Journal of Financial Economics 141 (2021) 968–980

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Who provides liquidity, and when?☆

Sida Li[a], Xin Wang[b], Mao Ye[a,c,*]

[a] University of Illinois at Urbana–Champaign, Gies College of Business, 340 Wohlers Hall, 1206 S. 6th Street, Champaign, IL 61820, USA
[b] Nanyang Business School, Nanyang Technological University, Division of Banking and Finance, Singapore 639798, Singapore
[c] National Bureau of Economic Research, 1050 Massachusetts Ave, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 2 January 2019
Revised 29 May 2020
Accepted 5 June 2020
Available online 29 April 2021

JEL classification:
G14
G18

Keywords:
High-frequency trading
Algorithmic trading
Tick size
Liquidity
Bid-ask spread

ABSTRACT

We model competition for liquidity provision between high-frequency traders (HFTs) and slower execution algorithms (EAs) designed to minimize investors' transaction costs. Under continuous pricing, EAs dominate liquidity provision by using aggressive limit orders to stimulate HFTs' market orders. Under discrete pricing, HFTs dominate liquidity provision if the bid-ask spread is binding at one tick. If the tick size (minimum price variation) is not binding, EAs choose between stimulating HFTs and providing liquidity to non-HFTs. Transaction costs increase with the tick size but can be negatively correlated with the bid-ask spread when all traders can provide liquidity.

Published by Elsevier B.V.

1. Introduction

In decades past, specialists on the New York Stock Exchange and dealers on Nasdaq provided liquidity to other traders by buying when other traders sell and selling when other traders buy. These traditional liquidity providers have almost disappeared in modern electronic markets (Clark-Joseph et al., 2017). Anyone can supply liquidity, but no one is obligated to provide it. Providing liquidity simply means posting a limit order (an offer to buy or sell at a specified price). A trade occurs when another trader (a liquidity demander) uses a market order to accept the terms of a posted offer. In the new ecosystem of voluntary liquidity supply, who provides liquidity and who demands liquidity, and when?

One hypothesis is that high-frequency traders (HFTs) become natural liquidity providers in modern electronic markets, because they incur lower operating costs

☆ For helpful comments and suggestions, we thank Ron Kaniel (the co-editor), Albert (Pete) Kyle (the referee), Hengjie Ai, Shmuel Baruch, Malcolm Baker, Dan Bernhardt, Hank Bessembinder, Eric Budish, Tarun Chordia, Thierry Foucault, Larry Glosten, Larry Harris, Katya Malinova, Jingyuan Mo, Joseph Noss, Maureen O'Hara, Monika Piazzesi, Veronika Pool, Neil Pearson, Barbara Rindi, Shrihari Santosh, Andriy Shkilko, Brian Weller, Chen Yao, Bart Yueshen, Marius Zoican, and seminar participants at the University of Rochester, University of California at Los Angeles, Texas A&M University, the University of Florida, and Washington University at St. Louis, as well as conference participants at the Society for Financial Studies (SFS) Cavalcade, the Carlson Junior Conference at the University of Minnesota, the New York University (NYU) Stern Market Microstructure Conference, the Second Sustainable Architecture for Finance in Europe (SAFE) Market Microstructure Conference, the Colorado Front Range Finance Seminar, the Bank of Canada–Laurier Market Structure conference, the Telfer Annual Conference on Accounting and Finance, the Wabash River Conference at Indiana University, the Smokey Mountain Conference at the University of Tennessee, the 2019 Financial Intermediation Research Society (FIRS) Conference, and the Conference on Financial Stability Implications of New Technology organized by the Federal Reserve Bank of Atlanta and Georgia State University. This research is supported by National Science Foundation grant 1,352,936 (jointly with the Office of Financial Research at the US Department of the Treasury) and National Science Foundation grant 1,838,183.

* Corresponding author.
E-mail address: maoye@illinois.edu (M. Ye).

https://doi.org/10.1016/j.jfineco.2021.04.020
0304-405X/Published by Elsevier B.V.

(Carrion, 2013), adverse selection costs (Hoffmann, 2014), and inventory costs (Brogaard et al., 2015; Aït-Sahalia and Sağlam, 2017). These cost advantages imply that HFTs should win the price competition in liquidity provision. Surprisingly, Yao and Ye (2018) find the opposite: Non-HFTs tend to quote more aggressive prices than HFTs. Who, then, are these non-HFTs and why can they undercut HFTs?

A more complex challenge is that no types of traders consistently dominate liquidity provision. For example, Yao and Ye (2018) and O'Hara et al. (2018) find that HFTs provide relatively more liquidity for low-priced stocks and that non-HFTs provide relatively more liquidity for high-priced stocks. Yao and Ye (2018) show that non-HFTs provide more liquidity as adverse selection risk increases. Therefore, who provides liquidity depends endogenously on security characteristics. Traditional market microstructure theory usually exogenously assigns who provides liquidity. Then, what drives the cross-sectional variations on who provides liquidity?

One key to addressing these questions is to explore uncharted territory: algorithmic traders who are not HFTs. To minimize transaction costs, buy-side institutions, such as mutual funds and pension funds, use computer algorithms extensively to execute their trades (Frazzini et al., 2018; O'Hara, 2015). Although execution algorithms (EAs) are key players in the financial ecosystem (Hasbrouck and Saar, 2013), they lack an independent identity in existing models. According to one view, financial markets include HFTs and everyone else, with the latter covering both sophisticated institutions and unsophisticated retail traders [see the survey by O'Hara (2015)]. According to the other view, algorithmic traders and HFTs are interchangeable [see the survey by Biais and Foucault (2014)]. The stark simplicity of these classifications has proved valuable and appropriate for studying basic foundational questions about machine-human interactions, but the same dichotomy prevents the study of interactions between different types of algorithms, and such machine-machine interactions are the key to understanding the current structure of financial markets.

Our model captures two fundamental differences between EAs and HFTs (Hasbrouck and Saar, 2013; O'Hara, 2015). First, EAs can use limit orders to provide liquidity, but their goal is to minimize transaction costs, not to profit from bid-ask spreads (Hasbrouck and Saar, 2013; Jones, 2013). Second, EAs are fast, but they are slower than HFTs (O'Hara, 2015). The differences in incentives and trading speed between algorithms are the main drivers of our model. EAs enjoy lower opportunity costs for providing liquidity because they must complete a trade. EAs can lose money when providing liquidity as long as the loss is lower than paying the bid-ask spread. We find that EAs always choose to provide liquidity at more aggressive prices than HFTs if pricing is continuous. In reality, pricing is discrete. For example, the US Securities and Exchange Commission (SEC) Regulation National Market Systems (Reg NMS) Rule 612 mandates a uniform tick size (minimum price variation) of 1 cent for any quote above $1. Discrete pricing increases the cost to EAs of narrowing the bid-ask spread. At the extreme, EAs cannot undercut HFTs at all if the bid-ask spread is binding at one tick. Under discrete pricing, we discover three types of equilibria, thereby offering testable predictions regarding who provides liquidity and when.

Our model contains N competitive HFTs who have no private value to trade. HFTs provide or demand liquidity to maximize their expected trading profits. Non-HFTs arrive at the market following a Poisson process, bringing inelastic demand to buy or sell one unit of a security. A fraction β of non-HFTs are EAs, and they can choose between limit and market orders to minimize transaction costs, and the remaining non-HFTs are market order traders (MOTs). When $\beta = 0$, our model degenerates into the Budish et al. (2015; BCS hereafter) framework, in which all non-HFTs must demand liquidity. MOTs have to pay a positive bid-ask spread even if the fundamental value of a security is public information. The positive bid-ask spread results from the sniping risk: The quote from an HFT can be sniped by $N-1$ equally fast HFTs during value jumps.

Once we allow non-HFTs to provide liquidity, they never demand liquidity when pricing is continuous, because the following strategy dominates demanding liquidity. Suppose that an EA buyer submits a limit order at price ε above the fundamental value. Her order immediately stimulates HFTs to submit market orders to collect ε as profit. The EA loses ε, but the loss is lower than the bid-ask spread when ε is sufficiently small. HFTs immediately accept a lower price than the ask price, because accepting an offer does not expose them to a sniping risk. Under continuous pricing, our model generates only one type of equilibrium, in which EAs provide liquidity to HFTs at the fundamental value and HFTs immediately accept the offer. The EA's limit order executes immediately like a market order and does not rest in the limit order book (LOB). HFTs, therefore, provide liquidity to MOTs.

When pricing becomes discrete, our model generates three types of equilibria because discrete pricing creates rents for both providing and demanding liquidity. When the sniping risk is very low relative to the tick size, the breakeven bid-ask spread can drop below one tick. The difference between the one tick–mandated bid-ask spread and the breakeven bid-ask spread becomes rents for providing liquidity, and speed allocates these rents to HFTs. In the first type of equilibrium, the queuing equilibrium, HFTs dominate liquidity provision because they can achieve time priority at the constrained one-tick bid-ask spread. This prediction is consistent with Yao and Ye (2018), who find that HFTs dominate liquidity provision when either the adverse selection risk is too low or the tick size is too large.

The second and the third type of equilibria occur when the breakeven spread for HFTs becomes wider than one tick. EAs can place limit orders within the spread, but they can no longer stimulate HFTs at the fundamental value. Instead, they need to cross the fundamental value to stimulate HFTs. The difference between the price that HFTs accept and the fundamental value is the rent for HFTs to demand liquidity and the loss for EAs to stimulate HFTs. EAs' choices then depend on whether such loss is larger than the loss from improving the breakeven bid-ask spread. Fig. 1 shows that the second and third type of equilibria rotates with parameter values. The second

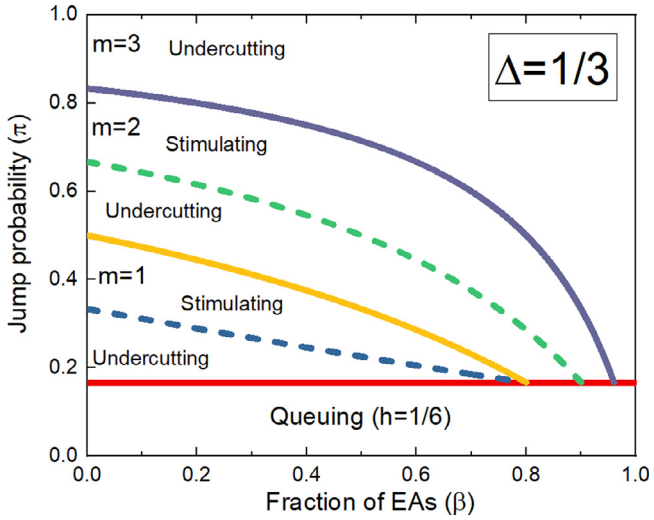

Fig. 1. Parameter ranges of the equilibria for the example of tick size $\Delta = \frac{1}{3}$. When sniping risks are very low, high-frequency traders (HFTs) quote a one-tick bid-ask spread and provide liquidity to all execution algorithms (EAs) and market order traders (MOTs) (Proposition 3, queuing equilibrium). When sniping risks are higher, HFTs widen their quoted spreads and provide liquidity to MOTs only. The regions in which HFTs quote differing spreads are separated by solid lines. Within each region, EAs choose between stimulating HFTs (Proposition 4, stimulating equilibrium) and undercutting HFTs (Proposition 5, undercutting equilibrium) depending on the parameters. When the tick size is not binding, the undercutting equilibrium alternates with the stimulating equilibrium with respect to π and β. m is the number of tick grids between the midpoint and the ask price, and h is the half bid-ask spread.

type of equilibrium, the undercutting equilibrium, arises when the breakeven ask (bid) price is close to the price grid below (above). As the loss incurred by improving the breakeven spread is small, EAs choose to undercut the breakeven spread and provide liquidity to other non-HFTs. The third type of equilibrium, the stimulating equilibrium, arises when the breakeven ask (bid) price is far from the price grid below (above). As the cost for improving the breakeven ask and bid prices is higher than stimulating HFTs, EAs choose to cross the midpoint to stimulate HFTs. HFTs then race to take liquidity to capture the rents from demanding liquidity. In the stimulating equilibrium, EAs provide liquidity to HFTs, and HFTs provide liquidity to MOTs because limit orders from EAs never stay in the LOB.

By opening the door to exploring the diversity of computer algorithms, our paper not only develops new predictions but also generates new perceptions. Machine-machine interaction blurs the distinction between providing and demanding liquidity. By definition, an EA provides liquidity because she uses a limit order, but her goal is to trigger immediate market orders from HFTs. In this sense, the EA demands liquidity because her stimulating limit order executes like a market order. HFTs use market orders, but they satisfy the EA's trading needs. In this sense, HFTs provide liquidity even though they take quotes from EAs.

After blurring the distinction between providing and demanding liquidity, this machine-machine interaction also makes it challenging to measure liquidity. Traditional liquidity measures, such as the bid-ask spread and depth, are based on orders resting on the LOB. Our model shows that the liquidity offered by the LOB can move in the opposite direction of true liquidity. Consider the case in which pricing is continuous. The bid-ask spread increases as the fraction of EAs increases because fewer non-HFTs take liquidity from HFTs. The average transaction costs that non-HFTs incur decline, however, because more EAs enjoy zero transaction costs. At the extreme, the bid-ask spread reaches its widest magnitude when all non-HFTs are EAs, but the market becomes infinitely liquid because non-HFTs always choose to stimulate HFTs when no transaction costs exist.

The discrete pricing channel of speed competition reconciles a number of contradictions between existing channels of speed competition and empirical facts, for both supplying liquidity and demanding liquidity. Regarding supplying liquidity, Carrion (2013), Hoffmann (2014), and Brogaard et al. (2015) show that speed reduces HFTs' intermediation costs, particularly adverse selection costs. Such reduced costs should give HFTs a competitive advantage in providing liquidity for stocks that are subject to higher adverse selection risk (Han et al., 2014). In addition, when the tick size is small, HFTs face less constraint in offering better prices and they should therefore crowd out liquidity provision by non-HFTs. Yao and Ye (2018) find, however, that an increase in adverse selection risk reduces HFTs' share in liquidity provision and that a small tick size crowds out liquidity provision from HFTs. The discrete pricing channel helps to reconcile these apparent contradictions. EAs can provide liquidity at better prices because they enjoy lower opportunity costs for doing so. Lower adverse selection risk or larger tick size, however, reduces the breakeven spread to below one tick and lead to speed competition at constrained prices.

Liquidity demand from HFTs usually has a negative connotation because, in existing models, HFTs typically adversely select liquidity providers when they demand liquidity (BCS; Foucault et al., 2017; Menkveld and Zoican, 2017). In our model, EAs use aggressive limit orders to stimulate HFTs. HFTs demand liquidity, but they do

not adversely select EAs. Instead, the transaction costs that EAs pay are lower when HFTs demand liquidity than when HFTs provide liquidity. This surprising prediction could explain why Latza et al. (2014) find that limit orders executed within 50 milliseconds after submission incur no adverse selection costs.

The closest paper to ours is BCS. We relax two assumptions made in BCS. First, because non-HFTs in BCS can use only market orders, the sniping risk leads to a positive bid-ask spread, motivating BCS to recommend frequent batch auctions. Our model shows that when all non-HFTs can choose between limit and market orders, transaction costs drop to zero once pricing is continuous. Second, we relax their assumption of continuous pricing and show that discrete pricing generates an arms race in speed. Therefore, BCS argue for a more discrete market with respect to time, while as an alternative, we posit a more continuous market with respect to pricing. Our insight undermines the rationale for increasing the tick size to 5 cents as proposed by the 2012 US Jumpstart Our Business Startups (JOBS) Act and the SEC's 2016 tick size pilot program. Proponents of increasing the tick size argue that a larger tick size increases liquidity, discourages HFTs, increases market-making profits, supports sell-side equity research, and, eventually, increases the number of initial public offerings (IPOs) (Weild et al., 2012). Our results show that an increase in the tick size reduces liquidity, encourages speed racing between HFTs, and allocates resources to latency reduction.

The paper proceeds as follows. Section 2 sets up the model. Section 3 presents the results when the price is continuous, and all non-HFTs use market orders. In Section 4 we allow some non-HFTs to use limit orders but keep price continuous and compare with the results in Section 3. In Section 5, we add one more realistic feature to our model: discrete pricing. Section 6 contains empirical predictions and policy implications. Section 7 concludes with a discussion for future research.

2. Model

We consider a continuous-time trading model with an infinite horizon. There is one security, and its fundamental value v_t is common knowledge, $t \in [0, \infty)$, with $v_0 = 0$.

There are two types of traders: HFTs and non-HFTs. All traders are risk-neutral, and there is no time discounting. There are N $(2 \le N \le \infty)$ HFTs always present in the market with the goal of maximizing expected trading profits.

Every non-HFT has an inelastic need to buy or sell one unit of a security. A fraction β of non-HFTs can choose between limit and market orders to minimize transaction costs. These are the EAs. The remaining fraction of $1-\beta$ non-HFTs, MOTs, use only market orders upon arrival.

As in BCS, non-HFTs arrive at the market at Poisson intensity λ_I, and the asset's fundamental value, v_t, evolves as a compound Poisson jump process with arrival rate λ_J. Therefore, at any time t, with probability $\pi = \frac{\lambda_J}{\lambda_I+\lambda_J}$ the next event is a value jump, and with probability $1-\pi$ the next event is an arrival of a non-HFT. During the value jump, the fundamental value of the security, v_t, increases to v_t+1 or decreases to v_t-1 with equal probability. Upon arrival, non-HFTs buy or sell with equal probability. Therefore, the probability that the next event is the arrival of a buy MOT is $\frac{1}{2}(1-\beta)(1-\pi)$.

The stock exchange operates as a continuous LOB. Each trade in the LOB requires a liquidity provider and a liquidity demander. The liquidity provider submits a limit order, which is an offer to buy or sell at a specified price and quantity. The liquidity demander accepts the price and quantity of a limit order. Execution precedence for liquidity providers follows price-time priority. Limit orders with higher buy or lower sell prices execute before less aggressive limit orders. For limit orders queuing at the same price, orders arriving earlier execute before later orders. HFTs are all equally fast, and their order messages (limit orders, market orders, or cancelations) are processed serially in random order if those messages arrive at the exchange at the same time.

As in BCS, v_t is common knowledge, but liquidity providers are subject to a sniping risk if they fail to update their stale quotes following value jumps. EAs are always subject to such a sniping risk because they are slower than HFTs, but an HFT is also subject to a sniping risk with a positive probability because other HFTs are equally fast. In equilibrium, sniping stale quotes are always profitable, and we allow an HFT to snipe her own quotes because it is economically equivalent to order cancelation.

The LOB contains all outstanding limit orders. Outstanding orders to buy are called bids, and outstanding orders to sell are called asks. In this paper, we focus on the highest bid and the lowest ask around v_t, which are called the best bid and ask. To simplify exposition, we assume that a limit order is canceled if it has no chance of trading with a non-HFT before the next jump occurs. This simplification rules out the trivial strategy of laying the book, that is, attempting to satiate all bid price levels below v_t-1 or all ask price levels above v_t+1, because such pricing levels involve no sniping risk at time t.

3. Continuous-pricing model without EAs

When $\beta = 0$, all non-HFTs use market orders, and our model essentially degenerates into the BCS framework. We use this section as a benchmark for evaluating, in Sections 4 and 5, the impact of allowing non-HFTs to provide liquidity.

Let h be the HFTs' quoted half bid-ask spread. We consider, without loss of generality, the expected payoff for an HFT's sell limit order at v_t+h if the HFT posts the first share at this price. When a buy event occurs, this sell order has the following expected payoff:

$$(1-\pi)\cdot h - \frac{N-1}{N}\pi\cdot(1-h) + \frac{1}{N}\pi\cdot 0. \tag{1}$$

Conditional on the occurrence of a buy event, with probability $1-\pi$ a non-HFT takes the limit order and the payoff is h, and with probability π v_t jumps upward by one and all HFTs race to snipe stale quotes on the ask side. HFTs are equally fast, so the probability that an HFT order is sniped by other HFTs is $\frac{N-1}{N}$. The loss for being sniped is $(v_t+1) - (v_t+h) = 1-h$.

An HFT's outside option beyond providing liquidity is to snipe the share when v_t jumps. The value for this outside option is zero when a non-HFT takes the share before the value jumps. When v_t jumps upward, each HFT has a $\frac{1}{N}$ chance of sniping the share and the payoff for the successful sniper is $(v_t+1)-(v_t+h)=1-h$. Therefore, the expected payoff for a sniper at v_t+h is

$$\frac{1}{N}\pi\cdot(1-h). \tag{2}$$

In equilibrium, HFTs should be indifferent between liquidity provision and stale quote sniping. Thus, the equilibrium half bid-ask spread is the value $h=h_0$, which solves

$$(1-\pi)\cdot h-\frac{N-1}{N}\pi\cdot(1-h)=\frac{1}{N}\pi\cdot(1-h). \tag{3}$$

The best bid and ask prices contain only one share because undercutting h_0 or quoting a second share at h_0 loses money. We summarize the equilibrium in Proposition 1.

Proposition 1. When non-HFTs cannot supply liquidity, the equilibrium half bid-ask spread under continuous pricing is $h_0=\pi$.

HFTs almost always maintain one unit in the LOB at the ask price v_t+h_0 *and one unit at the bid price* v_t-h_0.[1] *The bid and ask prices can belong to different HFTs.*

(i) *Upon arrival, non-HFTs demand liquidity from HFTs and pay* h_0.

(ii) *When* v_t *jumps up (down), all HFTs race to take stale ask (bid) quotes.*

The positive transaction cost h_0 results from the sniping risk (BCS).[2] Next, we show that non-HFTs can avoid such costs completely as long as they can use limit orders.

4. Continuous-pricing model with EAs

In this section, we relax only one assumption made in BCS. We allow a fraction of $\beta>0$ of non-HFTs to provide liquidity. EAs' objective function is to minimize expected transaction costs, and they can update their orders at any time.[3]

After we allow non-HFTs to provide liquidity, they never demand liquidity from HFTs under continuous pricing because crossing the midpoint strictly dominates paying HFTs a half-spread. When an EA submits a buy limit order at $v_t+\varepsilon$, the order immediately stimulates HFTs to demand liquidity because HFTs make a profit of ε by selling above the fundamental value. EA loses ε by providing liquidity, but the cost is lower if $\varepsilon<h$. Therefore, EAs never demand liquidity from HFTs.

The foregoing discussion reveals two new economic mechanisms. The first mechanism is the opportunity cost of liquidity provision. EAs can afford more aggressive limit order prices than HFTs because EAs enjoy a negative opportunity cost for providing liquidity. EAs' outside option for providing liquidity is to demand liquidity by paying h. Therefore, EAs can afford buy limit orders at price $v_t+\varepsilon$ as long as the loss is less than h. HFTs do not provide liquidity if they lose money. What is more, HFTs incur a positive opportunity cost for providing liquidity. When an HFT chooses to provide liquidity, she cannot profit from sniping the share, and the probability of sniping conditional on a value jump is $\frac{1}{N}$. This positive opportunity cost is equal to HFTs' reduced sniping cost relative to the cost to EAs.[4] Therefore, EAs can afford more aggressive quotes than HFTs for any of the model's parameter values.

The second mechanism, the make-take spread, captures the difference in prices between a trader's willingness to post an offer and her willingness to accept an offer. An HFT quoting an ask price of v_t+h would accept any buy limit price $v_t+\varepsilon$ $(\varepsilon\to 0)$. HFTs accept a worse price than the price they offer because accepting an order incurs no sniping risk. In our model, an HFT seller accepts any offer at or above v_t, but she quotes v_t+h. As a consequence, the EA's buy limit order at $v_t+\varepsilon$ executes immediately like a market order.

We discover the make-take spread because we allow traders to switch between providing and demanding liquidity at any time. Models with market makers, such as those used in Kyle (1985) and Glosten and Migrom (1985), exogenously assign the liquidity-provider and liquidity-demander roles. In LOB literature [Foucault et al. (2005), among others], traders can choose limit orders or market orders upon arrival, but they can no longer update their roles after the initial decision.

Proposition 2 characterizes the equilibrium. In the equilibrium, EAs always choose limit order prices at v_t and HFTs immediately demand liquidity once EAs submit such limit orders. EAs provide liquidity to HFTs, but the LOB contains no resting limit orders placed by the EAs. HFTs provide liquidity to MOTs by quoting one share at v_t+h_β and one share at v_t-h_β. The half-spread h_β equalizes the payoff of liquidity provision and stale quote sniping for HFTs, which is given by

$$\frac{(1-\beta)(1-\pi)}{(1-\beta)(1-\pi)+\pi}\cdot h-\frac{N-1}{N}\frac{\pi}{(1-\beta)(1-\pi)+\pi}\cdot(1-h)=\frac{1}{N}\frac{\pi}{(1-\beta)(1-\pi)+\pi}\cdot(1-h). \tag{4}$$

The left-hand side of Eq. (4) is the HFT seller's liquidity-provision profit conditional on a trade's occurring at the ask, and the right-hand side is a sniper's profit. The only difference between Eq. (4) and Eq. (3) is the factor $(1-\beta)$ [Eq. (4) degenerates into Eq. (3) when $\beta=0$]. We call Proposition 2 the stimulating equilibrium because EAs, who have an internal need to trade, use aggressive limit orders to stimulate HFTs to demand liquidity.

Proposition 2. ***(stimulating equilibrium).*** *With a tick size of zero and a positive fraction of EAs* $(\beta>0)$*, the equilibrium half bid-ask spread is* $h_\beta=\frac{\pi}{1-\beta(1-\pi)}$.

[1] HFTs' stale quotes can be sniped during value jumps, but HFTs immediately replenish the share around the new fundamental value.

[2] Aquilina, Budish, and O'Neill (2020) quantify return from the sniping.

[3] As our paper focus on costs led by the sniping risk, we assume away delay costs for EAs (Parlour, 1998; Foucault, 1999; Foucault, Kadan, and Kandel, 2005).

[4] During value jumps, an HFT is sniped with probability of $\frac{N-1}{N}$ and an EA is sniped with probability of one.

(i) HFTs almost always maintain one unit in the LOB at the ask price $v_t + h_\beta$ and one unit at the bid price $v_t - h_\beta$.
(ii) EAs submit limit orders at v_t when they arrive and all HFTs immediately demand liquidity at v_t.
(iii) When v_t jumps up (down), all HFTs race to take stale limit orders at the ask (bid) price.

In the existing literature, when HFTs demand liquidity, they usually adversely select other traders (BCS; Menkveld and Zoican, 2017; Foucault et al., 2017). Consequently, liquidity demand from HFTs often has negative connotations. Our model shows that HFTs can demand liquidity without adversely selecting other traders. Instead, transaction costs are lower for EAs when HFTs demand liquidity than when EAs demand liquidity from HFTs. Therefore, researchers and policymakers should not evaluate the welfare impact of HFTs based simply on whether they provide or demand liquidity.

Proposition 2 also shows that the definitions of providing and demanding liquidity blur when a machine interacts with another machine. Technically, EAs are liquidity providers in equilibrium because they use limit orders, but the goal of their limit orders is to immediately attract HFTs to submit market orders. Therefore, HFTs economically provide liquidity to EAs even though the HFTs use market orders.

The blurring of the distinction between providing and demanding liquidity has, in turn, a significant impact on how liquidity is measured. When more non-HFTs use limit orders, competitive HFTs receive fewer order flows and have to quote wider bid-ask spreads for the remaining market orders. Therefore, $h_\beta > h_0 > 0$. Despite an increase in the bid-ask spread, Corollary 1 shows that the total transaction costs for non-HFTs decrease, because EAs never pay the bid-ask spread. Let $\bar{C}(\beta)$ denote the weighted average transaction cost for all non-HFTs. We then have

$$\bar{C}(\beta) = \beta \cdot 0 + (1-\beta)\cdot h_\beta = \frac{(1-\beta)\pi}{1-\beta(1-\pi)}. \tag{5}$$

Corollary 1. The half-spread h_β strictly increases in β and $\bar{C}(\beta)$ strictly decreases in β. When $\beta \to 1$, $h_\beta \to 1$ and $\bar{C}(\beta) \to 0$.

Corollary 1 shows that the quoted bid-ask spread, a common measure of liquidity, can move in the opposite direction of the actual transaction costs when every trader can provide liquidity. As the proportion of EAs' β increases, the quoted bid-ask spread widens, but transaction costs fall. When all non-HFTs are EAs, HFTs' half bid-ask spreads widen to one, the maximum possible value jump size, but transaction costs zero out. The transaction costs drop because the bid-ask spread no longer represents all traders' interest in satisfying the trading needs of non-HFTs.

BCS show that continuous trading creates sniping risks and positive transaction costs for non-HFTs. Corollary 1 shows that their results no longer hold when all traders can provide liquidity. When all traders are EAs ($\beta = 1$), HFTs make zero profits in equilibrium and they have no economic incentive to invest in speed.[5] In the next section, we show that discrete pricing generates rents for both providing and demanding liquidity, thereby triggering an arms race in speed.

5. Discrete pricing

In this section, we add another realistic feature to our model: discrete pricing. In Section 5.1, we show that discrete pricing creates rents for providing liquidity. In Section 5.2, we show that discrete pricing also creates rents for demanding liquidity. These rents, in turn, destroy the unique type of equilibrium outlined in Section 4, in which EAs always provide liquidity to HFTs and HFTs always provide liquidity to MOTs. Discrete pricing generates three types of equilibria depending on parameter values, which then lead to cross-sectional and time-series predictions regarding who provides liquidity to whom.

We set the tick size $\Delta = \frac{1}{L}$. Therefore, the tick size decreases as L increases. For illustration purposes, we assume that $L = 1, 2, 3, \cdots$ and that v_0 is at the midpoint of a tick. Therefore, v_t is always at the midpoint of the two nearest ticks and the available price grids are $\{\cdots, v_t - \frac{3\Delta}{2}, v_t - \frac{\Delta}{2}, v_t + \frac{\Delta}{2}, v_t + \frac{3\Delta}{2}, \cdots\}$.

When v_t does not coincide with a pricing grid, EAs can no longer stimulate HFTs' market orders at zero cost. A positive cost for stimulating HFTs could lead EAs to choose limit orders that reside in the LOB. These resting orders could cause an explosion in the number of states in the LOB because infinitely many EAs will arrive in the future. To reduce the number of states, we make the following assumption for discrete pricing that is common in the LOB literature (Foucault et al., 2005).[6]

Assumption 1. Limit orders must be price-improving, that is, they must narrow the spread by at least one tick.

We relax Assumption 1 in the Online Appendix and find that the model's main intuition holds with dramatically greater mathematical complexity. Assumption 1 reduces the state of the LOB to 2^n, where n is the number of price levels between HFTs' best bid and ask prices. Assumption 1 is not binding under continuous pricing because the best bid and offer contains only one share when pricing is continuous. Therefore, we are able to compare the results obtained under discrete pricing with those obtained under continuous pricing.

Assumption 1 also rules out infinite loops between traders. For instance, without this assumption, an EA can submit an unprofitable undercutting order to force an incumbent EA's profitable limit order to cancel. Then, the entrant EA revises her undercutting order to the incumbent's

[5] When execution algorithms can use only market orders, Li and Ye (2021) show that bid-ask spread is still zero if execution algorithms can divide their demand of one share into a series of infinitesimal child orders. Therefore, sniping leads to positive spread under two conditions. First, investors cannot choose between market and limit orders. Second, if they can use market orders only, they face discrete lot size or other frictions that prevent them from slicing their orders.

[6] Goettler, Parlour, and Rajan (2005) allow limit orders to queue at the same price, but they have to rely on numerical solutions.

profitable price level. In return, the initial incumbent could want to fight back with the same strategy. By disallowing a trader to worsen the quotes, including her own quotes, our assumption rules out such trivial loops without affecting the model's main economic insights.

Assumption 2. $N = \infty$.

In Sections 3 and 4, we show that the number of HFTs does not affect the equilibrium bid-ask spread quoted by HFTs. An increase in N reduces the value of providing liquidity because it increases the probability of being sniped. An increase in N, however, reduces the value of sniping stale quotes by the same amount because each sniper is less likely to be successful. Therefore, N can affect sniping costs and opportunity costs, but it cannot affect the sum of these two costs. In turn, the equilibrium bid-ask spread does not depend on N as long as there is more than one HFT. To simplify the notation, we drop $\frac{N-1}{N}$ from our exposition by assuming that the number of HFTs is infinite. Consequently, the expected sniping profit for any share is zero, and an HFT provides liquidity as long as its expected profit is greater than zero.

5.1. Rents for providing liquidity and the queuing equilibrium

Consider the extreme case when $h_0 \to 0$, where the breakeven spreads in Propositions 1 and 2 are both close to zero. In that case, the bid-ask spread is binding at one tick and the tick size becomes pure rent for providing liquidity. As long as the breakeven bid-ask spread is smaller than one tick, the difference between the mandated one-tick minimum spread and the breakeven spread creates rents for providing liquidity, and the time priority rule allocates such rents to HFTs. EAs are not able to provide liquidity because they can neither win time priority nor place limit orders within the bid-ask spread. Therefore, a low sniping risk relative to the tick size leads to a queuing equilibrium, in which HFTs provide liquidity to both EAs and MOTs.

Proposition 3. (***queuing equilibrium***). *When tick size is* $\Delta = \frac{1}{L}$ *and* $\pi \le \frac{\Delta}{2}$ $(h_0 \le \frac{\Delta}{2})$,

(i) HFTs almost always maintain one share at the ask price $v_t + \frac{\Delta}{2}$ and one share at the bid price $v_t - \frac{\Delta}{2}$.
(ii) HFTs participate in two speed races: (a) the race to fill the queue when the depth at $v_t \pm \frac{\Delta}{2}$ becomes zero and (b) the race to snipe all stale quotes following a value jump.
(iii) All non-HFTs use market orders to trade upon arrival.

The queuing equilibrium has three features. First and most important, discrete pricing generates speed races to provide liquidity at $v_t \pm \frac{\Delta}{2}$ because the expected profit for providing liquidity on top of the book is higher than the expected profit for sniping the share. When the market opens, each HFT sends one sell limit order at $v_0 + \frac{\Delta}{2}$ and one buy limit order at $v_0 - \frac{\Delta}{2}$, and the winner of each race becomes a liquidity provider. When a non-HFT arrives and takes the order at $v_t + \frac{\Delta}{2}$ or $v_t - \frac{\Delta}{2}$, HFTs race to refill the order.[7] Following value jumps, HFTs race to provide liquidity at a half-spread of $\frac{\Delta}{2}$ around the new fundamental value.

Second, a discrete tick size forces EAs to use market orders because they cannot win the speed race. EAs never use market orders when pricing is continuous because they can use stimulating limit orders to achieve zero transaction costs (Proposition 2). When the tick size is discrete, EAs cannot submit stimulating orders at the fundamental value v_t. If the tick size is also binding, EAs cannot submit limit orders within the spread. EAs always use market orders under Proposition 3 because we assume that each price level can hold only one share. We show in the Online Appendix that the same intuition holds when each price level can hold more than one share. EAs never use market orders when pricing is continuous, but they use market orders when pricing is discrete because they can neither submit limit orders at the fundamental value v_t nor win speed races for top positions.

Third, a discrete tick size increases transaction costs for non-HFTs. The result is straightforward for EAs because they pay zero transaction costs under continuous pricing but they need to pay the bid-ask spread under discrete pricing.[8] The transaction costs for all non-HFTs also increase because the average transaction cost under discrete pricing is higher than $\bar{C}(0)$, and Corollary 1 shows that $\bar{C}(0)$ is greater than $\bar{C}(\beta)$ for any $\beta > 0$.

5.2. Rents for demanding liquidity: Stimulating and undercutting equilibria

When the tick size Δ decreases or sniping risk π increases, the breakeven spread for HFTs is larger than the tick size. As HFTs lose money by providing liquidity at $v_t \pm \frac{\Delta}{2}$, EAs are able to submit limit orders within HFTs' bid-ask spreads. Following a similar intuition expressed in Proposition 2, EAs would never use market orders. EAs can stimulate HFTs to demand liquidity if EAs submit buy limit orders at $v_t + \frac{\Delta}{2}$ or sell limit orders at $v_t - \frac{\Delta}{2}$. These stimulating limit orders strictly dominate market orders because $\frac{\Delta}{2}$ is lower than the half-spread. Discrete pricing, meanwhile, creates one new feature.

When pricing is continuous, stimulating limit orders offer the minimum possible transaction cost of zero. When pricing is discrete, a stimulating limit order costs $\frac{\Delta}{2}$, and a limit order that does not cross the midpoint can cost less. In Proposition 2, h_β is the breakeven spread for HFTs. We define m as the number of tick grid points strictly in the interval $(v_t,\ v_t + h_\beta)$ [or equivalently in the interval $(v_t - h_\beta, v_t)$]. Therefore, the tick grid immediately above $v_t + h_\beta$ is $v_t + (m + \frac{1}{2})\Delta$ and the tick immediately below $v_t + h_\beta$ is $v_t + (m - \frac{1}{2})\Delta$. Lemma 1 shows that $m - 2$ price

[7] Li, Ye, and Zheng (2021) provide empirical evidence on speed races between orders for front queue positions and estimate the economic gains from winning the races for front queue position.

[8] This result holds after we remove Assumption 1, which we do in the Online Appendix. Even when EAs use limit orders to queue after HFTs, the queue positions of EAs involve positive transaction costs. If not, HFTs would occupy those queue positions. Furthermore, an EA would use market orders if the queue were too long.

positions always cost more than $\frac{\Delta}{2}$, leaving only two price positions for further consideration. For an EA seller, these two remaining strategies are (1) submitting a sell order immediately below $v_t + h_\beta$ and (2) selling at $v_t + \frac{\Delta}{2}$, the price immediately above the midpoint.

Lemma 1. When $m > 1$, the strategy for stimulating HFTs strictly dominates all strategies except undercutting h_β by one or m ticks.

The intuition behind Lemma 1 is as follows. When an EA sells above $v_t + \frac{\Delta}{2}$, she can attract only MOTs, because EA buyers never accept an ask price above $v_t + \frac{\Delta}{2}$. Proposition 2 finds that the half-spread to break even by attracting all MOTs is h_β. An EA then loses more than one tick if she improves h_β by more than one tick, which is more costly than stimulating HFTs.

Next, we show that EAs never sell at $v_t + \frac{\Delta}{2}$ or buy at $v_t - \frac{\Delta}{2}$ unless $m = 1$. Consequently, our model generates only two types of equilibria depending on the parameter value. In Section 5.2.1, we present the stimulating equilibrium in which EAs sell at $v_t - \frac{\Delta}{2}$ or buy at $v_t + \frac{\Delta}{2}$. In Section 5.2.2, we present the undercutting equilibrium, in which EAs choose to improve the h_β to the closest tick. The results reported in Fig. 1 show that stimulating equilibria and undercutting equilibria alternate in parameter value and cover the whole parameter space outside the queuing equilibrium. Intuitively, an EA undercuts HFTs on the ask side if the breakeven ask price is less than half a tick away from the price grid immediately below and she chooses to stimulate HFTs if the breakeven ask price is more than half a tick away from the price grid immediately below.

5.2.1. Stimulating equilibrium under discrete pricing

In the stimulating equilibrium, an EA incurs a cost of $\Delta_v = \frac{\Delta}{2}$ to attract HFTs. Proposition 4 outlines the parameter space in which stimulating is the optimal strategy in equilibrium. A trader's strategy involves her response to all states of the LOB, even if such states do not appear in equilibrium. To conserve space and convey the main economic intuition, we describe only traders' best responses in the equilibrium path while deferring their off-equilibrium strategies to the proof.

*Proposition 4. (**stimulating equilibrium under discrete pricing**). When $m = 1$ and $\pi \geq \frac{\beta+2-\beta\Delta-\sqrt{(\Delta+1)^2\beta^2+(4-12\Delta)\beta+4}}{2\beta}$ or $m \geq 2$ and $h_\beta - (m-\frac{1}{2})\Delta \geq \Delta_v$,*

(i) HFTs almost always maintain one share at the ask price $v_t + (m+\frac{1}{2})\Delta$ and one share at the bid price $v_t - (m+\frac{1}{2})\Delta$.
(ii) EA buyers submit limit orders at $v_t + \frac{\Delta}{2}$ and EA sellers submit limit orders at $v_t - \frac{\Delta}{2}$.
(iii) HFTs participate in three speed races: (a) the race to snipe all stale quotes following value jumps, (b) the race to fill the queue when the depth at $v_t \pm (m+\frac{1}{2})\Delta$ becomes zero, and (c) the race to take the liquidity offered by EAs.

Proposition 4 is highly intuitive when $m \geq 2$. In the stimulating equilibrium, the cost of stimulating HFTs ($\frac{\Delta}{2}$) must be lower than the cost of undercutting h_β to the closest tick. When $m \geq 2$, undercutting h_β to the closest tick attracts MOTs and HFT snipers but not other EAs. Therefore, the EA faces the same order flows as that for HFTs in Proposition 2. The breakeven half-spread for providing liquidity to all MOTs is h_β. An EA then loses $h_\beta - (m-\frac{1}{2})\Delta$ when she narrows the spread to the closest tick. Therefore, the EA chooses to stimulate HFTs when $h_\beta - (m-\frac{1}{2})\Delta \geq \Delta_v$. The intuition holds similarly when $m = 1$, although the formula for the boundary becomes more complex, because the EA has to quote at $v_t \pm \frac{\Delta}{2}$ if she chooses to narrow the spread, and the quote now attracts other EAs.

In the stimulating equilibrium, EAs leave no limit orders on the book. Facing no competition in liquidity provision from EAs, HFTs provide liquidity to all MOTs. Even if the tick size is not binding, discrete pricing still creates rents for liquidity provision because HFTs now quote an ask price of $v_t + (m+\frac{1}{2})\Delta$ and a bid price of $v_t - (m+\frac{1}{2})\Delta$. As $(m+\frac{1}{2})\Delta > h_\beta$ by definition, HFTs still race to the top queue position to capture the rent created by the tick size.

The new insight from Proposition 4 is that the tick size generates rents and speed competition for demanding liquidity. When pricing is continuous, EAs can place limit orders at v_t. When pricing is discrete, EAs need to pay $\frac{\Delta}{2}$ more to attract HFTs, and HFTs race to demand liquidity to capture this rent. This race does not exist under continuous pricing because EAs leave no rents for HFTs. This race also does not exist under a queuing equilibrium because no price level exists at which to submit stimulating limit orders.

5.2.2. Undercutting equilibrium

In the undercutting equilibrium, the cost that EAs pay to sell at $v_t + (m-\frac{1}{2})\Delta$ or buy at $v_t - (m-\frac{1}{2})\Delta$ is less than $\frac{\Delta}{2}$. Therefore, the LOB can contain limit orders placed by EAs, and traders' strategies depend on the number of limit orders that EAs place on the LOB. We use $(i_t^1, i_t^2, \cdots, i_t^m, j_t^m, \cdots, j_t^2, j_t^1)$ to denote the state of the LOB, where $i_t^k, j_t^k \in \{0,1\}$ with $1 \leq k \leq m$ and $k \in \mathbb{N}^+$indicate the depth at prices $v_t + (m+\frac{1}{2}-k)\Delta$ and $v_t - (m+\frac{1}{2}-k)\Delta$, which are the prices k ticks below (above) HFT's ask (bid) in Proposition 4.

Fortunately, the proof of Proposition 5 shows that EAs never improve HFTs by more than one tick in equilibrium. For any equilibrium states of the LOB, EAs choose to sell at $v_t + (m-\frac{1}{2})\Delta$ or to buy at $v_t - (m-\frac{1}{2})\Delta$ if these price levels contain no other limit orders. In turn, $i_t^k = j_t^k = 0$ for all $2 \leq k \leq m$ always holds in equilibrium. Therefore, HFTs face four states of the LOB in equilibrium depending on the value of i_t^1 and j_t^1. Instead of $(i_t^1, 0, \cdots, 0, 0, \cdots, 0, j_t^1)$, for simplicity we use (i, j) to represent each equilibrium state of the LOB, where i represents the number of EAs who quote a half-spread of $(m-\frac{1}{2})\Delta$ on the ask side and j denotes the number of EAs who quote a half-spread of $(m-\frac{1}{2})\Delta$ on the bid side. The core of Proposition 5 characterizes HFTs' best response for each event and in each

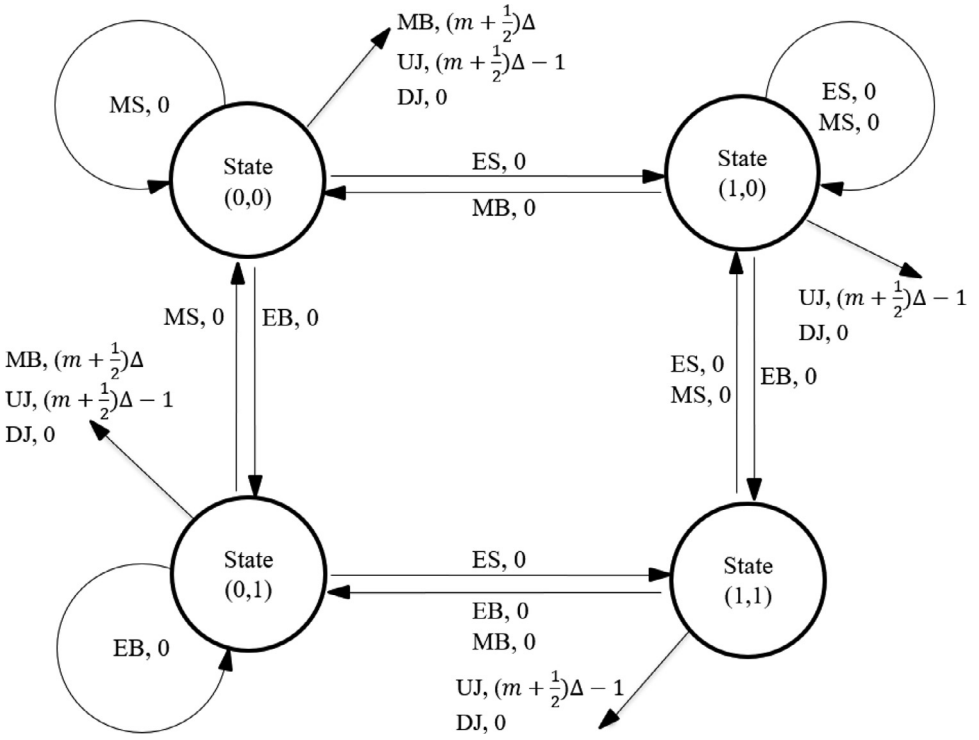

Fig. 2. Markov transition between limit order book (LOB) states and payoffs from the point of view of high-frequency trader (HFT) liquidity providers on the ask side. For instance, in the undercutting equilibrium when $m=1$, HFTs quote at $v_t \pm \frac{3\Delta}{2}$ and execution algorithms (EAs) can submit undercutting orders at $v_t \pm \frac{\Delta}{2}$. In state $(i,\ j)$, the number of undercutting EA sell orders at $v_t+\frac{\Delta}{2}$ is i, and the number of buy orders at $v_t-\frac{\Delta}{2}$ is j. EB and ES represent the arrival of EAs' buy and sell limit orders, MB and MS represent the arrival of market order traders' (MOTs') buy and sell market orders, and UJ and DJ denote upward and downward value jumps. The arrows between states represent state transitions, and arrows pointing toward the outside represent either order executions or cancelations. The number next to each event is the immediate payoff to HFTs from the event.

equilibrium state $(i,\ j)$. The four states of the LOB are

$$\begin{aligned}
&(0,0) \quad \textit{No limit order from EAs},\\
&(1,0) \quad \textit{An EA sell limit order at } v_t+\left(m-\frac{1}{2}\right)\Delta \textit{ only},\\
&(0,1) \quad \textit{An EA buy limit order at } v_t-\left(m-\frac{1}{2}\right)\Delta \textit{only, and}\\
&(1,1) \quad \textit{EA limit orders on both } v_t+\left(m-\frac{1}{2}\right)\Delta \textit{ and } v_t-\left(m-\frac{1}{2}\right)\Delta.
\end{aligned} \tag{6}$$

Denote the HFTs' expected value by supplying liquidity in state $(i,\ j)$ as $LP^{(i,j)}(m)$. Fig. 2 shows that $LP^{(i,j)}(m)$ depends on the expected value of all other states of the LOB. For example, consider $LP^{(0,0)}(m)$ for an HFT on the ask side of the LOB.[9]

(1) An EA buyer (EB) undercuts the bid side at $v_t-(m-\frac{1}{2})\Delta$ and changes $LP^{(0,0)}(m)$ to $LP^{(0,1)}(m)$.
(2) An EA seller (ES) undercuts the ask side at $v_t+(m-\frac{1}{2})\Delta$ and changes $LP^{(0,0)}(m)$ to $LP^{(1,0)}(m)$.
(3) An MOT buyer (MB) submits a buy market order and the HFT gains $(m+\frac{1}{2})\Delta$.
(4) An MOT seller (MS) submits a sell market order, HFTs race to fill the bid side immediately, and $LP^{(0,0)}(m)$ remains the same.
(5) In an upward value jump (UJ), the limit order on the ask side loses $1-(m+\frac{1}{2})\Delta$.
(6) In a downward value jump (DJ), the liquidity provider cancels the limit order, thereby changing $LP^{(0,0)}(m)$ to zero.

The first equation in Eq. (7) summarizes the value of $LP^{(0,0)}(m)$, which depends on the six types of events and the values for the other three states of the book. Similarly, the remaining three equations describe the value for state $LP^{(1,0)}(m)$, $LP^{(0,1)}(m)$, and $LP^{(1,1)}(m)$.

$$\begin{cases}
LP^{(0,0)}(m) = p_1\overline{LP}^{(0,1)}(m)+p_1\overline{LP}^{(1,0)}(m)+p_2\left(m+\frac{1}{2}\right)\Delta\\
\qquad +p_2LP^{(0,0)}(m)-p_3\left[1-\left(m+\frac{1}{2}\right)\Delta\right]+p_3\cdot 0\\
LP^{(1,0)}(m) = p_1\overline{LP}^{(1,1)}(m)+p_1LP^{(1,0)}(m)+p_2\overline{LP}^{(0,0)}(m)\\
\qquad +p_2LP^{(1,0)}(m)-p_3\left[1-\left(m+\frac{1}{2}\right)\Delta\right]+p_3\cdot 0\\
LP^{(0,1)}(m) = p_1LP^{(0,1)}(m)+p_1\overline{LP}^{(1,1)}(m)+p_2\left(m+\frac{1}{2}\right)\Delta\\
\qquad +p_2\overline{LP}^{(0,0)}(m)-p_3\left[1-\left(m+\frac{1}{2}\right)\Delta\right]+p_3\cdot 0\\
LP^{(1,1)}(m) = p_1\overline{LP}^{(0,1)}(m)+p_1\overline{LP}^{(1,0)}(m)+p_2\overline{LP}^{(0,1)}(m)\\
\qquad +p_2\overline{LP}^{(1,0)}(m)-p_3\left[1-\left(m+\frac{1}{2}\right)\Delta\right]+p_3\cdot 0
\end{cases} \tag{7}$$

where $p_1=\frac{\beta(1-\pi)}{2}$, $p_2=\frac{(1-\beta)(1-\pi)}{2}$, and $p_3=\frac{\pi}{2}$ are the probabilities that the next event is the arrival of an EA buyer (seller), the arrival of an MOT buyer (seller), and the

[9] HFTs make independent decisions on the bid and ask sides. HFTs' best response on the bid side can be characterized similarly. When the sniping risk is very high, such that $(m+\frac{1}{2})\Delta>1$, the HFTs lose zero in case (5) and always maintain a quote at $v_t\pm(m+\frac{1}{2})\Delta$. This happens only in the highest undercutting equilibrium, e.g., the $m=3$ case in Fig. 1.

upward (downward) jump of the fundamental value, respectively. We have $\overline{LP}^{(i,j)}(m) = max\{0, LP^{(i,j)}(m)\}$ because HFTs can simply choose not to submit limit orders or cancel existing limit orders once their expected values become negative. Proposition 5 summarizes the EAs' and HFTs' strategies in the undercutting equilibrium. To conserve space, we defer their off-equilibrium strategies to the proof of Proposition 5.

*Proposition 5. (**undercutting equilibrium**). When* $m = 1$ *and* $\frac{\Delta}{2} < \pi < \frac{\beta+2-\beta\Delta-\sqrt{(\Delta+1)^2\beta^2+(4-12\Delta)\beta+4}}{2\beta}$ *or when* $m \geq 2$ *and* $h_\beta - (m-\frac{1}{2})\Delta \in (0, \frac{\Delta}{2})$,

(i) EAs submit undercutting limit orders at price $v_t - (m-\frac{1}{2})\Delta$ *to buy or* $v_t + (m-\frac{1}{2})\Delta$ *to sell if no existing limit orders sit at that price level and they use stimulating orders otherwise.*

(ii) The HFTs' strategy is as follows:

(a) HFTs provide liquidity at $v_t \pm (m+\frac{1}{2})\Delta$ *if* $LP^{(i,j)}(m) \geq 0$ *at state* (i, j)*, and* $LP^{(i,j)}(m)$ *is determined by the transition matrix in Eq. (7).*

(b) HFTs race to snipe stale quotes from HFTs and EAs during value jumps.

(c) HFTs race to take stimulating limit orders from EAs.

As in Proposition 4, the boundary of the undercutting equilibrium for $m = 1$ involves a formula that differs from the formula involved in the case when $m > 1$. Intuitively, EAs provide liquidity to other EAs when $m = 1$, because quotes of $v_t \pm \frac{\Delta}{2}$ are aggressive enough to attract other EAs. When $m > 1$, the quote from an EA no longer attracts other EAs, because other EAs find that the cost of stimulating HFTs is lower. Therefore, EAs no longer trade with each other when $m > 1$, and the quotes from EAs attract only MOTs.

The undercutting equilibrium provides one explanation for the frequent addition and cancelation of HFTs' quotes (Hasbrouck and Saar, 2013; Biais and Foucault, 2014). In the undercutting equilibrium, HFTs' depth at their best quotes is not constant because they need to respond to EAs' undercutting orders. Therefore, HFTs can update their quotes even if the fundamental value does not change. For example, when $\pi = \frac{1}{3}$, $\beta = 0.6$, and $\Delta = \frac{1}{2}$, we have $m = 1$, $LP^{(0,j)}(1) > 0$, and $LP^{(1,j)}(1) < 0$.[10] HFTs provide liquidity at a half-spread of $\frac{3}{2}\Delta$ when there is no undercutting order, but the depth at a half-spread of $\frac{3}{2}\Delta$ becomes zero once an undercutting order establishes price priority over an HFT's order. If a market order executes against an undercutting order from an EA, HFTs again find that providing liquidity at a half-spread of $\frac{3}{2}\Delta$ is profitable and race to provide liquidity at such a spread.

6. Predictions and policy implications

By exploring the interactions between distinct types of trading algorithms, our paper not only rationalizes a number of puzzles in the literature but also generates new testable predictions. In Section 6.1, we summarize the predictions that are driven mainly by liquidity-providing non-HFTs. In Section 6.2, we summarize the predictions that are driven by discrete pricing. In Section 6.3, we discuss the policy implications of our paper.

6.1. Predictions driven by liquidity-providing non-HFTs

In Prediction 1, we posit that EAs tend to quote more aggressive prices than HFTs quote.

Prediction 1. **(price priority).** *Non-HFTs are more likely than HFTs to establish price priority in liquidity provision.*

Brogaard et al. (2015) and Yao and Ye (2018) find that non-HFTs are more likely than HFTs to establish price priority. Their results are puzzling because existing channels suggest that HFTs should quote more aggressive prices because they incur lower adverse selection costs [see Jones (2013) and Menkveld (2016) for surveys], lower inventory costs (Brogaard et al., 2015; Aït-Sahalia and Sağlam, 2017), and lower operational costs (Carrion, 2013). Our model shows that the opportunity cost of providing liquidity can reconcile this contradiction. EAs can afford to place more aggressive limit orders as long as they cost less to execute than market orders. Therefore, we show in Proposition 2 that EAs always quote more aggressive prices than HFTs do when pricing is continuous. Under discrete pricing, EAs also choose to establish price priority over HFTs as long as the tick size does not impose constraints that discourage EAs from undercutting HFTs (Propositions 4 and 5).

Prediction 2. **(negative correlation between the bid-ask spread and liquidity).** *Technology shocks that increase the fraction of EAs widen the bid-ask spread but reduce overall transaction costs.*

Black (1971, p. 30) describes a liquid market intuitively:

The market for a stock is liquid if the following conditions hold:

1) There are always bid and asked prices for the investor who wants to buy or sell small amounts of stock immediately.
2) The difference between the bid and asked prices (the spread) is always small.
3) An investor who is buying or selling a large amount of stock, in the absence of special information, can expect to do so over a long period of time at a price not very different, on average, from the current market price.

Conditions (1) through (3) were internally consistent when Black (1971) was published. At that time, most traders executed trades by paying the bid-ask spread to dealers or market makers. In the current market, every trader can use limit orders, and Conditions (1) through (3) could be internally inconsistent. As Proposition 2 implies, an increase in β widens the bid-ask spread because HFT market makers receive fewer non-HFT order flows. Meanwhile, the average transaction cost for non-HFTs falls. In the extreme case in which $\beta = 1$, the market becomes infinitely liquid because every trader pays zero transaction costs. At the same time, the bid-ask spread is at its widest.

[10] We solve these equations in the proof of Propositions 4 and 5.

Proposition 2 and Corollary 1 suggest that the definition of liquidity and the measure of liquidity should be updated for modern electronic markets. Prediction 2 derives directly from Corollary 1. One way to test Prediction 2 is to examine whether technology improvements for EAs can increase the bid-ask spread but reduce transaction costs for institutional traders (such as implementation shortfalls measured by ANcerno data).

6.2. Predictions driven by discrete pricing

When pricing is continuous, EAs always provide liquidity to HFTs, and HFTs always provide liquidity to MOTs. When pricing is discrete, who provides liquidity to whom depends on the parameter value, and this dependence generates cross-sectional and time series predictions regarding liquidity provision and demand.

Prediction 3. **(time priority versus price priority).** *HFTs crowd out liquidity provision by non-HFTs when the tick size is large.*

Prediction 3 derives from Proposition 3. Chordia et al., 644) worry that "HFTs use their speed advantage to crowd out liquidity supply when the tick size is small and stepping in front of standing limit orders is inexpensive." Yet Yao and Ye (2018) find that HFTs crowd out non-HFTs' liquidity supply when the tick size is large. Our paper provides the theoretical foundation for reconciling this contradiction. EAs can quote tighter bid-ask spreads than HFTs because EAs incur lower opportunity costs for providing liquidity. A large tick size prevents non-HFTs from establishing price priority over HFTs while helping HFTs establish time priority over non-HFTs. Yao and Ye (2018) find that the tick size is more likely to be binding for low-priced securities, for which a 1 cent uniform tick size leads to a larger relative tick size. They also find that HFTs provide a larger share of liquidity for low-priced securities. Ye et al., 2020 find that an increase in tick size crowds out share repurchases by firms because they cannot win the speed race in liquidity provision.[11] These results are consistent with Prediction 3.

In reality, the tick size is not the only source of constrained price competition. For example, the NYSE and Nasdaq offer rebates to liquidity providers. When the tick size is binding, the rebate to liquidity providers further widens the effective tick size and the cum-fee bid-ask spread (Chao et al., 2018). Technically, every trader can get the rebate for liquidity provision, but with discrete pricing, traders with high-speed capability are more likely to obtain the rebate, particularly when the tick size is binding.

Prediction 4. **(sniping and liquidity provision).** *An increase in sniping risk reduces the share of liquidity provided by HFTs.*

We obtain Prediction 4 by comparing Proposition 3 with Propositions 4 and 5. When the sniping risk is low, the binding bid-ask spread drives speed competition. If the incidence of sniping rises high enough, the spread is wider than one tick, allowing non-HFTs to undercut HFTs and reducing liquidity provision on the part of HFTs. One limitation of our model is that we consider only adverse selection led by sniping, but other types of adverse selection should provide the same economic mechanism. Generally, the breakeven bid-ask spread should be lower when the adverse selection risk is low. Once the breakeven spread falls below one tick, speed competition to achieve time priority should be more critical.

Prediction 4 differs significantly from predictions offered in the existing literature on HFTs. Prior studies typically model HFTs as traders who can access information more rapidly than other traders. Hoffmann (2014), Han et al. (2014), and Bongaerts and Van Achter (2020) find that HFTs incur lower sniping costs than non-HFTs. Therefore, an increase in the level of information should give HFTs a comparative advantage in liquidity provision.

Yao and Ye (2018) provide evidence consistent with Prediction 4. In the cross-section, an increase in adverse selection risk reduces the fraction of liquidity provided by HFTs. It would be interesting to test whether Prediction 4 holds in a time series, that is, whether, for a given security, HFTs provide a smaller fraction of liquidity when the sniping risk is high.

Prediction 5. (**speed competition over taking liquidity**). *Non-HFTs are more likely to provide liquidity at price levels that cross the midpoint (stimulating limit orders) than HFTs. HFTs are also more likely to demand liquidity from stimulating limit orders, but they do not adversely select these orders.*

In Proposition 4, EAs always choose to cross the midpoint to stimulate HFTs. Proposition 5 implies that an EA chooses to stimulate HFTs when another EA undercuts HFTs on the same side. HFTs never cross the midpoint because this strategy loses money, but EAs can cross the midpoint to stimulate HFTs as long as the loss is less than the cost of undercutting HFTs or using market orders.

Latza et al. (2014) find evidence consistent with Prediction 5. They classify a market order as "fast" if it executes against a standing limit order that is less than 50 milliseconds old. These fast market orders should come from HFTs. They also find that fast market orders often execute against limit orders that cross the midpoint and lead to virtually no permanent price impacts. Testing Prediction 5 more directly using data that include account information on traders could be fruitful.

6.3. Policy implications

Our paper offers policy implications for both HFTs and the tick size. For HFTs, BCS argue for a more discrete market in time, whereas we argue for a more continuous market in pricing. We show that when all non-HFTs are EAs, transaction costs are zero, and no incentive exists for HFTs to engage in speed competition. In this sense, our paper supports the Kyle and Lee (2017) vision of a fully continuous market.

[11] To prevent firms from inflating their share prices by outbidding other traders, SEC rule 10b-18 discourages firms from demanding liqudity in share repurchases.

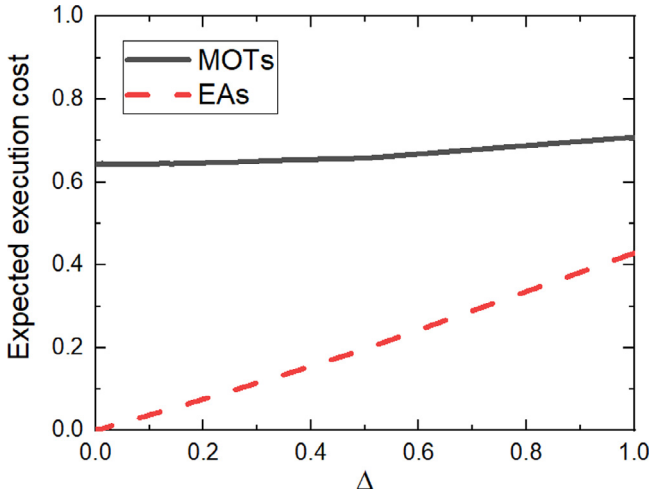

Fig. 3. Simulation of the impacts of tick sizes on the expected execution costs with $\beta \sim U(0,1)$, $\pi \sim U(0,1)$. For each Δ, we draw 100,000 random combinations of (β, π) and decide the equilibrium outcome (queuing, stimulating, or undercutting). We then calculate the expected execution cost of market order traders (MOTs) and execution algorithms (EAs) and take the average over all combinations of (β, π) drawn. Both EAs and MOTs see an increase in execution costs with respect to the tick size Δ. When $\Delta \to 0$, the execution cost of EAs goes to zero because they can use stimulating orders at the fundamental value.

On April 5, 2012, President Barack Obama signed the Jumpstart Our Business Startups Act. Section 106 (b) requires the SEC to examine the effects of tick sizes on initial public offerings. On October 3, 2016, the SEC implemented a pilot program to increase the tick size from 1 cent to 5 cents for twelve hundred common stocks. Proponents of the proposal argue that a larger tick size can improve liquidity (Weild et al., 2012). In Corollary 2 and Prediction 6, we posit that a larger tick size discourages non-HFTs from quoting their desired prices and increases execution costs.

Corollary 2. For all π, β, and Δ, $\bar{C}(\beta) < \bar{C}_\Delta(\beta)$, where $\bar{C}(\beta) \equiv (1-\beta)h_\beta$ is the average execution cost to non-HFTs under continuous pricing and $\bar{C}_\Delta(\beta)$ is the cost under discrete pricing.

Corollary 2 shows that, for any parameter, the transaction cost is higher under a discrete tick size than under continuous pricing because the tick size creates rents for demanding and supplying liquidity. To be sure, the tick size is never continuous in reality. When we compare large and small discrete tick sizes, we are not able to directly compare the formulas for any parameter values because of the complexity of the three equilibrium types. Instead, we draw our two parameter values β and π from a uniform distribution $[0, 1]$ and compute the expected transaction costs based on Propositions 3–5. The results reported in Fig. 3 show that the expected transaction cost increases with the tick size for both MOTs and EAs. The increase in the cost paid by MOTs reflects an increase in the bid-ask spread. An increase in the tick size leads to a larger increase in transaction costs incurred by EAs because a larger tick size not only increases the cost incurred by EAs when they demand liquidity but also increases those costs when they choose to stimulate HFTs.

Following Corollary 2 and Fig. 3, we derive our sixth prediction.

Prediction 6. . Discrete pricing leads to higher transaction costs for non-HFTs.

Empirically, Yao and Ye (2018) and Albuquerque et al. (2020) find evidence consistent with Prediction 6. Our model's prediction, along with their empirical evidence, shows that an increase in the tick size harms liquidity.

7. Conclusion

We provide the first model representing the behavior of algorithmic traders that are slower than HFTs. The interaction between these EAs and HFTs rationalizes several puzzles regarding who provides liquidity and when, as well as generates several new testable predictions. EAs incur lower opportunity costs than HFTs when providing liquidity. Therefore, EAs choose to provide liquidity at more aggressive prices if pricing is sufficiently continuous. A large tick size constrains price competition, creates rents for liquidity provision, and encourages speed competition to capture such rents through the time priority rule. A higher sniping risk increases the breakeven bid-ask spread relative to the tick size, which allows EAs to establish price priority over HFTs and reduces the share of liquidity provided by HFTs. All these predictions are consistent with the empirical findings of Yao and Ye (2018).

Our model also provides several new testable predictions. (1) EAs should not use market orders once the tick size becomes small enough relative to the bid-ask spread. (2) EAs are more likely than HFTs to provide liquidity at price levels that cross the midpoint, and these limit orders are more likely to be taken by HFTs almost immediately. (3) The bid-ask spread widens when technological shocks increase the proportion of EAs, but overall transaction costs decrease.

We find that a larger tick size increases transaction costs and drives an arms race in speed. These results challenge the rationale for the recent policy proposal that has increased the tick size to 5 cents. Thus, we encourage regulators to consider decreasing the tick size, particularly for liquid stocks.

Current policy debates over HFTs usually follow the binary classifications that pit fast versus slow traders or computers versus humans. This dichotomy reflects and affects the academic literature on HFTs. Our model shows that this policy debate should consider diversity within the class of machine traders, especially regarding machines that are slower than HFTs but faster than humans. For example, we find that EAs can cross the midpoint to stimulate HFTs to demand liquidity immediately, and the cost of stimulating HFTs is lower than the cost of paying the bid-ask spread offered by HFTs. Therefore, the impact of HFTs on liquidity and social welfare should not be evaluated based simply on whether they demand or provide liquidity. We also find that the bid-ask spread can move in the opposite direction of true liquidity.

EAs in our model make only execution decisions, and their incentives to buy or sell are exogenous. Some other algorithmic traders could use computers and machine-learning techniques to decide whether to buy or to sell. Therefore, fully assessing the diversity of algorithmic traders and their interactions is still in the early days. Just as insights into human behavior from the psychology literature spawned the field of behavioral finance, so insights into algorithmic behavior could prompt an analogous blossoming of research in algorithmic finance.

References

Aït-Sahalia, Y., Sağlam, M., 2017. High-frequency market making: optimal quoting. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2331613.

Albuquerque, R., Song, S., Yao, C., 2020. The price effects of liquidity shocks: a study of the SEC's tick size experiment. J Financ Econ 138 (3), 700–724.

Aquilina, M., Budish, E.B., O'Neill, P, 2020. Quantifying the high-frequency trading "arms race": a simple new methodology and estimates. Chicago Booth Research Paper 20-16.

Biais, B., Foucault, T., 2014. HFT and market quality. Bankers, Markets Investors 128 (1), 5–19.

Black, F., 1971. Toward a fully automated stock exchange, part I. Financial Anal J 27 (4), 28–35.

Bongaerts, D., Van Achter, M., 2020. Competition among liquidity providers with access to high-frequency trading technology. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2698702.

Brogaard, J., Hagströmer, B., Nordén, L., Riordan, R., 2015. Trading fast and slow: colocation and liquidity. Rev Financ Stud 28 (12), 3407–3443.

Budish, E., Cramton, P., Shim, J., 2015. The high-frequency trading arms race: frequent batch auctions as a market design response. Q J Econ 130 (4), 1547–1621.

Carrion, A., 2013. Very fast money: high-frequency trading on the Nasdaq. J Financ Markets 16 (4), 680–711.

Chao, Y., Yao, C., Ye, M., 2018. Why discrete price fragments US stock exchanges and disperses their fee structures. Rev Financ Stud 32 (3), 1068–1101.

Chordia, T., Goyal, A., Lehmann, B.N., Saar, G., 2013. High-frequency trading. J Financ Markets 16 (4), 637–645.

Clark-Joseph, A.D., Ye, M., Zi, C, 2017. Designated market makers still matter: evidence from two natural experiments. J Financ Econ 126 (3), 652–667.

Foucault, T., 1999. Order flow composition and trading costs in a dynamic limit order market. J Financ Markets 2 (2), 99–134.

Foucault, T., Kadan, O., Kandel, E., 2005. Limit order book as a market for liquidity. Rev Financ Stud 18 (4), 1171–1217.

Foucault, T., Kozhan, R., Tham, W.W., 2017. Toxic arbitrage. Rev Financ Stud 30 (4), 1053–1094.

Frazzini, A., Israel, R., Moskowitz, T., 2018. Trading costs. Social Science Research Network https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3229719.

Glosten, L.R., Milgrom, P.R., 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. J Financ Econ 14 (1), 71–100.

Goettler, R.L., Parlour, C.A., Rajan, U., 2005. Equilibrium in a dynamic limit order market. J Finance 60 (5), 2149–2192.

Han, J., Khapko, M., Kyle, A.S., 2014. Liquidity with high-frequency market making. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2416396.

Hasbrouck, J., Saar, G., 2013. Low-latency trading. J Financ Markets 16 (4), 646–679.

Hoffmann, P., 2014. A dynamic limit order market with fast and slow traders. J Financ Econ 113 (1), 156–169.

Jones, C.M., 2013. What do we know about high-frequency trading? Columbia Business School Research Paper, 13–11.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica: J Econometric Soc 53 (6), 1315–1335.

Kyle, A.S., Lee, J., 2017. Toward a fully continuous exchange. Oxford Rev Econ Policy 33 (4), 650–675.

Latza, T., Marsh, I.W., Payne, R., 2014. Fast aggressive trading. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2542184.

Li, S., Ye, M., 2021. The Trade-Off Between Discrete Pricing and Discrete quantities: Evidence from US-listed firms. Working paper. University of Illinois at Urbana-Champaign. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3763516.

Li, S., Ye, M., Zheng, M., 2021. Financial Regulation, Clientele Segmentation, and Stock Exchange Order Types. Working paper. University of Illinois at Urbana-Champaign. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3763455.

Menkveld, A.J., 2016. The economics of high-frequency trading: taking stock. Annu Rev Financ Econ 8, 1–24.

Menkveld, A.J., Zoican, M.A., 2017. Need for speed? Exchange latency and liquidity. Rev Financ Stud 30 (4), 1188–1228.

O'Hara, M., 2015. High-frequency market microstructure. J Financ Econ 116 (2), 257–270.

O'Hara, M., Saar, G., Zhong, Z., 2018. Relative tick size and the trading environment. Review Asset Pricing Stud 9 (1), 47–90.

Parlour, C.A., 1998. Price dynamics in limit order markets. Rev Financ Stud 11 (4), 789–816.

Weild, D., Kim, E., Newport, L., 2012. The trouble with small tick sizes: larger tick sizes will bring back capital formation, jobs, and investor confidence. Capital Markets Series. Grant Thornton, Chicago, IL.

Yao, C., Ye, M., 2018. Why trading speed matters: a tale of queue rationing under price controls. Rev Financ Stud 31 (6), 2157–2183.

Ye, M., Zheng, M., Li, X., 2020. Price Ceiling, Market Structure, and Payout Policies. Working paper. University of Illinois at Urbana-Champaign and Guangxi University of Finance and Economics. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727130.