprofitable. Panel (b) displays the sniper ratio $\frac{N_s^*}{N_r^*}$ and shows that for $\xi = 0$ and for large values of ν when $\xi = 1$, the relationship is linear: $\frac{N_s^*}{N_r^*} = \frac{2\nu}{c}$ because midpoint pegs are nonexistent for the indicated parameter region. The remaining panels show that both performance measures are constant when $\xi = 0$ for the same reason, and reach their maximal discrepancies for large ν when $\xi = 1$.

We conclude that, for a considerable range around the baseline value of ν, order protection has a powerful effect: it substantially increases equilibrium pegged orders and substantially reduces transactions costs (TC). Queuing costs (QC), however, increase. For example, as shown in panel (a) of Table 2, under the baseline calibration, order protection increases ω^* by 33%, decreases TC by 4.8%, but increases QC by 43%.

Figure 3 also shows what our model predicts for very low values of ν. When there are vanishingly few jumps in the fundamental asset value relative to investor order arrivals, protection becomes irrelevant and the equilibrium values and performance metrics are invariant

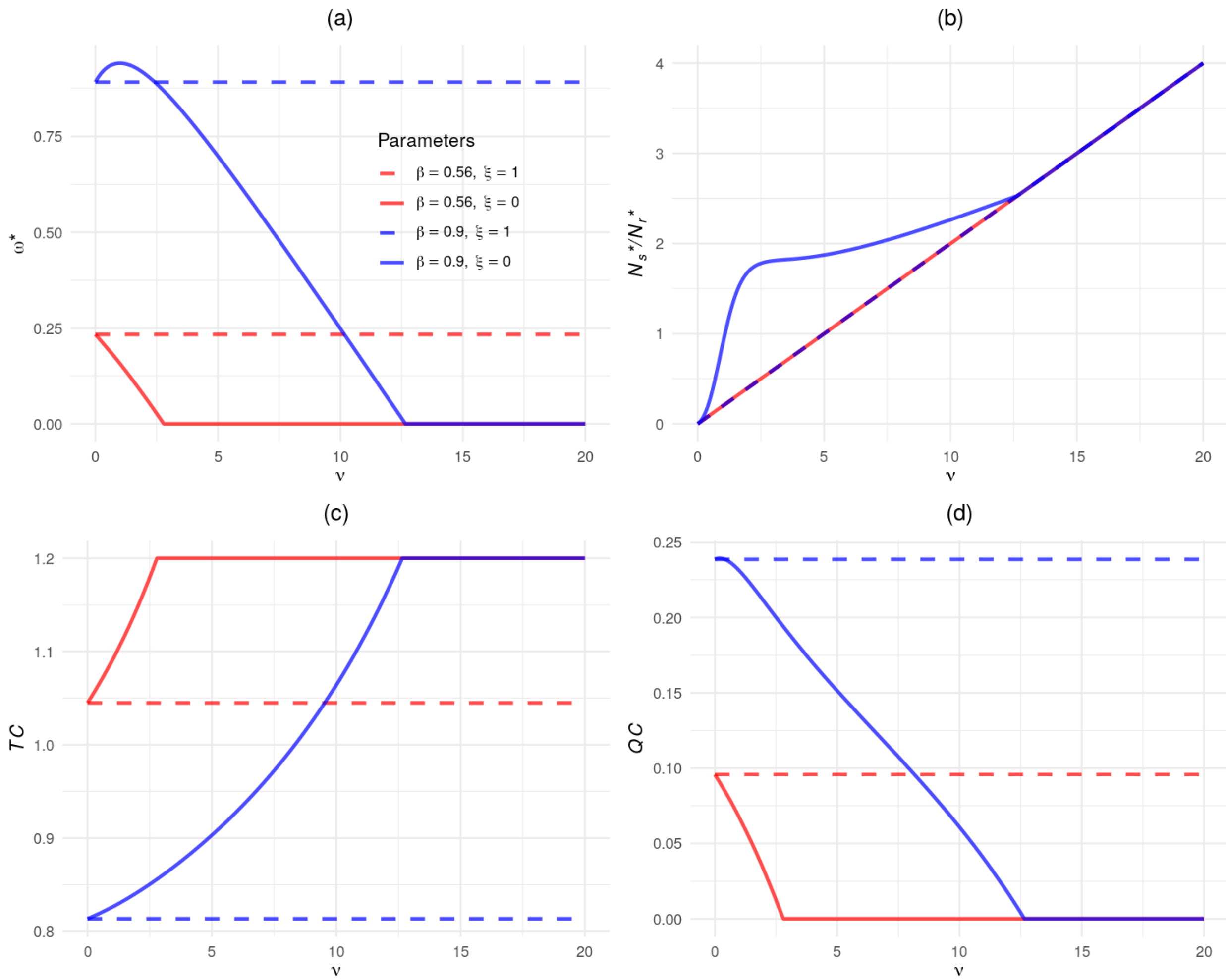

Figure 3: Impact of ν on equilibrium ratios and performance metrics. Other parameters are held fixed at baseline values, except that blue lines show impact when $\beta = 0.9$ instead of baseline $\beta = 0.56$. Dotted lines show values for $\xi = 0$ and solid lines for $\xi = 1$. Panel (a) shows the equilibrium fraction ω^* of pegged orders, panel (b) shows the equilibrium sniper ratio $\frac{N_s^*}{N_r^*}$, and panels (c) and (d) respectively show transactions cost and queuing cost performance metrics.

to $\xi \in \{0, 1\}$. Between $\nu = 0$ and the point where unprotected pegs disappear (e.g., $\nu \gtrsim 2.5$ in the baseline) the equilibrium ratios and the performance metrics are all monotonic, as one might expect, but with one surprising exception: the peg share ω^*.

Counterexample. A natural conjecture is that midpoint pegs are always more common when they are protected. The results of Appendix A show that this is true in the sense that removing protection decreases the mean peg queue length N_p^*. It is also true in the sense that, conditional on order imbalance k, removing protection impairs the profitability

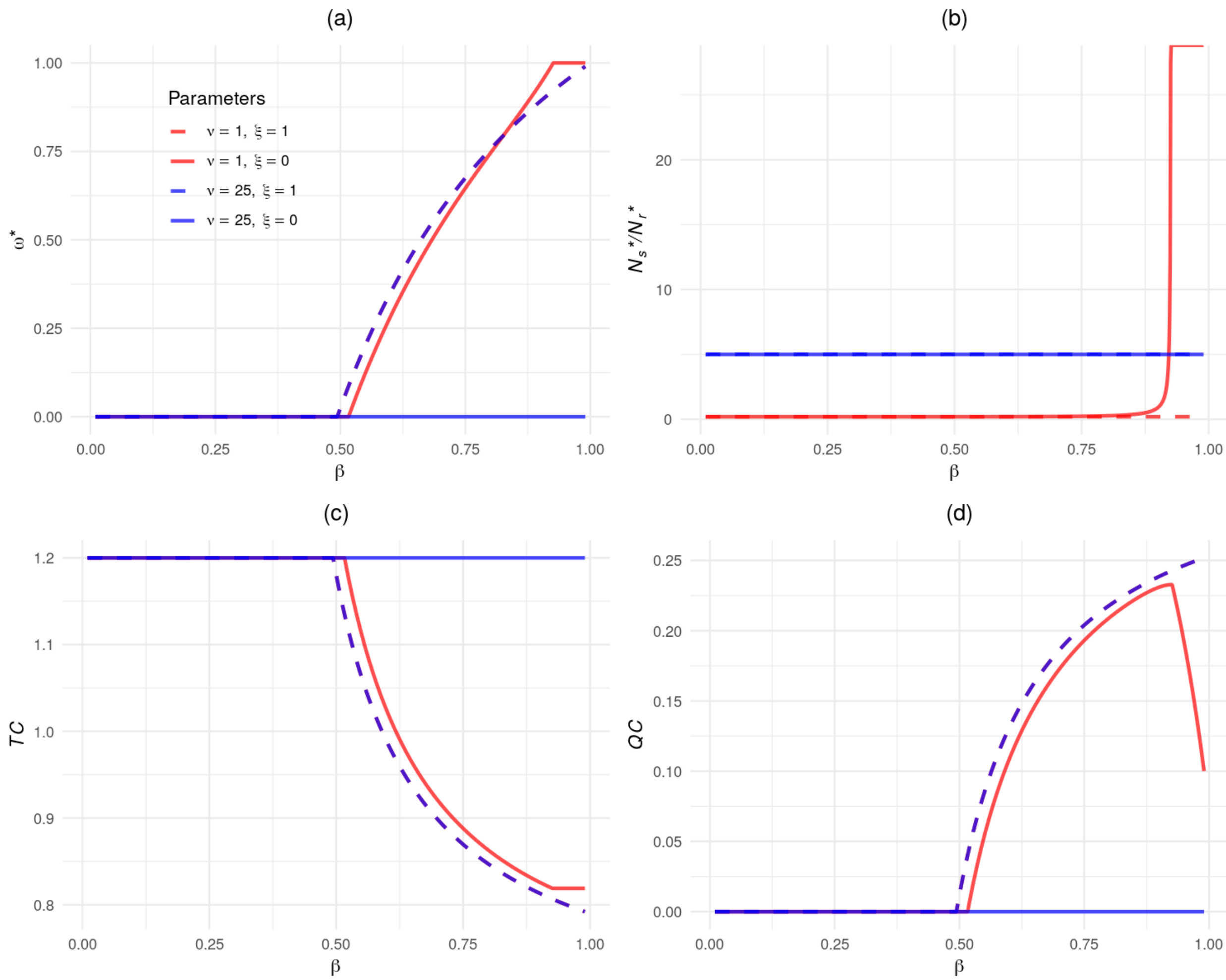

Figure 4: Impact of investor patience on equilibrium ratios and performance metrics. The horizontal axis is $\beta = \exp\left(-\frac{\delta}{\rho}\right)$. All other parameters are at baseline values for red lines, and all except $\nu = 25 = 0.5\rho$ for blue lines. Dotted lines show values for $\xi = 0$ and solid lines show $\xi = 0$.

of midpoint peg orders more than that of market orders and thus tends to reduce their equilibrium share. However, there is a subtle indirect effect that goes in the other direction: the distribution of queued orders shifts towards smaller imbalances, resulting in faster fills for midpoint peg orders. This reduces the sniping hazard and makes pegs more attractive.

Panel (a) of Figure 3 shows that the conjecture is false: that is, $\omega^* > \tilde{\omega}^*$ for very small values of ν when $\beta = 0.9$ (not for the baseline β). Evidently, when investors are sufficiently patient and sniping risk is sufficiently small, the indirect effect more than offsets the direct effects.

(a) Baseline				
Market	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.233	0.200	1.04	0.0958
$\xi = 1$	0.157	0.202	1.09	0.0671
Diff	0.0764	-0.00172	-0.0461	0.0287

(b) $\nu = 25$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.233	5.00	1.04	0.0958
$\xi = 1$	0.000	5.00	1.20	0.000
Diff	0.233	0.00	-0.155	0.0958

(c) $\rho = 1$, $\nu = 50$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.233	10.0	1.04	0.0958
$\xi = 1$	0.000	10.0	1.20	0.000
Diff	0.233	0.000	-0.155	0.0958

(d) $\beta = 0.1$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.000	0.200	1.20	0.000
$\xi = 1$	0.000	0.200	1.20	0.000
Diff	0.000	0.000	0.000	0.000

(e) $\beta = 0.9$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.892	0.200	0.813	0.239
$\xi = 1$	0.941	0.900	0.826	0.231
Diff	-0.0494	-0.700	-0.0127	0.00761

(f) $\nu = 0.1$, $\beta = 0.9$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.892	0.0200	0.813	0.239
$\xi = 1$	0.902	0.0264	0.815	0.239
Diff	-0.0107	-0.00641	-0.00113	-0.000580

(g) $c = 30$	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.233	0.0667	1.04	0.0958
$\xi = 1$	0.157	0.0672	1.09	0.0671
Diff	0.0764	-0.000573	-0.0461	0.0287

Table 2: Performance metrics at market equilibrium with ($\xi = 0$) and without ($\xi = 1$) order protection. Panel (a) corresponds to baseline parameters $c = 10$, $d = 0.18$, $\varphi = 1.8$, $\beta = 0.56$, $\rho = 50$ and $\nu = 1$; the remaining panels use specified deviations from the baseline case.

Figure 4 offers a more complete picture of how the impact of order protection depends on investor patience, β. For very low values (i.e., for very impatient investors), $\omega^* = \tilde{\omega}^* = 0$. Consequently (as with low ν in the previous figure) $\frac{N_s^*}{N_r^*} = \frac{2\nu}{c}$ and the performance metrics do not depend on whether pegs are protected for $\beta \lesssim 0.5$. However, for $\beta \gtrsim 0.5$, the protected equilibrium has $\omega^* > 0$, which is associated with lower transactions costs (TC), again at the expense of uniformly higher queuing costs (QC). Here again, unprotected markets also have a higher sniper ratio $\frac{N_s^*}{N_r^*}$.

Interestingly, panel (d) of Figure 4 shows that queuing costs *decrease* in the baseline unprotected case for very high values of β. This is a result of the fact that queuing costs are decreasing in the discount factor β^k, but increasing in the fraction of pegged orders $\omega^*(\beta)$. For most values of β, the second effect is stronger than the first, resulting in increasing queuing costs. However, for sufficiently large β, the first effect is stronger and queuing costs decline.

Table 2 reports specific equilibrium values and performance metric comparisons, beginning with baseline parameters in panel (a). An important implication of Equations (3.6) – (3.8) and (4.7a) – (4.7c) is that the cost of speed technology, c, affects equilibria only through the sniper mass, N_s^*; it does not impact the share of pegged orders or the mass of market makers. As a result, the performance measures do not vary with c although it does affect the sniper ratio, $\frac{N_s^*}{N_r^*}$. Table 2 confirms that order protection via messaging delay robustly (but not universally) improves performance metric TC but impairs QC.

6 Discussion

The ultimate source of profits for both proprietary traders and brokers in our model is the exogenous order flow from investors. Investor orders provide fee income to brokers, whose transactions subsequently provide income to proprietary traders who make markets via lit resting orders at the best bid and best offer. Some of that income is diverted to snipers, who transact with stale BBO orders immediately following a jump in the fundamental value. Intuitively, we have a food chain, with impatient investors' market orders sustaining regular limit orders, which sustain sniping.

A recent innovation at some exchanges offers investors/brokers an attractive new option: a hidden midpoint peg that is protected from snipers[5] and that executes at a better price. However, pegged orders incur an expected queuing cost that increases with the fraction ω of investors who choose pegs. Since pegged orders are hidden, traders can not observe the queue in advance, but in equilibrium they know its expected length and the resulting cost. When that queuing cost is sufficiently disadvantageous, investors (or their brokers) will resort to

[5]As explained at length earlier, the fact that the order is hidden does not protect it from sniping; protection rather comes from the peg, which automatically reprices it before (delayed) snipe orders arrive.

standard market orders, which execute against market makers' (lit) best bids and offers.

6.1 Testable predictions

Our model lays out the equilibrium consequences of the aforementioned tradeoffs, providing predictions that can be tested against laboratory and field data. The simplest version of the model is intended to capture the functioning of the new market format in calm conditions. It assumes a thick order book of slow (unprotected) regular orders at BBO, and assumes that midpoint pegged orders are protected from sniping. Key predictions include:

1. The mass of active market makers, N_r, and of snipers, N_s, will increase when the flow of investors, ρ, increases. Indeed, if the discount rate δ is proportional to ρ,[6] so that β and ω^* remain constant, then the equilibrium masses of both types of proprietary traders are directly proportional to ρ, as seen in equations (3.7) - (3.8).

2. The ratio ν/ρ captures the degree of market turbulence and the relative prevalence of sniping opportunities. Equations (3.7) - (3.8) show that an increase in this ratio will proportionately decrease the population mass of market makers, N_r, but (perhaps surprisingly) have no impact on the mass of snipers, N_s. Indeed, with the sole exception of N_s (which scales as $1/\rho$), all equilibrium expressions can be cast as functions of the ν/ρ ratio.

3. An increase in the cost of speed, c, will proportionately reduce the population mass of snipers, N_s, but will have no effect on the mass of market makers, N_r.

4. The fraction of impatient investors that transmit pegged orders, ω^*, is an increasing function of the discount factor, $\beta = \exp\left(-\frac{\delta}{\rho}\right) \in (0, 1]$.

What happens when midpoint orders are not protected from sniping ($\xi = 1$)? While this question is relevant to policy, it can only be studied in a laboratory environment as it is difficult to find two markets in the field that differ only in the trading format, and that differ for exogenous reasons. According to Propositions 4.1 and 4.2:

[6] This might be the case if impatience arises mainly from concerns about preemption by other investors.

1. Given a positive fraction of orders transmitted as pegs, ω, Equation (4.3) tells us that $\lambda < \omega$, i.e., the order imbalance is more tightly concentrated around zero when $\xi = 1$.

2. The equilibrium value of ω^* is smaller when $\xi = 1$ for a wide range of parameter values, including baseline parameters. However, as specified in Proposition 4.2, the inequality can go the other way for certain extreme parameter values.

3. In the usual case that $\tilde{\omega}^* < \omega^*$, Equations (3.7), (3.8), (4.7b), and (4.7c) show that the mass N_r of limit orders at BBO and the mass N_s of snipers will be larger when $\xi = 1$.

4. Most importantly, for parameter values in a large neighborhood of baseline, imposing midpoint peg order protection substantially lowers transactions costs, but increases queuing costs.

6.2 Future work

To isolate the impact of order protection, we have assumed a particular fee structure reflecting current practice at leading exchanges that offer order protection via messaging delay. Future work with small variants on the present model could investigate the impact of fee structure, with and without order protection.

How well do current results stand up when key simplifying assumptions are relaxed? For example, what happens when we relax Assumption A5 and allow liquidity adders (either of lit orders at BBO or of hidden midpoint pegs) to also purchase speed at flow cost $c > 0$? Appendix A.4 shows that such purchases are unprofitable in a large neighborhood of baseline parameters, but for some other parameter values adders would wish to purchase speed. Preliminary work indicates that closed-form solutions are no longer possible, but recursion techniques (in particular, the Erlang B model), which can be solved numerically, suggest no qualitative changes to current results. Preliminary work similarly suggests that relaxing Assumption A3 to allow a symmetric distribution of jumps (generalizing our distribution supported on ± 2) complicates the formulas but has little qualitative effect on the results.

Our model focuses on steady state equilibrium, but Section 3.6 and Appendix B consider dynamic strategies that seek to exploit short term information on the hidden order queue. We find surprisingly little scope for, or evidence of, such dynamic strategies for altering the choice between peg and market orders. Of course, this does not imply that dynamic strategies are ineffective for timing investor orders. Indeed, such strategies are the essence of high frequency trading, but (like Budish et al. (2015) and others) we have set such matters to one side by assuming an observable exogenous fundamental value and exogenous investor orders centered on it.

As an ambitious step in that direction, future researchers might replace our assumptions A3 and A4 by an exogenous and time-varying process of investor arrivals in which V is implicitly defined by balancing expected buy and sell order flows. This would bring back adverse selection, which we view as of first order importance in markets but do not see as interacting strongly with the market format differences we consider. Thus we conjecture that most of our qualitative results on order protection would still hold with unobservable fundamentals and with endogenous investor timing, but as yet we have little evidence on this question.

Another ambitious extension would be to relax Assumption A1 and to model competing exchanges, and possibly multiple securities. The NBBO and the fundamental value would be endogenous, given some appropriately specified overall investor demand that endogenously distributes itself across exchanges, assets and order types.

Empirical work need not wait for these theoretical extensions. In the laboratory, one could investigate whether human subjects in the broker role track $\hat{\omega}$ when the experimenter varies parameters such as (δ, φ, d), and whether human subjects in the proprietary trader role follow the comparative static predictions of the impact on (N_r, N_s) of the parameters (ν, ρ, c). Using field data, one might systematically examine the present model's comparative statics. We were encouraged to see that the empirical distribution of the hidden order queue so closely followed conformed to the discrete Laplace functional form predicted by our model, and hope that the present paper inspires new empirical and theoretical research.

Appendices

A Mathematical Details

The order queue has distribution function of the form $\mathbf{p} = (p_k)_{k\in Z} = (..., p_{-2}, p_{-1}, p_0, p_1, p_2, ...)$, where each $p_k \geq 0$ and $\sum_{k=-\infty}^{\infty} p_k = 1$. Let $\mathbf{P}$ be the set of all such distribution functions. This appendix analyzes how the investor arrival process induces dynamics on $\mathbf{P}$, and computes the steady states.

A.1 Protected Pegged Orders

Proposition 3.1. *Let $\omega \in (0,1)$ be the probability that an investor arrival on either side of the market results in a midpoint peg order. Given Assumptions A1-A5, there is a unique steady state distribution $\mathbf{q} = (q_k)_{k\in Z}$ of the order imbalance, with*

$$q_k = \left(\frac{1-\omega}{1+\omega}\right)\omega^{|k|}, \quad k \in \mathbb{Z}. \tag{3.1}$$

Proof. As noted in the text, an investor arrival generates a midpoint peg buy or sell order, or a market buy or sell order, with respective probabilities $\omega/2, \omega/2, (1-\omega)/2, (1-\omega)/2$. Recall also that either sort of sell (resp. buy) order generates a transition $k \to k+1$ (resp. $k \to k-1$) when $k < 0$ (resp. $k > 0$), but a new market sell (resp. buy) leaves k unchanged when $k \geq 0$ (resp. $k \leq 0$). Thus a single investor arrival transforms the distribution $\mathbf{p} \in \mathbf{P}$ to the distribution $\mathbf{Tp} = (Tp_k)_{k\in Z} \in \mathbf{P}$, where

$$Tp_k = \begin{cases} \frac{\omega}{2}p_{k+1} + \frac{1-\omega}{2}p_k + \frac{1}{2}p_{k-1}, & \text{if} \quad k < 0 \quad \text{(A.1a)} \\ \frac{1}{2}p_1 + (1-\omega)p_0 + \frac{1}{2}p_{-1}, & \text{if} \quad k = 0 \quad \text{(A.1b)} \\ \frac{1}{2}p_{k+1} + \frac{1-\omega}{2}p_k + \frac{\omega}{2}p_{k-1}, & \text{if} \quad k > 0. \quad \text{(A.1c)} \end{cases}$$

To verify that the distribution $\mathbf{q} \in \mathbf{P}$ defined in (3.1) is indeed a steady state, we begin with the case $k < 0$. Writing $B = \left(\frac{1-\omega}{1+\omega}\right)$ to simplify expressions, we substitute (3.1) into Equation (A.1a) to obtain

$$Tq_k = \frac{\omega}{2}B\omega^{-(k-1)} + \frac{1-\omega}{2}B\omega^{-k} + \frac{1}{2}B\omega^{-(k+1)}$$

$$= \left(\frac{1}{2} + \frac{1}{2}\right) B\omega^{-k} + \left(\frac{1}{2} - \frac{1}{2}\right) B\omega^{-(k+1)}$$

$$= B\omega^{-k} \tag{A.2}$$

$$= q_k. \tag{A.3}$$

Similarly one can use (A.1b) and (A.1c) to verify that $Tq_k = q_k$ when $k = 0$ and when $k > 0$. Hence $\mathbf{q}$ is a fixed point of the operator $\mathbf{T}$ defined on sequence spaces. To complete the proof, it remains only to verify that $\mathbf{q} = \mathbf{Tq} \in \mathbf{P}$, i.e., that it is indeed a probability distribution, i.e., that each $q_k \geq 0, \forall k \in \mathbb{Z}$ and $\sum_{k \in \mathbb{Z}} q_k = 1$. Nonnegativity is clear from inspection, while

$$\sum_{k \in \mathbb{Z}} q_k = B\left(1 + 2\sum_{k=1}^{\infty} \omega^k\right) = B\left(1 + 2\frac{\omega}{1-\omega}\right) = B\frac{1+\omega}{1-\omega} = BB^{-1} = 1. \quad \square$$

A.2 Dynamics

Away from the steady state, the dynamics are governed by the twisted-Toeplitz operator $\mathbf{T}$ defined by (A.1a) - (A.1c), and by the Poisson process of investor order arrivals. Since orders (buy and sell combined) arrive at rate 2ρ, the distribution $\mathbf{p}(t) \in \mathbf{P}$ evolves from any given initial distribution $\mathbf{p^o} \in \mathbf{P}$ according to the equation

$$\mathbf{p}(t) = e^{-2\rho t} \sum_{n=0}^{\infty} \frac{(2\rho t)^n}{n!} \mathbf{T}^n \mathbf{p^o}. \tag{A.4}$$

We are not aware of any explicit closed form expressions for the components $p_k(t)$, but it is straightforward to simulate them numerically using (A.4).

The data described in Appendix B are not point observations of $p_k(t)$, but instead are time-average observations of the form $A_k(t) = t^{-1} \int_0^t p_k(s)ds$. We obtain the predicted mean frequencies via Equation (A.4) in the following manner:

$$\begin{aligned} \mathbf{A}(t) &= \frac{1}{t} \int_0^t \mathbf{p}(s)ds \\ &= \frac{1}{t} \int_0^t e^{-2\rho s} \sum_{n=0}^{\infty} \frac{(2\rho s)^n}{n!} \mathbf{T}^n \mathbf{p}^o ds \end{aligned}$$

$$= \sum_{n=0}^{\infty} \mathbf{T}^n \mathbf{p}^o \frac{1}{2\rho t} \int_0^t \frac{(2\rho)^{n+1} s^n}{\Gamma(n)} e^{-2\rho s} ds$$

$$= \sum_{n=0}^{\infty} \frac{G(t; n+1, 2\rho)}{2\rho t} \mathbf{T}^n \mathbf{p}^o \tag{A.5}$$

where $G(t; n+1, 2\rho)$ is the CDF of the Gamma distribution with shape and rate parameters $(n+1, 2\rho)$. Again, Equation (A.5) can be used to generate theoretical predictions to any desired degree of accuracy.

A.3 Asymmetric Discrete Laplace Distribution

The asymmetric discrete Laplace distribution arises naturally from our model's dynamics, and was previously discovered by Kozubowski and Inusah (2006). In our notation, it has parameters $\omega_R, \omega_L \in (0, 1)$. For each integer $k \in Z$, the ADL(ω_R, ω_L) distribution assigns probability

$$p_k = \begin{cases} \dfrac{(1-\omega_R)(1-\omega_L)}{1-\omega_R\omega_L} \omega_L^{|k|} & \text{if} \quad k \leq 0 \qquad \text{(A.6a)} \\ \dfrac{(1-\omega_R)(1-\omega_L)}{1-\omega_R\omega_L} \omega_R^{|k|}, & \text{if} \quad k \geq 0. \qquad \text{(A.6b)} \end{cases}$$

It is straightforward to confirm that ADL$(\omega_R, \omega_L) \in \mathcal{P}$, i.e., that these probabilities are non-negative and sum to unity.

The steady state order queue in our basic model has the symmetric distribution ADL(ω, ω), where $p_0 = \frac{(1-\omega)^2}{1-\omega^2} = \frac{1-\omega}{1+\omega}$. By contrast, the order queue distribution immediately following an observed transaction at BB is $\mathbf{p^o} =$ ADL$(\omega, 0)$. Subsequently, the order distribution $\mathbf{p}(t)$ relaxes to the steady state as per Equation (A.4). Numerical exercises in Appendix B confirm that $\mathbf{p}(t)$ is very closely approximated by ADL$(\omega, \alpha(t)\omega)$ at time t, where $\alpha(t)$ is a strictly increasing function with $\alpha(0) = 0$ and $\alpha(\infty) = 1$. By symmetry, following an observed transaction at BO, the distribution is closely approximated by ADL$(\alpha(t)\omega, \omega)$. Below, we will refer to the side $K =$ R or L with $\omega_K = \alpha(t)\omega$ as the *light* side and the other side, $\omega_K = \omega$, as the *heavy* side.

Consequently, our model predicts that observed order queues t seconds after a transaction at either BB or BO will have light side ($k < 0$) frequencies

$$p_k^L(\alpha) = \frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2} \alpha^{|k|} \omega^{|k|} \tag{A.7}$$

for some $\alpha = \alpha(t) \in (0,1)$, and heavy side ($k \geq 0$) frequencies

$$p_k^H(\alpha) = \frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2}\omega^{|k|}. \tag{A.8}$$

Panel (b) of Figure 5 plots these frequencies for several values of t. Note that the frequencies lie along a triangle with apex at $k = 0$ and that the apex height decreases monotonically from $\ln(1-\omega)$ at $t = 0$ to an asymptote (as $t \to \infty$) of $\ln(1-\omega) - \ln(1+\omega)$. The light side (absolute) slope monotonically decreases from infinite to $-\ln\omega$, while the heavy side (absolute) slope remains constant at $-\ln\omega$.

A.4 Dynamic Strategic Implications

Recall the passive dynamic strategies mentioned in Section 3.6 when $\alpha(t) < 1$: (a) light-side investors place pegged orders, and (b) heavy-side investors place market orders. To compute the theoretical advantages of these strategies, note that by Equations (A.7) - (A.8):

$$\sum_{k=1}^{\infty} p_k^H(\alpha) = \frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2}, \qquad \sum_{k=0}^{\infty} p_k^H(\alpha) = \frac{(1-\alpha\omega)}{1-\alpha\omega^2} \tag{A.9}$$

$$\sum_{k=1}^{\infty} p_k^L(\alpha) = \frac{\alpha\omega(1-\omega)}{1-\alpha\omega^2}, \qquad \sum_{k=0}^{\infty} p_k^L(\alpha) = \frac{(1-\omega)}{1-\alpha\omega^2}. \tag{A.10}$$

Thus

$$\begin{aligned}
\pi_P^L(\alpha) &= (\varphi - d)\left[\sum_{k=1}^{\infty} p_k^H(\alpha) + \sum_{k=0}^{\infty} p_k^L(\alpha)\beta^{k+1}\right] \\
&= (\varphi - d)\left[\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \beta\frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2}\sum_{k=0}^{\infty}(\alpha\beta\omega)^k\right] \\
&= (\varphi - d)\left[\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\alpha\beta\omega)(1-\alpha\omega^2)}\right],
\end{aligned} \tag{A.11}$$

while

$$\begin{aligned}
\pi_M^L(\alpha) &= (\varphi - d)\sum_{k=1}^{\infty} p_k^H(\alpha) + (\varphi - 1)\sum_{k=0}^{\infty} p_k^L(\alpha) \\
&= (\varphi - d)\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \frac{(\varphi-1)(1-\omega)}{1-\alpha\omega^2}.
\end{aligned} \tag{A.12}$$

Using the fact that $(\varphi - 1) = (\varphi - d)\frac{\beta(1-\omega)}{1-\beta\omega}$ (see Equation (3.9)), it follows that strategy (a) has payoff advantage

$$\Delta\pi^L(\alpha) = \pi_P^L(\alpha) - \pi_M^L(\alpha)$$

$$= \frac{(\varphi - d)\beta(1-\omega)(1-\alpha\omega)}{(1-\alpha\beta\omega)(1-\alpha\omega^2)} - \frac{(\varphi - d)\beta(1-\omega)(1-\omega)}{(1-\beta\omega)(1-\alpha\omega^2)}$$
$$= (\varphi - d)\frac{\beta(1-\omega)}{1-\alpha\omega^2}\left[\frac{1-\alpha\omega}{1-\alpha\beta\omega} - \frac{1-\omega}{1-\beta\omega}\right], \tag{A.13}$$

The factor in brackets is zero at $\alpha = 1$, which yields $\Delta\pi^L(1) = 0$, as must be the case in equilibrium. Inspection of the same factor shows that the payoff advantage is positive over the relevant parameter range when $\alpha \in [0, 1)$. It is is maximal at $\alpha = 0$, where Equation (A.13) reduces to

$$\Delta\pi^L(0) = (\varphi - d)\omega(1-\beta)\frac{\beta(1-\omega)}{1-\beta\omega}. \tag{A.14}$$

Recall that steady-state market-order equilibrium profit is

$$\pi^L(1) = \left[\frac{\varphi - d}{1+\omega}\right]\frac{\omega(1-\beta\omega) + \beta(1-\omega)}{1-\beta\omega}. \tag{A.15}$$

Hence strategy (a) obtains proportional advantage of at most

$$\frac{\Delta\pi^L(0)}{\pi^L(1)} = \frac{\beta\omega(1-\beta)(1-\omega)(1+\omega)}{\omega(1-\beta\omega) + \beta(1-\omega)}. \tag{A.16}$$

For reasonable parameters, this maximal proportional advantage is small; Appendix B shows that realistic estimates (taking into account the evolution of $\alpha(t)$) are far below this upper bound.

As for strategy (b), the advantage of a peg over a market order for a heavy side investor, conjectured to be negative, is actually

$$\Delta\pi^H(\alpha) = \pi^H_P(\alpha) - \pi^H_M(\alpha)$$
$$= (\varphi - d)\frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\beta\omega)(1-\alpha\omega^2)} - (\varphi - 1)\frac{(1-\alpha\omega)}{(1-\alpha\omega^2)}$$
$$= (\varphi - d)\left[\frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\beta\omega)(1-\alpha\omega^2)} - \frac{\beta(1-\omega)}{(1-\beta\omega)}\frac{(1-\alpha\omega)}{(1-\alpha\omega^2)}\right] = 0. \tag{A.17}$$

A.5 Unprotected Pegged Orders

When midpoint peg orders are not protected from sniping, the queue state k is affected by two additional events – up and down jumps in the fundamental – as well as the four types of investor arrivals. The stationary distribution then has the same general form as in the protected case, but is a more complicated function of the exogenous parameters.

Proposition 4.1. *Let $\omega \in (0,1)$ be the probability that an investor arrival on either side of the market results in a midpoint peg order. Given assumptions A1-A5b, there is a unique steady state distribution $\mathbf{q} = (q_k)_{k \in Z}$ of the order imbalance, with*

$$q_k = \left(\frac{1-\lambda}{1+\lambda}\right)\lambda^{|k|}, \quad k \in \mathbb{Z}, \tag{4.2}$$

where

$$\lambda = \frac{1}{2}\left(1 + \frac{\xi\nu}{\rho} + \omega\right) - \frac{1}{2}\sqrt{\left(1 + \frac{\xi\nu}{\rho} + \omega\right)^2 - 4\omega} \qquad \in (0, \omega], \tag{4.3}$$

and the variable $\xi = 0$ (resp. $\xi = 1$) indicates that pegged orders are protected from sniping (resp. are not protected).

Proof. Conditional on the event being an investor order, the probabilities of the four kinds of orders are unchanged from the previous proposition, so the unconditional probabilities are attenuated by the probability $R = \frac{\rho}{\rho + \xi\nu} \leq 1$ that the event is an investor order rather than a jump in the fundamental value. Each upwards jump causes a direct transition $k \to 0$ when $k \geq 0$ and $\xi = 1$, and otherwise has no effect on k, while each downwards jump similarly causes $k \to 0$ when $k \leq 0$. Thus the probability that $k > 0$ will remain unchanged is now the probability of a market sell plus the probability of a downward jump, $P_{ms} + P_{dj} = \frac{R(1-\omega)}{2} + \frac{1-R}{2} = \frac{1-Rw}{2}$; the same probability applies to the symmetric case $k < 0$.

The updating operator T is now defined for $k \neq 0$ by:

$$Tp_k = \begin{cases} \frac{R\omega}{2}p_{k+1} + \frac{1-R\omega}{2}p_k + \frac{R}{2}p_{k-1}, & \text{if} \quad k < 0; \quad \text{(A.18a)} \\ \frac{R}{2}p_{k+1} + \frac{1-R\omega}{2}p_k + \frac{R\omega}{2}p_{k-1}, & \text{if} \quad k > 0. \quad \text{(A.18b)} \end{cases}$$

As for $k = 0$, note that Tp_0 is the sum of the R-attenuated RHS of Equation (A.1b), plus the probability mass arising from direct transitions to $k = 0$ following jumps in the fundamental. That mass is $\frac{1-R}{2}\left(\sum_{k\geq 0} p_k + \sum_{k \leq 0} p_k\right) = \frac{1-R}{2}(1 + p_0)$. Hence

$$\begin{aligned} Tp_0 &= R\left(\frac{1}{2}p_1 + (1-\omega)p_0 + \frac{1}{2}p_{-1}\right) + \frac{1-R}{2} + \frac{1-R}{2}p_0 \\ &= \frac{1-R}{2} + \frac{R}{2}p_1 + \frac{1+R-2R\omega}{2}p_0 + \frac{R}{2}p_{-1} \end{aligned} \tag{A.19}$$

Note that Equations (A.18a) - (A.19) reduce to Equations (A.1a) - (A.1c) when $\xi = 0$, which implies that $R = 1$.

To find a steady state for the general operator T defined above, first set $Tp_0 = p_0$ in Equation (A.19) to obtain

$$0 = \frac{1-R}{2} + \frac{R}{2}p_1 + \frac{R-1-2R\omega}{2}p_0 + \frac{R}{2}p_{-1}. \tag{A.20}$$

In steady state, the symmetry of the market ensures $p_1 = p_{-1}$, so Equation (A.20) yields

$$p_1 = p_{-1} = (\omega + \frac{r}{2})p_0 - \frac{r}{2}, \tag{A.21}$$

where $r = \frac{\xi\nu}{\rho}$. Next, to find the steady state equation for $k > 0$, set $Tp_k = p_k$ in Equation (A.18b) and simplify to obtain

$$p_{k+1} - (1 + r + \omega)\, p_k + \omega p_{k-1} = 0, \quad \text{for } k > 0. \tag{A.22}$$

Equation (A.22) is a linear second order homogeneous difference equation, whose general solution (see, e.g., Sargent 1979 pp. 177 ff) takes the form

$$p_k = a_1\lambda_1^k + a_2\lambda_2^k, \tag{A.23}$$

for $k > 0$ where

$$\lambda_1 = \frac{1}{2}(1 + r + \omega) + \frac{1}{2}\sqrt{(1 + r + \omega)^2 - 4\omega} \tag{A.24}$$

$$\lambda_2 = \frac{1}{2}(1 + r + \omega) - \frac{1}{2}\sqrt{(1 + r + \omega)^2 - 4\omega}, \tag{A.25}$$

are the roots of the quadratic equation

$$\lambda^2 - (1 + r + \omega)\,\lambda + \omega = 0. \tag{A.26}$$

The discriminant $(1 + r + \omega)^2 - 4\omega$ is bounded above by $(1 + r + \omega)^2$ and bounded below by $(1 + r + \omega)^2 - 4\omega(1 + r) = (1 + r - \omega)^2$ for all $\nu, \rho > 0$ and $\xi, \omega \in [0, 1]$. It follows that $\lambda_1 \geq 1$; to ensure that $p_k \to 0$ as $k \to \infty$, we therefore must have $a_1 = 0$ in Equation (A.23). The same bounds establish that $\lambda_2 \in (0, \omega] \subset (0, 1)$. Note that $r = 0$ and $\lambda_2 = \omega$ when $\xi = 0$.

Set $\lambda = \lambda_2$, consistent with Equation (4.3) of the proposition. Given the symmetry of our model, it is natural to conjecture (and straightforward to check) that

$$p_k = a_2\lambda^{|k|} \tag{A.27}$$

also satisfies the steady state equation

$$p_{k-1} - (1 + r + \omega)\, p_k + \omega p_{k+1} = 0, \quad \text{for } k < 0, \tag{A.28}$$

obtained from Equation (A.18a). To ensure that Equation (A.27) defines a probability distribution, we choose a_2 so that

$$1 = \sum_{k=-\infty}^{\infty} p_k = a_2 + 2a_2 \sum_{k=1}^{\infty} \lambda^k = a_2 \left[1 + 2\frac{\lambda}{1-\lambda}\right] = a_2 \left[\frac{1+\lambda}{1-\lambda}\right], \tag{A.29}$$

Hence, $a_2 = p(0) = \frac{1-\lambda}{1+\lambda}$. Inserting this into Equation (A.27) and relabeling the left hand side as $\tilde{q}_k$, we obtain the desired expression, Equation (3.1). We have already verified that it satisfies the steady state equations for $k > 0$ and $k < 0$. The last step is to verify that it satisfies the steady state Equation (A.20) for $k = 0$, or equivalently Equation (A.21), i.e., that $p_1 - (\omega + \frac{r}{2})p_0 + \frac{r}{2} = 0$ when $p_k = \frac{1-\lambda}{1+\lambda}\lambda^{|k|}$. Indeed,

$$\begin{aligned}
\frac{1-\lambda}{1+\lambda}\lambda - \left(\omega + \frac{r}{2}\right)\frac{1-\lambda}{1+\lambda} + \frac{r}{2} &= (1+\lambda)^{-1}\left[\lambda - \lambda^2 - w - \frac{r}{2} + \lambda w + \lambda\frac{r}{2} + \frac{r}{2} + \lambda\frac{r}{2}\right] \\
&= (1+\lambda)^{-1}[-\lambda^2 + (1 + r + w)\lambda - w] \\
&= (1+\lambda)^{-1}[0] = 0,
\end{aligned}$$

where the penultimate step uses Equation (A.26), which defines λ. □

Corollary A.1. *Given parameters ν, ρ and ξ, the steady-state fraction of brokers choosing to place midpoint peg orders is*

$$\omega = \lambda + \xi \left[\frac{\lambda}{1-\lambda}\right] \frac{\nu}{\rho}, \tag{A.30}$$

where λ is the steady state value determined in Proposition 4.1.

Proof The result is obtained by solving for ω in Equation (A.26).

Remark. Clearly ω is strictly increasing in λ for the relevant parameter values, so its inverse function $\lambda(\omega|\xi = 1, \nu, \rho)$ exists and is also strictly increasing.

A.6 Market Equilibrium

Proposition 4.2. *Under assumptions A1 - A5b and parameter restrictions $\varphi \geq 1 > d \geq 0$, $\beta = \exp\left(-\frac{\delta}{\rho}\right) \in (0,1)$, and $c, \rho, \nu > 0$, there is a unique market equilibrium, with*

$$\tilde{\omega}^* = \tilde{\lambda} + \xi \left(\frac{\tilde{\lambda}}{1-\tilde{\lambda}}\right) \frac{\nu}{\rho} \tag{4.7a}$$

$$N_r^* = \frac{\rho}{\nu}\left(\frac{1-\tilde{\omega}^*}{1+\tilde{\lambda}}\right) \qquad \textit{and} \tag{4.7b}$$

$$N_s^* = 2\nu\frac{N_r^* + 2\xi N_p^*}{c} = \frac{2\rho}{c}\left(\frac{1-\tilde{\omega}^*}{1+\tilde{\lambda}}\right) + \frac{4\xi\nu}{c}\frac{\tilde{\lambda}}{1-\tilde{\lambda}^2} \tag{4.7c}$$

where

$$\tilde{\lambda} = \frac{1}{2\beta^2\rho(1-d)}\Bigg[\beta^2\rho(\varphi-d) - \beta\xi\nu(\varphi+1) - \beta\rho(\varphi+d-2) - \Big(\big(\beta\xi\nu(\varphi+1) + \beta\rho(\varphi+d-2) - \beta^2\rho(\varphi-d)\big)^2 - 4\beta^2\rho(1-d)\left(\beta\rho(\varphi-d) - \rho(\varphi-1) - \xi\nu(\varphi-1+2\beta)\right)\Big)^{1/2}\Bigg]_+ \tag{4.8}$$

and $N_p^* = \dfrac{\tilde{\lambda}}{1-\tilde{\lambda}^2}$. (4.9)

Proof. Equating expected surplus $\pi_p = \pi_m$ for pegged and market orders in Equations (4.4) and (4.5) yields the following quadratic expression in λ:

$$\frac{\beta\rho(1-\lambda)(\varphi-d)}{\rho+\xi\nu-\beta\rho\lambda} - \frac{2\beta(1-\lambda)}{1-\beta\lambda} + \frac{2\beta\rho(1-\lambda)}{\rho+\xi\nu-\beta\rho\lambda} = \varphi - 1. \tag{A.31}$$

Solving for λ via the usual quadratic formula results in two solutions. Appendix A.5 shows that the condition $\lambda < 1$ requires the larger (smaller) solution to satisfy

$$(1-\beta)(\varphi-1)(\nu+\rho(1+\beta)) < 0(> 0) \tag{A.32}$$

The parameter restrictions ensure that the left-hand-side of Equation (A.32) is nonnegative, so the relevant solution involves the negative discriminant, which is written in Equation (4.8). Equation (4.7a) follows from Corollary A.1 above.

Equation (4.9), which gives the expected number of pegged orders vulnerable to sniping, N_p^*, is obtained as follows:

$$N_p^* = \sum_{k=-\infty}^{0} 0q_k + \sum_{k=1}^{\infty} kq_k = \frac{1-\tilde{\lambda}}{1+\tilde{\lambda}}\sum_{k=1}^{\infty} k\,\tilde{\lambda}^k = \left(\frac{1-\tilde{\lambda}}{1+\tilde{\lambda}}\right)\frac{\tilde{\lambda}}{(1-\tilde{\lambda})^2} = \frac{\tilde{\lambda}}{1-\tilde{\lambda}^2}. \tag{A.33}$$

Equations (4.7b) and (4.7c) follow by substituting (4.7a) and (4.9) into Equations (3.5) and (4.6), setting them equal to zero in accordance with the equilibrium condition, and solving for N_r and N_s. Note that the denominators of Equations (3.7) and (4.7b) differ, since the probability of investor market orders finding no contra-side liquidity at the midpoint is $\sum_{k=0}^{\infty} q_k = \frac{1}{1+\lambda}$ when $\xi = 1$.

The notation $[\cdot]_+$ in (4.8) means that $\tilde{\lambda}$, and hence Equation (4.7a), are truncated below at 0. For parameters such that the truncation binds, the same logic as in the previous proposition shows that profit inequalities imply that $\tilde{\omega}^* = 0$. For the other boundary case, Appendix A.3 explains why the market equilibrium value of ω is always < 1. $\square$

A.7 Limiting Case $\omega \to 1$

Suppose the model parameters are chosen so that $\omega = 1$. Then Equation (3.2) says that $\pi_p = \frac{\varphi - d}{2}$. That is, with probability $\frac{1}{2}$ there is a contra-side queue and a pegged order executes immediately, yielding surplus $\varphi - d$. When there is no contra-side queue, the pegged order joins an arbitrarily long queue and has zero present value.

Formulas such as Equation (3.2) may not convey the intuition behind this result. To better understand it, consider the limiting distribution $q_k(\omega)$ in Equation (3.1) as $\omega \to 1$. Up to a multiplicative normalizing constant, the probability $\omega^{|k|}$ approaches unity for any fixed k. More precisely, for any large but fixed integer K and centered sequence $\mathcal{K} = (-K, -K+1, ..., -1, 0, 1, ..., K-1, K)$, each queue length $k \in \mathcal{K}$ has probability $q_k < \frac{1}{2K+1}$ as $\omega \to 1$. Thus, in the limit we have an improper distribution on Z, in which the probability "leaks out to $\pm\infty$". The result is an infinite expected wait time and zero present value.

When $\omega = 1$, Equation (3.3) gives $\pi_m = \frac{\varphi - d}{2} + \frac{\varphi - 1}{2} = \pi_p + \frac{\varphi - 1}{2} \geq \pi_p$. That is, as usual, the market order gets the same fill as a peg when there is a contra-side queue, but if there is not, the market order is filled profitably (at the BBO) and so dominates a pegged order. Thus, the equal profit condition always fails when $\omega = 1$ (and $\varphi > 1$), and so $\omega = 1$ is never part of a market equilibrium. The logic applies equally to protected and unprotected midprice orders. Of course, the deep-book-at-BBO assumption does not make sense in this case, unless the BBO orders are routed from other exchanges (see Appendix C.2).

A.8 Makers and Speed Purchases

Under assumption A5, no makers purchase speed. In equilibrium, when would it be profitable for a single maker to deviate from this decision?

Purchasing speed enables a maker to escape N_s fast snipers with probability $1/(N_s+1)$, because each speedy trader is as likely as any other to be have her order processed first. Since a slow maker's flow sniping losses are 2ν (because both lit bids and offers at BBO are vulnerable), her expected flow gain from purchasing speed is $\frac{2\nu}{N_s+1}$, while the flow cost is c. Recalling Equation (3.8), we see that this deviation is not worthwhile if

$$\frac{2\nu}{N_s+1} \leq c \iff 2\nu \leq (N_s+1)c \qquad \text{(A.34)}$$
$$\iff \nu \leq \rho\,\frac{1-\omega}{1+\omega} + \frac{c}{2}.$$

At baseline, $\nu = 1$ and $\rho\frac{1-\omega}{1+\omega} + \frac{c}{2} \approx 50\,\frac{0.77}{1.23} + \frac{10}{2} \approx 36$, so such a deviation is indeed highly unprofitable.

A.9 Unit Sniping Limitation

Assumption A1 ("... one indivisible unit at a time") can be construed as limiting each sniper to at most one snipe per jump in the fundamental value, V. We do not adopt that interpretation in the text, but note here the impact it would have on our model. Assuming a one-unit sniping capacity, the factor $\frac{N_r}{N_s}$ in Equation (3.4) is replaced by $\min\left\{1, \frac{N_r}{N_s}\right\}$, and Equations (3.5) requires a similar modification. In that case, when $2\nu < c$, snipers necessarily earn negative profit, so in equilibrium there are $N_s = 0$ snipers and $N_r = +\infty$ market makers. However, in the less expensive sniping case $2\nu \geq c$, the formulas in the text are unaffected by the alternative interpretation of A1.

B Dynamics

We now analyze potential strategic behavior of investors immediately following publicly observed transactions at best bid or offer. As noted in Section 3.6, such events reveal information about the state of the midpoint order queue. We utilize data provided by IEX, similar to that used in Section 5 to estimate the discrete Laplace parameter, ω. Specifically, the data consist of total time (in nanoseconds, during December 2016) spent in each state of the midpoint peg order queue, k, following transactions at BBO, measured over discrete windows of time. Those time windows are 100 microseconds, 100 milliseconds, 1 second, 10 seconds, and 60 seconds. As the data is split on a BB/BO conditioning variable, and since the data are symmetric across that variable, we focus our analysis on strategic behavior following BB transactions.

Panel (a) Figure 5 depicts the empirical distributions (fraction of time) of midpoint pegs for each time window, following BB transaction events. To account for the small volume of

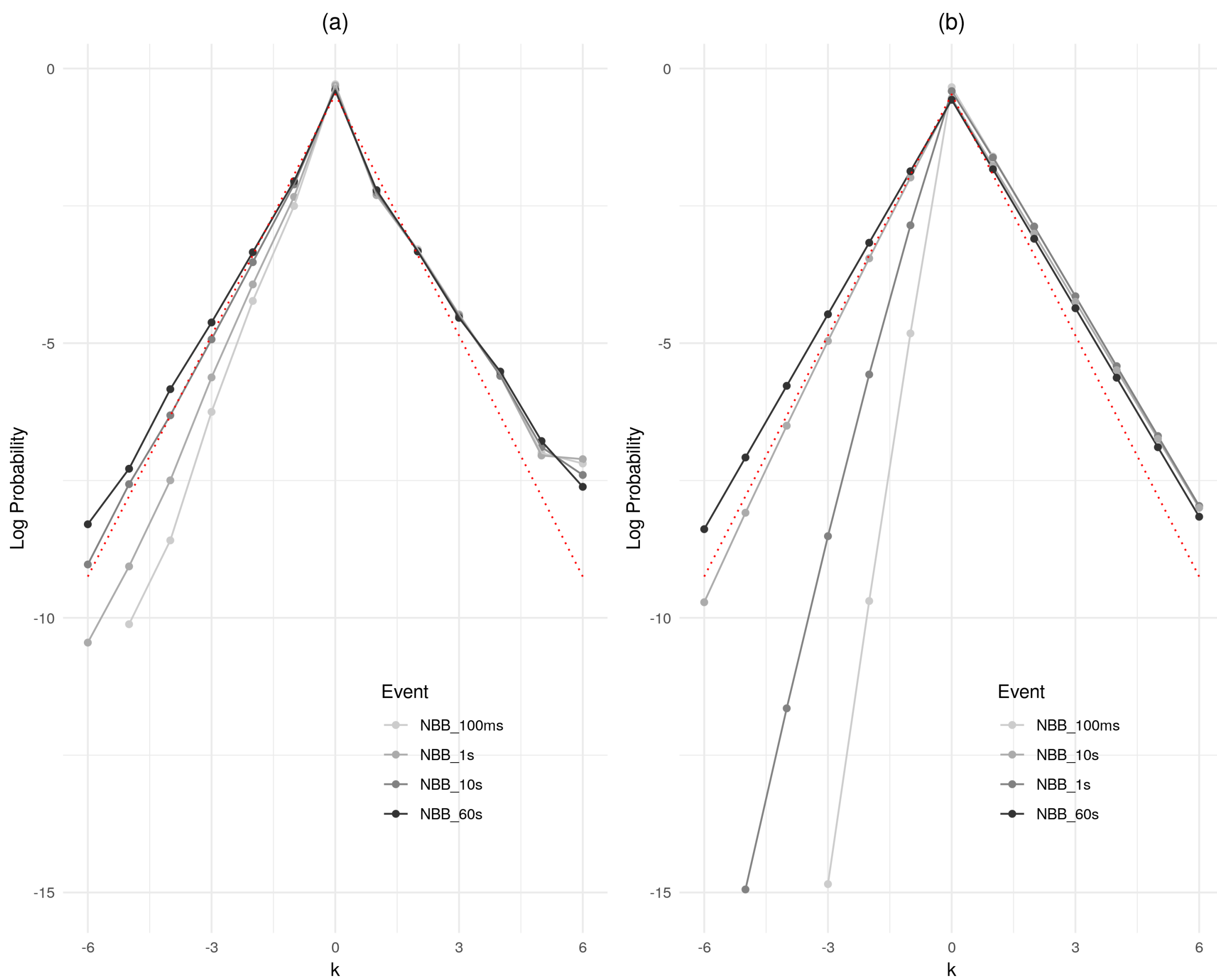

Figure 5: Dynamic response to BB transaction, empirical (Panel a) and theoretical (Panel b). The dotted red lines indicate the steady state distribution for $\omega = .23$.

residual midpoint bid orders remaining in the queue due to size restrictions[7] at the moment of transaction events, we establish the 100 microsecond midpoint bid distribution as baseline and use it to normalize the empirical distributions at the other horizons. Panel (b) shows the theoretical distribution for the same time horizons, obtained by iterating on Equation (A.5), with $\rho = 50$.

Aggregating the data for each time horizon, we are able to infer that there are approximately 30,000 BB events in the data. These events, however, are not independent. Using the 62 days of SPY transaction data associated with the quotations referenced in Section D.4[8] (for calibration of the jump intensity), we estimate the autocorrelation functions (ACFs) of inter-trade durations on a daily basis. The 62 autocorrelation functions are depicted in Figure 6, along with 0.025, 0.5, and 0.975 quantiles (dotted lines) at each lag. The transaction data exhibits clear persistance for long horizons – during the sample period, there was roughly 1 transaction every 1.5 seconds, which means that the 100-lag horizon of Figure 6 corresponds to roughly 2.5 minutes of clock time. Following the methodology of Thiebaux and Zwiers (1984), we use the ACF values $\{\rho_i\}_{i=1}^{M}$, truncated at $M = 1000$ lags, to estimate an effective sample size (ESS) ratio for each day,

$$\zeta_t = \frac{N_t}{N_t^e} = 1 + 2\sum_{i=1}^{N_t-1} \rho_i, \tag{B.1}$$

where N_t^e and N_t represent the effective and actual sample sizes (respectively) for day t. Taking the median of ESS ratios, we find $median\{\zeta_t\} \approx 115$. We thus approximate the effect number of BB transaction events as $N^e = \frac{30{,}000}{115} \approx 260$.

Denoting $K = |k|$, Kozubowski and Inusah (2006) show that the maximum likelihood estimators of the asymmetric discrete Laplace parameters are

$$\hat{\omega}_L = \begin{cases} \frac{2\bar{K}^-(1+\bar{K})}{1+2\bar{K}^- - \bar{K} + \sqrt{1+4\bar{K}^-\bar{K}^+}}, & \text{if } \bar{K} \geq 0 \\ \frac{\hat{\omega}_R - \bar{K}(1-\hat{\omega}_R)}{1-\bar{K}(1-\hat{\omega}_R)}, & \text{otherwise,} \end{cases} \tag{B.2}$$

[7] For December 2016, 0.866% (resp. 0.281%) of the time both bid and offer m-pegs (resp. d-pegs) coexisted, due mainly to size restrictions preventing a fill against a contraside order at midpoint

[8] The data provided by IEX is not transaction-level resolution. However, given the tight correlation of transactions across U.S. equities exchanges, we believe that autocorrelation functions estimated with Nasdaq transaction are a good surrogate for transaction autocorrelations in all markets, including IEX.

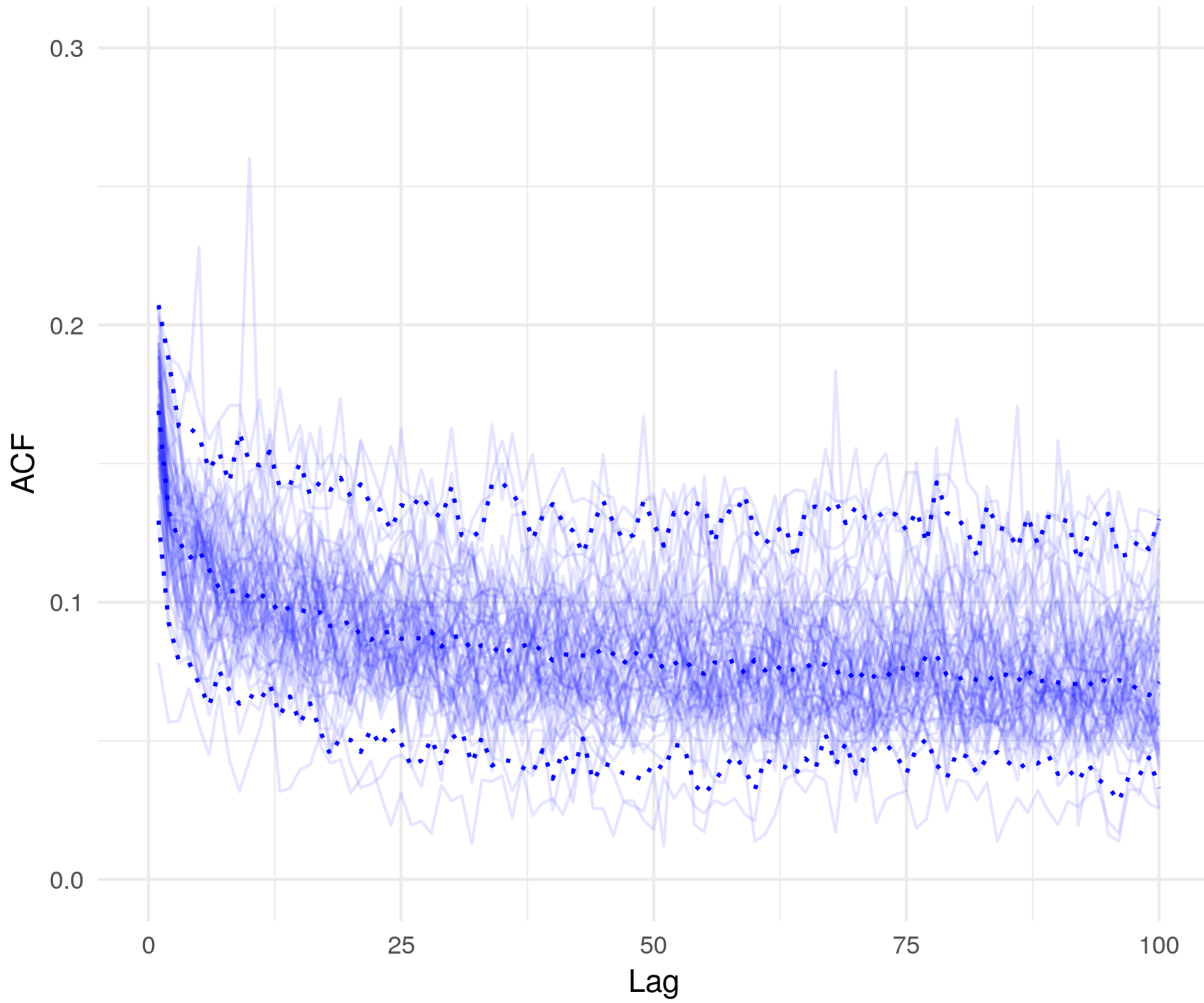

Figure 6: Daily autocorrelation functions of inter-trade durations for SPY transactions on the Nasdaq exchange, 16, Jun 2014 – 11 Sep, 2014. Dotted lines represent 0.025, 0.5, and 0.975 quantiles at each lag.

$$\hat{\omega}_R = \begin{cases} \frac{2\bar{K}^+(1-\bar{K})}{1-2\bar{K}^+\bar{K}+\sqrt{1+4\bar{K}^-\bar{K}^+}}, & \text{if } \bar{K} \leq 0 \\ \frac{\hat{\omega}_L+\bar{K}(1-\hat{\omega}_L)}{1+\bar{K}(1-\hat{\omega}_L)}, & \text{otherwise,} \end{cases} \tag{B.3}$$

$$\Sigma_{MLE} = \frac{\omega_L\omega_R(1-\omega_L)(1-\omega_R)}{1+\omega_L\omega_R} \begin{bmatrix} \frac{(1-\omega_R)(1-\omega_R\omega_L^2)}{\omega_L(1-\omega_L)^2} & 1 \\ 1 & \frac{(1-\omega_L)(1-\omega_L\omega_R^2)}{\omega_R(1-\omega_R)^2} \end{bmatrix}, \tag{B.4}$$

where $\bar{K} = \sum_{k\in\mathcal{K}} w_k\,|k|$, $\bar{K}^- = \sum_{k\in\mathcal{K}} \mathbb{1}(k \leq 0) w_k\,|k|$, and $\bar{K}^+ = \sum_{k\in\mathcal{K}} \mathbb{1}(k \geq 0) w_k\,|k|$, and where $\mathcal{K} = \{-6, \ldots, 0, \ldots, 6\}$.

We use the delta method to estimate $\hat{\alpha} = \hat{\omega}_L/\hat{\omega}_R$ for each of the four time windows in the data provided by IEX. The point estimates are represented by blue dots in Figure 7. We also compute 95% asymptotic error bands, scaling the asymptotic variance by the effective sample size and expected number of investor arrivals during each time window, $2\rho N^e$. The 95% intervals are represented by blue diamonds and vertical blue dotted lines. Additionally, we obtain theoretical probabilities of the order queue distribution by iterating on Equation (A.5) for a fine sequence of $t \in (0, 60)$ seconds, and likewise obtain point estimates and asymptotic errors for the theoretical distribution. The theoretical point estimates are depicted as a solid black line in Figure 7, and the 95% interval estimtes are depicted as black dotted lines. For instances where the confidence bands span zero (i.e. for low values of t), we truncate the lower bounds at zero. Given the remarkably large overlap in confidence intervals, we conclude that the ADL is a good approximation both for our model and the data.

The foregoing analysis suggests that, indeed, publicly observed transactions at BBO provide useful strategic information to investors. To understand the economic importance of this information, we estimate the proportional increase in profit that a single investor could earn during the 60 seconds following all BBO transactions in December 2016. Specifically, we integrate Equation A.13 over $t \in (0, 60)$ seconds, substuting our theoretical point estimates, $\hat{\alpha}(t)$, in Figure 7. The proportional increase in profit, relative to equilibrium profit in Equation (3.3), for 30,000 60-second periods is 0.56% (0.0056).

C Institutional Information

C.1 Exchanges Imposing Delay

Several exchanges impose messaging delays on their systems. On May 16, 2017, nearly a year after the SEC approval of IEX to operate as a national securities exchange, NYSE American (formerly NYSE MKT) received similar approval to impose a 350 microsecond delay to all inbound and outbound messages in its system. Much like IEX, the delay protects non-displayed pegged orders, which includes a discretionary pegged order type (nearly identical to the IEX discretionary peg) that was approved by the SEC in June, 2016.

Several months later, the Chicago Stock Exchange (CHX) also received approval to impose a 350 microsecond delay, a decision that was later stayed by the SEC (and which is

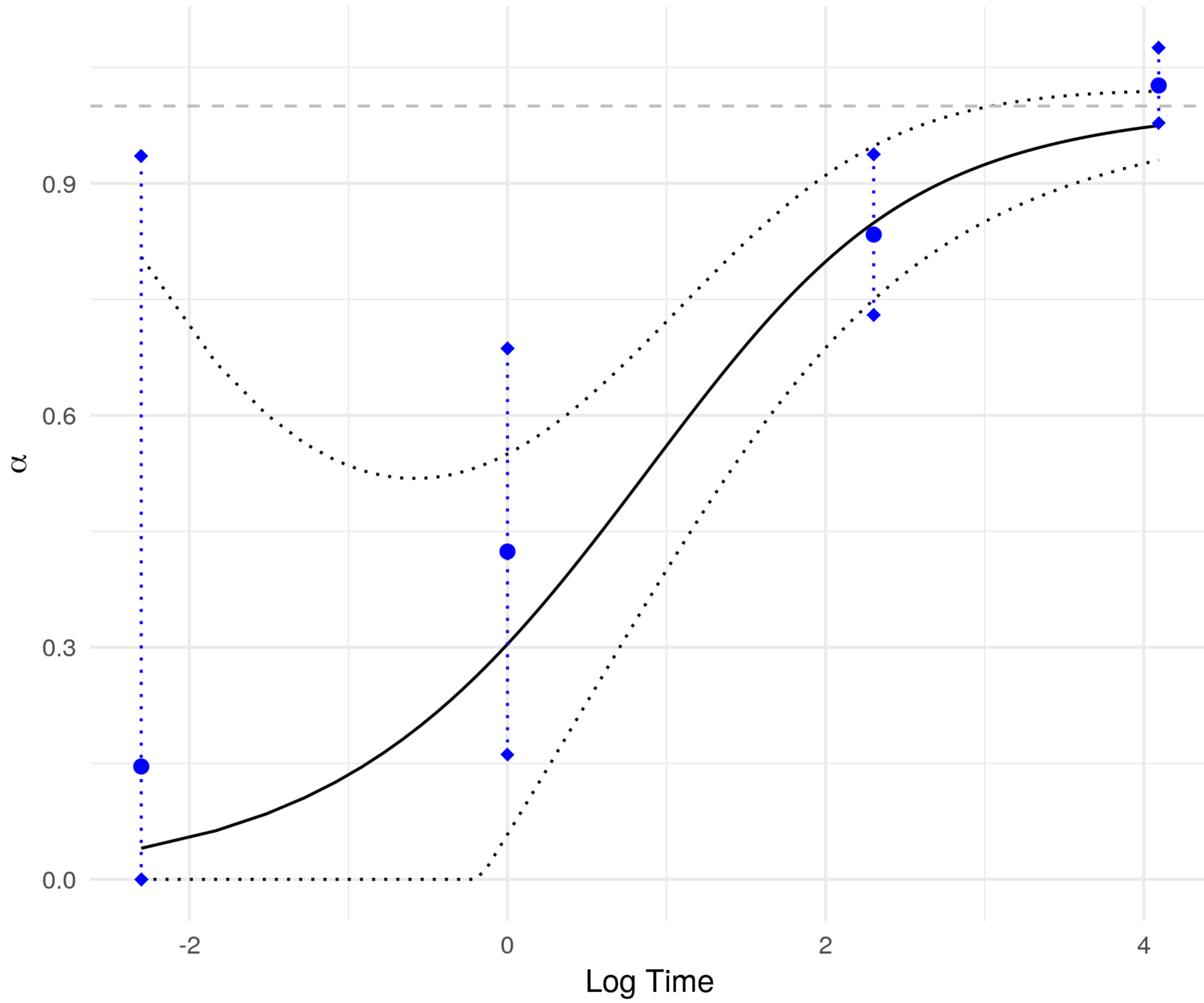

Figure 7: Decay function, theoretical vs empirical. Theoretical prediction from Equation (A.5) for $\rho = 50$ (solid black line) with 95% asymptotic confidence bands (black dotted lines), vs estimates from the December 2016 IEX data (blue dots) and 95% asymptotic confidence bands (blue diamonds and dotted lines).

pending approval at the time of this writing). Unlike the predecessor systems mentioned above, the CHX messaging delay would protect all pegged orders, not only those that are non-displayed. This system, referred to as Liquidity Enhancing Access Delay (LEAD), would also allow limit and cancel orders sent by specially designated market makers to be exempt from the delay. To obtain LEAD market maker status, traders would be subject to specific month-to-month liquidity provision and transaction requirements.

Unlike the foregoing systems, TSX Alpha, launched in September 2015, imposes a longer, random delay of 1 – 3 milliseconds. Like CHX, the TSX messaging delay protects all pegged orders. Additionally, "post-only" limit orders are not subject to the delay. Post-only orders enter the order book as traditional limit orders, but in the event that they cross a standing quotation, they are either repriced (less aggressively) or cancelled. TSX Alpha also uses an inverted taker-maker fee structure, issuing a rebate ($0.0010) to traders taking liquidity and charging fees ($0.0014 – $0.0016 for post-only limits and $0.0013 – $0.0014 for non-post-only limits) to traders providing liquidity. As a result, traders may surpass the delay by paying an explicit fee to the exchange.

C.2 Order Routing

In accordance with Regulation National Market System (Reg NMS), all exchanges in the United States route orders to protected quotations at other exchanges when those quotations offer price improvement. The IEX router does this both at initial receipt of an order, and at periodic intervals for orders resting on the book. The latter feature is referred to as resweep. To be eligible for such protection, orders must be designated as "routable", whereas "nonroutable" orders are sent directly to the IEX book and are not eligible for resweep.

The order book and router are distinct components of the IEX system. After passing through the initial 350 microsecond point-of-presence delay, nonroutable orders are sent directly to the IEX order book, whereas routable orders are sent to the router. The IEX order router then disseminates these latter orders to all national market systems (including their own) following a proprietary routing table. Messages that are passed between the IEX order book and router are subject to an additional one-way 350 microsecond delay. As a result, routable orders that are sent to the IEX order book experience a cumulative delay of 700 microseconds before queuing behind other orders in the system. No additional delay is enforced between the IEX router and external exchanges.

As noted in Section 2, routable orders constitute only 15% of IEX trading volume and represent traders that use the IEX router as an access point to the national market system. The remaining, nonroutable volume, represents trading interest intended to capture incentives of the IEX market design.

C.3 Pegged Order Types

Section 2 lists the three types of pegged orders at IEX. Midpoint pegs rest at the midpoint of NBBO, whereas primary pegs are booked in the hidden order queue one price increment (typically $0.01) below (above) NBB (NBO), and are promoted to transact at NBB or NBO if sufficient trading interest arrives at those prices. Discretionary pegs combine the benefits of these first two: when entering the order book, they check the NBBO midpoint for contra-side interest, but in the absence of such interest, are pegged to NBB or NBO and are queued behind other hidden orders at those prices. Further, in the event that contra-side interest subsequently arrives at the NBBO midpoint, discretionary peg orders can be promoted to transact at the midpoint. If no such interest arrives, discretionary pegs are treated as typical hidden NBBO orders.

Table 1 shows that midpoint trading constitutes a little more than 60% of volume, discretionary peg trading accounts for 37% of volume and 89% of discretionary pegs are transacted at the midpoint. The implication is that midpoint volume is nearly evenly split between midpoint and discretionary pegs. Primary pegs and discretionary pegs transacted at BBO each account for 5% or less of reported volume. Thus, while there is a distinction between midpoint and discretionary peg orders, in practice nearly all discretionary peg orders transact at midpoint. For this reason, we reduce the decision space for order types in our model to a simple midpoint peg.

Table 1 also reports small volume statistics for seemingly incongruous trades: (1) midpoint orders that transact at BBO and (2) hidden nonroutable orders (not pegs) that transact at midpoint. The first case occurs when midpoint pegs are booked with a limit price constraint which binds after subsequent movements in the NBBO. In such instances, an order that originally rested at midpoint might later rest and transact at BBO. The second case occurs under nuanced conditions where the NBBO is more than a single price increment wide or when the IEX BBO is wider than the NBBO (which may be a single increment). In such instances, the NBBO may coincide with the IEX midpoint or the hidden order at IEX may be subject to a special midpoint price constraint[9] and later transact with contra-side

[9] When the IEX BBO is wider than the NBBO and a nonroutable hidden order enters the order book with a limit that would otherwise be passed on to another exchange displaying NBBO, the order is booked at the NBBO midpoint and may be promoted to transact at the NBBO at a later time. For example, suppose the

orders at midpoint.

C.4 Crumbling Quote

The volume statistics for midpoint pegs in Table 1 show that proprietary firms are three times more likely to act as liquidity removers at midpoint (7.16% of volume) than as liquidity adders (2.12% of volume). This is indicative of opportunistic stale-quote arbitrage in advance of movements in the NBBO. Despite the fact that the IEX delay is intended to combat such exploitative activities, the company has reported an increase in anticipatory trading: midpoint quotes being removed at unfavorable prices immediately prior to changes in the NBBO (Bishop, 2017). This trading is almost certainly a result of improved probabilistic modeling of NBBO liquidity shifts by fast traders.

In an effort to further protect pegged orders from adverse selection, IEX has developed the "crumbling quote signal": a model that forecasts changes in the NBBO (the crumbling quote) and temporarily prevents primary and discretionary peg orders from exercising discretion at their potentially more aggressive prices in order to minimize their exposure to anticipatory traders. That is, when the crumbling quote signal is on, discretionary pegs do not transact at midpoint and primary pegs do not transact at BBO. Midpoint pegs do not receive protection from the crumbling quote signal.

While we view the crumbling quote signal as an important innovation to the IEX market design, we have excluded it from our model in order to focus attention on the primary role of the speed bump and its interaction with pegged order types. We consider study of the crumbling quote signal, however, to be a valuable direction for future work.

NBBO is \$10.00 × \$10.01 and the IEX order book is \$10.00 × \$10.02 when a nonroutable hidden buy order arrives with a limit of \$10.01. The order will be booked at \$10.005 and will later transact at \$10.01 if a sell limit arrives at that price. Alternatively, it may transact with midpoint pegs, discretionary pegs, or market orders at midpoint.

D Calibration

First we explain how our baseline parameter values connect with available market data. Then we collect details on short-run order queue dynamics after a BBO transaction.

D.1 Investor Fraction ω^*

Recall that ω is the fraction of investor orders transmitted as midpoint pegs, and that Equation (3.3) tells us that, for $\xi = 0$, the steady state probability that there is a contra-side order resting at midprice is

$$P = \sum_{k=1}^{\infty} q_k = \frac{\omega}{1+\omega}, \tag{D.1}$$

since q_k is from the discrete Laplace distribution with parameter ω. Using data given to us by IEX, as described in Section 5, we estimate ω via maximum likelihood, via Equation (5.1):

$$\hat{\omega} = \frac{|\bar{k}|}{1 + \sqrt{1 + |\bar{k}|^2}} = 0.23., \tag{5.1}$$

and make this the baseline value.

D.2 Cost of Speed, Midprice Transaction Fee and Investor Surplus

At the time of this writing, one of the premier microwave transmission services, McKay Brothers LLC, offers low latency data services for 8 select ETFs (such as SPY) for \$3,100 per month. This translates to $\$3100/(8 \times 8190) = \$0.047$ or approximately $c = 10$ half-spreads per symbol, per minute.

The IEX fee for transacting at the mid price is \$0.0009. As a single price unit in our model is equivalent to \$0.005, we set $d = 0.18$ price units.

We define φ as the surplus for the marginal investor with impatience β^* (defined below). Such an investor is just willing to transmit a market order at unit cost (0.5 spreads or pennies) in addition to the direct fee, b, of \$0.003 – \$0.005 (an approximation reported to us by practitioners) per share. The direct fee is equivalent to 0.6 – 1 half-spreads, so $\varphi \approx 1 + 0.8 = 1.8$ half-spreads.

D.3 Discount Factor

Suppose each investor i has private impatience parameter $\beta_i \in [0,1]$, drawn independently from a given distribution $F(\beta)$. In practice, investors choose from a long menu of broker algorithms for placing orders, and their choices partially reveal their values of β_i.

In our model, investors only choose between midpoint pegs and market orders, implying a threshold, $\tilde{\beta}$, such that more patient investors (those with $\beta_i > \tilde{\beta}$) choose pegs and less patient investors choose market orders. Thus, given $\tilde{\beta}$, a fraction $\omega = 1 - F(\tilde{\beta})$ of the orders are transmitted as pegs.

Our steady state distribution of order imbalances (Proposition 3.1) implies a distribution of waiting times, and thus expected investor profits $\pi_i(\theta|\omega, \beta_i)$, for order types $\theta \in \{\text{peg}, \text{mkt}\}$. By maximizing over θ (choosing the preferred order type) we obtain a new threshold $\tilde{\beta}'$. The result is a map $M : [0,1] \to [0,1], \quad \tilde{\beta} \mapsto \tilde{\beta}'$.

Lemma D.1. *If the distribution F is continuous, then the mapping M, defined above, has a unique fixed point $\beta^* \in [0,1]$.*

Proof sketch. M is continuous and monotone decreasing, so the conclusion follows from the intermediate value theorem.

This result allows us to infer β^* from our calibration of ω^* and the other parameters: given vector (ω^*, φ, d) we use the equal profit condition for the marginal investor, Equation (3.9), to solve

$$\begin{aligned}\beta^* &= \frac{\varphi - 1}{\varphi - d - \omega^*(1-d)} \\ &= \frac{1.8 - 1}{1.8 - 0.18 - 0.23 \times (1 - 0.18)} \\ &= 0.56. \end{aligned} \tag{D.2}$$

Substituting $\rho = 50$ (determined below) into the relation $\beta^* = \exp(-\delta/\rho)$ we arrive at $\delta = -\rho \log(\beta^*) \approx 30$.

D.4 Arrival Intensities

Table 3 reports quantiles of the distribution interarrival times (measured in nanoseconds) of submitted orders for the S&P 500 exchange traded fund (ticker SPY) at IEX during December, 2016. The distibution is broken down across several classifications. Investors

	Order Type								
Quantile	All	Buys	Sells	D-peg	M-peg	P-peg	Other	Agency	Proprietary
0.05	25172	89444	87947	2951559	920231	24185	20647	352325	20576
0.10	115862	299044	317757	16750721	4968059	1195898	96937	1393549	100501
0.25	1191343	3978118	5160039	308253568	31011051	71938785	1063259	16961911	1151426
0.50	13556983	37724572	29199640	1410139781	**404316291**	599993574	13477677	**248146134**	13558630
0.75	99397865	263438679	208858208	4012264653	1592582790	2022289139	109005668	1022688976	103098092
0.90	393317550	799390256	745466029	5044062194	4244552991	4965817914	462395201	2664485573	439602722
0.95	712558170	1345011543	1300901230	12282108869	6835804561	8038904707	872135984	4133503605	825720463

Table 3: Quantiles of the distribution of interarrival times (measured in nanoseconds) of submitted orders for the S&P 500 exchange traded fund (ticker SPY) at IEX during December, 2016.

in our model place midpoint pegs and market orders. Unfortunately, none of the values in Table 3 correspond precisely with investor orders in our model: the interarrival times for midpoint pegs include those of investors and proprietary traders, whereas the interarrival times for investor orders (labelled 'Agency') include all order types. We thus approximate the intensity of investor arrivals in the following manner. Table 1 reveals that investors (agencies) account for 65.7% of all midpoint peg transactions. Using this value to extrapolate the share of order submission in Table 3, we find that the median time between investor arrivals is approximately $404/0.657 = 615$ milliseconds, which equates to 1.62 arrivals per second or roughly 100 arrivals per minute on both sides of the market. This suggests a baseline arrival rate for each side of the market of $\rho = 100/2 = 50$. A similar approximation, using the median interarrival time of 248 milliseconds for agency orders, and the share of volume in Table 1 for the agency order types in our model, yields a nearly identical value of ρ.

To calibrate ν we utilize SPY quotation data at Nasdaq, which, given its liquidity and overall market share, is a good surrogate for the SPY NBBO. Our sample covers the period 16 June – 11 September, 2014. There are 26,216,524 quotations in the 62-day period, which comprises 1,450,800,000 milliseconds during trading hours, or approximately 1 quote every 55 milliseconds. Defining a jump as any midpoint price change of at least \$0.01 which is not reversed over the subsequent period of four quotations[10], or 220 milliseconds, resulted in an average of 733 jumps per day, 1.88 jumps per minute, or one jump every 32 seconds. As with the investor arrival intensity parameter, ν represents the intensity of jumps on one side of the market. Thus, we set our baseline calibration to be $\nu = 1$.

Combining the values of ρ and ν, our baseline measures suggest $\frac{\nu}{\rho} \approx \frac{1}{50}$, or that the intensity of investor arrivals is about 50 times that of jumps.

[10]We also considered shorter post-jump intervals, with little change in total counts. Additionally, we applied a different methodology which counted midpoint price changes over non-overlapping intervals of fixed lengths (100,200,300,400 milliseconds) and found the jump counts to be quite stable across methodologies and interval choice.

References

Aldrich, E. M. and Lee, S. (2016), "Relative Spread and Price Discovery," *Working Paper*, 1–35.

Baldauf, M. and Mollner, J. (2016), "Fast Traders Make a Quick Buck: The Role of Speed in Liquidity Provision," *Working Paper*, 1–56.

Bershova, N. and Rakhlin, D. (2012), "High-Frequency Trading and Long-Term Investors: A View from the Buy Side," *Working Paper*.

Bishop, A. (2017), "The Evolution of the Crumbling Quote Signal," *IEX White Paper*, 1–30.

Breckenfelder, J. (2013), "Competition Between High-Frequency Traders, and Market Quality," *Working Paper*.

Brogaard, J. and Garriott, C. (2017), "High-Frequency Trading Competition," *Working Paper*.

Brogaard, J., Hendershott, T., Hunt, S., and Ysusi, C. (2014), "High-frequency trading and the execution costs of institutional investors," *Financial Review*, 49, 345–369.

Brogaard, J., Hendershott, T., and Riordan, R. (2017), "Price Discovery without Trading: Evidence from Limit Orders," *Working Paper*, 1–51.

Brolley, M. and Cimon, D. (2018), "Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays," *Working Paper*, 1 – 38.

Budish, E., Cramton, P., and Shim, J. (2015), "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, 130, 1547–1621.

Buti, S., Consonni, F., Rindi, B., Wen, Y., and Werner, I. M. (2015), "Sub-Penny and Queue-Jumping," *Working Paper*, 1–52.

Buti, S., Rindi, B., and Werner, I. M. (2017), "Dark pool trading strategies, market quality and welfare," *Journal of Financial Economics*, 124, 244–265.

Chen, H., Foley, S., and Ruf, T. (2017), "The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets," *Working Paper*.

Copeland, T. E. and Galai, D. (1983), "Information Effects on the Bid-Ask Spread," *The Journal of Finance*, 38, 1457–1469.

Du, S. and Zhu, H. (2017), "What is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*.

Easley, D., Lopez de Prado, M. M., and O'Hara, M. (2012), "Flow Toxicity and Liquidity in a High-frequency World," *Review of Financial Studies*, 25, 1457–1493.

Foucault, T. (1999), "Order flow composition and trading costs in a dynamic limit order market," *Journal of Financial Markets*, 2, 99–134.

Fox, M. B., Glosten, L. R., and Rauterberg, G. V. (2015), "The New Stock Market: Sense and Nonsense," *Duke Law Journal*, 65, 191–277.

Glosten, L. R. and Milgrom, P. R. (1985), "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 14, 71–100.

Hagströmer, B., Nordén, L., and Zhang, D. (2014), "How Aggressive Are High-Frequency Traders?" *Financial Review*, 49, 395–419.

Hasbrouck, J. and Saar, G. (2013), "Low-latency trading," *Journal of Financial Markets*, 16, 646–679.

Hirschey, N. (2017), "Do high-frequency traders anticipate buying and selling pressure?" *Working Paper*.

Hoffmann, P. (2014), "A dynamic limit order market with fast and slow traders," *Journal of Financial Economics*, 113, 156–169.

Inusah, S. and Kozubowski, T. J. (2006), "A discrete analogue of the Laplace distribution," *Journal of Statistical Planning and Inference*, 136, 1090–1102.

Jovanovic, B. and Menkveld, A. J. (2015), "Middlemen in Limit Order Markets," *Working Paper*.

Kozubowski, T. J. and Inusah, S. (2006), "A skew laplace distribution on integers," *Annals of the Institute of Statistical Mathematics*, 58, 555–571.

Kyle, A. S. and Lee, J. (2017), "Toward a Fully Continuous Exchange," *Working Paper*.

Lewis, M. (2015), *Flash Boys: A Wall Street Revolt*, New York: W. W. Norton & Company.

Malinova, K., Park, A., and Riordan, R. (2014), "Do retail traders suffer from high frequency traders?" *Working Paper*.

Menkveld, A. J. and Zoican, M. A. (2017), "Need for speed? Exchange latency and liquidity," *Review of Financial Studies*, 30, 1188–1228.

O'Hara, M. (2015), "High frequency market microstructure," *Journal of Financial Economics*, 116, 257–270.

Pisani, B. (2016), "SEC gives its blessing to the IEX's 'speed bump' trading," *http://www.cnbc.com/2016/06/17/sec-gives-its-blessing-to-the-iexs-speed-bump-trading.html*.

Plott, C., Roll, R., Seo, H., and Zhao, H. (2018), "Tick Size, Price Grids and Market Performance: Stable Matches as a Model of Market Dynamics and Equilibrium," *Working Paper*.

SEC (2014), "Equity Market Structure Literature Review Part II: High Frequency Trading," *Staff Report*.

Thiebaux, H. and Zwiers, F. (1984), "thiebaux-etal-1984.pdf," *Journal of Climate and Applied Methodology*, 23.

Wah, E., Hurd, D. R., and Wellman, M. P. (2015), "Strategic Market Choice: Frequent Call Markets vs. Continuous Double Auctions for Fast and Slow Traders," *Working paper*.

Werner, I., Wen, Y., Rindi, B., Consonni, F., and Buti, S. (2015), "Tick Size: Theory and Evidence," *Working Paper*, 1–60.

Zhang, S. and Riordan, R. (2011), "Technology and Market Quality: The Case of High Frequency Trading," *ECIS 2011 Proceedings*.