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

July 11, 2018

Abstract

Several financial exchanges recently introduced messaging delays (e.g., the 350 microsecond delay at 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 features a discrete price grid and pegged orders, and is solved in closed form. It shows how messaging delay can indeed lower transactions costs but typically increases queuing costs. Some of the model's many comparative static predictions are testable in the field and others are testable in a laboratory environment.

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

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 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 Experimental Finance Workshop at the Max Plank Institute at Bonn and the Expbot Workshop in Luxembourg (June, 2018). 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), 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.[1]

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 uniform 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

[1] 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.

(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 Budish et al. (2015), but requires a discrete price grid and the distinction between lit (publicly displayed) and hidden orders. Both are features of Buti et al. (2017), who 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 describes order types, the traditional continuous double auction (CDA) format, and the variation of CDA 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. Section 4 obtains parallel results for CDA markets while also checking robustness by relaxing some of the more restrictive assumptions. Section 5 presents comparative statics results and introduces performance metrics to analyze 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 Appendices A, B and C.

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 (CLOB). 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. 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, 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 small) orders can be placed arbitrarily close together, and (b) 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[2] 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 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 B.2 shows 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.

[2]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.

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.

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

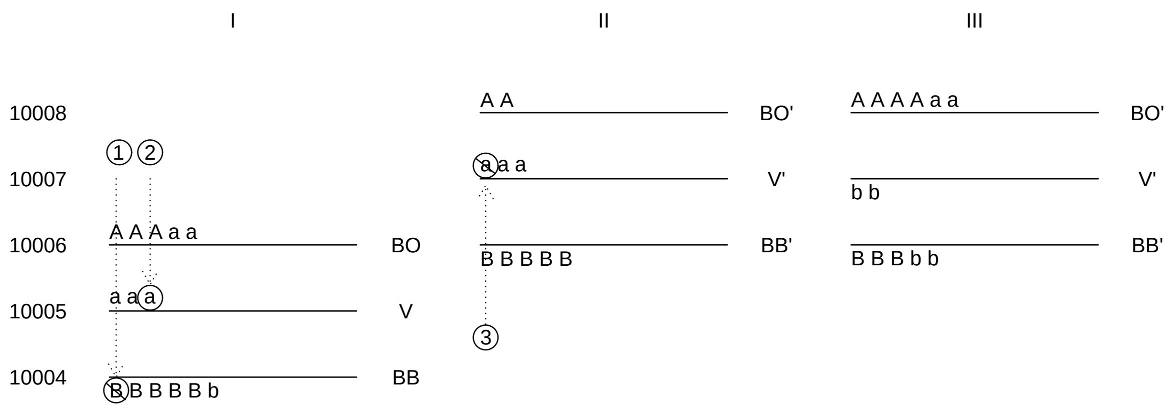

Figure 1: Example states in IEX market. Uppercase (resp. lowercase) denotes lit (resp. hidden) orders, B/b for buy and A/a for sell; those to the left have higher priority at that price. Panel I: initial state is $k = 2$ and $V = 10005$ half-pennies (i.e., $V = \$50.025$ per share); event (1) is a market 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.

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 $q : \mathbb{Z} \to (0,\infty)$ 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 symmetric around a sharp peak at zero, and drops off 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., to 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

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

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

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

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 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) 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. $\square$

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. The fraction of pegged orders, however, 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 number 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 number 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 number of snipers is inversely related to the cost of fast communication technology.

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 $\tilde{\tau}_{fast} = \tau_{fast} + \eta > \tau_{SIP}$. Thus, in principle there is a threshold messaging delay, η^*, that separates the equilibrium in Section 3 from that in a traditional CDA. To compress notation, define the composite binary 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.

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. Midpoint peg bids are treated the same way as offers.

The steady state distribution of order imbalance is different than in the protected case, because sniping now induces transitions $k \to 0$. The following proposition, proved in Appendix A using queuing theory and difference equations, generalizes the stationary distribution of k 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 $\tilde{q} : \mathbb{Z} \to (0, \infty)$ of the order imbalance, with*

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

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).

Note that (4.3) collapses to $\lambda = \omega$ in the protected case $\xi = 0$.

We proceed as before to obtain the equilibrium value of λ, and thus ω.

Peg. Following Proposition 4.1 and the preceding discussion, one can verify that Equation (3.2), the expected net surplus for a midpoint peg order, generalizes to:

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

Market order. An investor choosing a market order will earn the same expected net surplus as in Equation (3.3) with $\tilde{q}_k$ replacing q_k:

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

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

Sniper profit. 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

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

Market maker profit. The conditions for market makers are unchanged, so (3.5) still characterizes their profitability. The only difference is in the equilibrium fraction of investors choosing pegged orders, as we now show.

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

$$N_p^* = \frac{\tilde{\lambda}}{1-\tilde{\lambda}^2} \qquad \textit{and} \tag{4.8}$$

$$\begin{aligned}\tilde{\lambda} = \frac{1}{2\beta^2\rho(1-d)} \Bigg[&\beta^2\rho(\varphi-d) - \beta\xi\nu(\varphi+1) - \beta\rho(\varphi+d-2) \\ &- \Big(\big(\beta\xi\nu(\varphi+1) + \beta\rho(\varphi+d-2) - \beta^2\rho(\varphi-d)\big)^2 \\ &\quad - 4\beta^2\rho(1-d)\left(\beta\rho(\varphi-d) - \rho(\varphi-1) - \xi\nu(\varphi-1+2\beta)\right) \Big)^{1/2} \Bigg]_+ .\end{aligned} \tag{4.9}$$

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:

$$\frac{\beta\rho(1-\lambda)(\varphi-d)}{\rho+\xi\nu-\beta\rho\lambda} - \frac{2\beta(1-\lambda)}{1-\beta\lambda} + \frac{2\beta\rho(1-\lambda)}{\rho+\xi\nu-\beta\rho\lambda} = \varphi - 1. \tag{4.10}$$

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

$$(1-\beta)(\varphi-1)(\nu+\rho(1+\beta)) < 0 (> 0) \tag{4.11}$$

The parameter restrictions ensure that the left-hand-side of Equation (4.11) is nonnegative, so the relevant solution involves the negative discriminant, which is written in Equation (4.9). Equation (4.7a) follows from Corollary A.1 in Appendix A.

Equation (4.8), which gives the expected number of pegged orders vulnerable to sniping, N_p^*, is obtained as follows:

$$N_p^* = \sum_{k=-\infty}^{0} 0\tilde{q}_k + \sum_{k=1}^{\infty} k\tilde{q}_k = \frac{1-\tilde{\lambda}}{1+\tilde{\lambda}}\sum_{k=1}^{\infty} k\,\tilde{\lambda}^k = \left(\frac{1-\tilde{\lambda}}{1+\tilde{\lambda}}\right)\frac{\tilde{\lambda}}{(1-\tilde{\lambda})^2} = \frac{\tilde{\lambda}}{1-\tilde{\lambda}^2}. \tag{4.12}$$

Equations (4.7b) and (4.7c) follow by substituting (4.7a) and (4.8) 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 (4.9) means that $\tilde{\lambda}$ and hence (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. □

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

$$\hat{\lambda} = 1 - \xi \frac{(\varphi + 1)}{(1 - d)} \frac{\nu}{\rho}, \tag{4.13}$$

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

Proof. When $\beta = 1$, Equation (4.10) simplifies to

$$(\varphi - d)(1 - \lambda)\rho + 2(1 - \lambda)\rho = (\varphi + 1)(\rho + \xi\nu - \rho\lambda) \tag{4.14}$$

from which (4.13) follows. □

5 Comparative statics

The impact of offering midpoint peg protection can be determined via Proposition 4.2. For any admissible parameter vector, the predicted impact on the order queue and the number of snipers and market makers is obtained by subtracting the equilibrium expressions evaluated at $\xi = 1$ from those evaluated at $\xi = 0$. For such exercises, it helps to have a common starting point, or baseline parameter vector. A casual look at financial market data, summarized in Appendix B, leads us to baseline values $(c, d, \varphi, \beta, \rho, \nu) = (10, 0.18, 1.8, 0.80, 7, 1)$.

5.1 Performance metrics

What are the welfare implications of delayed messaging? Does protection enhance market performance? We take the investor's perspective on these questions, 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 focus on two performance metrics relevant to investor welfare.

Transaction cost. Investors pay brokerage fee b, which is typically 0.6 to 1.0 half spread (\$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 1 half spread in the form of worse execution price. Thus for a market order, the per-share mean transaction cost is

$$TC = b + d \cdot P + 1 \cdot (1 - P) = b + \frac{1 + d\lambda}{1 + \lambda}. \tag{5.1}$$

In market equilibrium, TC will be the same for either type of order when $\omega > 0$, so equation (5.1) 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,

$$\begin{aligned} QC &= \omega \sum_{k=0}^{\infty} \tilde{q}_k \left[1 - \left(\frac{\beta\rho}{\rho + \xi\nu}\right)^{k+1}\right] \\ &= \frac{\omega}{1+\lambda} - \frac{\omega\beta\rho}{\rho + \xi\nu} \left(\frac{1-\lambda}{1+\lambda}\right) \sum_{k=0}^{\infty} \left(\frac{\beta\rho\lambda}{\rho + \xi\nu}\right)^k \\ &= \frac{\omega}{1+\lambda} \left(\frac{\rho + \xi\nu - \beta\rho}{\rho + \xi\nu - \beta\rho\lambda}\right). \end{aligned} \tag{5.2}$$

We see no natural way to combine the two metrics into an overall welfare measure; their relative importance would seem to 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.)

5.2 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.

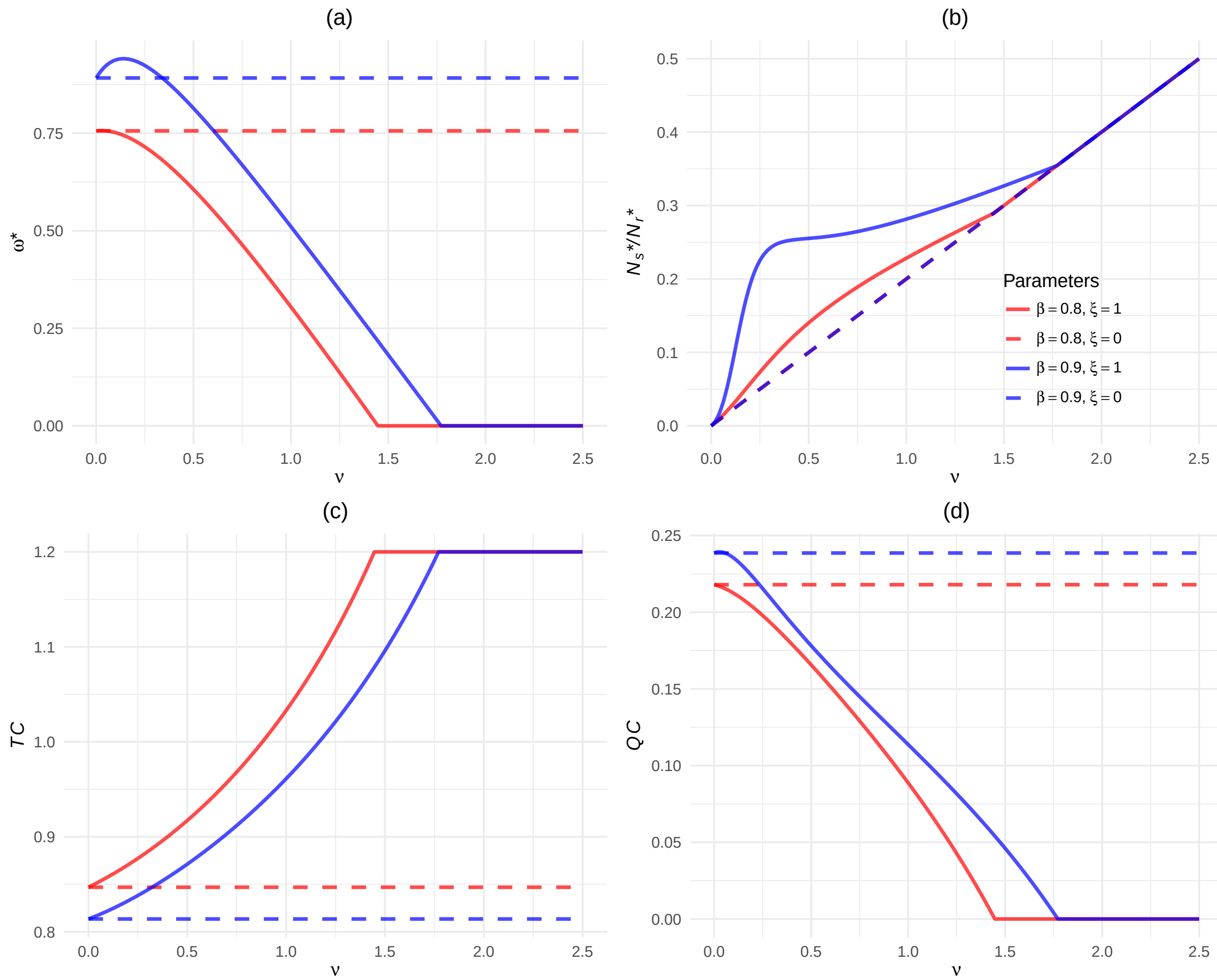

Figure 2: Impact of ν on equilibrium ratios and performance metrics. Other parameters are held fixed at baseline values, except that blue lines show impact when $\beta = 0.9$ instead of baseline $\beta = 0.8$. Dotted lines show values for $\xi = 0$ and solid lines for $\xi = 1$. Panel (a) shows the equilibrium 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.

Figure 2 depicts those equilibrium ratios and performance metrics as we vary the fundamental jump arrival rate $\nu \in (0, 2.5)$ with investor arrival rate ρ held constant at its baseline value of 7. Panel (a) shows that with protection, the equilibrium share ω^* of pegged orders is independent of the jump rate ν; the horizontal dashed red line shows that it remains at its baseline value 0.756 and the dashed blue line shows that it is a bit higher when investors are more patient. The solid lines show that when protection is removed, $\xi = 1$, midpoint pegs disappear for $\nu > 1.45$ in the baseline, and for a somewhat higher value 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 increases equilibrium pegged orders from minority to majority share and substantially reduces transactions costs (TC). Queuing costs (QC), however, increase to a moderate value of approximately 0.25.

Figure 2 also shows what our model predicts for very low values of ν. When there are vanishingly few jumps in the fundamental relative to investor order arrivals, protection becomes irrelevant and we get the same equilibrium values and performance metrics with $\xi = 1$ as with $\xi = 0$. Between $\nu = 0$ and the point where unprotected pegs disappear (e.g., $\nu > 1.45$ 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 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 2 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, for some extreme parameter values, the indirect effect more than offsets the direct effects.

Figure 3 offers a more complete picture of how the impact of order protection depends

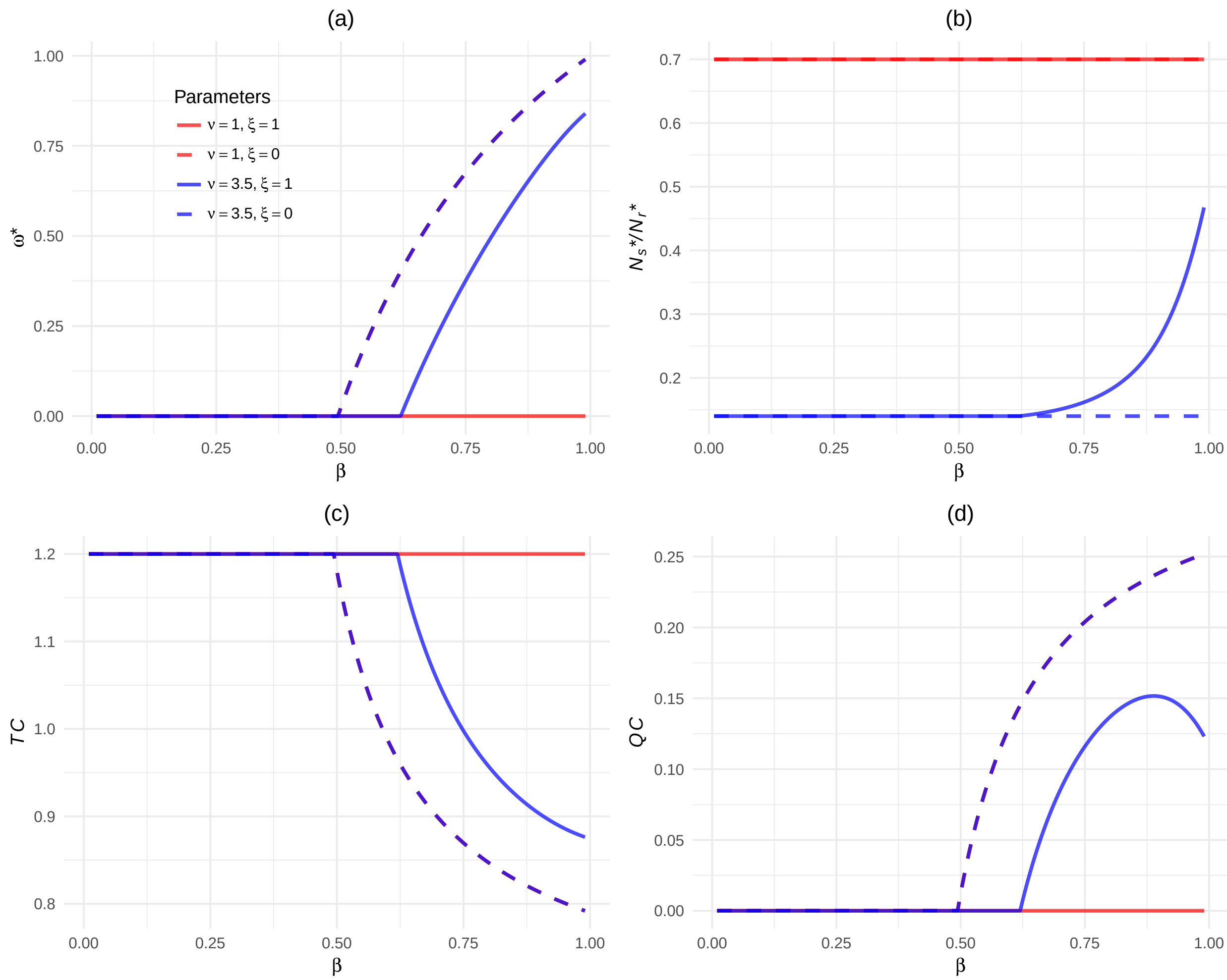

Figure 3: Impact of investor patience on equilibrium ratios and performance metrics. The horizontal axis is $\beta = \exp\left(-\frac{\delta}{\rho}\right)$. All other parameters are at baseline values for red lines, and all except $\nu = 3.5 = 0.5\rho$ for blue lines. Dotted lines show values for $\xi = 0$ and solid lines show $\xi = 0$.

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) in this range we have $\frac{N_s^*}{N_r^*} = \frac{2\nu}{c}$ and the performance metrics do not depend on whether pegs are protected. 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 3 shows that queuing costs *decrease* in the 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

(a) Baseline

Market	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.756	0.200	1.45	0.218
$\xi = 1$	0.305	0.228	1.61	0.0888
Diff	0.451	-0.0283	-0.161	0.129

(b) $\nu = 3.5$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.756	0.700	1.45	0.218
$\xi = 1$	0	0.700	1.80	0
Diff	0.756	0	-0.35	1.45

(c) $\rho = 1$, $\nu = 7$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.756	1.40	1.45	0.218
$\xi = 1$	0	1.40	1.80	0
Diff	0.756	0	-0.353	0.218

(d) $\beta = 0.1$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0	0.200	1.80	0
$\xi = 1$	0	0.200	1.80	0
Diff	0	0	0	0

(e) $\beta = 0.99$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.892	0.200	1.41	0.239
$\xi = 1$	0.511	0.282	1.52	0.114
Diff	0.381	-0.082	-0.11	0.125

(f) $\nu = 0.07$, $\beta = 0.9$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.892	0.0140	1.41	0.239
$\xi = 1$	0.931	0.0400	1.41	0.238
Diff	-0.039	-0.0260	0	0.001

(g) $c = 30$

	ω^*	N_s^*/N_r^*	TC	QC
$\xi = 0$	0.756	0.0670	1.45	0.218
$\xi = 1$	0.305	0.0760	1.61	0.0888
Diff	0.451	-0.0090	-0.16	0.129

Table 2: Performance metrics at market equilibrium with ($\xi = 0$) and without ($\xi = 1$) order protection. Panel (a) corresponds to baseline parameters $c = 10$, $d = 0.18$, $\varphi = 1.8$, $\beta = 0.8$, $\rho = 7$ and $\nu = 1$; the remaining panels use specified deviations from the baseline case.

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, $\frac{N_s^*}{N_r^*}$. The Table 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[3] 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 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:

[3] 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.

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,[4] 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:

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

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 there are other parameter values for which adders would wish to purchase speed. Preliminary work so far 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.

We estimate queuing costs using the steady state distribution. A broker probably could gain a competitive advantage by estimating the current expected queuing cost conditional on the most recent transaction stream, e.g., executions at midpoint vs BBO. The value of such information would seem to decay rapidly, and we conjecture that order placement strategies using it would not have first-order impact on our comparative statics predictions. It is, of course, an empirical question whether those predictions work despite this and other simplifications we have made.

An ambitious extension would replace 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 issues that we (following Budish et al. (2015) and others) have evaded by assuming an observable, exogenous fundamental value. We believe that adverse selection is of first order importance in understanding market microstructure, but do not see it interacting strongly with the market format differences we consider. Thus we conjecture that most of our qualitative results on order protection would still hold, but as yet we have little evidence on this question.

It might also be of interest to examine formally what happens when the grid size goes to zero in Assumption A2. In that limit it might not be reasonable to assume resting limit orders at best bid and offer, and it is not obvious to us how to define a peg in this setting.

Of greater practical interest 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. Even contemplating such a model highlights that fact that protection via messaging delay can not be a complete solution to perceived problems caused by high-frequency trading, since the protection relies on viewing the NBBO established in markets that are not delayed.

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 examine the order imbalance distribution and the present model's comparative statics. We hope that the present paper encourages such new empirical and theoretical research.

Appendices

A Distribution of Order Imbalance

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 $q : \mathbb{Z} \to (0,\infty)$ 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 a new pegged sell (resp. buy) order generates a transition $k \to k+1$ (resp. $k \to k-1$), while 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 no transition.

Thus an arrival updates a negative imbalance probability $p(k|k<0)$ to $p'(k|k<0) = \frac{\omega}{2}p(k+1|k<0) + \frac{1-\omega}{2}p(k|k<0) + \frac{1}{2}p(k-1|k<0)$; the first two terms arise from p and m buy orders respectively, and the third term from any sell order. In steady state $p'(k|k<0)$ is equal to the pre-update value $p(k|k<0)$. Thus we obtain the following steady state equation for negative imbalance, together with analogous equations for a zero imbalance and a positive imbalance:

$$p(k|k<0) = \frac{\omega}{2}p(k+1|k<0) + \frac{1-\omega}{2}p(k|k<0) + \frac{1}{2}p(k-1|k<0) \tag{A.1}$$

$$p(0) = \frac{1}{2}p(1) + (1-\omega)p(0) + \frac{1}{2}p(-1) \tag{A.2}$$

$$p(k|k>0) = \frac{1}{2}p(k+1|k>0) + \frac{1-\omega}{2}p(k|k>0) + \frac{\omega}{2}p(k-1|k>0). \tag{A.3}$$

Substituting $p(k|k<0) = q_k$ in (3.1) for all $k < 0$, and writing $B = \left(\frac{1-\omega}{1+\omega}\right)$ to reduce notation, we verify directly that Equation (A.1) holds:

$$\begin{aligned}
B\omega^{-k} &= \frac{\omega}{2}B\omega^{-(k-1)} + \frac{1-\omega}{2}B\omega^{-k} + \frac{1}{2}B\omega^{-(k+1)} \\
&= \left(\frac{1}{2}+\frac{1}{2}\right)B\omega^{-k} + \left(\frac{1}{2}-\frac{1}{2}\right)B\omega^{-(k+1)} \\
&= B\omega^{-k}.
\end{aligned} \tag{A.4}$$

Verifying that $q_k = p(k|k \geq 0)$ in (3.1) satisfies (A.2) and (A.3) is similarly straightforward. Any other solution of the steady state equations (A.1) - (A.3) must be proportional to q, so to prove the proposition it suffices to verify that q is a probability distribution on $\mathbb{Z}$:

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

A.2 Unprotected Pegged Orders

When midpoint peg orders are not protected from sniping, the set of possible events increases to six: the four types of investor arrivals, in addition to increasing and decreasing jumps in the fundamental. 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 $\tilde{q} : \mathbb{Z} \to (0,\infty)$ of the order imbalance, with*

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

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. The relative probabilities of the four investor events are unchanged from the previous proposition, but to accommodate the directional jumps in the fundamental value, those four probabilities all shrink by the factor $\frac{\rho}{\rho+\xi\nu}$. An upwards jump has probability $\frac{1}{2}\frac{\nu}{\rho+\nu}$ and causes the transition $k \to 0$ when $k > 0$ and $\xi = 1$, and otherwise has no effect. A downwards jump has the same probability and causes the transition $k \to 0$ when $k < 0$ and $\xi = 1$, and otherwise has no effect.

With those modifications, the equations parallel to (A.1) - (A.3) that define the steady state distribution become:

$$p(k|k < -1) = \frac{\omega}{2}\frac{\rho}{\rho+\xi\nu}p(k+1|k < -1)$$

$$+\left(\frac{1}{2}-\frac{\omega}{2}\frac{\rho}{\rho+\xi\nu}\right)p(k|k<-1)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(k-1|k<-1) \tag{A.5}$$

$$\begin{aligned}
p(0) &= \sum_{k\neq 0}\frac{1}{2}\frac{\xi\nu}{\rho+\xi\nu}p(k)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(1)+\left(1-\omega\frac{\rho}{\rho+\xi\nu}\right)p(0)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(-1)\\
&= \frac{1}{2}\frac{\xi\nu}{\rho+\xi\nu}-\frac{1}{2}\frac{\xi\nu}{\rho+\xi\nu}p(0)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(1)+\left(1-\omega\frac{\rho}{\rho+\xi\nu}\right)p(0)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(-1)\\
&= \frac{1}{2}\frac{\xi\nu}{\rho+\xi\nu}+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(1)+\left(1-\frac{\rho\omega+(1/2)\xi\nu}{\rho+\xi\nu}\right)p(0)+\frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(-1)
\end{aligned} \tag{A.6}$$

$$\begin{aligned}
p(k|k>1) &= \frac{1}{2}\frac{\rho}{\rho+\xi\nu}p(k+1|k>1)\\
&\quad+\left(\frac{1}{2}-\frac{\omega}{2}\frac{\rho}{\rho+\xi\nu}\right)p(k|k>1)+\frac{\omega}{2}\frac{\rho}{\rho+\xi\nu}p(k-1|k>1).
\end{aligned} \tag{A.7}$$

By symmetry, $p(1)=p(-1)$, so the equation for $p(0)$ implies

$$p(1)=p(-1)=\left(\omega+\frac{\xi\nu}{2\rho}\right)p(0)-\frac{\xi\nu}{2\rho}. \tag{A.8}$$

Solving Equations (A.5) - (A.8) for $p(k|k<0)$ and $p(k|k>0)$, we find

$$p(k+1|k>0)=\left(1+\frac{\xi\nu}{\rho}+\omega\right)p(k|k>0)-\omega p(k-1|k>0) \tag{A.9}$$

$$p(k-1|k<0)=\left(1+\frac{\xi\nu}{\rho}+\omega\right)p(k|k<0)-\omega p(k+1|k<0). \tag{A.10}$$

Equations (A.9) and (A.10) are linear second order homogeneous difference equations, whose general solution takes the form

$$p(k)=a_1\lambda_1^{|k|}+a_2\lambda_2^{|k|}, \tag{A.11}$$

where

$$\lambda_1=\frac{1}{2}\left(1+\frac{\xi\nu}{\rho}+\omega\right)+\frac{1}{2}\sqrt{\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2-4\omega} \tag{A.12}$$

$$\lambda_2=\frac{1}{2}\left(1+\frac{\xi\nu}{\rho}+\omega\right)-\frac{1}{2}\sqrt{\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2-4\omega}, \tag{A.13}$$

are the roots of the quadratic equation

$$\lambda^2-\left(1+\frac{\xi\nu}{\rho}+\omega\right)\lambda+\omega=0. \tag{A.14}$$

The discriminant $\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2-4\omega$ is bounded above by $\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2$ and bounded below by $\left(1+\frac{\xi\nu}{\rho}+\omega\right)^2-4\omega(1+\frac{\xi\nu}{\rho})=\left(1+\frac{\xi\nu}{\rho}-\omega\right)^2$ for all $\nu,\rho>0$ and $\xi,\omega\in[0,1]$. As a result, $\lambda_1\geq 1$ and, as required by Equation (4.3) of the proposition, $\lambda\equiv\lambda_2\in(0,\omega)$. It is easily seen that $\lambda=\omega$ when $\xi=0$.

From the boundary condition $p(k)\to 0$ as $k\to\infty$, we see that $a_1=0$ since $\lambda_1\geq 1$. Consequently, Equation (A.11) implies that $p(0)=a_2$. Enforcing the summability constraint for a probability distribution, we find:

$$1=\sum_{k=-\infty}^{\infty}p(k)=a_2+2a_2\sum_{k=1}^{\infty}\lambda^k=a_2\left[1+2\frac{\lambda}{1-\lambda}\right]=a_2\left[\frac{1+\lambda}{1-\lambda}\right], \tag{A.15}$$

Hence, $a_2=p(0)=\frac{1-\lambda}{1+\lambda}$ and from Equation (A.11) we obtain the desired expression (3.1). $\square$

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

$$\omega=\lambda+\xi\left[\frac{\lambda}{1-\lambda}\right]\frac{\nu}{\rho}, \tag{A.16}$$

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

Proof The result is obtained by solving for ω in Equation (A.14).

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.3 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 (3.2) may not convey the intuition behind this result. To better understand it, consider the limiting distribution $q_k(\omega)$ in (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 $q_k < \frac{1}{2K+1}$ in the limit 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.4 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

$$\begin{aligned} \frac{2\nu}{N_s + 1} \leq c &\iff 2\nu \leq (N_s + 1)c \\ &\iff \nu \leq \rho \frac{1 - \omega}{1 + \omega} + \frac{c}{2}. \end{aligned} \tag{A.17}$$

At baseline, $\nu = 1$ and $\rho \frac{1-\omega}{1+\omega} + \frac{c}{2} \approx 7 \frac{.25}{1.75} + \frac{10}{2} = 6$, so such a deviation is indeed highly unprofitable.

A.5 Unit Sniping Limitation

Assumption A1 ("... one indivisible unit at a time") can be construed as limiting each sniper to at most one snipe per V jump. 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 (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 Baseline Model Calibration

Here we explain how our baseline parameter values connect with available market data.

B.1 Investor Fraction ω^*

Recall that ω is the fraction of investor orders transmitted as midpoint pegs and for $\xi = 0$

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

is the probability that there is a contra-side order resting at midprice (see, e.g., Equation (3.3)). A new investor order is represented in Table 1 in one of three ways:

1. With probability P, an agency will remove liquidity at midprice.
2. With probability $\omega(1-P)$, an agency will add liquidity at midprice.
3. With probability $(1-\omega)(1-P)$, an agency will remove liquidity at BBO.

Conditional on an agency order, these probabilities sum to unity; unconditionally (given the presence of proprietary traders) the sum of probabilities (.2081, .2784, .0744, respectively) is 0.5609. As a result,

$$\frac{0.2081}{0.5609} = P = \frac{\omega}{1+\omega} \implies \omega \approx 0.59 \tag{B.2}$$

$$\frac{0.2784}{0.5609} = \omega(1-P) = \frac{1}{1+\omega} \implies \omega \approx 1.01 \tag{B.3}$$

$$\frac{0.0744}{0.5609} = (1-\omega)(1-P) = \frac{1-\omega}{1+\omega} \implies \omega \approx 0.77. \tag{B.4}$$

$$\tag{B.5}$$

The average of these values is 0.79, so we choose baseline parameters that yield $\omega \approx 0.75$.

B.2 Midprice Transaction Fee

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.

B.3 Investor Surplus

We define φ as the surplus for the marginal investor with impatience β^* (defined below). Such an investor is 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.

B.4 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 and canceling 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 $M : [0, 1] \to [0, 1]$, $\quad \hat{\beta} \mapsto \hat{\beta}'$.

Lemma B.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

$$\begin{aligned}\beta^* &= \frac{\varphi - 1}{\varphi - d - \omega^*(1-d)} \\ &= \frac{1.8 - 1}{1.8 - 0.18 - 0.75 \times (1 - 0.18)} \\ &= 0.79.\end{aligned} \tag{B.6}$$

Substituting $\rho = 12.82$ (determined below) into the relation $\beta^* = \exp(-\delta/\rho)$ we arrive at $\delta = -\rho \log(\beta^*) \approx 3$.

B.5 Arrival Intensities

Table 3 reports the S&P 500 exchange traded fund (ticker SPY) transaction volume at IEX for the month of December, 2016. As we consider the asset in our model to be similar to a highly liquid asset such as SPY, we use the volume statistics in Table 3 to calibrate the investor arrival intensity parameter, ρ. The total number of SPY shares traded by Agency Removers (across all order types) is 5,312,493 and the shares traded by Agency Adders via only Midpoint and Discretionary pegs is 5,186,025, resulting in a total volume of 10,498,518 shares by the equivalent of investors in our model. Since our model considers an investor arrival to be a unit transaction, and since a unit transaction at IEX is 100 shares, there were a total of $10,498,518/100 \approx 105,000$ investor arrivals during the 21 trading days, or $21 \times 6.5 \times 60 = 8190$ trading minutes of December, 2016, resulting in $105,000/8190 = 12.82$ investor arrivals per minute or roughly 1 arrival every 4.68 seconds. Since ρ represents the arrival rate of investors on one side of the market, these values suggest a baseline value of $12.82/2 = 6.41$. For simplicity, we set $\rho = 7$.

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

	Other Nonroutable				Primary Peg			
	Hidden		Lit		Hidden		Lit	
	BBO	Mid	BBO	Mid	BBO	Mid	BBO	Mid
Agency Remover	489370	80279	2506659	0	0	0	0	0
Prop Remover	2262796	230826	1201430	0	0	0	0	0
Agency Adder	146504	24141	2030384	0	263913	0	0	0
Prop Adder	14655	1041	5978539	0	1882565	0	0	0
	Midoint Peg				Discretionary Peg			
	Hidden		Lit		Hidden		Lit	
	BBO	Mid	BBO	Mid	BBO	Mid	BBO	Mid
Agency Remover	0	1644581	0	0	0	591604	0	0
Prop Remover	0	565548	0	0	0	2200	0	0
Agency Adder	99915	1228941	0	0	1800154	2057015	0	0
Prop Adder	0	4802	0	0	3100	4490	0	0

Table 3: IEX SPY volume for December 2016 by order type and transaction price. Excludes routable orders and transactions in locked or crossed market conditions.

not reversed over the subsequent period of four quotations[5], 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 $\frac{\nu}{\rho} \approx \frac{1}{7}$, or that the intensity of investor arrivals is about seven times that of jumps.

[5]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.

B.6 Cost of Speed

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.

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 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[6] and later transact with contra-side orders at midpoint.

[6] 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 NBBO is \$10.00 $\times$ \$10.01 and the IEX order book is \$10.00 $\times$ \$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.

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.

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.

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

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*.

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*.