Order Protection through Delayed Messaging

Eric M. Aldrich[*] Daniel Friedman[†]

October 3, 2021

Abstract

Several financial exchanges (e.g., IEX and NYSE American) recently introduced messaging delays 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 examines the dynamics of midpoint pegged order queues and finds their steady states. It shows hows how messaging delays can protect pegged orders and improve investor welfare, but typically increase queuing costs. Recently available field data shows that the empirical distribution of queued pegged orders is highly leptokurtotic and resembles the discrete Laplace distribution predicted by the model.

Keywords: High-frequency trading, continuous double auction, pegged orders, IEX.

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

Acknowledgments:
The authors gratefully acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 741409); the results reflect only the authors' views; the ERC is not responsible for any use that may be made of the information it contains. We are also indebted to the Center for Analytical Finance (CAFIN) at UCSC for seed funding of the Experimental High-Frequency Trading project of which this paper is a part, and to Dan Aisen and Ben Connault for access to data and for insight into the IEX system. The editorial team of this journal was exceptionally helpful in improving the exposition. We also benefited from comments by Peter Bossaerts, Peter Cramton, Tomasz Kozubowski, Pete Kyle, Natalia Lazzati, Kristian Lopez Vargas, John Musacchio, Axel Ockenfels, Ioanid Rosu, Nirvikar Singh and Harold Widom, and from participants at the 2018 Bonn MPI Experimental Finance Workshop, the 2018 Luxembourg Expbot Workshop, the May 2020 Australia - New Zealand economics workshop, and Economics Department seminars at the Universities of Essex, Maryland, Monash, and Virginia Commonwealth.

[*] Supply Chain Optimization Technologies Forecasting, Amazon.com; Email: ealdrich@gmail.com

[†] Departments of Economics, University of Essex and University of California Santa Cruz; 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 algorithms now involved in 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., Budish et al., 2015; Du and Zhu, 2017; 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: Investors Exchange (IEX), NYSE American, Thomson Reuters, Electronic Broking Services (EBS), and TSX Alpha all have imposed delays on 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 investors under the traditional continuous market format? Does a messaging delay help ordinary investors, and does it have unintended consequences? The present paper contributes to the growing theoretical literature that addresses such questions. Our model spotlights the consequences of imposing a uniform delay on all new orders when a particular type of resting "pegged" order is repriced without delay. The model thus captures the essential elements of HFT-inspired reforms at IEX and NYSE American, which are closely related to reforms at TSX Alpha and other exchanges.

Our model of exchange trading has three types of participants. *Investors* arrive at random times and each decides whether to place a market order that transacts immediately, or to place a pegged order at the midpoint price between the best bid and best offer. A pegged buy order transacts immediately with a queued pegged sell order if any, and otherwise is is queued behind

[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. We should add that in the present paper, "protection" henceforth refers to attempts to prevent HFT traders from taking advantage of slower traders in a given exchange, and not to passing orders across exchanges.

older pegged buy orders; pegged sell orders are analogous. The midpoint pegged queue is "hidden," i.e., not visible to participants. *Market makers* are always present. They place lit (visible to all) orders at the best bid (BB) and best offer (BO), and earn half a minimal price increment from an investor when hit by a market order. *Snipers* pay a cost to access HFT technology ("speed"), and earn profits at the expense of market makers and (potentially) investors by picking off stale orders whenever the fundamental value (BBO midpoint) jumps. The exchange can impose a messaging delay to protect investors' pegged orders from snipers.

We focus on steady state equilibrium of this dynamic stochastic model, characterized by (a) the distribution of the hidden order queue which makes investors indifferent between market orders and pegged orders, and (b) the mass of market makers and the mass of snipers that drive the profit of both types of participants to zero. It turns out that the distribution in (a) has the discrete Laplace functional form, and so has "fat tails." That distribution leads to closed form expressions in (b). Those expressions in turn offer predictions on how messaging delays and other structural parameters impact investor welfare and the sustainable masses of market makers and snipers.

The model thus yields a wealth of predictions that can be tested against laboratory or field data. For example, a messaging delay that fully 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) higher investor welfare overall, but partially offset by (e) higher queuing cost. The model also identifies parametric conditions under which some of these effects are diminished or even reversed.

Popular accounts of financial market reforms involving a messaging delay (e.g., Lewis, 2015; Pisani, 2016) have focused on IEX's 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; see Khapko and Zoican (2020) for a laboratory confirmation. However, that delay allows the exchange to have a timely view of the National Best Bid and Offer (NBBO) — an aggregation of price quotes across all public exchanges — and to automatically reprice pegged orders before predatory orders arrive at the matching engine. Thus slow traders' queued pegged orders are protected from fast traders who attempt to "snipe" stale orders following a NBBO jump.

Pegged orders are available on all national securities exchanges in the United States. They are typically "hidden," i.e., not shown in the publicly available limit order book. Exchanges 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 to 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.

Previous Literature. Despite their essential role in protecting investors at exchanges featuring a uniform messaging delay, pegged orders have not been analyzed in previous literature of which we are aware. 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 Zhu (2014), Buti et al. (2015) and Buti et al. (2017) in examining dark trading, but we examine the consequences *within* an exchange that also maintains a lit order book.

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 environment by representing ordinary investors as exogenous order flows that balance in expectation around an exogenous fundamental value V. Active participants – market makers and snipers – observe V, which jumps at random times, and can purchase speed. Each market maker sets a bid and ask price centered at V, choosing from a continuous range. Our model retains many of these simplifications but requires a discrete price grid in order to model pegged orders properly and to analyze timing and position in a queue of orders tied at a given price. We also need to distinguish between lit (publicly displayed) and hidden orders.

Some previous articles have considered hidden orders, albeit in a different context. Advancing on predecessors such as Hendershott and Mendelson (2000) and Degryse et al. (2009), Zhu (2014) develops a tractable two-period model in which informed traders and liquidity traders choose in period 1 between placing market orders in a lit exchange or in a dark pool where all orders are hidden. In the exchange, a zero profit market maker chooses a continuous spread centered at the posterior expected price, while the dark pool crosses orders at that midprice. Typically buy and sell orders in the dark pool do not balance at that price; excess heavy side traders each incur a fixed delay cost and transact in period 2. In equilibrium, opening the dark pool improves price discovery but widens the spread at the exchange and increases delay costs. In a similar vein, Buti et al. (2017) find that opening a dark trading venue, with trading at the midpoint between best bid and offer, increases fill rates at a pre-existing lit market, but reduces liquidity, increases spreads and reduces overall welfare. 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.

Like the present paper, Brolley and Cimon (2020) investigate the impact of exchange-imposed messaging delays, but from a very different perspective. Their 3-period model in the spirit of Zhu (2014) focuses on how protecting market makers in one venue influences informed investors' choices between competing venues. By contrast, we look at a single venue where investors may be able to protect their own orders. Baldauf and Mollner (2020) also develop a multi-venue model and show that exempting order cancellations from a random delay leads to smaller spreads and more informed trading. Chen et al. (2017) likewise study a randomized order processing delay introduced at the TSX Alpha exchange in 2015, and find that asymmetrically favoring liquidity-providing orders (i.e. not subjecting them to the delay) results in a deterioration of market quality.

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. (2019). 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 net benefits of aggressive and passive HFT are often estimated to be positive overall, but usually with the acknowledgment of non-negligible costs, e.g., Brogaard and Garriott (2019), Hasbrouck and Saar (2013), Bershova and Rakhlin (2013), and Breckenfelder (2013). The findings in Hirschey (2020) suggest that HFT behavior provides a net improvement to liquidity, but increases costs to non-HFT traders.

Hu (2019) investigates empirically the impact of order protection on market quality and price discovery. Exploiting heterogeneity in the transition of IEX ticker symbols to exchange trading (where they receive quote protection), Hu (2019) finds that trading costs and adverse selection decrease, while measures of price discovery are mixed. He concludes that although such 'slow' exchanges may contribute little to price discovery, they are feasible alternatives to regulatory intervention aimed at improving overall market quality.

Roadmap. Section 2 presents the ingredients of our dynamic market model, e.g., two types of random events and three types of participants, as well as our equilibrium concepts. Section 3 reports closed-form expressions for the equilibrium outcomes, and shows how they are obtained in instructive special cases. It also shows how the equilibrium responds to changes in the degree of pegged order protection, and to changes in other structural parameters. Section 4 develops the practical implications of our model. Using recently obtained data on hidden orders, it shows that the empirical distribution appears to be remarkably close to a discrete Laplace distribution, as predicted by our model. Section 4 then collects predictions of the impact of order protection and other structural features of markets. Section 5 offers a concluding discussion, noting some promising avenues for future empirical and theoretical work. Longer formal proofs are collected in Appendix A. An Online Supplement reports a calibration exercise and numerical comparative statics (sections B and C); dynamic extensions of the model and empirics (section D); more mathematical details (E); and connections to current financial practice (F).

2 Model

We now present a model of a continuous double auction with messaging delays that may protect midpoint pegs against sniping. The model ignores some important complications in order to focus sharply on others.

2.1 Market and Prices

The first assumption ignores trading odd lots or multiple lots at fragmented exchanges, enabling a tighter focus on messaging delays and pegged orders.
A1: The market is a single financial exchange allowing trade of indivisible single units of a single asset.

Virtually all financial markets have a minimal price increment, typically a penny per share. Theoretical models nevertheless often assume a continuous price interval, but that would be problematic for us because (i) it is not clear how to define pegs when (possibly tiny) orders can be priced arbitrarily close together, and (ii) messaging speed is far less important when ties on price are rare, as with a continuous price space. Hence our model assumes:
A2: Prices lie on a discrete, uniform grid $\mathcal{P} = 1, 2, \ldots, \hat{P}$. The grid step size represents half of the minimum price increment (e.g., half a penny per share).

Our market does not operate in a vacuum; we idealize the NBBO midpoint (established mainly in other exchanges) as the "fundamental value" V of the asset, which moves exogenously. For our purposes it is natural to assume that V is publicly observable as in Budish et al. (2015), and that it is equally likely to jump up or down by the minimum price increment as in, e.g., Foucault (1999) or Hoffmann (2014).[2] Thus our model assumes

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-step (i.e., a minimal price increment) upward jump and one-sided rate ν of a two-step downward jump.

2.2 Order types, participants and choices

Online Supplement F discusses a large variety of order types, visible ("lit") or otherwise ("hidden") that are currently available on major financial exchanges. That discussion justifies streamlining our model to just three order types:

r: single unit $r$egular lit limit orders that rest at best bid $(V-1)$ and best offer $(V+1)$.

p: single unit midpoint $p$eg hidden limit orders that enter at price V.

m: single unit $m$arket orders that remove liquidity at the midpoint if it is occupied by contra-side orders and otherwise remove contra-side r orders.

Online Supplement F also reviews how exchanges classify participants as agencies and proprietary traders, and how these classifications map to the three participant types in our theoretical model: investors, market makers, and snipers. The model focuses on investors, and their choice between pegged and market orders. Specifically, it assumes that

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

a. each unit demand has gross surplus $\varphi > 1$ and impatience is captured by continuously compounded discount rate $\delta > 0$.

b. Each investor chooses whether to transmit her demand via a market order (m) or a midpoint peg (p).

Assumptions A3 and A4 taken together say that the fundamental value V always equilibrates investor supply and demand, even while experiencing exogenous shifts. Assumption A4a is a

[2] Of course, 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 focus on matters of greater interest, we assume that such modifications are negligible because we are operating far away from the extremes.

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). The restriction $\varphi > 1$ ensures that investors are willing to transact at BB or BO. A4b defines investors' choice set, and is discussed further in Section 2.5 below.

Market makers add liquidity via regular limit orders (r). With highly liquid securities in mind, the definition of r assumes that these orders rest at BBO, and our notion of equilibrium presumes that the BBO queues are not empty. To that end, we assume
A4c. Whenever a new investor arrives, there is always at least one r ask at best offer, BO $= V + 1$, and at least one r bid at best bid, BB $= V - 1$.

As explained in Section 2.5 below, snipers remove liquidity: immediately following a jump in V, they send market orders to transact with stale orders resting in the order book.

2.3 Events, states, and transitions.

We model an ongoing exchange in continuous time, punctuated randomly by two types of events: investor arrivals and jumps in V. These events trigger transitions in the state of the market. Given the strong assumption A4c (the lit order book is always deep), there is no need to track changes in the lit order book, but we must keep track of transitions in the hidden order book at midpoint.

To do so, recall that in any continuous double auction (CDA) market, any new order that meets or crosses a contra-side resting order transacts immediately with the oldest such order at the best price; e.g., a bid that equals or exceeds the best (lowest) ask in the book transacts at that ask price. More specifically, we assume
A5a: An investor's market order executes immediately at price V against the oldest contra-side midpoint peg if any are queued; otherwise, it executes immediately against a market maker's contra-side r order resting at BBO $= V \pm 1$.
A5b: An investor's pegged order also executes immediately against the oldest contra-side pegged order if any; otherwise, it rests behind existing same-side pegs at V.

Since buy and sell p orders execute against each other, 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, an integer $k < 0$ indicates precisely $-k > 0$ midpoint peg buy orders resting in the hidden order book and no sell orders, $k > 0$ indicates k resting midpoint peg sell orders and no buy orders, and $k = 0$ indicates an empty queue at the midpoint price V. See Figure 1.

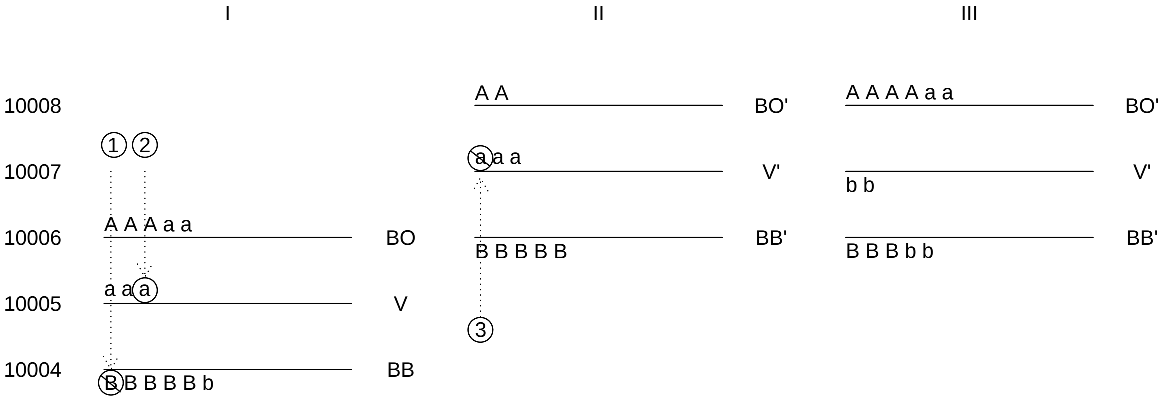

Figure 1: Example market states and transitions. Uppercase (resp. lowercase) denotes lit (resp. hidden) orders, B/b for bids and A/a for asks; those to the left have higher priority at that price. Panel I: initial state is $k = 2$ and $V = 10005$ half-pennies (i.e., $V = \$50.025$ per share); event (1)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

New investor arrivals can trigger transitions in k. Let ω denote the fraction of investors who transmit midpoint peg orders, with the remaining fraction, $1 - \omega$, transmitting market orders. Assumption A4 tells us that a new arrival generates a midpoint peg bid or ask with probability $\omega/2$ each, or a market bid or ask with probability $(1 - \omega)/2$ each. A new pegged ask (resp. bid) always generates a transition $k \to k+1$ (resp. $k \to k-1$). A new market ask (resp. bid) generates a transition $k \to k+1$ when $k < 0$ (resp. $k \to k-1$ when $k > 0$) and otherwise executes at BBO with no change in k.

2.4 Timing

The other type of event, a jump in V, can also trigger transitions in the order imbalance k. To see how, we now detail our model's stylized timing of events and actions; Online Supplement F summarizes actual timing practices in today's financial markets.

Jumps in the NBBO (represented in our model as V) are registered at the Securities Information Processor (SIP), which communicates with our exchange with fixed latency $\tau_{SIP} > 0$. With that latency, all resting pegged orders in our exchange automatically adjust in parallel fashion. Traders' messages to the exchange have default latency τ_{slow}, but traders can reduce their latency by renting or purchasing state-of-art communication facilities. The model allows traders (snipers) to acquire fast technology with latency $\tau_{fast} \in (0, \tau_{slow})$ at rental cost $c > 0$ per unit time.

The exchange imposes an additional delay $\eta \geq 0$ so that traders experience overall latency either $\tau_{fast} + \eta$ or $\tau_{slow} + \eta$ on new orders. To avoid trivialities, we assume that $\tau_{slow} + \eta > \tau_{SIP}$; otherwise sniping is infeasible. In the traditional CDA format $\eta = 0$, while exchanges featuring order protection choose $\eta > 0$ so that $\tau_{fast} + \eta > \tau_{SIP}$. For the sake of greater generality, we assume that η is a random variable independently realized at each V jump event, and define the composite parameter

$$\xi = \text{Prob}[\tau_{fast} + \eta \leq \tau_{SIP}]. \tag{2.1}$$

When $\xi = 0$, pegged orders are fully protected from sniping and are guaranteed to jump in parallel with V before any new messages (e.g., snipes) reach the exchange. When $\xi = 1$, pegged orders are not protected. Intermediate values of ξ represent the probability that all midpoint pegs are vulnerable at any given jump in V.

2.5 Payoffs

Assumption A4b says that investors choose between midpoint peg orders and market orders. The payoff consequences are as follows.

Market order. Let q_k denote the probability of order imbalance k. With probability $\sum_{k=-\infty}^{-1} q_k$ a market sell order will execute immediately against a contra-side midpoint peg and earn φ. With complementary probability $\sum_{k=0}^{\infty} q_k$, that market order will execute immediately against an r order which (by A5a and A4c) is 1 step away from V, and so it earns $\varphi - 1$. Thus the payoff (expected net surplus) for a market sell order is

$$\pi_m = \varphi \sum_{k=-\infty}^{-1} q_k + (\varphi - 1) \sum_{k=0}^{\infty} q_k. \tag{2.2}$$

A market buy order has the same payoff except that q_k is replaced by q_{-k}.

Peg. By A5b, a new midpoint peg sell order executes immediately with a contra-side midpoint peg order if any are present, resulting in the same first term as in equation (2.2). Otherwise, when the imbalance is $k \geq 0$, the new order is placed at position $k+1$ and the payoff is reduced by expected delay and sniping hazard. To compute the delay cost, first note that for state $k = 0$ (empty order queue) with $\xi = 0$ (no sniping hazard), the new peg sell order will execute against the next investor buy order, which arrives at Poisson rate $\rho > 0$. With discount rate $\delta > 0$, the realized discount factor is $e^{-\delta t}$ if that order happens to arrive with delay t, so the appropriate discount factor is its expected value, $\beta = \int_0^\infty e^{-\delta t} \rho e^{\rho t} dt = \frac{\rho}{\rho+\delta} < 1$. A more intricate calculation in Online Supplement E.2 verifies that the discount factor is indeed β^{k+1} when the hidden order book was at state $k \geq 0$.

Sniping hazard has two effects. It reduces the expected delay: e.g., when $k = 0$ the pegged sell will execute against either the next investor buy (rate ρ) or the next snipe buy (rate $\xi\nu$). Hence when $k \geq 0$ the discount factor is $\hat{\beta}^{k+1}$, where $\hat{\beta} = \int_0^\infty e^{-\delta t}(\rho + \xi\nu)e^{(\rho+\xi\nu)t} dt = \frac{\rho+\xi\nu}{\rho+\xi\nu+\delta} < 1$. The other effect, of course, is that sniping produces a non-discounted loss, and by A3 that loss is 2 price steps. Given existing state $k \geq 0$, the conditional probability of not being sniped is $\left(\frac{\rho}{\rho+\xi\nu}\right)^{k+1}$, the probability that $k + 1$ investor buy orders arrive before the next upwards jump in V. Thus the payoff for a midpoint peg sell order is

$$\pi_p = \varphi \left[\sum_{k=-\infty}^{-1} q_k + \sum_{k=0}^{\infty} q_k \hat{\beta}^{k+1} \right] - 2 \sum_{k=0}^{\infty} q_k \left[1 - \left(\frac{\rho}{\rho + \xi\nu} \right)^{k+1} \right] \tag{2.3}$$

A peg buy order has the same payoff except that q_k is replaced by q_{-k}.

Our model also includes snipers and market makers. For analytical clarity, we regard them as distinct types although in actuality both activities could be undertaken by the same firm. To assign profits, we assume that (like investors) they are risk neutral and (unlike investors) they are patient and have no intrinsic trading surplus. We adopt the natural and customary convention that they cash out their inventory at V', the (possibly new) BBO midpoint at the time. The first two parts of our final assumption state this more formally.

A6a: Market makers reverse each transaction immediately at V', and thereby earn a profit of 1 (resp., -1) step from each transaction with investors (resp. snipers).

A6b: Snipers have the capacity to accept all stale orders at BBO and (when not protected) at midpoint. They reverse all transactions immediately at V'.

A6c: Snipers purchase speed but market makers and investors do not.

A6d: Investors never place orders at own-side BBO.

A6c-d are restrictions on action sets that simplify our analysis. Although there is evidence that market makers pay for speed services (see, e.g., Brogaard et al., 2015), there is similar evidence that they do not pay for the highest tier of speed services offered by exchanges (see Shkilko and Sokolov, 2020). Our simplifying assumption A6c reflects this speed differential; see Oline Supplement E.4 for additional discussion. The last part of Appendix A characterizes wide regions of parameter space for which restriction A6d does not bind.[3]

Our model does not fully pin down market scale, and here integer constraints seem unhelpful. Therefore we treat the numbers N_s and N_r of snipers and market makers as real numbers (population masses) rather than as integers (population counts).

Snipers trade off the flow cost, c, of buying speed against profits from sniping stale r and unprotected p orders following a jump in the fundamental value V. Thus there are $N_r + \xi N_p$ potential targets, where N_p is the expected number of stale pegs, e.g., $N_p = \sum_{k=0}^{\infty} kq_k$ for an up jump in V. By assumptions A3 and A6, 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 price step, while the profit for sniping a midpoint peg is $|V' - V| = 2$ steps. Since opportunities arrive at each side of the market at rate ν, the expected profit flow for each of the N_s snipers is

$$\pi_s = 2\nu \frac{N_r + 2\xi N_p}{N_s} - c. \tag{2.4}$$

Market makers place r orders at the BBO, trading off the one price step gain of transacting with a market order against a possible one step loss to a sniper. With probability $\sum_{k=0}^{\infty} q_k$ an investor market order ask (or $\sum_{k=-\infty}^{0} q_k$ for a bid) encounters no contra-side liquidity at the midpoint and thus reaches a market maker. Since these orders arrive at each side of the market at rate $(1 - \omega)\rho$ and earn 1 step profit each, the expected gross profit flow to be divided among the N_r makers is $(1 - \omega)\rho[\sum_{k=0}^{\infty} q_k + \sum_{k=-\infty}^{0} q_k] = (1 - \omega)\rho[1 + q_0]$. With sniping opportunities occurring on each side of the market at rate ν, each costing each market maker 1 price step, those makers get per capita net payoff flow

$$\pi_r = \frac{(1 - \omega)\rho}{N_r}[1 + q_0] - 2\nu. \tag{2.5}$$

Parameter summary. Table 1 below summarizes notation, classifying variables as structural parameters, observable outcomes, and other endogenous variables. The structural parameter vector $(\xi, \nu, \rho, \delta, c, \varphi)$ is called *admissible* if $\xi \in [0, 1], \varphi > 1$, and $c, \nu, \rho, \delta > 0$.

2.6 Equilibrium

Our model is intrinsically dynamic. Appendix A.1 writes out how the order imbalance distribution $\mathbf{q} = \{q_k\}_{k \in Z}$ evolves over time when a given fraction ω of investor orders are transmitted as pegs.

[3] The "thick market" assumption A4c, that makers already have orders resting at BBO, implies that investors placing an own-side BBO order will always face nontrivial delay. This helps explain why A6d typically does not bind.

Variable	Description
	Structural parameters
$\xi \in [0,1]$	Probability that pegs are vulnerable (= 0)
$\nu > 0$	Innovation rate in fundamental value (= 1)
$\rho > 0$	Investor arrival rate (= 50)
$\delta > 0$	Investor time discount rate (= 50)
$c > 0$	Flow cost of speed (= 10)
$\varphi > 1$	Investor gross surplus (= 1.8)
	Other exogenous variables
$\beta \in [0,1]$	discount factor; $\hat{\beta}$ includes sniping hazard
$V = 2n+1$	Fundamental value (supported on odd numbered price steps)
	Endogenous outcomes
$\omega \in [0,1)$	Fraction of investor orders transmitted as pegs
$q_k \in [0,1]$	Probability that peg imbalance is $k \in Z$
$\lambda \in [0,1)$	Discrete Laplace parameter for imbalance distribution
$N_r > 0$	Mass of market makers
$N_s > 0$	Mass of snipers
$N_p > 0$	Expected absolute peg imbalance
$TC > 0$	Expected trading cost
$W > 0$	Welfare = investor surplus net of TC

Table 1: Symbols, range and description for variables used in the model. For structural parameters, $(= x)$ in description indicates the baseline value.

To find equilibrium in our model, the first step is to find a steady state distribution of the order imbalance. In the next section we will show that there is a unique steady state, and it can be written explicitly in terms of ω and exogenous parameters.

Given that steady state distribution, the second step is to find an equilibrium value of ω and the other endogenous variables N_r and N_s. The formal definition is as follows.

Definition 1. *Given an exogenous flow cost of speed $c > 0$, discount rate $\delta > 0$, arrival rates $\rho, \nu > 0$ for investors and fundamental value innovations, and investor gross surplus $\varphi > 1$, the vector (ω^*, N_r^*, N_s^*) constitutes a* market equilibrium *if*

1. *at the steady state queue distribution for $\omega^* \in (0,1)$ (resp. $\omega^* = 0$), a midpoint peg order has the same payoff as (resp. no more payoff than) a market order, and*

2. *with $N_s^* \geq 0$ snipers and $N_r^* \geq 0$ market makers, these trader types both earn zero profit.*

The idea behind the first equilibrium condition is that investors will increase the fraction $\omega \in (0,1)$ of pegged orders whenever the payoff difference $\pi_p - \pi_m$ is positive, and decrease ω when the difference is negative. Hence expected (net discounted) surplus should be equal at an interior steady state, while at $\omega^* = 0$ we should have $\pi_p \leq \pi_m$. Online Supplement E.3 shows that $\omega^* = 1$ is ruled out when $\delta > 0$. The second equilibrium condition arises from the reasonable assumption that there are no substantial barriers to entry or exit for market makers and snipers.

3 Results

We shall now see that our model has a unique equilibrium, and it can be written out explicitly in terms of the model's structural parameters. Simple instructive proofs are included here, while the longer proofs are collected in Appendix A.

3.1 Steady state imbalance

Proposition 1. *Let $\omega \in (0,1)$ be the probability that investors transmit a midpoint peg order, let $\xi \in [0,1]$ be the probability that those orders are vulnerable to sniping, and let $\frac{\nu}{\rho} > 0$ be the event rate ratio of jumps in V relative to investor arrivals. Given assumptions A1-A6, there is a unique steady state distribution $\mathbf{q} = (q_k)_{k \in Z}$ of the order imbalance, with*

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

where

$$\lambda = \frac{1}{2}\left(1 + \frac{\xi\nu}{\rho} + \omega\right) - \frac{1}{2}\sqrt{\left(1 + \frac{\xi\nu}{\rho} + \omega\right)^2 - 4\omega} \qquad \in (0, \omega]. \tag{3.2}$$

Equation (3.1) tells us that the steady state distribution is discrete Laplace, a distribution introduced in Inusah and Kozubowski (2006): symmetric and unimodal at $k = 0$ and decreasing at exponential rate as $|k|$ increases. The distribution has a single parameter $\lambda \in (0, \omega]$, which in our model is a function of structural parameters (ρ, ν, ξ) as well as of the endogenous fraction ω of investor orders that are transmitted as midpoint peg orders. The discrete Laplace distribution is very leptokurtotic — an empty queue is more likely, a short queue is less likely, and a long queue is much more likely than under Gaussianity. For example, suppose that $\lambda = 0.25$. Then the steady state probability of an empty queue is $q_0 = 0.60$, and of a queue of length $|k| = 5$ is $q_k \approx 5.9 \times 10^{-4}$. A Gaussian distribution with the same mean (0) and variance (0.6) has density less than 0.52 at $k = 0$ and less than 4.6×10^{-10} at $k = \pm 5$.

The special case with full protection, $\xi = 0$, is of interest in its own right.

Corollary 1. *When midpoint pegs are fully protected ($\xi = 0$), the steady state imbalance in Proposition 1 reduces to*

$$q_k = \left(\frac{1-\omega}{1+\omega}\right)\omega^{|k|}, \quad k \in \mathbb{Z}. \tag{3.3}$$

The corollary is trivial since expression (3.2) collapses to $\lambda = \omega$ when $\xi = 0$, but Appendix A offers an elementary direct proof to help build intuition for the more general case. That proof begins with the state transition audit noted at the end of Section 2.3, and verifies that, in order to balance outgoing transitions $k \to k \pm 1$ against incoming transitions $k \pm 1 \to k$, the queue probabilities must decline in $|k|$ at geometric rate ω. This directly implies the discrete Laplace distribution with parameter ω.

When $\xi > 0$, successful sniping will augment single unit (± 1) transitions with multi-unit transitions $k \to 0$. The intuition is now less clear, and the formal proof of Proposition 1 requires queuing theory and second order difference equations. Nevertheless, the steady state distribution again turns out to be discrete Laplace.

3.2 Market equilibrium

Proposition 2. *Under assumptions A1 - A6, for every admissible structural parameter vector there is a unique market equilibrium, with*

$$\omega^* = \lambda^* + \left(\frac{\lambda^*}{1-\lambda^*}\right)\frac{\xi\nu}{\rho} \tag{3.4a}$$

$$N_r^* = \frac{\rho}{\nu}\left(\frac{1-\omega^*}{1+\lambda^*}\right) \qquad \textit{and} \tag{3.4b}$$

$$N_s^* = 2\nu\frac{N_r^* + 2\xi N_p^*}{c} = \frac{2\rho}{c}\left(\frac{1-\omega^*}{1+\lambda^*}\right) + \frac{4\xi\nu}{c}\frac{\lambda^*}{1-(\lambda^*)^2} \tag{3.4c}$$

where

$$\lambda^* = \frac{1}{2\rho(\rho+\xi\nu)}\Bigg[\xi\nu(\rho-\xi\nu) + \rho\left((1-\varphi)\delta + 2\rho\right) - \Big(\left(\xi\nu(\xi\nu-\rho) + \rho\left((\varphi-1)\delta - 2\rho\right)\right)^2 - 4\rho(\rho+\xi\nu)\left(\rho(\rho+\delta(1-\varphi)) - \xi\nu(\xi\nu+\delta(1+\varphi))\right)\Big)^{1/2}\Bigg]_+ \tag{3.5}$$

$$\textit{and } N_p^* = \frac{\lambda^*}{1-(\lambda^*)^2}. \tag{3.6}$$

We use notation $[x]_+ = \max\{0, x\}$. The empirical implications of this Proposition are developed in Sections 3.3 and 3.4 below. Again, the special case is instructive.

Corollary 2. *With $\xi = 0$, the conclusions of Proposition 2 reduce to*

$$\omega^* = \left[1 - (\varphi - 1)\frac{\delta}{\rho}\right]_+, \tag{3.7}$$

with

$$N_r^* = \frac{\rho}{\nu} q_0 \text{ and } N_s^* = \frac{2\rho}{c} q_0 \text{ where } q_0 = \left(\frac{1-\omega^*}{1+\omega^*}\right). \tag{3.8}$$

Equation (3.7) shows that, not surprisingly, the equilibrium fraction of midpoint peg orders, ω^*, is insensitive to sniping risk (as captured by the fundamental innovation rate ν) when those orders are protected. That equation shows that a positive fraction of investors will then choose pegged orders when $\varphi < \frac{\delta+\rho}{\delta} = (1-\beta)^{-1}$. The intuition is that a very large gross surplus or very heavy (i.e., small) discount factor will lead to prohibitive queuing costs. The same intuition suggests, and equation (3.7) confirms, that equilibrium ω is negatively related to gross surplus φ and to the discount rate δ relative to the investor arrival rate ρ. Equation (3.8) shows that the equilibrium mass of market makers is inversely proportional to the risk of sniping, ν, and is more than directly proportional (via the impact on q_0) to the arrival rate of investor orders, ρ. By the same token, the equilibrium mass of snipers is more than proportional to ρ and is inversely proportional to c, the cost of fast communication technology. Perhaps surprisingly, the mass of snipers also is insensitive to the arrival rate of sniping opportunities (again, when pegs are fully protected). This is due to the offsetting equilibrium decrease in the mass N_r^* of sniping targets as ν increases.

A direct proof of Corollary 2 proceeds as follows. Apply the first market equilibrium condition to profit expressions (2.2) and (2.3), recalling that first terms cancel.[4] Also recall that here $\xi = 0$, so $\hat{\beta} = \beta = \frac{\rho}{\rho+\delta}$ and steady state $q_k = \left(\frac{1-\omega}{1+\omega}\right)\omega^{|k|}$. Thus

$$\begin{aligned}
\pi_p = \pi_m \quad &\iff \quad \varphi\beta\left[\sum_{k=0}^{\infty} q_k \beta^k\right] = (\varphi - 1)\left[\sum_{k=0}^{\infty} q_k\right] \\
&\iff \quad \varphi\beta\frac{1-\omega}{1+\omega}\left[\frac{1}{1-\beta\omega}\right] = \frac{\varphi-1}{1+\omega}\left(\frac{1-\omega}{1-\omega}\right) \\
&\iff \quad \frac{1-\omega}{1+\omega}\left[\frac{\varphi}{\beta^{-1}-\omega}\right] = \frac{\varphi-1}{1+\omega} \\
&\iff \quad (1-\omega)\varphi = (\varphi-1)(\beta^{-1}-\omega) \\
&\iff \quad \omega = \varphi - \beta^{-1}(\varphi-1) = 1 - (\varphi-1)\frac{\delta}{\rho}.
\end{aligned} \tag{3.9}$$

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 $\omega < 1$ in (3.9) for all admissible parameters. To obtain equation (3.8), apply

[4] Since market orders and pegs both transact immediately against resting contra-side pegs. It follows that (see Online Appendix D) that a simple market order dominates the intuitively appealing dynamic strategy of first submitting a pegged order and then, if it doesn't execute immediately, substituting a market order.

the market equilibrium condition $\pi_r = 0$ to Equation (2.5), and solve for N_r when $\xi = 0$, recalling that at steady state $q_0 = \frac{1-\omega}{1+\omega}$. Then solve $\pi_s = 0$ for N_s using Equation (2.4). □

The method for finding equilibrium for the general case $\xi \in [0, 1]$ is the same as for the special case, but the expressions are more complicated because $\pi_p = \pi_m$ now yields an expression in λ that is quadratic rather than linear as in (3.9). See Appendix A for details.

3.3 Comparative statics

The model's empirical implications arise mainly from its comparative statics — how equilibrium outcomes respond to changes in the degree of protection and other structural parameters. Therefore we investigate the signs of the partial derivatives $y_\theta \equiv \frac{\partial y}{\partial \theta}$ of equilibrium values $y = \omega, N_r, N_s$ of potentially observable variables, with respect to parameters $\theta = \xi, \nu, \delta, \rho, \varphi, c$. Because of its central role, we begin with the comparative statics of λ^*.

Proposition 3. *For all admissible parameter vectors, λ^* in equation (3.5) is [a.] strictly decreasing in ξ, i.e., $\lambda^* > 0 \implies \lambda^*_\xi < 0$; [b.] strictly decreasing in δ, i.e., $\lambda^* > 0 \implies \lambda^*_\delta < 0$; [c.] decreasing in ν, i.e., $\lambda^*_\nu \leq 0$; [d.] strictly increasing in ρ if $\xi = 0$, i.e., $\xi = 0 \ \wedge \ \lambda^* > 0 \implies \lambda^*_\rho > 0$; [e.] strictly decreasing in φ, i.e., $\lambda^* > 0 \implies \lambda^*_\varphi < 0$; and [f.] independent of c, i.e., $\lambda^*_c = 0$.*

The formal proof in the internal Appendix uses standard calculus techniques. Since λ is the discrete Laplace parameter for the distribution of pegged orders, this Proposition predicts how those distributions will vary with the degree of protection and other structural parameters. It is also key to our main comparative static results:

Proposition 4. *For all admissible parameters vectors,*

a. the equilibrium value of ω in equation (3.4a) is [strictly] decreasing in δ and in φ, and when midpoint pegs are fully protected, it is [strictly] increasing in ρ [when $\omega > 0$]. It is independent of c, and the net impacts of changes in ξ, ν and (for vulnerable pegs) ρ are ambiguous.

b. the equilibrium value of N_r in equation (3.4b) is strictly increasing in δ and in φ and is independent of c. When midpoint pegs are fully protected, it is increasing in ν and ρ. The net impacts of changes in ξ, ν and ρ are ambiguous when $\xi > 0$.

c. the equilibrium value of N_s in equation (3.4c) is strictly decreasing in c; when midpoint pegs are fully protected, it is strictly increasing in δ and in φ, and independent of ν. The net impacts of changes in ξ, ν and ρ are ambiguous when $\xi > 0$.

3.4 Welfare

We take the investor's perspective in assessing market performance. A technical reason is that the other participants in our model earn zero profit in equilibrium. More fundamentally, one can argue that the social value of financial markets lies in serving investors, not in extracting revenue from them. Thus welfare and market performance is captured in our model by transactions costs, broadly construed as shortfalls in investor payoff.[5]

To compute that shortfall, first consider the case where an investor transmits a market order. With steady state probability $\sum_{k=1}^{\infty} q_k = \frac{\lambda}{1+\lambda}$ there is a resting contra-side pegged order; the market order then executes immediately at midpoint and the investor receives full payoff φ with shortfall 0. With complementary probability $\frac{1}{1+\lambda}$ the market order executes at BBO and the investor receives net payoff $\varphi - 1$, i.e., a payoff shortfall of 1. Writing $\pi_m = \varphi - TC$, the expected shortfall, or transaction cost, for a market order is therefore

$$TC = \frac{1}{1+\lambda}. \tag{3.10}$$

If an investor instead transmits a midpoint pegged order, it still executes immediately and gives full payoff with probability $\frac{\lambda}{1+\lambda}$, but otherwise it goes into the midpoint queue and incurs the delay and sniping costs captured in equation (2.3). Let queuing cost QC denote that expected shortfall in investor payoff, so $\pi_p = \varphi - QC$.

Of course, in equilibrium $\pi_m = \pi_p$, so $QC = TC$. This justifies defining equilibrium welfare as

$$W = \varphi - TC = \varphi - \frac{1}{1+\lambda}. \tag{3.11}$$

Its comparative statics are as follows.

Proposition 5. *For all admissible parameters vectors, equilibrium welfare W in equation (3.11) is decreasing in vulnerability to sniping ξ, investor impatience δ, and fundamental innovation rate ν. With full protection ($\xi = 0$), welfare increases in investor arrival rate ρ.*

The proof is straightforward. Equation (3.11) shows that W is a smooth strictly increasing function of λ, and that the parameters mentioned affect W only via λ. Therefore the chain rule implies for those θ's that W_θ has the same sign as λ_θ given in Proposition 3. □

The formal analysis supports the following economic intuition. An increase in investor surplus φ or impatience δ or sniping hazard $\xi\nu$ will increase the queuing cost for a peg order relative to

[5] We continue to ignore brokerage fees and exchange-imposed fees for adding and removing liquidity; see Online Appendix F.5 for a discussion of current practice and how it can be incorporated into our model.

the fixed 1 step transaction cost for a market order, and therefore will tend to lower the fraction ω of pegs and the corresponding DL parameter λ. An increase in investor arrival rate ρ will tend to have the opposite effect, since new investors burn faster through the preexisting queue when they arrive more frequently. That reduces queuing costs and increases the fraction of pegs.

These effects govern equilibrium welfare W, as noted in the proof just given. For example, an increase in sniping hazard $\xi\nu$ will increase the fraction of market orders, hence increase transaction costs and reduce welfare. As can be seen in the formal analysis, when $\xi > 0$ so that pegged orders may be sniped, an increase in ρ has additional effects that will partially (or perhaps fully) offset its queuing cost impact on λ and W. Likewise, as can be seen from equations (3.4b - 3.4c), several structural parameters have direct impact on the masses of makers and snipers; those direct effects in some cases reinforce, and in other cases may partially or fully offset, the effects via λ. Proposition 4 summarizes the net directional impact.

Does anyone benefit from the investor shortfall TC? In equilibrium, the fraction $1 - \omega$ of TC is transferred to market makers, who transfer it all to snipers, who fully dissipate it in buying speed. Of the remaining fraction ω, some will transfer directly to snipers when pegs aren't fully protected, and the rest is deadweight queuing cost. Thus, one way or another, the shortfall is fully dissipated.

4 Empirical implications

We now check the consistency of the theoretical steady state distribution with recently available data, and then list baseline values of structural parameters that are consistent with that and other data. That baseline enables us to sharpen predictions that potentially can be tested in laboratory and field data.

4.1 Calibration

As detailed in Online Supplement D.4, we recently obtained data on the hidden queue states for an actively traded security with fully protected midpoint pegs. Observations with $|k| \leq 6$ constitute over 99.97% of the 4.9×10^{14} nanoseconds of available data; we ignore the noisy minuscule fractions of observations with $|k| \geq 7$. Black dots in Figure 2 plot natural logs of the observed fractions.

To estimate the equilibrium discrete Laplace parameter $\omega^* = \lambda^*$, we use the maximum likelihood estimator of Inusah and Kozubowski (2006),

$$\hat{\omega} = \frac{K}{1 + \sqrt{1 + K^2}}, \tag{4.1}$$

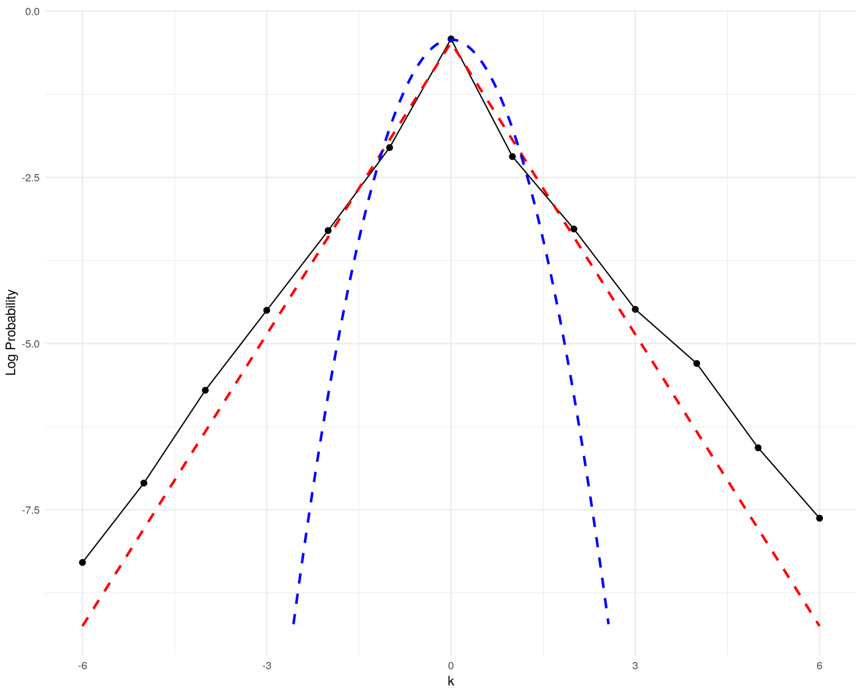

Figure 2: Pegged order queue (log) distributions: empirical (connected black dots), estimated discrete Laplace (red dashed lines), and estimated Gaussian (blue dashed curve). Data are from all trading hours of security SPY (a popular exchange-traded fund) on Investors' Exchange during December 2016.

where $K = \sum_{k \in \mathcal{K}} w_k |k|$ and where $\mathcal{K} = \{-6, \ldots, 0, \ldots, 6\}$.[6] We find $\hat{\omega} = 0.23$ and depict the corresponding discrete Laplace (DL) distribution in Figure 2 as a dashed red line. The figure also shows the Gaussian distribution obtained by minimizing w_k-weighted squared errors (dashed blue curve). Consistent with our theoretical model, the DL distribution appears to be a far better description of the data than a Gaussian distribution. (Indeed, apart from a slight excess of time spent at $k = 0$, the empirical distribution seems an almost perfect symmetric DL.) We emphasize that our model was circulated before we obtained the pegged order data. The model was not designed to fit any known distribution; rather, the DL distribution emerged as a byproduct of our model primitives and equilibrium concept.

The estimated value of ω helps us obtain baseline values for the model's structural parameters, as explained in Online Supplement B.

Counterexample. A natural conjecture is that midpoint pegs are more common when they are

[6] This estimator, which can also be written as $\frac{\sqrt{1+K^2}-1}{K}$, can also be derived by inverting equation (3.6) and noting that by definition $K = 2N_p$.

better protected. It is true that better protection (lower ξ) increases the number of resting pegged orders N_p^*, as can be seen in the proof of Proposition 4. However, more resting pegs implies increased queuing cost, which will partially offset the increased profitability due to reduced sniping hazard. Conceivably the offset could be more than 100%, falsifying the conjecture.

Figure 3 in the Online Supplement investigates. The first two panels show that ω_ξ is negative as conjectured when all parameters are near baseline values. However, when δ is very small (below 10, about 20% of its baseline value) or ρ is quite large (about 2.5 times baseline) then that partial derivative turns positive and so the conjecture fails. Apparently the indirect effect (longer midpoint queues) more than offsets the direct effect (increased profitability due to reduced sniping hazard) when investors are sufficiently patient and numerous.[7]

The lower panels in Figure 3 show the impact of parameters ξ, δ and ν on the comparative statics of ω and λ with respect to ρ, and thus, in conjunction with the proof of Proposition 2, largely resolve the ambiguous comparative statics for N_r and N_s. For example, the direct impact of ρ on N_r is positive, and at baseline so are the indirect impacts via λ and ω.

4.2 Testable predictions

The foregoing analysis yields the following predictions that potentially can be tested against data gathered in the lab or observed in the field.

1. Peg vulnerability. An increase in the probability ξ that pegs can be sniped will, other things equal,

a. reduce the DL parameter λ, e.g., shrink the mean length N_p of the hidden order queue;

b. increase investors' per-unit transaction costs TC and decrease welfare W;

c. reduce the fraction ω of orders transmitted as midpoint pegs, except when relative impatience $\frac{\delta}{\rho}$ is quite small; and

d. typically increase the masses N_r, N_s of both market makers and snipers, with exceptions inherited from [c.]

As noted in Online Supplement C, due to a symmetry in key equations, when $\xi > 0$ these predictions also apply word-for-word to an increase in the fundamental innovation rate ν.

[7] Does a very small innovation rate ν also undermine the conjecture? Not so. The third top panel of Figure 3 shows, and analytic expressions confirm, that ω_ξ remains negative and converges to zero as $\nu \downarrow 0$.

2. Relative impatience. An increase in investor impatience δ will, other things equal,

a. reduce λ and ω;

b. increase the mass N_r of market makers and typically also the mass N_s of snipers; and

c. increase investors' per-unit transaction costs TC and decrease welfare W.

As noted in Online Supplement C, the predicted directional impacts of a decrease in investor arrival rate ρ are exactly the same with full order protection ($\xi = 0$), and also when other parameters are in a large neighborhood of baseline. However, the impact of ρ becomes ambiguous when it is very large relative to δ and pegs are vulnerable to sniping. The same caveat applies to the impact of δ on N_s in item b.

3. Speed cost. An increase in the cost of speed c that does not invalidate Assumption A6c will, other things equal,

a. have no impact on λ, ω, or N_r;

b. decrease the mass N_s of snipers; and

c. have no impact on investors' per-unit transaction costs TC or welfare W.

5 Discussion

The ultimate source of profits for all participants in our model is the exogenous order flow from investors. Investors have an intrinsic desire to buy or sell captured in the parameter φ. A market maker earns income when investors' market orders transact against her lit orders resting at the best bid and best offer. Some of that income is diverted to snipers, who snap up stale BBO orders immediately following a jump in the fundamental value. Intuitively, we have a food chain, with investors' market orders sustaining regular limit orders, which sustain sniping.

A recent innovation at some exchanges offers investors an alternative to market orders: a hidden midpoint peg that is protected from snipers[8] and that executes at a better price. However,

[8] As elaborated earlier, the fact that the order is hidden does not protect it from sniping; protection rather comes from the peg, which automatically reprices before (delayed) snipe orders arrive. It is worth emphasizing that the messaging delay required to fully protect pegs (about 0.00035 seconds as noted in the popular press) is negligible compared to the delay between investor arrivals, which creates a one step increment in queuing cost. Online Supplement D.4 notes that the median one-side arrival delay is about $2 \times 0.615 \approx 1.2$ seconds in our data, or about 3000 times larger than the messaging delay.

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

Our model predicts the direct and indirect impacts of those tradeoffs. We derive analytic expressions for equilibrium outcome variables, and these expressions imply that a greater degree of protection due to delayed messaging generally increases the equilibrium fraction of pegs and reduces the masses of snipers and market makers. Contrary to the conclusions of some previous models (see Online Appendix E.5), greater protection in our model always increases investor welfare in equilibrium. The analytic expressions offer numerous predictions about the impact of other structural parameters.

How well do those results stand up when key simplifying assumptions are relaxed? Our model rules out investors competing with market makers to place orders at best bid and offer, and rules out market makers placing midpoint pegs. The last paragraph of Appendix A shows that those simplifications are without loss of generality for wide subsets of parameter space. The model also rules out liquidity adders (either of lit orders at BBO or of hidden midpoint pegs) purchasing speed technology. Online Appendix E.4 shows that such purchases are unprofitable in a large neighborhood of baseline parameters, but at some more distant parameters adders would wish to purchase speed. Closed-form solutions no longer seem possible in such cases, but numerical exercises using recursion techniques (in particular, the Erlang B model), so far 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. As noted in Online Supplement F.5, straightforward extensions of the present model accommodate fees for placing and removing lit and hidden orders, and thus help investigate fee structure impact, with and without order protection.

Our model focuses on steady state equilibrium, but Online Supplements D.3 and D.4 consider dynamic strategies that seek to exploit short term information on the hidden order queue. We find surprisingly little scope for, or evidence of, dynamic strategies that alter the choice between peg and market orders. Of course, this does not imply that dynamic strategies are ineffective for timing investor orders. Dynamic timing strategies surely are an important part of high frequency trading, but (like Budish et al. (2015) and others) we have set them to one side by assuming exogenous investor orders centered on an observable exogenous fundamental value.

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

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

Empirical work need not wait for these theoretical extensions. In the laboratory, one could investigate whether human subjects in the investor role track $\hat{\omega}$ when the experimenter varies parameters such as (δ, φ), and whether human subjects in the proprietary trader role follow the comparative static predictions of how the structural parameters (ν, ρ, c) impact outcomes (N_r, N_s, W). Following up on our finding that the empirical distribution of the hidden order queue conforms closely to the predicted discrete Laplace functional form, future studies might seek additional field data to test our model's comparative static predictions. We hope that the present paper inspires such new empirical and theoretical research.

A Mathematical Details

This appendix contains formal proofs not included in the main text.

The order queue has distribution function of the form $\mathbf{p} = (p_k)_{k \in Z} = (..., p_{-2}, p_{-1}, p_0, p_1, p_2, ...)$, where each $p_k \geq 0$ and $\sum_{k=-\infty}^{\infty} p_k = 1$. Let $\mathbf{P}$ be the set of all such distribution functions. Poisson event processes induce dynamics on $\mathbf{P}$, and we seek to characterize their steady states. We begin with a self-contained proof of Corollary 1, the special case with fully protected midpoint pegs ($\xi = 0$), where the only relevant event process is investor arrivals.

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

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

To verify that the distribution $\mathbf{q} \in \mathbf{P}$ defined in (3.3) is indeed a steady state, we need to show that $Tq_k = q_k \ \ \forall k \in \mathbb{Z}$. First consider the case $k < 0$, and write $B = \left(\frac{1-\omega}{1+\omega}\right)$ to simplify expressions. Substitute Equation (3.3) into (A.1a) to obtain

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

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

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

When midpoint peg orders are not protected from sniping, the queue state k is affected by two additional events – up and down jumps in the fundamental – as well as the four types of investor arrivals covered in the preceding proof. We shall now see that the resulting stationary distribution has the same functional form (DL, for discrete Laplace) as in the protected case, but the DL parameter is a more complicated function of the structural parameters.

Proof of Proposition 1. Conditional on the event being an investor order, the probabilities of the four kinds of orders are unchanged from those given in (A.1a-c) above, so the unconditional probabilities are attenuated by the probability $R = \frac{\rho}{\rho + \xi\nu} \leq 1$ that the event is an investor order rather than an unprotected jump in the fundamental value. Each upwards jump with probability

ξ causes a direct transition $k \to 0$ when $k \geq 0$, and otherwise has no effect on k. Similarly, each downwards jump with probability ξ causes $k \to 0$ when $k \leq 0$. Thus the probability that $k > 0$ will remain unchanged is now the probability of a market sell plus the probability of an unprotected downward jump, $P_{ms} + P_{dj} = \frac{R(1-\omega)}{2} + \frac{1-R}{2} = \frac{1-R\omega}{2}$; the same probability applies to the symmetric case $k < 0$. The updating operator T is now defined for $k \neq 0$ by:

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

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

$$\begin{aligned} Tp_0 &= R\left(\frac{1}{2}p_1 + (1-\omega)p_0 + \frac{1}{2}p_{-1}\right) + \frac{1-R}{2} + \frac{1-R}{2}p_0 \\ &= \frac{1-R}{2} + \frac{R}{2}p_1 + \frac{1+R-2R\omega}{2}p_0 + \frac{R}{2}p_{-1} \end{aligned} \qquad \text{(A.3)}$$

Note that Equations (A.2a) - (A.3) reduce to Equations (A.1a) - (A.1c) when $\xi = 0$ so $R = 1$. To find a steady state for the general operator T just defined, first set $Tp_0 = p_0$ in Equation (A.3) to obtain

$$0 = \frac{1-R}{2} + \frac{R}{2}p_1 + \frac{R-1-2R\omega}{2}p_0 + \frac{R}{2}p_{-1}. \qquad \text{(A.4)}$$

In steady state, the symmetry of the market ensures $p_1 = p_{-1}$, so Equation (A.4) yields

$$p_1 = p_{-1} = (\omega + \frac{r}{2})p_0 - \frac{r}{2}, \qquad \text{(A.5)}$$

where $r = \frac{\xi\nu}{\rho}$. Next, to find the steady state equation for $k > 0$, set $Tp_k = p_k$ in Equation (A.2b) and simplify to obtain

$$p_{k+1} - (1 + r + \omega)\, p_k + \omega p_{k-1} = 0, \quad \text{for } k > 0. \qquad \text{(A.6)}$$

Equation (A.6) is a linear second order homogeneous difference equation, whose general solution (see, e.g., Sargent 1979 pp. 177) takes the form

$$p_k = a_1 \lambda_1^k + a_2 \lambda_2^k, \qquad \text{(A.7)}$$

for $k > 0$ where

$$\lambda_1 = \frac{1}{2}(1 + r + \omega) + \frac{1}{2}\sqrt{(1 + r + \omega)^2 - 4\omega} \qquad \text{(A.8)}$$

$$\lambda_2 = \frac{1}{2}\left(1+r+\omega\right) - \frac{1}{2}\sqrt{\left(1+r+\omega\right)^2 - 4\omega}, \tag{A.9}$$

are the roots of the quadratic equation

$$\lambda^2 - \left(1+r+\omega\right)\lambda + \omega = 0. \tag{A.10}$$

The discriminant $\left(1+r+\omega\right)^2 - 4\omega$ is bounded above by $\left(1+r+\omega\right)^2$ and bounded below by $\left(1+r+\omega\right)^2 - 4\omega(1+r) = \left(1+r-\omega\right)^2$ for all $\nu, \rho > 0$ and $\xi, \omega \in [0,1]$. It follows that $\lambda_1 \geq 1$; to ensure that $p_k \to 0$ as $k \to \infty$, we therefore must have $a_1 = 0$ in Equation (A.7). The same bounds establish that $\lambda_2 \in (0, \omega] \subset (0,1)$. Note that $r = 0$ and $\lambda_2 = \omega$ when $\xi = 0$.

Set $\lambda = \lambda_2$, consistent with Equation (3.2) of the proposition. Given the symmetry of our model, it is natural to conjecture (and straightforward to check) that

$$p_k = a_2 \lambda^{|k|} \tag{A.11}$$

also satisfies the steady state equation

$$p_{k-1} - \left(1+r+\omega\right)p_k + \omega p_{k+1} = 0, \quad \text{for } k < 0, \tag{A.12}$$

obtained from Equation (A.2a). To ensure that Equation (A.11) defines a probability distribution, we choose a_2 so that

$$1 = \sum_{k=-\infty}^{\infty} p_k = a_2 + 2a_2 \sum_{k=1}^{\infty} \lambda^k = a_2\left[1 + 2\frac{\lambda}{1-\lambda}\right] = a_2\left[\frac{1+\lambda}{1-\lambda}\right], \tag{A.13}$$

Hence, $a_2 = p_0 = \frac{1-\lambda}{1+\lambda}$. Inserting this into Equation (A.11) and relabeling the left hand side as q_k, we obtain the desired expression, Equation (3.3). We have already verified that it satisfies the steady state equations for $k > 0$ and $k < 0$. The last step is to verify that it satisfies the steady state Equation (A.4) for $k = 0$, or equivalently Equation (A.5), i.e., that $p_1 - (\omega + \frac{r}{2})p_0 + \frac{r}{2} = 0$ when $p_k = \frac{1-\lambda}{1+\lambda}\lambda^{|k|}$. Indeed,

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

where the penultimate step uses Equation (A.10), which defines λ. □

Proof of Proposition 2. Substitute into (2.3) the steady state expressions for q_k, sum the geometric series and simplify to obtain

$$\pi_p = \frac{\varphi}{1+\lambda}\left[\lambda + \frac{(1-\lambda)(\rho+\xi\nu)}{(1-\lambda)(\rho+\xi\nu)+\delta}\right] - \frac{2}{1+\lambda}\left[\frac{\xi\nu}{(1-\lambda)\rho+\xi\nu}\right] \tag{A.14}$$

Likewise, in steady state

$$\pi_m = \frac{\varphi\lambda}{1+\lambda} + \frac{\varphi - 1}{1+\lambda}. \tag{A.15}$$

To find market equilibrium, set $\pi_p = \pi_m$, cancel the identical terms $\frac{\varphi\lambda}{1+\lambda}$, cancel the common factor $\frac{1}{1+\lambda}$ and cross multiply to obtain the following quadratic equation for λ:

$$\begin{aligned}(1-\lambda)^2[\rho(\rho+\xi\nu)] - (1-\lambda)[\xi\nu(\rho+\xi\nu) + (\varphi-1)\rho\delta] - [\xi\nu\delta(\varphi+1)] = 0 \\ \Rightarrow \rho(\rho+\xi\nu)\lambda^2 + (\xi\nu(\xi\nu-\rho) + \rho\left((\varphi-1)\delta - 2\rho\right))\lambda \\ +\rho(\rho+\delta(1-\varphi)) - \xi\nu(\xi\nu+\delta(1+\varphi)) = 0.\end{aligned} \tag{A.16}$$

The relevant root is written out in Equation (3.5). Evaluating Equation (A.16) at $\lambda = 0$, we see that that the region of parameter space where $\lambda = 0$ is bounded by the locus

$$\rho(\rho+\delta(1-\varphi)) = \xi\nu(\xi\nu+\delta(1+\varphi)). \tag{A.17}$$

Equation (3.6), which gives the expected number of pegged orders potentially vulnerable to sniping, N_p^*, is obtained as follows:

$$N_p^* = \sum_{k=-\infty}^{0} 0q_k + \sum_{k=1}^{\infty} kq_k = \frac{1-\lambda}{1+\lambda}\sum_{k=1}^{\infty} k\,\lambda^k = \left(\frac{1-\lambda}{1+\lambda}\right)\frac{\lambda}{(1-\lambda)^2} = \frac{\lambda}{1-\lambda^2}. \tag{A.18}$$

Equations (3.4b) and (3.4c) follow by substituting (3.4a) and (3.6) into Equations (2.5) and (2.4), setting them equal to zero and solving for N_r and N_s. The notation $[\cdot]_+$ in (3.5) means that λ is truncated below at 0. For parameters such that the truncation binds, the same logic as in the Corollary (protected case) shows that $\tilde{\omega}^* = 0$. For the other boundary case, Online Appendix E.3 explains why the market equilibrium value of ω is always < 1. □

Proof of Proposition 3. Let $x(\theta)$ be defined implicitly by the identity $F(x,\theta) = 0$. Totally differentiate to get $F_x x_\theta + F_\theta = 0$, so $x_\theta = -\frac{F_\theta}{F_x}$. Apply that standard technique to $F = Q$, where for parameters $\theta = \rho, \xi$, etc.,

$$Q(x,\theta) = [\rho(\rho+\xi\nu)]x^2 - [\xi\nu(\rho+\xi\nu) + (\varphi-1)\rho\delta]x - \xi\nu\delta(\varphi+1) \tag{A.19}$$

is the quadratic expression in $x = 1-\lambda$ from equation (A.16). Thus

$$\lambda_\theta = \frac{Q_\theta}{Q_x}. \tag{A.20}$$

The denominator $Q_x > 0$ at the relevant (positive) root, since Q is an upward-opening parabola in x given admissible structural parameters. Hence the sign of λ_θ is the sign of the numerator Q_θ.

To prove part b of the Proposition, let $\theta = \delta$ and note that

$$\begin{aligned} Q_\delta &\equiv \frac{\partial Q}{\partial \delta} = -x(\varphi - 1)\rho - \xi\nu(\varphi + 1) \\ &= -(1-\lambda)(\varphi - 1)\rho - \xi\nu(\varphi+1) < 0. \end{aligned} \tag{A.21}$$

Likewise, part e holds since $Q_\varphi = -\rho\delta x - \xi\nu\delta < 0$. Part f follows from the fact that c does not appear in Q, so $Q_c = 0$.

Part a similarly follows from $Q_\xi = x^2\rho\nu - x\rho\nu - 2x\xi\nu^2 - \nu\delta(\varphi+1) < 0$, since the last two terms are negative and, for $x \in (0, 1]$, the first two terms together are $-x(1-x)\rho\nu \leq 0$. Likewise, part c follows from $Q_\nu = x^2\rho\xi - x\rho\xi - 2x\xi^2\nu - \xi\delta(\varphi + 1) \leq 0$, since the last two terms are nonpositive and the first two terms are $-x(1-x)\rho\xi \leq 0$. Indeed, in the protected case $\xi = 0$, we have $Q_\nu = 0$ so $\lambda = \omega$ is unaffected by ν. In partial (or no) protection cases $\xi > 0$, the same expressions show that $\lambda_\nu < 0$. Finally for part d, cancel the common factor x and rearrange slightly, and Q_ρ reduces to $A = 2x\rho - \lambda\xi\nu - (\varphi - 1)\delta$. When $\xi = 0$ (full protection, so $\lambda = \omega$), equation (3.9) tells us that $x\rho = (\varphi - 1)\delta$, so here A reduces to $x\rho > 0$, i.e., in this case, $\lambda_\rho > 0$. □

Proof of Proposition 4. For part (a), recall that, by equation (3.4a), the fraction of midpoint pegs is

$$\omega = \lambda + \left(\frac{\lambda}{1-\lambda}\right)\left(\frac{\xi\nu}{\rho}\right). \tag{A.22}$$

Thus the derivative with respect to arbitrary parameter θ is

$$\begin{aligned} \omega_\theta &= \lambda_\theta + \left(\frac{(1-\lambda)\lambda_\theta + \lambda\lambda_\theta}{(1-\lambda)^2}\right)\left(\frac{\xi\nu}{\rho}\right) + \left(\frac{\lambda}{1-\lambda}\right)\frac{\partial}{\partial\theta}\left(\frac{\xi\nu}{\rho}\right) \\ &= \left(1 + \frac{\xi\nu}{(1-\lambda)^2\rho}\right)\lambda_\theta + \left(\frac{\lambda}{1-\lambda}\right)\frac{\partial}{\partial\theta}\left(\frac{\xi\nu}{\rho}\right). \end{aligned} \tag{A.23}$$

For $\theta = \delta, \varphi, c$, the last partial derivative and hence the second term in (A.23) are zero. Since the coefficient of λ_θ is positive in the first term, we see that in these cases ω_θ has the same sign as λ_θ, so the desired results follow from Proposition 3. The partial derivative and hence the second term in (A.23) are positive for $\theta = \xi, \nu$ and negative for $\theta = \rho$, but the first term typically has the opposite sign in these cases so here the signs of ω_θ are ambiguous. However, when $\xi = 0$ the second term disappears and the coefficient of λ_θ in the first term is 1.0, so the remaining assertion in (a) follows from the result for $\theta = \rho$ in Proposition 3.

For part (b), note that N_r in equation (3.4b) is decreasing in both ω and λ. Since δ, c and φ don't appear elsewhere in (3.4b), the desired result for these parameters follows from the chain rule and Proposition 3. Offsetting effects again render ambiguous the sign of $(N_r)_\theta$ for the remaining parameters. For part (c), note that N_s in equation (3.4c) is inversely proportional to c, and when $\xi = 0$ it is directly proportional to N_r. Thus the desired results for δ and φ follow from the previous paragraph. $\square$

We now investigate where the order type constraints in Section 2.5 are not binding, i.e., where they are harmless in the context of our model.

Market makers never use pegged orders or market orders. Since intrinsic payoff φ is zero for market makers, trading at midpoint can never be profitable, and it generates twice the loss of a lit BBO order when sniped. Hence for market makers, p orders are dominated by r orders. Placing a market order guarantees a loss of 1, and so is also dominated for all admissible parameter vectors.

Investors never place orders at own-side BBO. By Proposition 2, an investor's expected profit from a market order is $\pi_m = \varphi - \frac{1}{1+\lambda}$. Her profit from placing an own-side BBO limit order is $\varphi + 1$ discounted by the expected delay. By axiom A4c and strict price/time priority, that order will be placed behind $n \geq 1$ existing BBO orders, so the discount factor is $\hat{\beta}^{n+1} = [\frac{\rho+\xi\nu}{\rho+\xi\nu+\delta}]^{n+1}$. Hence market orders yield higher expected payoff than BBO orders iff

$$\varphi - \frac{1}{1+\lambda} > (\varphi+1)\hat{\beta}^{n+1} \iff \varphi(1-\hat{\beta}^{n+1}) > \hat{\beta}^{n+1} + \frac{1}{1+\lambda}$$

$$\iff \varphi > \frac{\hat{\beta}^{n+1}}{1-\hat{\beta}^{n+1}} + \frac{1}{(1+\lambda)(1-\hat{\beta}^{n+1})}. \tag{A.24}$$

The last term in (A.24) is awkward to expand due to the complicated formula for λ. However, since $\lambda \in [0, 1)$, a sufficient condition is $\varphi > \frac{\hat{\beta}^{n+1}}{1-\hat{\beta}^{n+1}} + \frac{1}{1-\hat{\beta}^{n+1}}$ or

$$\varphi > \frac{1+\hat{\beta}^{n+1}}{1-\hat{\beta}^{n+1}} = \frac{(\rho+\xi\nu+\delta)^{n+1} + (\rho+\xi\nu)^{n+1}}{(\rho+\xi\nu+\delta)^{n+1} - (\rho+\xi\nu)^{n+1}}. \tag{A.25}$$

Since $\hat{\beta} < 1 < \varphi$, the first inequality in (A.25) is more easily satisfied in thicker markets (i.e., with larger n) and is automatically satisfied for n sufficiently large. At the other extreme, for the minimum value $n = 1$ consistent with axiom A4c, (A.25) simplifies to

$$\varphi - 1 > \frac{(\rho+\xi\nu)^2}{(\rho+\xi\nu)\delta + 0.5\delta^2}; \tag{A.26}$$

Imposing that parameter restriction guarantees that investors will not place own-side BBO orders.[9]

[9] For our baseline parameters, even for $n = 1$ the RHS of (A.24) is $\frac{1}{3} + \frac{16}{15} = 1.4$, comfortably below the baseline value $\varphi = 1.8$. The cruder inequality (A.26) is also satisfied at baseline: $1.80 - 1 > \frac{2}{3}$.

References

Aldrich, E. and Lee, S. (2018), “Relative spread and price discovery,” *Journal of Empirical Finance*, 48, 81–98.

Baldauf, M. and Mollner, J. (2020), “High-Frequency Trading and Market Performance,” *Journal of Finance*, 75, 1495–1526.

— (2021), “Fast traders make a quick buck: The role of speed in liquidity provision,” *Journal of Financial Markets*, 1–22.

Bershova, N. and Rakhlin, D. (2013), “High Frequency Trading and Long-Term Investors: A View from the Buy-Side,” *Journal of Investment Strategies*, 2, 3–47.

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. (2019), “High-Frequency Trading Competition,” *Journal of Financial and Quantitative Analysis*, 54, 1469–1497.

Brogaard, J., Hagstromer, B., Norden, L., and Riodan, R. (2015), “Trading Fast and Slow: Colocation and Liquidity,” *Review of Financial Studies*, 1–18.

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. (2019), “Price Discovery without Trading: Evidence from Limit Orders,” *Journal of Finance*, 74, 1621–1658.

Brolley, M. and Cimon, D. A. (2020), “Order-flow segmentation, liquidity, and price discovery: The role of latency delays,” *Journal of Financial and Quantitative Analysis*, 55, 2555–2587.

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.

Degryse, H., Van Achter, M., and Wuyts, G. (2009), “Dynamic order submission strategies with competition between a dealer market and a crossing network,” *Journal of Financial Economics*, 91, 319–338.

Du, S. and Zhu, H. (2017), "What is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*, 84, 1606–1651.

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.

Hendershott, T. and Mendelson, H. (2000), "Crossing Networks and Dealer Markets: Competition and Performance," *The Journal of Finance*, 55, 2071–2115.

Hirschey, N. (2020), "Do High-Frequency Traders Anticipate Buying and Selling Pressure?" *Management Science*, 1–25.

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

Hu, E. (2019), "Intentional Access Delays, Market Quality, and Price Discovery: Evidence from IEX Becoming an Exchange," *Working Paper.*

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

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

Khapko, M. and Zoican, M. (2020), "Do speed bumps curb low-latency investment? Evidence from a laboratory market," *Journal of Financial Markets.*

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

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

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

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

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

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

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

Plott, C., Roll, R., Seo, H., and Zhao, H. (2019), "Tick size, price grids and market performance: Stable matches as a model of market dynamics and equilibrium," *Games and Economic Behavior*, 118, 7–28.

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

Shkilko, A. and Sokolov, K. (2020), "Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity, and Trading Costs," *Journal of Finance*, 75, 2899–2927.

Thiebaux, H. and Zwiers, F. (1984), "The Interpretation and Estimation of Effective Sample Size," *Journal of Climate and Applied Methodology*, 23, 800–811.

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.

Yang, L. and Zhu, H. (2020), "Back-Running: Seeking and Hiding Fundamental Information in Order Flows," *Review of Financial Studies*, 33, 1484–1533.

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

Zhu, H. (2014), "Do dark pools harm price discovery?" *Review of Financial Studies*, 27, 747–789.

Online Supplement for Order Protection through Delayed Messaging

This online supplement contains background information for "Order Protection through Delayed Messaging" together with extensions of the model and miscellaneous details.

B Baseline Parameters

We first explain how the baseline parameter values in Table 1 connect with available market data.

B.1 Investor Fraction ω^*

Recall that ω is the fraction of investor orders transmitted as midpoint pegs, and that Equation (2.2) tells us that, for $\xi = 0$, the steady state probability that there is a contra-side order resting at midprice is

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

since q_k is from the discrete Laplace distribution with parameter ω. In Section D.4 below, we apply a maximum likelihood estimator to the data provided to us by IEX and obtain

$$\hat{\omega} = \frac{|\bar{k}|}{1+\sqrt{1+|\bar{k}|^2}} = 0.23, \tag{4.1}$$

which we take as value of ω^* at baseline parameters.

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

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

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

B.3 Discount Factor

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

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

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

Lemma 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) we use the equal profit condition for the marginal investor, Equation (3.9), to solve

$$\beta^* = \frac{\varphi - 1}{\varphi - \omega^*} \approx 0.5. \tag{B.2}$$

Substituting $\rho = 50$ (determined below) into the relation $\beta^* = \frac{\rho}{\rho + \delta}$ we arrive at $\delta = \rho(1/\beta^* - 1) = 50$.

B.4 Arrival Intensities

Table 2 reports quantiles of the distribution interarrival times (measured in nanoseconds) of submitted orders for the S&P 500 exchange traded fund (ticker SPY) at IEX during December, 2016. The distibution is broken down across several classifications. Investors in our model place midpoint pegs and market orders. Unfortunately, none of the values in Table 2 correspond precisely with investor orders in our model: the interarrival times for midpoint pegs include those of investors and proprietary traders, whereas the interarrival times for investor orders (labelled 'Agency') include all order types. We thus approximate the intensity of investor arrivals in the following manner.

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

Table 2: Quantiles of the distribution of interarrival times (measured in nanoseconds) of submitted orders for the S&P 500 exchange traded fund (ticker SPY) at IEX during December, 2016.

Table 3 reveals that investors (agencies) account for 65.7% of all midpoint peg transactions. Using this value to extrapolate the share of order submission in Table 2, we find that the median time between investor arrivals is approximately $404/0.657 = 615$ milliseconds, which equates to 1.62 arrivals per second or roughly 100 arrivals per minute on both sides of the market. This suggests a baseline arrival rate for each side of the market of $\rho = 100/2 = 50$. A similar approximation, using the median interarrival time of 248 milliseconds for agency orders, and the share of volume in Table 3 for the agency order types in our model, yields a nearly identical value of ρ.

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

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

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

C Numerical Comparative Statics

We gain insight into the comparative static properties of our model by investigating numerically what happens when we vary structural parameters, one at a time, away from their baseline values.

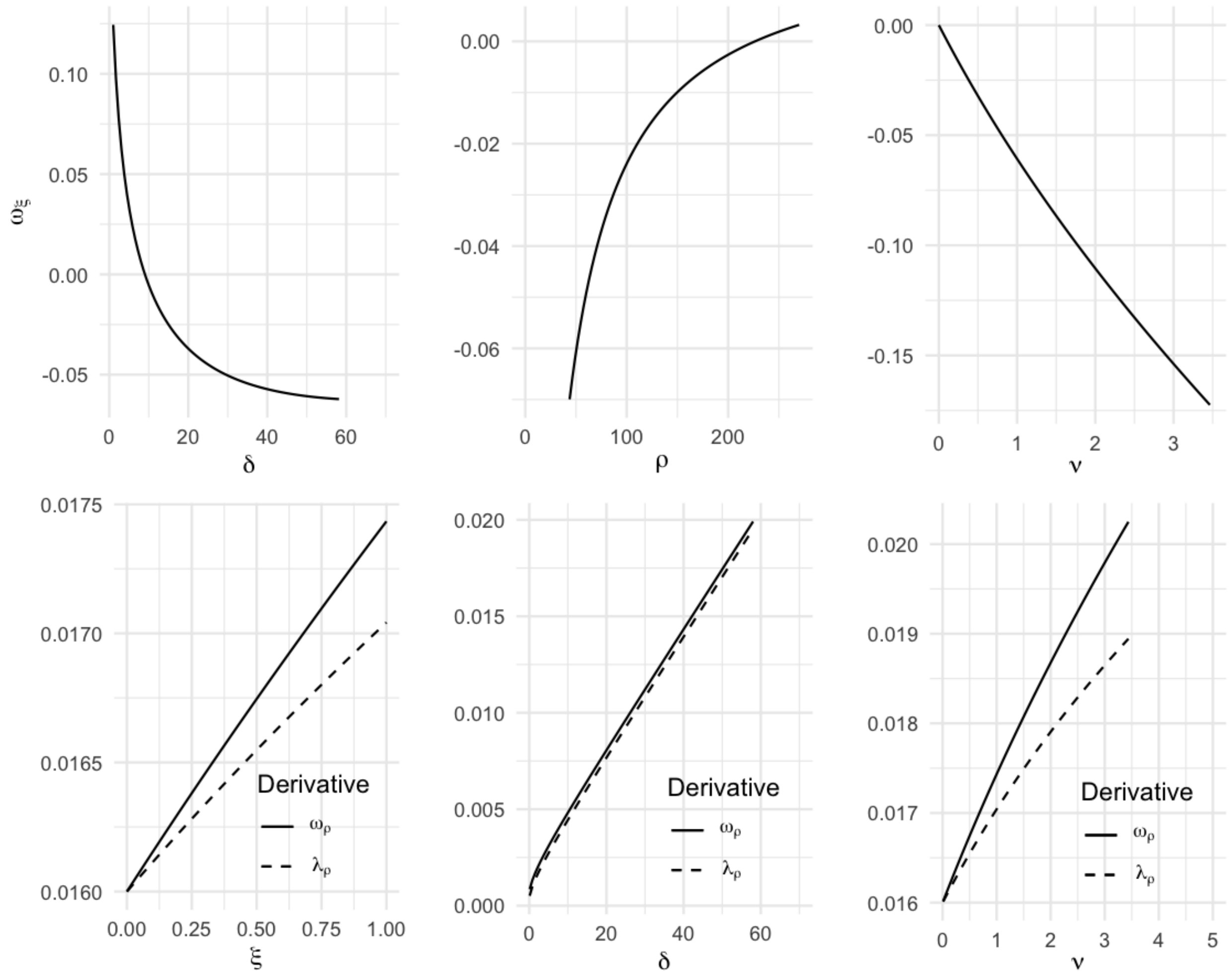

Figure 3: Numerical comparative statics of equilibrium peg fraction ω. Top (resp. bottom) panels show how impact of peg vulnerability ξ (resp. of investor arrival rate ρ) varies with impatience δ, investor arrival rate ρ and fundamental innovation rate ν, other parameters at baseline (resp. $\xi = 1$ in the last two bottom panels.)

Figure 3 sheds some light on key ambiguous comparative statics. As suggested in the Counterexample in Section 4.1, the upper left and center panels show that ω indeed can increase in vulnerability ξ (i.e., we have $\omega_\xi > 0$) for very small values of δ and large values of ρ. The upper right panel shows that, for other parameters at baseline values, ω_ξ remains negative for all admissible values of ν. That panel suggests that $\omega_\xi = 0$ when $\nu = 0$; to confirm this analytically, note that if $\nu = 0$ or $\xi = 0$, then equation (A.22) implies that $\omega_\xi = \lambda_\xi$ while (A.19-A.20) imply that $\lambda_\xi = 0$.

Since ν and ξ appear in those equations only via their product, the partial derivative ω_ν is the same (up to a positive multiplicative factor) as that graphed for ω_ξ in those two panels. In particular, ω decreases in the fundamental value innovation rate ν over the entire admissible range of ν and ξ, except (as noted earlier) when δ is very small and ρ is large.

The lower panels of Figure 3 show that the impact of investor arrival rate ρ on the equilibrium peg fraction ω closely tracks its impact on DL parameter λ. Both impacts are comfortably positive, but fall to near zero (other structural parameters at baseline) as investors become infinitely patient ($\delta = 0$).

D Dynamic Strategies

We now extend the model to analyze dynamic strategies that investors could use when choosing whether to submit a peg or a market order. To do so, in Section D.1 we obtain continuous time expressions $\mathbf{p}(t)$ for how the queue distribution evolves away from steady state. In Section D.2 we introduce a more general parametric class of distributions than DL. With these preliminaries in hand, Section D.3 analyzes the profitability of appealing dynamic strategies, and Section D.4 reports empirical results.

D.1 Stochastic State Dynamics

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

$$\mathbf{p}(t) = e^{-2\rho t} \sum_{n=0}^{\infty} \frac{(2\rho t)^n}{n!} \mathbf{T}^n \mathbf{p^o}. \tag{D.1}$$

We are not aware of any explicit closed form expressions for the components $p_k(t)$, but it is straightforward to simulate them numerically using (D.1).

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

$$
\begin{aligned}
\mathbf{A}(t) &= \frac{1}{t}\int_0^t \mathbf{p}(s)ds \\
&= \frac{1}{t}\int_0^t e^{-2\rho s}\sum_{n=0}^{\infty}\frac{(2\rho s)^n}{n!}\mathbf{T}^n\mathbf{p}^o ds \\
&= \sum_{n=0}^{\infty}\mathbf{T}^n\mathbf{p}^o\frac{1}{2\rho t}\int_0^t \frac{(2\rho)^{n+1}s^n}{\Gamma(n)}e^{-2\rho s}ds \\
&= \sum_{n=0}^{\infty}\frac{G(t;n+1,2\rho)}{2\rho t}\mathbf{T}^n\mathbf{p}^o \qquad \text{(D.2)}
\end{aligned}
$$

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

D.2 Asymmetric Discrete Laplace Distribution

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

$$
p_k = \begin{cases} \dfrac{(1-\omega_R)(1-\omega_L)}{1-\omega_R\omega_L}\omega_L^{|k|} & \text{if } \; k \le 0 \qquad \text{(D.3a)} \\ \dfrac{(1-\omega_R)(1-\omega_L)}{1-\omega_R\omega_L}\omega_R^{|k|}, & \text{if } \; k \ge 0. \qquad \text{(D.3b)} \end{cases}
$$

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

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

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

$$
p_k^L(\alpha) = \frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2}\alpha^{|k|}\omega^{|k|} \qquad \text{(D.4)}
$$

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

$$p_k^H(\alpha) = \frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2}\omega^{|k|}. \tag{D.5}$$

Panel (b) of Figure 4 plots these frequencies for several values of t. Note that the frequencies lie along a triangle with apex at $k = 0$ and that the apex height decreases monotonically from $\ln(1-\omega)$ at $t = 0$ to an asymptote (as $t \to \infty$) of $\ln(1-\omega) - \ln(1+\omega)$. The light side (absolute) slope monotonically decreases from infinite to $-\ln\omega$, while the heavy side (absolute) slope remains constant at $-\ln\omega$.

D.3 Contingent actions

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

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

Below we investigate both intuitive conjectures. In brief, we find that strategy (b) doesn't work. Placing a market order when the midpoint queue is known to be heavier than usual does not increase (or decrease) profit in market equilibrium. The mathematically informed intuition is that the posterior expected queuing cost is the same in case (b) as it is when the investor's own order fails to execute at midpoint, and that posterior expectation is compatible with market equilibrium as in Proposition 2.

To a limited extent we will vindicate the intuition behind strategy (a). Indeed, if the investor were able to react immediately to the empty queue, we will see that strategy (a) would increase profit above that earned in equilibrium (as given, e.g., in equation (2.2)) by the proportion $\frac{\beta\omega(1-\beta)(1-\omega)(1+\omega)}{\omega(1-\beta\omega)+\beta(1-\omega)}$. However, for reasonable parameters, this maximal proportional advantage is small. Our best numerical estimate is that the expected advantage (taking into account the lag until investors arrive, and relaxation towards the steady state distribution) is less than 1%. Also, available data shows no evidence of such strategies; see Section D.4.

To begin the investigation, suppose that $\alpha(t) < 1$. By Equations (D.4) - (D.5) we have

$$\sum_{k=1}^{\infty} p_k^H(\alpha) = \frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2}, \qquad \sum_{k=0}^{\infty} p_k^H(\alpha) = \frac{(1-\alpha\omega)}{1-\alpha\omega^2} \tag{D.6}$$

$$\sum_{k=1}^{\infty} p_k^L(\alpha) = \frac{\alpha\omega(1-\omega)}{1-\alpha\omega^2}, \qquad \sum_{k=0}^{\infty} p_k^L(\alpha) = \frac{(1-\omega)}{1-\alpha\omega^2}. \tag{D.7}$$

Thus

$$\begin{aligned} \pi_P^L(\alpha) &= (\varphi-d)\left[\sum_{k=1}^{\infty} p_k^H(\alpha) + \sum_{k=0}^{\infty} p_k^L(\alpha)\beta^{k+1}\right] \\ &= (\varphi-d)\left[\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \beta\frac{(1-\omega)(1-\alpha\omega)}{1-\alpha\omega^2}\sum_{k=0}^{\infty}(\alpha\beta\omega)^k\right] \\ &= (\varphi-d)\left[\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\alpha\beta\omega)(1-\alpha\omega^2)}\right], \end{aligned} \tag{D.8}$$

while

$$\begin{aligned} \pi_M^L(\alpha) &= (\varphi-d)\sum_{k=1}^{\infty} p_k^H(\alpha) + (\varphi-1)\sum_{k=0}^{\infty} p_k^L(\alpha) \\ &= (\varphi-d)\frac{\omega(1-\alpha\omega)}{1-\alpha\omega^2} + \frac{(\varphi-1)(1-\omega)}{1-\alpha\omega^2}. \end{aligned} \tag{D.9}$$

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

$$\begin{aligned} \Delta\pi^L(\alpha) &= \pi_P^L(\alpha) - \pi_M^L(\alpha) \\ &= \frac{(\varphi-d)\beta(1-\omega)(1-\alpha\omega)}{(1-\alpha\beta\omega)(1-\alpha\omega^2)} - \frac{(\varphi-d)\beta(1-\omega)(1-\omega)}{(1-\beta\omega)(1-\alpha\omega^2)} \end{aligned}$$

$$= (\varphi - d)\frac{\beta(1-\omega)}{1-\alpha\omega^2}\left[\frac{1-\alpha\omega}{1-\alpha\beta\omega} - \frac{1-\omega}{1-\beta\omega}\right], \tag{D.10}$$

The factor in brackets is zero at $\alpha = 1$, which yields $\Delta\pi^L(1) = 0$, as must be the case in equilibrium. Inspection of the same factor shows that the payoff advantage is positive over the relevant parameter range when $\alpha \in [0, 1)$. It is is maximal at $\alpha = 0$, where Equation (D.10) reduces to

$$\Delta\pi^L(0) = (\varphi - d)\omega(1-\beta)\frac{\beta(1-\omega)}{1-\beta\omega}. \tag{D.11}$$

Recall that steady-state market-order equilibrium profit is

$$\pi^L(1) = \left[\frac{\varphi - d}{1+\omega}\right]\frac{\omega(1-\beta\omega) + \beta(1-\omega)}{1-\beta\omega}. \tag{D.12}$$

Hence strategy (a) obtains proportional advantage of at most

$$\frac{\Delta\pi^L(0)}{\pi^L(1)} = \frac{\beta\omega(1-\beta)(1-\omega)(1+\omega)}{\omega(1-\beta\omega)+\beta(1-\omega)}. \tag{D.13}$$

For reasonable parameters, this maximal proportional advantage is small; Section D.4 shows that realistic estimates (taking into account the evolution of $\alpha(t)$) are far below this upper bound.

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

$$\begin{aligned}\Delta\pi^H(\alpha) &= \pi_P^H(\alpha) - \pi_M^H(\alpha)\\ &= (\varphi - d)\frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\beta\omega)(1-\alpha\omega^2)} - (\varphi - 1)\frac{(1-\alpha\omega)}{(1-\alpha\omega^2)}\\ &= (\varphi - d)\left[\frac{\beta(1-\omega)(1-\alpha\omega)}{(1-\beta\omega)(1-\alpha\omega^2)} - \frac{\beta(1-\omega)}{(1-\beta\omega)}\frac{(1-\alpha\omega)}{(1-\alpha\omega^2)}\right] = 0.\end{aligned} \tag{D.14}$$

D.4 Empirical Dynamics

Our data consist of total time (measured in nanoseconds) during December 2016 market hours on the Investors' Exchange (IEX) that the midpoint peg plus discretionary peg order queues for the heavily-traded security SPY (an ETF that tracks the S&P 500 index of US equities) spent in each possible state k. We combine the discretionary and midpoint peg state times, since 89% of discretionary peg transactions occur at midpoint (see Table 3 and Section F.7), and confine attention to state values $|k| \leq 6$, which account for 99.97% of the total trading time (4.912751×10^{14} nanoseconds) during December 2016.

Here we analyze potential strategic behavior of investors immediately following publicly observed transactions at best bid or offer. As noted in Section D.3, such events reveal information about the state of the midpoint order queue. Therefore we focus on subsets of data immediately following transactions at BBO, measured over discrete windows of time. Those time windows are 100 microseconds, 100 milliseconds, 1 second, 10 seconds, and 60 seconds. The data are split on a BB/BO conditioning variable, and (as expected) are symmetric across that variable. For concision we focus our analysis on strategic behavior following BB transactions.

Panel (a) Figure 4 depicts the empirical distributions (fraction of time) of midpoint pegs for each time window, following BB transaction events. To account for the small volume of residual midpoint bid orders remaining in the queue due to size restrictions[11] at the moment of transaction events, we establish the 100 microsecond midpoint bid distribution as baseline and use it to normalize the

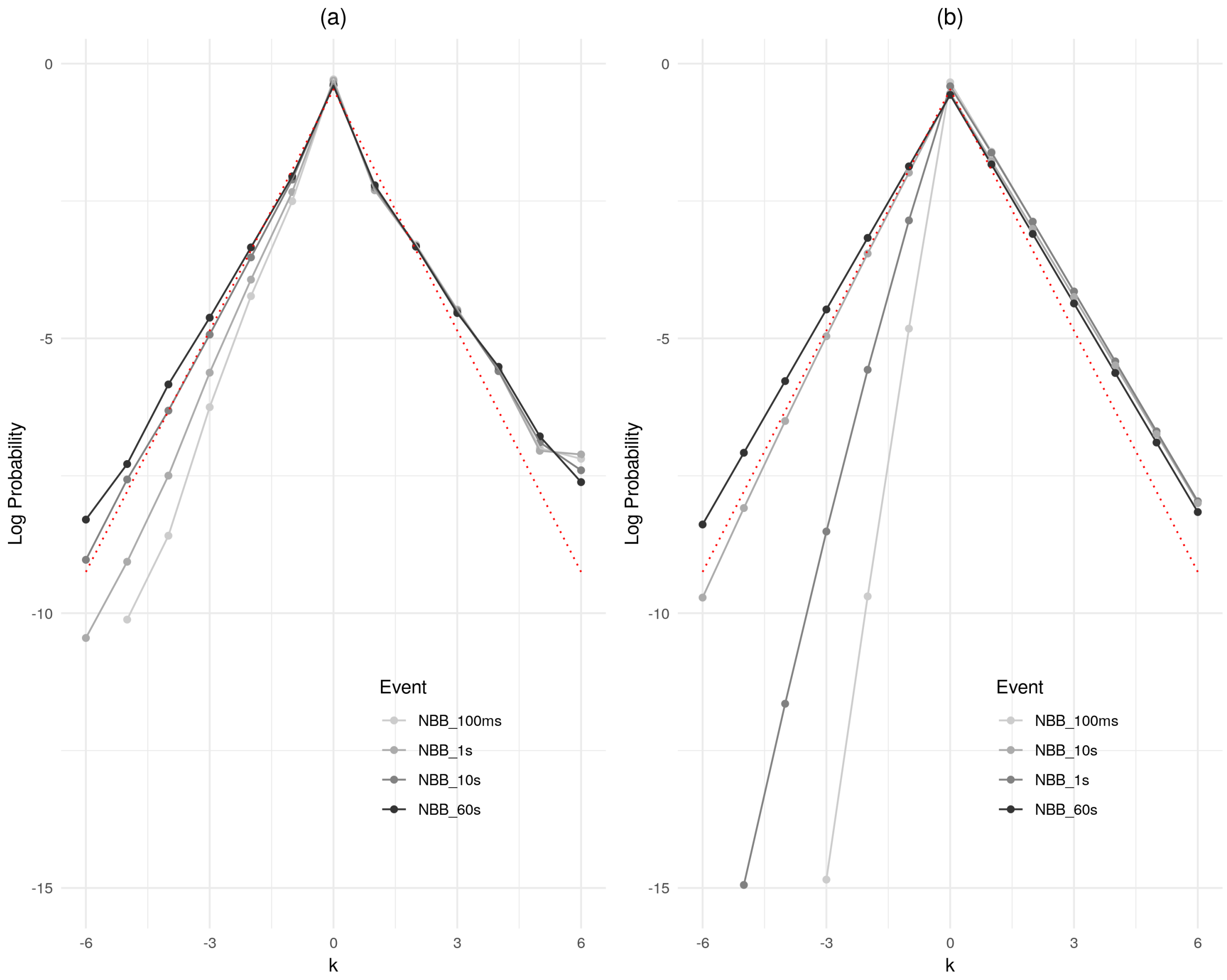

Figure 4: Dynamic response to BB transaction, empirical (Panel a) and theoretical (Panel b). The dotted red lines indicate the steady state distribution for $\omega = .23$.

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

empirical distributions at the other horizons. Panel (b) shows the theoretical distribution for the same time horizons, obtained by iterating on Equation (D.2), with $\rho = 50$.

Aggregating the data for each time horizon, we are able to infer that there are approximately 30,000 BB events in the data. These events, however, are not independent. Using the 62 days of SPY transaction data associated with the quotations referenced in Section B.4[12] (for calibration of the jump intensity), we estimate the autocorrelation functions (ACFs) of inter-trade durations on a daily basis. The 62 autocorrelation functions are depicted in Figure 5, along with 0.025, 0.5, and 0.975 quantiles (dotted lines) at each lag. The transaction data exhibits clear persistance for long horizons – during the sample period, there was roughly 1 transaction every 1.5 seconds, which means that the 100-lag horizon of Figure 5 corresponds to roughly 2.5 minutes of clock time. Following the methodology of Thiebaux and Zwiers (1984), we use the ACF values $\{\rho_i\}_{i=1}^{M}$, truncated at $M = 1000$ lags, to estimate an effective sample size (ESS) ratio for each day,

$$\zeta_t = \frac{N_t}{N_t^e} = 1 + 2\sum_{i=1}^{N_t - 1} \rho_i, \tag{D.15}$$

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

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

$$\hat{\omega}_L = \begin{cases} \frac{2\bar{K}^-(1+\bar{K})}{1+2\bar{K}^- - \bar{K} + \sqrt{1+4\bar{K}^-\bar{K}^+}}, & \text{if } \bar{K} \geq 0 \\ \frac{\hat{\omega}_R - \bar{K}(1-\hat{\omega}_R)}{1-\bar{K}(1-\hat{\omega}_R)}, & \text{otherwise,} \end{cases} \tag{D.16}$$

$$\hat{\omega}_R = \begin{cases} \frac{2\bar{K}^+(1-\bar{K})}{1-2\bar{K}^+ + \bar{K} + \sqrt{1+4\bar{K}^-\bar{K}^+}}, & \text{if } \bar{K} \leq 0 \\ \frac{\hat{\omega}_L + \bar{K}(1-\hat{\omega}_L)}{1+\bar{K}(1-\hat{\omega}_L)}, & \text{otherwise,} \end{cases} \tag{D.17}$$

$$\Sigma_{MLE} = \frac{\omega_L \omega_R (1-\omega_L)(1-\omega_R)}{1+\omega_L\omega_R} \begin{bmatrix} \frac{(1-\omega_R)(1-\omega_R\omega_L^2)}{\omega_L(1-\omega_L)^2} & 1 \\ 1 & \frac{(1-\omega_L)(1-\omega_L\omega_R^2)}{\omega_R(1-\omega_R)^2} \end{bmatrix}, \tag{D.18}$$

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

We use the delta method to estimate $\hat{\alpha} = \hat{\omega}_L/\hat{\omega}_R$ for each of the four time windows in the

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

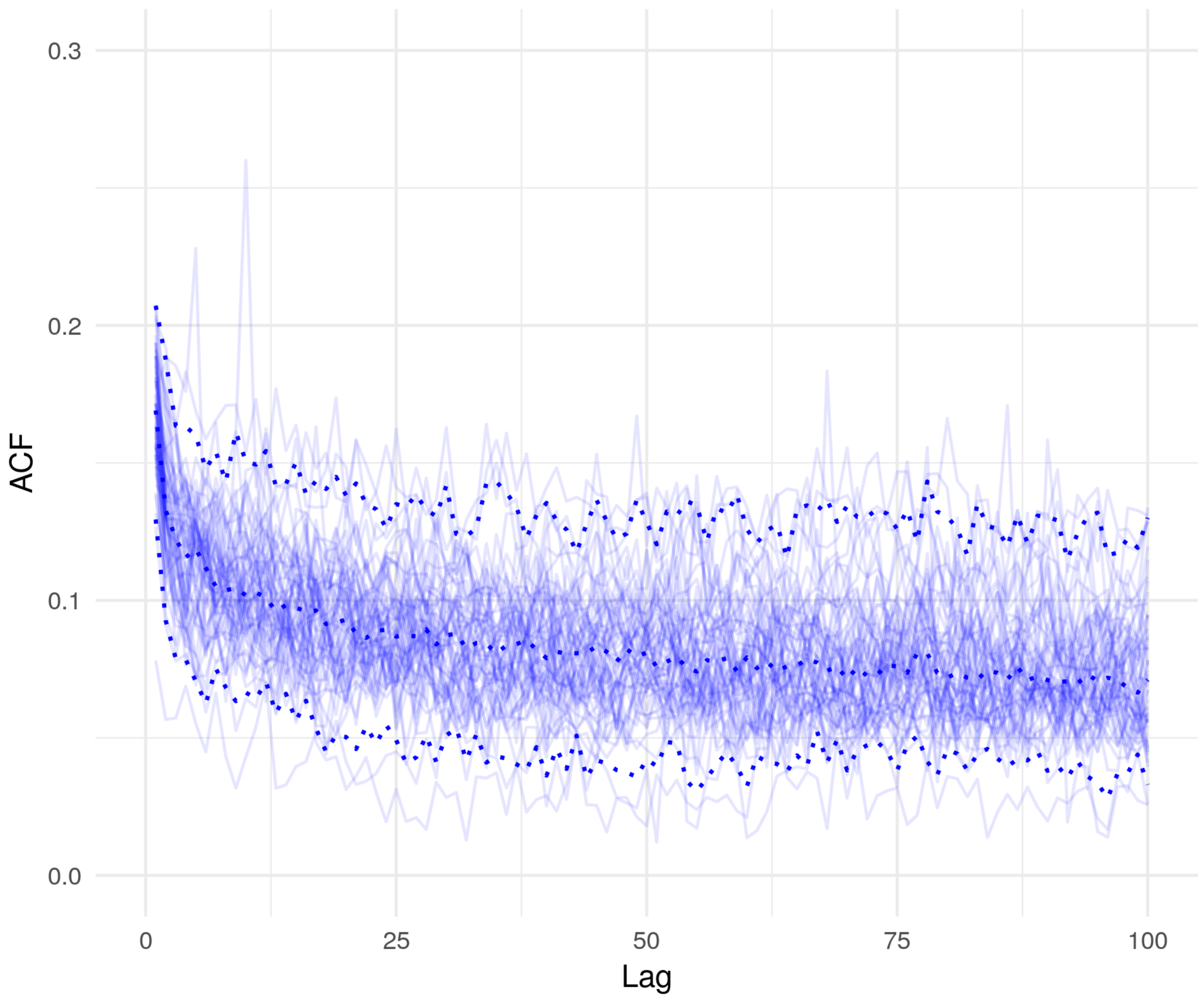

Figure 5: Daily autocorrelation functions of inter-trade durations for SPY transactions on the Nasdaq exchange, 16, Jun 2014 – 11 Sep, 2014. Dotted lines represent 0.025, 0.5, and 0.975 quantiles at each lag.

data provided by IEX. The point estimates are represented by blue dots in Figure 6. We also compute 95% asymptotic error bands, scaling the asymptotic variance by the effective sample size and expected number of investor arrivals during each time window, $2\rho N^e$. The 95% intervals are represented by blue diamonds and vertical blue dotted lines. Additionally, we obtain theoretical probabilities of the order queue distribution by iterating on Equation (D.2) for a fine sequence of $t \in (0, 60)$ seconds, and likewise obtain point estimates and asymptotic errors for the theoretical distribution. The theoretical point estimates are depicted as a solid black line in Figure 6, and the 95% interval estimtes are depicted as black dotted lines. For instances where the confidence bands span zero (i.e. for low values of t), we truncate the lower bounds at zero. Given the remarkably large overlap in confidence intervals, we conclude that the ADL is a good approximation both for our model and the data.

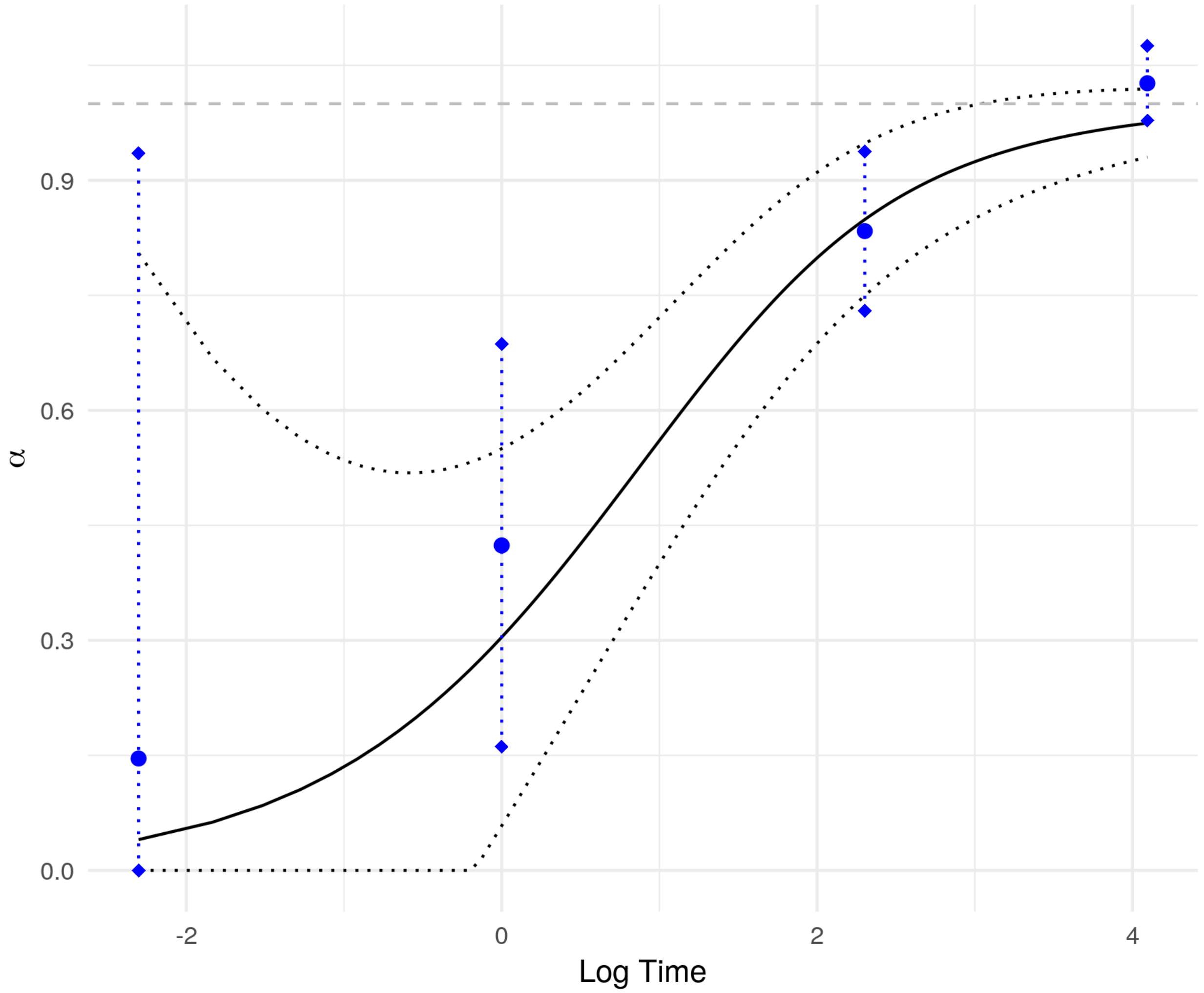

Figure 6: Decay function, theoretical vs empirical. Theoretical prediction from Equation (D.2) for $\rho = 50$ (solid black line) with 95% asymptotic confidence bands (black dotted lines), vs estimates from the December 2016 IEX data (blue dots) and 95% asymptotic confidence bands (blue diamonds and dotted lines).

The foregoing analysis suggests that, indeed, publicly observed transactions at BBO provide useful strategic information to investors. To understand the economic importance of this information, we estimate the proportional increase in profit that a single investor could earn during the 60 seconds following all BBO transactions in December 2016. Specifically, we integrate Equation D.10 over $t \in (0, 60)$ seconds, substuting our theoretical point estimates, $\hat{\alpha}(t)$, in Figure 6. The proportional increase in profit, relative to equilibrium profit in Equation (2.2), for 30,000 60-second periods is 0.56% (0.0056).

E Miscellaneous Mathematical Details

Here we collect tangential results alluded to at various points in the text.

E.1 Obtaining ω from λ

Corollary 3. *Given parameters ν, ρ and ξ, the steady-state fraction of investors choosing to place midpoint peg orders is*

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

where λ is the steady state value determined in Proposition 1.

Proof. The result is obtained by solving for ω in Equation (A.10).

Remark. Clearly ω is strictly increasing in λ for admissible parameter values, so its inverse function $\lambda(\omega|\xi,\nu,\rho>0)$ exists and is also strictly increasing.

E.2 Discounting

Recall that Section 2.5 showed that the discount factor for an exponentially distributed delay time is $\beta = \int_0^\infty e^{-\delta t}\rho e^{\rho t}dt = \frac{\rho}{\rho+\delta} < 1$, where $\delta > 0$ is the impatience rate and $\rho > 0$ is the one-sided Poisson investor arrival rate. Here we establish formally the intuitive result that if one has to wait for $k = 1, 2, 3, ...$ investors to arrive, the discount factor is indeed β^k.

One can establish the desired result by mathematical induction, or by direct calculation. The latter begins with the Poisson probability

$$F(t) = 1 - e^{-\rho t}\sum_{n=0}^{k-1}\frac{(\rho t)^n}{n!} \tag{E.2}$$

that there will be at least k investor arrivals by time $t > 0$. The associated density is known to be $f(t) = \frac{t^{k-1}\rho^k}{(k-1)!}e^{-\rho t}$, i.e., Gamma with with parameters $\alpha = k$ and $\beta = 1/\rho$, as can be verified by differentiating E.2. The desired discount factor is the expected value of $e^{-\delta t}$ for this distribution of t:

$$\mathrm{E}\left[e^{-\delta t}\right] = \int_{t=0}^{\infty} e^{-\delta t}\frac{t^{k-1}\rho^k}{(k-1)!}e^{-\rho t}dt = \left(\frac{\rho}{\rho+\delta}\right)^k \int_{t=0}^{\infty}\frac{t^{k-1}(\rho+\delta)^k}{(k-1)!}e^{-(\rho+\delta)t}dt = \left(\frac{\rho}{\rho+\delta}\right)^k = \beta^k. \tag{E.3}$$

E.3 Limiting Case $\omega \to 1$

Consider the fully protected case $\xi = 0$, and recall from equation (3.3) that $q_k = \left(\frac{1-\omega}{1+\omega}\right)\omega^{|k|}$ is the probability of queue state k in steady state. As $\omega \to 1$, we have $q_k \to 0$, and indeed $\sum_{k=-K}^{K} q_k \to 0$ for any finite positive integer K. Thus in the limit we have an improper distribution on Z, in which the probability "leaks out to $\pm\infty$". The expected wait time therefore gets arbitrarily large and the present value gets arbitrarily close to zero as $\omega \to 1$. With equal probability in the limit, a pegged order either executes immediately or languishes at the end of an arbitrarily long queue, and so has payoff $\pi_p = \frac{\varphi}{2}$. By contrast, the payoff to a market order remains at $\pi_m = \frac{\varphi}{2} + \frac{\varphi-1}{2} > \pi_p$. Thus, since admissible $\varphi > 1$, the equal profit condition fails (as does the relevant Kuhn-Tucker condition $\pi_m \leq \pi_p$.) We conclude that $\omega = 1$ is never part of a market equilibrium when $\xi = 0$. To the extent that lesser peg protection decreases ω (Section 4.1 notes a caveat), the same is true a fortiori when $\xi > 0$.

E.4 Makers and Speed Purchases

Under assumption A5, market makers do not purchase speed. In equilibrium, when would it be profitable for a single maker to deviate, and violate that restriction?

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. Focusing for simplicity on the fully protected case $\xi = 0$, and using equation (3.8), we see that this deviation is not worthwhile if

$$\begin{aligned} \frac{2\nu}{N_s+1} \leq c \iff 2\nu &\leq (N_s+1)c \\ \iff \nu &\leq \rho\,\frac{1-\omega}{1+\omega} + \frac{c}{2}. \end{aligned} \tag{E.4}$$

At baseline, $\nu = 1$ and $\rho\frac{1-\omega}{1+\omega} + \frac{c}{2} \approx 50\,\frac{0.77}{1.23} + \frac{10}{2} \approx 36$, so such a deviation is indeed highly unprofitable. Clearly the same conclusion holds for a large neighborhood around the baseline parameters.

Might purchasing speed enable a market maker to transact more frequently? In our model, the answer is essentially no. A new market order always executes against the oldest of the $n > 1$ contra-side resting r orders at BBO. Even if the maker who held that resting order were able to renew it instantaneously, it would still enter the queue at position n, as would (absent a rare intervening event) a renewed order by a slow maker.

A rigorous analysis of queue dynamics when some proper subset of makers placing BBO lit limit orders (and a proper subset of traders who place midpoint pegged orders) have purchased speed is beyond the scope of the present model. Notes on how to pose and numerically solve such a dynamic model are available on request from the authors.

E.5 Reconciling Welfare Results

As noted in the introductory literature survey, and as suggested in the concluding discussion of our paper, some authors model an exchange-imposed delay η as directly costly to impatient investors. Such direct costs, if sufficiently high, might overturn our result that greater protection never impairs investor welfare W.

Readers might note that η dropped out of our list of structural variables in Table 1; it was subsumed into ξ, the probability that pegs are vulnerable to sniping. Footnote 7 suggests ignoring the direct cost of η on the grounds that, in practice, it is three or four orders of magnitude smaller than the delays associated with resting in the midpoint order queue.

Readers interested in constructing a more comprehensive model that includes both sorts of delay cost might proceed as follows. Replace our expression of investor welfare, $W = \varphi - \frac{1}{1+\lambda}$ by the more general expression $\tilde{W} = [\varphi - \frac{1}{1+\lambda}]e^{-\delta\eta}$. Replace the structural parameter ξ by an appropriate smooth monotonic function $f(\eta)$ where $f(0) = 1$ and $f(\eta) = 0$ for all $\eta \geq \bar{\eta} > 0$. Then replace comparative statics such as W_ξ by expressions such as $\tilde{W}_\eta$.

We have not yet pursued that idea for three reasons. First, it occurred to us only very recently, inspired by the final round of comments of a helpful referee. Second, we are not sure how to calibrate $f(\eta)$ other than to say that, empirically, it seems that $\bar{\eta} < 0.00035/60 = 0.0000583$ minutes. Third, that empirical bound suggests that we were justified in ignoring the direct cost of the exchange-imposed delay η, because the multiplicative factor $e^{-\delta\eta}$ lies in the narrow interval $[e^{-\delta\bar{\eta}}, e^{-\delta 0}] \subset [0.997, 1]$ at baseline $\delta = 50$.

F Institutional Information

F.1 Basics

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. We then present summary data from the first exchange to implement that format, and use that data to motivate elements of the model introduced in Section 2.

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

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 exchanges with less aggressive quotations to pass orders on to exchanges with better bids or offers.

Most 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.

F.2 Delayed messaging format

A variant of the CDA market format implemented by IEX and by NYSE American 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.

Details regarding common pegged order types in use at IEX and NYSE American are in Section F.7.

F.3 Some Data

Table 3 reports transaction volume statistics at IEX during the month of December 2016.[13] 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). 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

[13] In response to our request, the IEX made these data available to us with the understanding that we will make them available to other academics.

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

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 3 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 investor clients.

3. Midpoint and discretionary peg orders transacting at midpoint: 47.0% of volume. Our model will attribute this to investor clients.

Following is a similar breakdown for proprietary firms.

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 pool this order flow with midpoint orders transmitted by agencies on behalf of their clients.

F.4 Exchanges Imposing Delay

Messaging delays in a variety of forms are seeing increasing adoption at financial exchanges worldwide. Examples include the Investors Exchange (IEX), New York Stock Exchange (NYSE) American, Toronto Stock Exchange (TSX) Alpha, Aequitas Neo, Electronic Broking System (EBS), Thomson Reuters, London Metal Exchange (LME), Chicago Board Options Exchange (CBOE) EDGA, Intercontinental Exchange (ICE), Eurex, and the Moscow Exchange. Broadly speaking these messaging delays can be described as either deterministic or random, and as either symmetric or asymmetric.

Our model allows a random delay η, realized independently whenever there is a jump in the fundamental value, such that the probability that midpoint pegs are protected is $\xi \in (0, 1)$. Of course, the model also allows a deterministic delay corresponding to $\xi = 0$ or to $\xi = 1$. Delays in our model are symmetric in that they apply equally to all traders and all new orders, but there is an asymmetry in that repricing of resting pegged orders occurs with no exchange-imposed delay.

IEX and NYSE American are the first and most prominent examples of exchanges that impose a deterministic and symmetric messaging delay of 350 microseconds on all new orders (but not on repricing pegs). As such delays involved an apparent violation of Regulation NMS Rule 611 (the "Order Protection" rule), which obliged exchanges to "immediately" pass orders to exchanges with more competitive prices, the SEC modified the interpretation of "immediacy" in a June 2016 decision. By explicitly allowing a buffer of one millisecond for the order protection rule, the SEC accommodated the IEX and NYSE American messaging delays, which allowed for their subsequent approval as national securities exchange June 2016 and May 2017, respectively.

Unlike the foregoing systems, TSX Alpha, launched in September 2015, imposes a longer, random delay of 1 – 3 milliseconds on all pegged orders, both displayed and nondisplayed. It is asymmetric in that "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 bypass the delay by paying an explicit fee to the exchange.

The "latency floor" mechanisms of the EBS and Thomas Reuters currency exchanges are different from those above and involve an intricate form of randomized delay. In brief, a new order triggers a batching period of random length (e.g. one, two, or three milliseconds) during which all subsequent orders are collected and randomized by trader ID. The orders are then processed in a round-robin procedure, beginning with the first order of each trader, and subsequently the second order of each trader, and so on. The delay is largely symmetric in that all traders and most orders are treated in the same (random) way; the asymmetric exception is that all cancellations are processed first in each batch.

F.5 Fees

Trading fees have interesting interactions with messaging delays. Traditional fee structures are maker-taker, in which passive (making) orders receive rebates and aggressive (taking) orders pay fees. However, a few exchanges use an "inverted" taker-maker fee structure in which passive orders pay fees when executed and aggressive orders received rebates. As noted above, TSX Alpha uses an inverted fee structure to allow traders to "pay" for protection.

As fee structures at exchanges are heterogeneous and often quite complex (dependent on a variety order types and trader statuses) we choose for our model to be agnostic of fees. However, some readers may be interested in modeling the impact of altering the fee structure. We now outline straightforward extensions of our model can incorporate some simple fee structures.

Consider, for example, the fee structure recently employed by IEX: all orders that add (make) or remove (take) non-displayed liquidity are charged $0.0009 and all orders that add or remove displayed liquidity are charged $0.0003. To accommodate such a fee structure, let d be the difference between non-displayed and displayed fees, so in this example $d = 0.0006$. Rewrite our equations (2.2) and (2.3) as follows:

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

$$\pi_p = (\varphi - d) \left[\sum_{k=-\infty}^{-1} q_k + \sum_{k=0}^{\infty} q_k \hat{\beta}^{k+1} \right] - 2 \sum_{k=0}^{\infty} q_k \left[1 - \left(\frac{\rho}{\rho + \xi\nu} \right)^{k+1} \right]. \tag{F.2}$$

That is, the terms representing payoffs from pegged (non-displayed) orders are now multiplied by $\varphi - d$ instead of φ, while the other terms, representing payoffs from regular (displayed) orders are unchanged. It is then straightforward (though occasionally messy) to propagate these modifications through derivations of market equilibrium and thus to study how this fee structure affects market outcomes.

F.6 Order Routing

In accordance with Regulation National Market System (Reg NMS), all exchanges in the United States route orders to quotations at other exchanges when those quotations offer price improvement. For example, 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 F.3, routable orders constitute only 12% 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.

F.7 Pegged Order Types

Pegged order types come in many flavors. In this section, we provide details for the most common pegged order types in use 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 3 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 3 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[14]

[14] 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 × \$10.01 and the IEX order book is \$10.00 × \$10.02 when a nonroutable hidden buy order arrives with a limit of \$10.01. The order will be booked at \$10.005 and will later transact at \$10.01 if a sell limit arrives at that price. Alternatively, it may transact with midpoint pegs, discretionary pegs, or market orders at midpoint.

and later transact with contra-side orders at midpoint.

A recent addition to the set of IEX order types is the Discretionary Limit Order, or D-Limit, that was approved by the SEC on August 26, 2020. The D-Limit is automatically repriced to a less aggressive price (one price increment below the best bid or above the best offer) when the crumbling quote indicator (described in the next section) is activated. Otherwise, the D-Limit is identical to a regular limit order.

What would be the impact on equilibrium predictions if our model included this new order type? It would not be attractive to our investors who, as explained in A.12, are too impatient to use limit orders that rest at BBO. Our model's market makers might use them, balancing the reduction in sniping costs (when the crumbling quote indicator forecasts correctly) against the reduction in market order executions due losing position in the BBO queue when responding to crumbling quote false alarms. Our Assumption A4c (deep market at BBO) evidently implies that including this new limit order type would have no impact on equilibrium λ or ω (i.e., on steady state midpoint order queues) and their comparative statics. Including D-limit orders could shift equilibrium N_r and N_s (masses of market makers and snipers), but we don't anticipate any change in the model's qualitative predictions. On the other hand, D-Limit orders could affect extensions of the model that drop A4c and examine BBO dynamics explicitly. To the extent that D-Limit orders displace r-orders, the BBO spread would tend to widen when the crumbling quote indicator is active. In an expanded model relaxing Assumption A6c, the availability of D-limit orders might reduce market makers' demand for speed.

F.8 Crumbling Quote

The volume statistics for midpoint pegs in Table 3 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.