NBER WORKING PAPER SERIES

WHO PROVIDES LIQUIDITY, AND WHEN?

Sida Li
Xin Wang
Mao Ye

Working Paper 25972
http://www.nber.org/papers/w25972

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2019

We thank Hengjie Ai, Shmuel Baruch, Malcolm Baker, Dan Bernhardt, Hank Bessembinder, Eric Budish, Tarun Chordia, Thierry Foucault, Katya Malinova, Maureen O'Hara, Monika Piazzesi, Veronika Pool, Neil Pearson, Barbara Rindi, Shri Santosh, Andriy Shkilko, Brian Weller, Chen Yao, Bart Yueshen, and Marius Zoican for helpful comments. We are also grateful for input from participants in conferences at the University of Rochester, UCLA, Texas A&M University, the University of Florida, and Washington University at St. Louis, as well as at the Carlson Junior Conference at the University of Minnesota, the NYU Stern Market Microstructure Conference, the 2nd 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, and the Smokey Mountain Conference at the University of Tennessee for their helpful suggestions. This research is supported by National Science Foundation grant 1352936 (jointly with the Office of Financial Research at U.S. Department of the Treasury) and National Science Foundation grant 1838183. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Who Provides Liquidity, and When?
Sida Li, Xin Wang, and Mao Ye
NBER Working Paper No. 25972
June 2019
JEL No. G1,G12

ABSTRACT

We model competition for liquidity provision between high-frequency traders (HFTs) and slower execution algorithms designed to minimize transaction costs for buy-side institutions (B-Algos). Under continuous pricing, B-Algos 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 is not binding, B-Algos choose between stimulating HFTs and providing liquidity to other non-HFTs. Flash crashes arise under certain parameter values. Transaction costs can be negatively correlated with the bid–ask spread when all traders can provide liquidity.

Sida Li
Department of Finance,
Gies School of Business
University of Illinois, Urbana-Champaign
1206 South Sixth Street
Champaign, IL 61820
sidali3@illinois.edu

Mao Ye
University of Illinois, Urbana-Champaign
343K Wohlers Hall
1206 South Sixth Street
Champaign, IL 61820
and NBER
maoye@illinois.edu

Xin Wang
Nanyang Technological University
S3-B1B-60
50 Nanyang Ave
Singapore 639798
xin.wang@ntu.edu.sg

1. Introduction

To minimize their transaction costs, buy-side institutions, such as mutual funds and pension funds, use computer algorithms extensively to execute their trades (Frazzini, Israel, and Moskowitz 2014; O'Hara 2015). These buy-side algorithmic traders (B-Algos) differ from high-frequency traders (HFTs) in two fundamental ways (Hasbrouck and Saar 2013; Jones 2013; O'Hara 2015). First, B-Algos may use limit orders to provide liquidity, but their goal is to minimize transaction costs rather than to profit from the bid–ask spread. Second, B-Algos are faster than humans but slower than HFTs (O'Hara 2015). Although buy-side institutions are major players in financial markets, their trading algorithms have no independent identity in existing models. From one point of view, financial markets include HFTs and everyone else, where the latter includes both sophisticated institutions and unsophisticated retail traders (see the survey by O'Hara [2015]). From the other point of view, algorithmic traders and HFTs are interchangeable (see the survey by Biais and Foucault [2014]). In this paper, we offer the first theoretical study of B-Algos by examining how they interact with HFTs and humans.

In an environment populated by HFTs, B-Algos, and humans, who provides liquidity? Who demands liquidity, and when? Answering these questions is important because traditional liquidity providers, such as New York Stock Exchange (NYSE) specialists and NASDAQ dealers, almost disappear in modern electronic markets (Clark-Joseph, Ye, and Zi 2017). Everyone *can* provide liquidity, but no one is *obligated to* provide liquidity. We examine how this new environment of voluntary liquidity provision and demand reaches equilibrium.

In our model, HFTs and two types of non-HFTs (B-Algos and humans) trade a security in a dynamic limit-order book (LOB). All traders are risk-neutral. A liquidity provider in the LOB submits limit orders (offers to buy or sell a stock at a specified price and quantity), and a liquidity

demander accepts a limit order using a market order. HFTs have no private value to trade, as they simply provide or demand liquidity to maximize their expected profits from trading. Non-HFTs arrive at the market following a Poisson process, bringing inelastic demand to buy or sell one unit of a security. Some of the non-HFTs are B-Algos, and they can choose between limit and market orders to minimize transaction costs; the remaining traders are humans, who use only market orders.

Our model includes one security whose fundamental value is public information. However, liquidity providers are subject to sniping risks (Budish, Cramton, and Shim 2015; BCS hereafter), because they may fail to cancel stale quotes during value jumps. B-Algos are always sniped during value jumps, and HFTs can reduce the probability of being sniped by $\frac{1}{N}$, where N is the number of equally fast HFTs. HFTs in our model, like those in BCS, quote a positive bid–ask spread because of the sniping risk. Surprisingly, we find that B-Algos always quote better prices than HFTs as long as the price in a given trade is continuous enough, even though B-Algos face greater exposure to sniping risks.

Opportunity cost explains why B-Algos can afford more aggressive limit orders than HFTs. In our model, B-Algos have an inelastic need to trade. If they do not use limit orders to provide liquidity, they must use market orders and pay the bid–ask spread. Therefore, B-Algos incur a negative opportunity cost for providing liquidity, and they choose limit orders as long as their expected costs are lower in value than their market orders. HFTs, on the other hand, do not have to trade, and their speed advantage leads to a positive opportunity cost for providing liquidity. An HFT who posts a limit order surrenders the profit from sniping the share once it becomes stale. Interestingly, although higher speed reduces HFTs' sniping costs, it increases their opportunity costs by the same amount. Adding the sniping and opportunity costs together, B-Algos incur lower

overall costs for liquidity provision, and they always choose to provide liquidity when the price grid is continuous enough. Our prediction is consistent with that of Brogaard et al. (2015), who find that non-HFTs quote tighter bid–ask spreads than HFTs (Table A1), but non-HFTs usually quote only on one side of the market.

In the equilibrium under continuous pricing, B-Algos place limit orders at fundamental values. These limit orders execute immediately at zero transaction cost, because they immediately stimulate market orders from HFTs. Therefore, B-Algos can achieve minimum possible transaction costs even though the bid–ask spread is positive. This surprising result is driven by the following mechanism. Because HFTs can provide and demand liquidity at any time, they have one price they are willing to offer and another they are willing to accept. For example, an HFT offers an ask price to sell above the fundamental value, because the ask price is subject to sniping risk. The same HFT accepts an offer to buy at the fundamental value, because demanding liquidity entails no sniping risk. A B-Algo buyer never pays HFTs' ask price, because she can induce HFTs to respond to her offer immediately with a price at or above the fundamental value.

In BCS, non-HFTs can use only market orders, and the sniping risk leads to a positive bid–ask spread, motivating BCS to recommend frequent batch auctions as an alternative market design. Our model with continuous pricing shows that when all non-HFTs can choose between limit and market orders, transaction costs drop to zero despite the sniping risk. Also, when all traders can provide liquidity, the bid–ask spread can move in the direction that runs opposite to the true level of liquidity. When we increase the fraction of B-Algos, the bid–ask spread widens, because fewer non-HFTs bear the sniping risk. The transaction cost decreases, however, because more non-HFTs enjoy zero transaction costs. The market becomes infinitely liquid when all non-HFTs are B-Algos, but the bid–ask spread reaches its widest magnitude.

Under continuous pricing, our model results in only one type of equilibrium for any parameter value, in which B-Algos provide liquidity to HFTs at the fundamental value while HFTs provide liquidity to humans. Next, we add discrete pricing to our model to reflect the tick size (minimum price variation) of one cent imposed by the U.S. Securities and Exchange Commission's (SEC's) Regulation National Market Systems (Reg NMS) Rule 612, and to evaluate the recent policy initiative that has increased the tick size from one cent to five cents. We find that discrete pricing generates rents for both providing and demanding liquidity. Such rents lead in turn to four types of equilibria, depending on the sniping risk and the fraction of B-Algos, thereby generating cross-sectional and time-series predictions regarding who provides and who demands liquidity.

First, discrete pricing generates rents for liquidity provision because it prevents the bid–ask spread from reaching its break-even level. Such rents are most apparent when sniping risk is low relative to the tick size. In that case, the break-even bid–ask spread drops below one tick and the difference between a one-tick mandated bid–ask spread and the break-even bid–ask spread drives a speed race to capture the rents. In our first type of equilibrium, queuing equilibrium, B-Algos cannot undercut the price for HFTs because of the binding tick size, and HFTs dominate liquidity provision through time priority. Yao and Ye (2018) find empirically that HFTs dominate liquidity provision when either the adverse selection risk is too low or the tick size is too large, which is consistent with the queuing equilibrium.

Second, discrete pricing also creates rents for liquidity demand. As sniping risk increases, the break-even spread for HFTs becomes wider than one tick, allowing B-Algos to submit limit orders at more aggressive prices. In that case, however, B-Algos can no longer submit limit orders at precisely the fundamental value unless it coincides with a price tick. To stimulate market orders from HFTs, B-Algos have to cross the fundamental value, and the difference between the price of

a stimulated order and the fundamental value then generates a speed race between HFTs to demand liquidity. In this type of equilibrium, stimulating equilibrium, B-Algos provide liquidity to HFTs; HFTs provide liquidity to humans because limit orders from B-Algos do not stay in the LOB. Also, because discrete pricing destroys the possibility of stimulating HFTs at a minimum possible cost of zero, under certain parameter ranges B-Algos may find it is less costly to submit limit orders that do not cross the midpoint.[4] In our third type of equilibrium, undercutting equilibrium, B-Algos choose to provide liquidity to other non-HFTs instead of stimulating HFTs.

In the final type of equilibrium, crash equilibrium, HFTs submit orders outside the maximum value of the jump when the sniping risk or the fraction of non-HFTs who are B-Algos is too high. In the crash equilibrium, HFTs effectively quit liquidity provision through limit orders, because most liquidity demand for their quotes comes from sniping. Despite a dramatic increase in HFTs' bid–ask spreads, the transaction costs for B-Algos do not increase relative to the stimulating equilibrium, because a B-Algo can still use stimulating orders to attract market orders from HFTs. A crash equilibrium, however, imposes a threat to traders who use only market orders. The concern is most severe following value jumps. For example, an upward jump may remove all quotes from B-Algos on the ask side. If a human submits a market order after the jump, her order would hit the price quoted by HFTs and lead to an extreme transaction price. A crash equilibrium provides a possible interpretation of flash crashes, which are sharp price movements in one direction followed by quick reversion (Biais and Foucault, 2014). There are certainly other drivers of flash crashes (Kirilenko et al. 2017; Kyle and Obizhaeva 2016), but our mechanism offers the following unique predictions: 1) Flash crashes are as likely to make prices go up as they are to

[4] The phrase "certain parameter ranges" refers to a level of sniping risk that just forces HFTs to quit liquidity provision at a given price. Such price level, however, may still attract limit orders from B-Algos as long as they lose less money than stimulating limit orders. We analytically solve for the range in Proposition 4.

make prices go down, 2) sophisticated traders can avoid extreme execution price by using limit orders during flash crashes, 3) flash crashes are more likely to occur when an initial value jump clears the limit order book, and 4) flash crashes are less likely to occur when the share of trading by B-Algos is either too low or too high. When there are too few B-Algos, most non-HFTs demand liquidity from HFTs, so HFTs do not need to quote very wide spreads. When there are too many B-Algos, HFTs need to quote very wide spreads but B-Algos never hit such spreads.

In the existing literature, information drives arms races in speed.[5] Speed competition driven by discrete pricing works in the opposite direction. In the absence of information, the break-even spread is zero, which generates maximum rents for racing to the top of the liquidity provision queue. This new channel of speed competition reconciles the contradiction between existing channels of speed competition and the empirical facts. 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 a higher adverse-selection risk (Han, Khapko and Kyle 2014). In turn, HFTs should dominate liquidity provision when the tick size is small because constraints that prevent them from offering better prices are less binding (Chordia et al. 2013). Yao and Ye (2018) find, however, that an increase in the adverse-selection risk reduces HFTs' share in liquidity provision. Yao and Ye (2018) and O'Hara, Saar, and Zhong (2018) find that a reduction in the tick size reduces HFTs' share in liquidity provision. Our model helps us to reconcile these contradictions. First, lower adverse-selection risk reduces the break-even spread below one tick and drives speed competition at constrained prices. Second, a large tick size drives speed

[5] On the one hand, speed can reduce adverse-selection costs for liquidity providers and improve liquidity; on the other hand, speed can allow HFTs to adversely select other traders, which has a detrimental effect on liquidity (see Jones [2013], Biais and Foucault [2014], and Menkveld [2016] for surveys). Our model also incorporates these two types of speed competition, but the main driver of speed competition in our model is discrete pricing.

competition because it raises the spread above the break-even level.

The closest paper to ours is BCS. We relax two assumptions made in BCS. First, we allow non-HFTs to use limit orders. By taking the initial step of modeling sophisticated non-HFTs, we not only develop new predictions but also generate new perceptions. Liquidity demand from HFTs once had a negative connotation because, in existing models, HFTs typically adversely select liquidity providers when they demand liquidity (BCS; Foucault, Kozhan, and Tham, 2017; Menkveld and Zoican, 2017). In our model, B-Algos can use aggressive limit orders to prompt HFTs to demand liquidity, which involves no adverse-selection costs. Instead, B-Algos reduce their transaction costs by stimulating HFTs to demand liquidity. This may help to explain why Latza, Marsh, and Payne (2014) find that limit orders executed within 50 milliseconds after submission incur no adverse-selection costs.

Second, BCS consider continuous pricing, arguing for a more discrete market with respect to time, with frequent batch auctions. We consider discrete pricing and argue for a more continuous market with a lower tick size. We question the rationale for increasing the tick size to five cents, as proposed by the 2012 U.S. Jumpstart Our Business Startups (JOBS) Act. 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, Kim, and Newport 2012). Our results show that an increase in the tick size reduces liquidity, encourages speed racing between HFTs, and allocates resources to latency reduction.

2. The Benchmark Model

In this section, we set up and solve a model similar to that utilized in BCS. This section serves as

a benchmark against which to evaluate the impact of allowing non-HFTs to provide liquidity.

2.1 Setup of the Benchmark Model

We consider a continuous-time model with an infinite horizon. All the random variables in our model are mutually independent. Our model has one security, whose fundamental value, v_t, evolves as a compound Poisson jump process at arrival rate λ_J, where t runs continuously on $[0, \infty)$. $v_0 = 0$ and jumps by $J = d$ or $-d$ with equal probability, where $d > 0$. v_t is common knowledge.

Limit-Order Book with Continuous Pricing. The stock exchange operates as a continuous LOB. As in BCS, pricing is continuous in the benchmark model, and we consider the impact of discrete pricing in Section 4. 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. Following value jumps, liquidity providers are subject to an adverse-selection risk if they fail to update stale quotes before liquidity demanders snipe them. Execution precedence for liquidity providers follows the 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. The LOB contains all outstanding limit orders. Outstanding orders to buy are called bids, and outstanding orders to sell are called asks. The highest bid and lowest ask are called the best bid and ask (offer) (BBO), and the difference between them is the bid–ask spread.

Two Types of Traders. The benchmark model includes two types of traders: HFTs and non-HFTs. All traders are risk-neutral and there is no time-discounting. Non-HFTs arrive at the market at

Poisson intensity λ_I. Each Non-HFT has an inelastic need to buy or sell one unit of the security at equal probability. As in BCS, we assume non-HFTs can use only market orders, and we relax this assumption in sections 3 and 4. N ($2 \leq N \leq \infty$) HFTs receive no private value from trading. They always present at the market with the goal of maximizing profits from trading, and they can place or take limit orders at any time t. We assume that a HFT's ability to place additional orders is not affected by her exiting orders. For example, all HFTs can snipe stale quotes following value jumps and sniping one's own share is economically equivalent to order cancellation. HFTs are all equally fast, and when multiple HFT order messages (limit orders, market orders, or cancellations) reach the exchange at the same time, they are processed serially in a random order.

Comments. Our benchmark model is essentially the same as that utilized in BCS, except that we simplify the jump size to d or $-d$ so that we can solve our model analytically when the price is discrete. This simplification leads to a trivial strategy of quoting a bid–ask spread around any possible future fundamental value $v_t \pm Kd$ where $K \in \mathbb{N}^+$. We rule out such a possibility by focusing on the BBO around the current fundamental value v_t. Specifically, throughout the paper we assume that a limit order is cancelled if it has no chance to trade with a non-HFT before next jump occurs.[6] As in BCS, to focus exclusively on the sniping risk we also assume that there is no inventory cost for HFTs to hold an asset and there is no asymmetric information on the asset's fundamental value.

2.2 Solution to the Benchmark Model

[6] A trader can certainly submit a limit order far away from the current price and wait for the price to approach it. We argue that the trader's transaction cost depends only on the market condition *when* the price approaches the limit order. For example, consider a \$50 buy order for a security trading at \$100. Conditional on the order's execution, the expected fundamental value would be around \$50, and the transaction cost would be a matter of a few ticks. When discrete pricing kicks in and execution priority matters, "allowing all traders to submit orders far away from the market" is equivalent to "letting all traders compete on execution priority once the market jumps to the price," but the latter allows us to track fewer orders. In other words, slow traders lose execution priority with or without this assumption.

As in BCS, the benchmark model needs to characterize only the HFTs' strategy. Let s be the bid-ask spread. We consider, without loss of generality, the expected payoff for an HFT's limit sell order at $v_t + \frac{s}{2}$, $LP\left(\frac{s}{2}\right)$.

$$LP\left(\frac{s}{2}\right) = \frac{\lambda_I/2}{\lambda_I+\lambda_J} \cdot \frac{s}{2} + \frac{\lambda_I/2}{\lambda_I+\lambda_J} \cdot LP\left(\frac{s}{2}\right) + \frac{N-1}{N}\frac{\lambda_J/2}{\lambda_I+\lambda_J} \cdot \left(\frac{s}{2} - d\right) + \frac{\lambda_J/2}{\lambda_I+\lambda_J} \cdot 0 \quad (1)$$

When non-HFTs demand liquidity only, a sell limit order from an HFT faces four types of events. At probability $\frac{\lambda_I/2}{\lambda_I+\lambda_J}$, the next event is a buy order from a non-HFT, which leads to a profit of $\frac{s}{2}$ to the liquidity provider. At probability $\frac{\lambda_I/2}{\lambda_I+\lambda_J}$, a non-HFT sell order arrives, which does not affect $LP\left(\frac{s}{2}\right)$ on the ask side, because HFTs immediately restore the previous state of the LOB by refilling the bid side. At probability $\frac{\lambda_J/2}{\lambda_I+\lambda_J}$, v_t jumps upward by d, and all HFTs race to snipe stale quotes on the ask side. The conditional probability of being sniped by other HFTs is $\frac{N-1}{N}$. The payoff of being sniped by other traders is $(v_t + \frac{s}{2}) - (v_t + d) = \frac{s}{2} - d$. When v_t jumps downward, the liquidity provider cancels the order, and the payoff is zero.

The solution for equation (1) is:

$$LP\left(\frac{s}{2}\right) = \frac{\lambda_I}{\lambda_I+2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{N-1}{N}\left(d - \frac{s}{2}\right) \quad (2)$$

Equation (2) reveals an additional intuition regarding the expected payoff for providing liquidity. With probability $\frac{\lambda_I}{\lambda_I+2\lambda_J}$, a non-HFT takes the limit order, and the payoff is $\frac{s}{2}$; with probability $\frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{N-1}{N}$, the limit order is sniped by other HFTs, and the payoff is $\left(\frac{s}{2} - d\right)$; with the remaining probability of $\frac{\lambda_J}{\lambda_I+2\lambda_J}(1 + \frac{1}{N})$, the limit order is cancelled, and the payoff is zero.

The outside option to provide liquidity at $v_t + \frac{s}{2}$ for one share is to potentially snipe the share when v_t jumps. The value for this outside option, $SN\left(\frac{s}{2}\right)$, is zero when a non-HFT takes the share, and it remains $SN\left(\frac{s}{2}\right)$ when a non-HFT seller takes liquidity on the opposite side. Each sniper has a $\frac{1}{N}$ chance of sniping the share when v_t jumps upward, and the payoff for the successful sniper is $d - \frac{s}{2}$. When v_t jumps downward, the value to the sniper becomes zero because we assume that the liquidity provider cancels the order. Therefore,

$$SN\left(\frac{s}{2}\right) = \frac{\lambda_I/2}{\lambda_I+\lambda_J} \cdot 0 + \frac{\lambda_I/2}{\lambda_I+\lambda_J} \cdot SN\left(\frac{s}{2}\right) + \frac{1}{N}\frac{\lambda_J/2}{\lambda_I+\lambda_J} \cdot \left(d - \frac{s}{2}\right) + \frac{\lambda_J/2}{\lambda_I+\lambda_J} \cdot 0 \tag{3}$$

The solution for equation (3) is:

$$SN\left(\frac{s}{2}\right) = \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{1}{N}\left(d - \frac{s}{2}\right) \tag{4}$$

In equilibrium, HFTs should be indifferent between liquidity provision and stale-quote sniping. Thus, equating (2) and (4) solves the equilibrium bid–ask spread s_1^*. As in BCS, the best bid and ask prices contain only one share, because undercutting s_1^* or quoting a second share at s_1^* loses money. We summarize the equilibrium as follows:

Proposition 1 (BCS 2015). *With a zero tick size and with non-HFTs demanding liquidity only), the equilibrium bid–ask spread is* $s_1^* = \frac{2\lambda_J}{\lambda_I+\lambda_J} d$.[7]

(i) At almost all times t,[8] *HFTs always maintain one unit in the LOB at the ask price* $v_t +$

[7] If we reduce the lot size to $\frac{1}{l}$ and allow B-Algos to slice their orders to l consecutive child orders, HFTs can quote tighter bid–ask spreads with reduced lot sizes. $s \to 0$ when $l \to \infty$.

[8] HFTs' stale quotes may be sniped during value jumps, but the status of the LOB restores immediately around the new fundamental value.

$\frac{s_1^*}{2}$ *and one unit at the bid price* $v_t - \frac{s_1^*}{2}$*. The bid and ask prices may belong to different HFTs.*

(ii) Upon arrival, non-HFTs take liquidity from HFTs and pay $\frac{s_1^*}{2}$ *as transaction costs.*

(iii) When v_t *jumps up (down), all HFTs race to take stale limit orders at the ask (bid) price.*

The key result in Proposition 1 is that $s_1^* > 0$ and thus all non-HFTs pay a positive transaction cost even without fundamental uncertainty or inventory costs to HFTs for providing liquidity. This is essentially why BCS suggest frequent batch auctions to curb sniping. In the next section, we show non-HFTs are able to use limit orders to completely avoid the sniping cost.

3. Continuous Pricing Model with B-Algos

In this section, we relax only one assumption that is made in BCS. We allow a fraction of $\beta > 0$ non-HFTs to provide liquidity, but this seemingly small variation significantly changes the results reported in BCS.

3.1 First Extension: Three Types of Traders

From this section forward, we assume a fraction β of non-HFTs can choose between limit orders and market orders. We call them B-Algos. We call the remaining fraction of $1 - \beta$ non-HFTs as humans. Our model allows β to equal one, in which case our model includes only HFTs and B-Algos. When $\beta = 0$, our model degenerates into the benchmark model. We consider B-Algos to be execution desks of mutual funds or hedge funds, or brokers who represent the funds' order flows. According to O'Hara (2015), B-Algos are slower than HFTs in reality, and B-Algos

are also slower than HFTs in our model. B-Algos' object function is to minimize the expected transaction costs of fulfilling their trading needs. As our paper focuses on costs led by the sniping risk, we assume away the delay cost (Parlour 1998; Foucault 1999; Foucault, Kadan, and Kandel 2005) for B-Algos. We allow B-Algos to update their orders at any time, although we require a B-Algo to maintain her order within $v_t \pm d$ before her trading need is satisfied.

When non-HFTs can provide liquidity, the four types of possible upcoming events in BCS expand to six types of events with corresponding possibilities:

$$\begin{cases} \frac{\beta\lambda_I/2}{\lambda_I+\lambda_J} & \text{New B-Algo sells (BS)} \\ \frac{\beta\lambda_I/2}{\lambda_I+\lambda_J} & \text{New B-Algo buys (BB)} \\ \frac{(1-\beta)\lambda_I/2}{\lambda_I+\lambda_J} & \text{New human sells (HS)} \\ \frac{(1-\beta)\lambda_I/2}{\lambda_I+\lambda_J} & \text{New human buys (HB)} \\ \frac{\lambda_J/2}{\lambda_I+\lambda_J} & \text{Value jumps up (UJ)} \\ \frac{\lambda_J/2}{\lambda_I+\lambda_J} & \text{Value jumps down (DJ)} \end{cases} \tag{5}$$

As in BCS, only the ratio between the arrival intensity of the security's value jumping over non-HFTs matters, and an order's liquidity-provision revenue does not depend on the market velocity (Kyle and Obizhaeva, 2016). Therefore, we define $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ and the six probabilities become $\frac{\beta/2}{\kappa+1}$, $\frac{\beta/2}{\kappa+1}$, $\frac{(1-\beta)/2}{\kappa+1}$, $\frac{(1-\beta)/2}{\kappa+1}$, $\frac{\kappa/2}{\kappa+1}$, and $\frac{\kappa/2}{\kappa+1}$, respectively.

3.2 Solution with Liquidity-providing non-HFTs

BCS assume that non-HFTs demand liquidity only. We find that non-HFTs never demand liquidity if they are not forced to do so. We show this result by contradiction.

Suppose that B-Algos demand liquidity from HFTs and that HFTs quote a sell price of

$v_t + \frac{s}{2}$. Then $\frac{s}{2}$ must be strictly greater than 0; otherwise, HFTs lose money by providing liquidity. A B-Algo who wants to buy then pays $v_t + \frac{s}{2}$ by demanding liquidity from HFTs. A strictly dominant strategy for B-Algos is to submit a buy limit order at price $v_t + \varepsilon$, where $\varepsilon > 0$ and can be arbitrarily small. Because the price of this buy limit order is above the fundamental value v_t, the order immediately stimulates HFTs to demand liquidity. The HFT who successfully takes the liquidity gains ε; the B-Algo loses ε by providing liquidity, but the cost is lower if $\varepsilon < \frac{s}{2}$. Therefore, we prove that B-Algos never demand liquidity from HFTs.

The previous proof uncovers two economic mechanisms that are new to the literature. The first mechanism, the make–take spread, captures the difference in prices between a trader's willingness to post and her willingness to accept an offer. Most market microstructure models do not include make–take spreads because liquidity providers and liquidity demanders cannot switch roles. 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 studies on LOB (Foucault et al. [2005] among others), traders can choose a limit order or a market order upon arrival, but they can no longer update their roles after the initial decision. When traders are free to use limit and market orders at every point in time, our model shows that they have one price level that they are willing to offer and another price level that they are willing to accept. The divergence of these two price levels comes from the sniping risk. For example, a trader accepts a lower price to sell than the price at which she offers to sell because a sell limit order is subject to the sniping risk, whereas a sell market order is not. Because sniping is the only source of adverse selection in our model, the make–take spread for HFTs happens to be half of the bid–ask spread.[9] An HFT

[9] If B-Algos' orders contain private information or if HFTs incur inventory costs, HFTs may not take liquidity from limit orders priced at v_t and the make–take spread would be less than half of the bid–ask spread.

quoting an ask price of $v_t + \frac{s}{2}$ would accept a limit buy price of v_t. Therefore, a B-Algo buyer can use an aggressive limit order at price $v_t + \varepsilon$ ($\varepsilon \to 0$) to save the half-spread. The limit order at $v_t + \varepsilon$ executes like a market order because of immediate execution.

The second mechanism, the opportunity cost of liquidity provision, explains why B-Algos can afford more aggressive limit order prices than HFTs. HFTs incur lower adverse selection costs for providing liquidity. During value jumps, an HFT is sniped at a probability of $\frac{N-1}{N}$, whereas a B-Algo is sniped at a probability of one. An HFT, however, incurs a positive opportunity cost for liquidity provision. An HFT cannot profit from sniping a share once she provides liquidity for the share, and the probability of sniping conditional on a value jump is $\frac{1}{N}$. The positive opportunity cost exactly offsets the reduced sniping cost. B-Algos, on the other hand, enjoy negative opportunity costs for liquidity provision. B-Algos have to execute their trades. The outside option of providing liquidity is to demand liquidity by paying $\frac{s}{2}$. Therefore, B-Algos can afford a buy limit order at price $v_t + \varepsilon$ but HFTs cannot.

Proposition 2 characterizes the equilibrium. In the equilibrium, B-Algos always choose limit-order prices at v_t and HFTs always immediately demand liquidity from B-Algos. The LOB effectively contains only one state: HFTs quote one share at $v_t + \frac{s_2^*}{2}$ and one share at $v_t - \frac{s_2^*}{2}$, and the LOB contains no limit order from B-Algos. In summary, B-Algos provide liquidity to HFTs, and HFTs provide liquidity to humans. s_2^* equalizes the payoff of liquidity provision and stale-quote sniping for HFTs.

Proposition 2 (Stimulating Equilibrium). *With zero tick size and a positive fraction of B-Algos ($\beta > 0$), the equilibrium bid–ask spread* $s_2^* = \frac{2\lambda_J}{(1-\beta)\lambda_I + \lambda_J} d$.

(i) At almost all times t, HFTs always maintain one unit in the LOB at ask price $v_t + \frac{s_2^*}{2}$ and one unit at bid price $v_t - \frac{s_2^*}{2}$.

(ii) B-Algos submit limit orders at v_t when they arrive, and all HFTs immediately take liquidity from B-Algos.

(iii) When v_t jumps up (down), all HFTs race to take stale limit orders at the ask (bid) price.

We call Proposition 2 the "Stimulating Equilibrium" because B-Algos, who have an internal need to trade, can "stimulate" HFTs to trade with them. Proposition 2 uncovers the existence of liquidity beyond that contained in displayed limit orders, in the sense that sophisticated traders can attract market makers by submitting aggressive limit orders. From an HFT's perspective, providing liquidity with limit orders is costly because sniping risks must be priced in. Market orders, on the other hand, are not subject to sniping risks.

In the existing literature, when HFTs demand liquidity, they usually adversely select other traders (BCS; Menkveld and Zoican 2017; Foucault, Kozhan, and Tham 2017). Consequently, HFTs' liquidity demands often have negative connotations. Our model shows that HFTs can demand liquidity without adversely selecting other traders. Instead, the transaction cost is lower for B-Algos when HFTs demand liquidity than when B-Algos demand liquidity from HFTs. Therefore, researchers and policymakers should not evaluate the welfare impact of HFTs simply based on whether they provide or demand liquidity.

The equilibrium spread s_2^* has two interesting features. First, like s_1^*, s_2^* is independent of the number of HFTs. This result is a consequence of the opportunity cost of providing liquidity. 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 adverse selection 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.

Second, $s_2^* > s_1^* > 0$, which means that humans pay more when B-Algos can use limit orders. When more non-HFTs use limit orders, HFTs have to quote wider bid–ask spreads for the remaining market orders. In this sense, B-Algos reduce their transaction costs at the expense of humans. Corollary 1 shows that the total transaction costs for non-HFTs decreases as β increases. That is, transaction costs for B-Algos decrease more than transaction costs for humans increase. Therefore, an increase in β increases overall market liquidity and benefits B-Algos while reducing liquidity for humans. We denote $\bar{C}(\beta)$ as the weighted average transaction cost for B-Algos and humans.

Corollary 1. *s_2^* strictly increases in β and $\bar{C}(\beta)$ strictly decrease in β. When $\beta \to 1$, $s_2^* \to 2d$ and $\bar{C}(\beta) \to 0$.*

The quoted bid–ask spread is a common measure of liquidity. Corollary 1 shows that this measure can be misleading when every trader can provide liquidity. As β increases, the quoted bid–ask spread widens, but transaction costs fall. When all non-HFTs are B-Algos, HFTs' bid–ask spreads widen to $2d$ but transaction costs zero out.

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 $\beta = 1$, however, HFTs make zero profits in equilibrium, and they have no economic

incentive to invest in speed. In the next section, we show that discrete pricing generates rents for both providing and demanding liquidity, thereby triggering arms races for speed.

4. Discrete Pricing

In this section we add another realistic feature to our model: discrete pricing. In Section 4.1 we show that discrete pricing creates rents for providing liquidity. In Section 4.2 we show that discrete pricing also creates rents for demanding liquidity. These rents, in turn, destroy the unique type of equilibrium outlined in Section 3, in which B-Algos always provide liquidity to HFTs and HFTs always provide liquidity to humans. Discrete pricing generates four types of equilibrium depending on parameter values, which then leads to cross-sectional and time-series predictions regarding who provides liquidity to whom.

For illustration purposes, we set the pricing grid as $\left\{\ldots, -\frac{\boldsymbol{3d}}{\boldsymbol{4}}, -\frac{\boldsymbol{d}}{\boldsymbol{4}}, \frac{\boldsymbol{d}}{\boldsymbol{4}}, \frac{\boldsymbol{3d}}{\boldsymbol{4}}, \ldots\right\}$. Therefore, the tick size is $\frac{d}{2}$ and v_t remains at the midpoint of the two nearest ticks after the fundamental value jumps. The intuition applied in this section, however, holds for any discrete tick size as long as B-Algos are not always able to achieve zero transaction costs by submitting limit orders at v_t. Then, B-Algos may choose limit orders that reside in the LOB and the state of the LOB can explode as infinitely many B-Algos arrive. To reduce the number of states, we make the following assumptions that are common in the LOB literature with discrete pricing. It is worth noticing that *none of these assumptions is binding when pricing is continuous.* Therefore, we are able to compare the results under discrete pricing as well as under continuous pricing based on these assumptions.

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

Assumption 1 implies that no traders can queue after existing orders at the same price. Assumption 1 does not offer a binding constraint for equilibrium under continuous pricing, in which the best bid and offer contains only one share. We introduce Assumption 1 here to reduce the state of the LOB to 2^n, where n is the number of price levels within the bid–ask spread. If we relax the assumption that traders can queue up to q shares, we need to track $(q+1)^n$ states of the LOB. The case for $q > 1$ only increases mathematical complexity without offering any additional intuitions.[10] We assume limit orders must be price-improving because we are tracking the best price. Assumption 1 is common in the LOB literature. For instance, Foucault, Kadan, and Kandel (2005) make the same assumption to reduce number of states of the LOB.[11]

Assumption 2: $N = \infty$

In Sections 2 and 3 we show that the number of HFTs does not affect the equilibrium bid–ask spread quoted by HFTs. To simplify the analysis, we assume that the number of HFTs is infinite to drop $\frac{N-1}{N}$ from our exposition and proofs. Consequently, the ex-ante expected sniping profit for any share is zero, and an HFT provides liquidity as long as its expected profit is greater than zero.

Assumption 3: *Non-spoofing: A trader cannot submit limit orders that they aim to cancel*

[10] The queue of B-Algos is finite because the execution probability associated with later queue positions is so low, and the sniping risk is so high, that B-Algos prefer using market orders or limit orders with better prices. Tracking the finite queue, however, can be complex as the state of the LOB depends on the random arrival of previous B-Algos.

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

before executing them.

We prevent both HFTs and B-Algos from spoofing, defined by the Dodd-Frank Act of 2010 as "bidding or offering with the intent to cancel the bid or offer before execution."[12] Without this assumption, spoofing arises endogenously in our model because the tick size creates rents for resting limit orders. However, such rents may turn negative if an incumbent order loses execution priority to an undercutting order. Thus, the spoofer can submit a non-profitable undercutting order to force an incumbent's profitable limit order to cancel, occupy the incumbent's position, and then cancel the spoofing order. We rule out spoofing because it is illegal, and it is also not the focus of our paper.

4.1. Rents for Providing Liquidity and Queuing Equilibrium

Consider the extreme case when $\kappa \equiv \frac{\lambda_J}{\lambda_I} = 0$, where the break-even spread in Propositions 1 and 2 are both zero. Then, the bid–ask spread is binding at one tick, and the tick size becomes pure rent for providing liquidity. These rents generate speed races for providing liquidity. A similar intuition holds when the break-even bid–ask spread is smaller than one tick. The difference between the mandated one-tick minimum spread and the break-even spread creates rents for providing liquidity, and the time-priority rule allocates such rents to HFTs. B-Algos 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 the queuing equilibrium, in which HFTs provide liquidity to both B-Algos and humans.

[12] 7 U.S.C.A. § 6c(a)(5)(C)

The key to characterizing the queuing equilibrium is to find the parameter value when the tick size is binding. Because of symmetry, we illustrate only the ask side of the LOB. Consider HFTs' expected profits for providing liquidity at $v_t + \frac{d}{4}$, $LP\left(\kappa, \frac{d}{4}\right)$:

$$LP\left(\kappa, \frac{d}{4}\right) = \frac{1}{2\kappa+2} \cdot \frac{d}{4} + \frac{1}{2\kappa+2} \cdot LP\left(\kappa, \frac{d}{4}\right) + \frac{\kappa}{2\kappa+2} \cdot \left(\frac{d}{4} - d\right) + \frac{\kappa}{2\kappa+2} \cdot 0 \tag{6}$$

Because B-Algos do not provide liquidity under Assumption 1, four types of events can change the status of the LOB: 1) At probability $\frac{1}{2\kappa+2}$, the next event is a non-HFT buy order, and it leads to a profit of $(v_t + \frac{d}{4}) - v_t = \frac{d}{4}$; 2) at probability $\frac{1}{2\kappa+2}$, a non-HFT sell order arrives, and it does not affect $LP(\kappa, \frac{d}{4})$ on the ask side, because HFTs immediately refill the bid side and restore the previous state of the LOB; 3) at probability $\frac{\kappa}{2\kappa+2}$, the fundamental value v_t jumps upward to $v_t + d$, and the HFT sell limit order is sniped; the payoff from being sniped is $(v_t + \frac{d}{4}) - (v_t + d) = -\frac{3d}{4}$ and we no longer have the term $\frac{N-1}{N}$ here because we assume $N = \infty$; and 4) when v_t jumps downward, the liquidity supplier cancels the order and joins the race to provide liquidity at a new BBO and the payoff is zero. Also, we write liquidity provision revenue as a function of both the half-spread and κ because what really matters for HFTs' liquidity provision revenue is the arrival intensity of the security's value-jumping over non-HFTs. The former imposes costs on liquidity provision, while HFTs make profits from the latter. The solution for equation (6) is:

$$LP\left(\kappa, \frac{d}{4}\right) = \frac{1}{2\kappa+1}\frac{d}{4} - \frac{\kappa}{2\kappa+1}\frac{3d}{4} \tag{7}$$

$LP\left(\kappa, \frac{d}{4}\right) \geq 0$ when $\kappa \leq \frac{1}{3}$, which is the region in which HFTs sustain a one-tick bid–ask spread.

Proposition 3. (**Queuing Equilibrium**) *When Tick Size is* $\frac{d}{2}$ *and* $\kappa \leq \frac{1}{3}$:

(i) At almost all times t, HFTs maintain one share at the ask price $v_t + \frac{d}{4}$ and one share at the bid price $v_t - \frac{d}{4}$.

(ii) HFTs participate in two speed races: (a) the race to fill the queue when the depth at $v_t \pm \frac{d}{4}$ becomes zero; (b) the race to pick off all stale quotes following a value jump.

(iii) All non-HFTs use market orders to trade upon arrival.

The new feature of the queuing equilibrium is the race to provide liquidity at $v_t \pm \frac{d}{4}$. When the market opens, each HFT sends two limit orders: one sell limit order at $v_0 + \frac{d}{4}$ and one buy limit order at $v_0 - \frac{d}{4}$. When a non-HFT arrives and takes the order at $v_t + \frac{d}{4}$ or $v_t - \frac{d}{4}$, HFTs race to refill the order. Following value jumps, HFTs race to provide liquidity at a half-spread of $\frac{d}{4}$ around the new fundamental value.

4.2. Rents for Demanding Liquidity and Stimulating, Undercutting, and Crash Equilibriums

When $\kappa > \frac{1}{3}$, the value of providing liquidity at $v_t \pm \frac{d}{4}$ becomes negative. Therefore, HFTs no longer quote a bid–ask spread at one binding tick. Once B-Algos are able to place limit orders within HFTs' BBO, one result immediately emerges following a similar intuition expressed in Proposition 2. B-Algos would never use market orders, because a stimulating buy limit order at $v_t + \frac{d}{4}$ or a stimulating sell limit order at $v_t - \frac{d}{4}$ strictly dominates market orders. Discrete pricing, however, creates two new features that are not present with continuous pricing.

First, the tick size generates rents for demanding liquidity. When pricing is continuous, B-Algos can place limit orders at v_t. When pricing is discrete, B-Algos have to place buy limit orders at $v_t + \frac{d}{4}$ and sell limit orders at $v_t - \frac{d}{4}$ to attract HFTs. A rent of $\frac{d}{4}$ for market orders drives speed competitions for demanding liquidity.

Second, when pricing is continuous, B-Algos can achieve a minimum possible transaction cost by stimulating HFTs to demand liquidity. When pricing is discrete, a stimulating limit order incurs a transaction cost of $\frac{d}{4}$. Therefore, B-Algos may find it optimal to submit limit orders that do not cross the midpoint.

These two new features generate three types of equilibria depending on the parameter values. Figure 1 illustrates the intuition underlying these three types of equilibria, and we will characterize the exact boundary that divides them in this subsection. Intuitively, when κ rises slightly above $\frac{1}{3}$, HFTs retreat to $v_t \pm \frac{3d}{4}$ because $v_t \pm \frac{d}{4}$ loses money. B-Algos, however, are willing to submit limit orders at $v_t \pm \frac{d}{4}$ as long as they lose less than $\frac{d}{4}$ in value, the cost of stimulating limit orders. We call this type of equilibrium undercutting equilibrium, in which B-Algos provide liquidity to non-HFTs. As κ further increases, the cost of using limit orders at $v_t \pm \frac{d}{4}$ increases but the cost of using stimulating limit orders remains at $\frac{d}{4}$. Thus, B-Algos submit limit orders at $v_t + \frac{d}{4}$ to buy or at $v_t - \frac{d}{4}$ to sell to stimulate HFTs to demand liquidity when κ is higher than the undercutting equilibrium but below the short-dashed line. As κ increases above the short-dashed line, HFTs lose money even when they quote at $v_t \pm \frac{3d}{4}$. Therefore, they choose to quote at $v_t \pm \frac{5d}{4}$. We call this type of equilibrium the crash equilibrium because HFTs effectively quit liquidity provision by quoting a spread outside the jump size d. The share in traders of B-Algos, β, plays a role that is similar to that played by κ, because an increase in β effectively reduces future market order flows and thereby increases the probability of being sniped for limit orders.

[Insert Figure 1 about here]

We present the three types of equilibrium in the order of their complexity. In Subsection

4.2.1 we present the stimulating equilibrium under discrete pricing, in Subsection 4.2.2 we present the undercutting equilibrium, and in Subsection 4.2.3 we present the crash equilibrium.

4.2.1 Stimulating Equilibrium under Discrete Pricing

In this subsection, we characterize the stimulating equilibrium, in which B-Algos use buy (sell) limit orders at $v_t + \frac{d}{4}$ ($v_t - \frac{d}{4}$) to stimulate market orders from HFTs, and HFTs choose to submit limit orders at $v_t \pm \frac{3d}{4}$ to provide liquidity to humans. Proposition 4 summarizes the stimulating equilibrium.

Proposition 4. (**Stimulating Equilibrium under Discrete Pricing**) *When the tick size is* $\frac{d}{2}$ *and* $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa \leq 3(1-\beta)$:

(i) At almost all times t*, HFTs maintain one share at the ask price* $v_t + \frac{3d}{4}$ *and one share at the bid price* $v_t - \frac{3d}{4}$.

(ii) B-Algo buyers submit limit orders at $v_t + \frac{d}{4}$ *and B-Algo sellers submit limit orders at* $v_t - \frac{d}{4}$.

(iii) *HFTs participate in three speed races: (a) the race to pick off all stale quotes following value jumps, (b) the race to fill the queue when the depth at* $v_t \pm \frac{3d}{4}$ *becomes zero, and (c) the race to take the liquidity offered by B-Algos.*

Part 3 of Proposition 4 reveals a new type of speed competition: racing to be the first to take the liquidity offered by stimulating limit orders. This race does not exist under continuous pricing, because B-Algos leave no rents for HFTs. This race also does not exist under queuing equilibrium, because there is no price level at which to submit stimulating limit orders. Stimulating

equilibrium retains the two speed races discussed above. As occurs under queuing equilibrium, HFTs still race for the top queue positions. Even if the tick size is not binding, discrete pricing still creates rents for liquidity provision, because the bid–ask spread quoted by HFTs would be wider than the break-even spread, unless in the knife-edge case they happen to be identical. Finally, under stimulating equilibrium and all the equilibria in our model, HFTs always race to snipe stale quotes.

4.2.2 Undercutting Equilibrium

In this subsection, we characterize the undercutting equilibrium, in which HFTs quote $v_t + \frac{3d}{4}$ to sell and $v_t - \frac{3d}{4}$ to buy, and B-Algos quote $v_t + \frac{d}{4}$ to sell and $v_t - \frac{d}{4}$ to buy when these price levels contain no other limit orders. An undercutting equilibrium occurs when $\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$. If $\kappa \leq \frac{1}{3}$, the sniping risk is so low that HFTs will find it profitable to quote a binding one tick. If $\kappa \geq \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, the sniping risk is so high that B-Algos will find it optimal to use stimulating limit orders at a cost of $\frac{d}{4}$.

The undercutting equilibrium includes an additional feature that is not present under the previously presented equilibria. Because B-Algos leave no resting limit orders under any of those equilibria, the LOB as referenced in previous sections effectively contains only one state and HFTs immediately restore the unique equilibrium state after any event. After B-Algos leave limit orders on the book, the LOB contains four states under simplifying Assumption 1. We define the state of the LOB as (i, j), where i represents the number of B-Algos' limit orders on the same side of the LOB, and j denotes the number of B-Algos' limit orders on the opposite side of the LOB. For example, for a trader who wants to buy, i represents the number of B-Algos' limit orders on the bid side, and j represents the number of B-Algos' limit orders on the ask side. Then,

(0,0) No limit order from B-Algos
(1,0) A B-Algo limit order on the same side
(0,1) A B-Algo limit order on the opposite side
(1,1) B-Algo limit orders on both sides

The core of Proposition 5 characterizes HFTs' strategies in each state and for each event. Denote the payoff of an HFT who supply liquidity at state (i,j) as $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)$. HFTs will quote a price at $v_t \pm \frac{3d}{4}$ if $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right) \geq 0$. $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)$, in turn, depends on all the payoffs of all other states in the book, because the six types of events outlined in equation (5) transit the LOB from one state to the other. Figure 2 illustrates the dynamics of such transitions. To take one example, consider $LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$ for an HFT on the ask side of the LOB:[13]

1) A B-Algo buyer undercuts the bid side at $v_t - \frac{d}{4}$ and changes $LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$ to $LP^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right)$.
2) A B-Algo seller undercuts the ask side at $v_t + \frac{d}{4}$ and changes $LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$ to $LP^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$.
3) A human buyer submits a market buy order and the HFT gains $\frac{3d}{4}$.
4) A human seller submits a sell market order, HFTs race to fill the bid side immediately, and $LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$ remains the same.
5) In an upward value jump, the limit order on the ask side loses $\frac{d}{4}$.
6) In a downward value jump, the liquidity provider cancels the limit order, thereby changing $LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)$ to zero.

[13] HFTs make independent decisions on bid and ask sides. The state (i,j) for one side is state (j,i) for the other side.

[Insert Figure 2 about here]

The following equation system summarizes the dynamics described above:

$$\begin{cases} LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right) = p_1\overline{LP}^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_1\overline{LP}^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2\frac{3d}{4} + p_2LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3\cdot 0 \\ LP^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) = p_1\overline{LP}^{(1,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_1LP^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2\overline{LP}^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2LP^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3\cdot 0 \\ LP^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right) = p_1LP^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_1\overline{LP}^{(1,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2\frac{3d}{4} + p_2\overline{LP}^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3\cdot 0 \\ LP^{(1,1)}\left(\kappa,\beta,\frac{3d}{4}\right) = p_1\overline{LP}^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_1\overline{LP}^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2\overline{LP}^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_2\overline{LP}^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3\cdot 0 \end{cases} \quad (8)$$

where $p_1 = \frac{\beta\lambda_I/2}{\lambda_I+\lambda_J} = \frac{\beta}{2+2\kappa}$, $p_2 = \frac{(1-\beta)\lambda_I/2}{\lambda_I+\lambda_J} = \frac{1-\beta}{2+2\kappa}$, and $p_3 = \frac{\lambda_J/2}{\lambda_I+\lambda_J} = \frac{\kappa}{2+2\kappa}$ are the probabilities that the next event is the arrival of a B-Algo buyer (seller), the arrival of a human buyer (seller), and the upward (downward) jump of the fundamental value, respectively. $\overline{LP}^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right) = \max\{0, LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)\}$ reflects the fact that HFTs can simply choose not to submit a limit order or cancel an existing limit order once the payoff becomes negative.

Equation system (8) comprises four equations. Each equation shows the payoff for HFTs' liquidity provision under each state (i,j), which depends on the payoffs of other states. Depending on the next arrival event, the state of the LOB will change, as will the HFT's liquidity-provision profits, at $v_t \pm \frac{3d}{4}$.

Proposition 5. (**Undercutting Equilibrium**): *When the tick size is* $\frac{d}{2}$ *and* $\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, *the equilibrium is characterized as follows:*

1. *B-Algos who intend to buy (sell) submit limit orders at price* $v_t - \frac{d}{4}$ $(v_t + \frac{d}{4})$ *if no existing limit orders sit at that price level, or buy (sell) limit orders at price* $v_t + \frac{d}{4}$ $(v_t - \frac{d}{4})$ *otherwise.*
2. *HFTs' strategy:*

a. *HFTs provide liquidity at* $v_t + \frac{3d}{4}$ *or* $v_t - \frac{3d}{4}$ *if* $LP^{(i,j)}\left(\kappa, \beta, \frac{3d}{4}\right) \geq 0$ *at state* (i, j). $LP^{(i,j)}\left(\kappa, \beta, \frac{3d}{4}\right)$ *is determined by the transition matrix (8).*

b. *HFTs race to snipe stale quotes from HFTs and B-Algos during value jumps.*

c. *HFTs race to take stimulating limit orders from B-Algos.*

In undercutting equilibrium, B-Algos submit undercutting limit orders that rest in the LOB. Their goal is to provide liquidity to other non-HFTs, though HFTs may snipe the undercutting limit orders following during value jumps. Though B-Algos have higher sniping costs, they incur lower opportunity costs than HFTs when providing liquidity. Therefore, their resting limit orders can have more aggressive price than the quotes from HFTs.

The undercutting equilibrium offers one new feature relative to previous equilibria. In undercutting equilibrium, HFTs' depth at their best quotes is not constant, because they need to respond to B-Algos' undercutting orders. Therefore, HFTs may update their quotes even if the fundamental value does not change at all. For example, when $\kappa = 0.5$, $\beta = 0.6$, we have $LP^{(0,j)}\left(\kappa, \beta, \frac{3d}{4}\right) > 0$ and $LP^{(1,j)}\left(\kappa, \beta, \frac{3d}{4}\right) < 0$.[14] HFTs provide liquidity at a half spread of $\frac{3d}{4}$ when there is no undercutting order, but the depth at a half spread of $\frac{3d}{4}$ becomes zero once an undercutting order changes the book state to $(1, j)$. If a market order executes against the undercutting order from the B-Algo, the state of the LOB change back to $(0, j)$. HFTs again find that providing liquidity at a half spread of $\frac{3d}{4}$ profitable and races to provide liquidity at such spread. Therefore, the undercutting equilibrium provide one channel to explain the frequent addition and cancellation of HFTs' quotes (Biais and Foucault [2014]; Hasbrouck and Saar [2013]).

4.2.3 Crash Equilibrium

[14] We analytically solve them in the proof of Proposition 5.

In this subsection, we show that when the sniping risk is high or when the share of trading by B-Algos is large, HFTs cannot make profits at any price level within the jump size. They must then quote a bid–ask spread that is wider than the jump size and thus effectively quit providing liquidity. Proposition 6 shows that HFTs retreat to $v_t \pm \frac{5d}{4}$ when the jump size is d. Corollary 2 generalizes the intuition of Proposition 6 by allowing larger jump sizes. We call this equilibrium the crash equilibrium because it provides an intuition for flash crashes, defined by Biais and Foucault (2014) as sharp price movements in one direction followed by quick reversion.

Proposition 6 shows that, when the jump size is d, HFTs quote $v_t \pm \frac{5d}{4}$ when κ is greater than $\max\left\{3(1-\beta), \frac{1}{3}\right\}$. Therefore, a B-Algo seller can choose from four price levels: $v_t + \frac{3d}{4}$, $v_t + \frac{d}{4}$, $v_t - \frac{d}{4}$, or $v_t - \frac{3d}{4}$. The proof of Proposition 6 shows that a B-Algo seller uses only two price levels: $v_t + \frac{3d}{4}$ and $v_t - \frac{d}{4}$. Here we offer the intuition underlying this result.

Selling at $v_t - \frac{d}{4}$ strictly dominates $v_t - \frac{3d}{4}$ because both price levels immediately prompt HFTs to demand liquidity, and $v_t - \frac{d}{4}$ has a lower cost of $\frac{d}{4}$. Selling at $v_t - \frac{d}{4}$ also strictly dominates $v_t + \frac{d}{4}$ because we show in Proposition 4 that B-Algos prefer $v_t - \frac{d}{4}$ to $v_t + \frac{d}{4}$ when $\kappa > \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, and the sniping risk is $\kappa > \max\left\{3(1-\beta), \frac{1}{3}\right\} > \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$ in the crash equilibrium. In summary, a stimulating limit order at price $v_t - \frac{d}{4}$ strictly dominates both a more aggressive stimulating limit order at $v_t - \frac{3d}{4}$ and a regular limit order at $v_t + \frac{d}{4}$ for a B-Algo seller.

Finally, sell limit orders at $v_t + \frac{3d}{4}$ dominate stimulating limit orders at $v_t - \frac{d}{4}$. Quotes at

$v_t + \frac{3d}{4}$ lose $\frac{d}{4}$ in value jumps but have strictly positive profits when a human takes the quote, so its expected transaction cost is lower than $\frac{d}{4}$. Therefore, B-Algo sellers start with a quote at $v_t + \frac{3d}{4}$ when the price level does not contain an order; otherwise, they use stimulating limit orders. We summarize these results in Proposition 6.

Proposition 6 (**Crash Equilibrium**). *When tick size is* $\frac{d}{2}$ *and* $\kappa > \max\left\{3(1-\beta), \frac{1}{3}\right\}$:

(i) At almost all times t*, HFTs maintain one share at the ask price* $v_t + \frac{5d}{4}$ *and one share at the bid price* $v_t - \frac{5d}{4}$.

(ii) A B-Algo buyer submits a buy limit order at price $v_t - \frac{3d}{4}$ *and a B-Algo seller submits a sell limit order at price* $v_t + \frac{3d}{4}$ *if the price levels have no limit orders. Otherwise, B-Algo buyers submit stimulating buy limit orders at price* $v_t + \frac{d}{4}$*, and B-Algo sellers submit stimulating sell limit orders at price* $v_t - \frac{d}{4}$.

(iii) HFTs participate in three speed races: (a) the race to pick off all stale quotes following value jumps; (b) the race to fill the order at $v_t \pm \frac{5d}{4}$*; (c) the race to take the stimulating liquidity offered by B-Algos at* $v_t \pm \frac{d}{4}$*;*

One main feature of a crash equilibrium is large variations in transaction prices. A human buyer may pay $v_t + \frac{5d}{4}$ if she hits the quote given by HFTs. A B-Algo buyer, however, pays at most $v_t + \frac{d}{4}$, which is the same price she pays under the stimulating equilibrium. Therefore, an increase in sniping risk κ has a much smaller impact on B-Algos than on humans. Corollary 2 shows that the size of a jump also has a very limited impact on a transaction price for B-Algos.

Corollary 2. (Crash Equilibrium with Wider Jump Sizes) *For any jump size* $J = md$ *where* $m \in \mathbb{N}^+$, *when* $\kappa > max\{\frac{1}{4m-1}, (4m-1)(1-\beta)\}$, *HFTs quote one share at* $v_t \pm \left(m + \frac{1}{4}\right) d$. *B-Algos submit regular buy limit orders at* $v_t - \left(m - \frac{1}{4}\right) d$ *and sell limit orders at* $v_t + \left(m - \frac{1}{4}\right) d$ *when the price level is available. Otherwise B-Algos submit stimulating buy limit orders at* $v_t + \frac{d}{4}$ *and sell limit orders at* $v_t - \frac{d}{4}$.

An increase in the potential jump size further increases the buy price for a human, but the buy price for B-Algos remains bounded by $v_t + \frac{d}{4}$. Therefore, a flash crash occurs when a (naïve) market order hits the quotes of an HFT. In our model, we use flash crashes to refer to both the market wide flash crashes such as the flash crash on May 6, 2010, in which Dow Jones plunged 998.5 points, and the 18,520 individual stock mini-flash crashes (Johnson et al. 2013). Flash crashes led by the mechanism outlined in Proposition 6 and Corollary 2 have the following two features. First, Flash crashes are equally likely to go up than to go down. Second, B-Algos can still trade at low transaction costs around flash crashes, although they are more likely to use stimulating limit orders to attract HFTs.

Regarding the first prediction, Nanex, the firm that invented the concept of the mini-flash crash, finds that mini-flash crashes are equally likely to be upward or downward. Indeed, even during the flash crash on May 6, 2010, in which the Dow Jones plunged 998.5 points, some stocks, including Sotheby's, Apple Inc., and Hewlett-Packard, increased in value to over $100,000 in price (SEC, 2010). More broadly, although the stock market experienced a market-wide flash crash, the

treasury market experienced a flash rally on October 15, 2014.[15] Therefore, traditional theories explaining market crashes, such as that posited by Huang and Wang (2009), cannot explain this symmetric pattern, because these theories work well for crashes but not for rallies.

The second prediction implies that sophisticated traders can still trade in both directions during flash crashes and flash rallies without incurring large transaction costs, because such flash crashes and flash rallies are driven by naïve market orders. This prediction, which is also unique for our model, has not been tested. However, Federal Reverse Board Governor Lael Brainard mentioned a contradictory dynamic in market-turmoil episodes:[16] trading activity continues despite a wider spread and lower depth. She then points out that "the dynamic nature of liquidity provision by high-speed market makers makes static measures of liquidity, such as posted bid–ask spreads and market depth, less useful." Our model provides one possible way to reconcile this contradiction: when market turmoil occurs, sophisticated traders' trading costs can be much lower than the displayed bid–ask spread.

The time-series patterns of mini-flash crashes reported in Brogaard et al. (2018) support our mechanism. They show the following pattern. Ten seconds before a mini-flash crash, HFTs demand liquidity from non-HFTs. At the time of a mini-flash crash, HFTs supply liquidity to non-HFTs, but at a much wider bid–ask spread. The authors also find that liquidity provision during the mini-flash crash is profitable. This evidence is consistent with the theoretical mechanism for mini-flash crashes in our model: (1) slightly before a mini-flash crash, HFTs snipe limit orders

[15] On July 13, 2015 a joint staff report was released on the findings of the Treasury flash rally by the U.S. Department of the Treasury, the Board of Governors of the Federal Reserve System, the Federal Reserve Bank of New York, the U.S. Securities and Exchange Commission, and the U.S. Commodity Futures Trading Commission.

[16] Brainard, L. (2018). "The Structure of the Treasury Market: What Are We Learning?" *The Evolving Structure of the U.S. Treasury Market Fourth Annual Conference* Hosted by the Federal Reserve Bank of New York, New York, New York.

from other traders; (2) a mini-flash crash occurs when a market order hits HFTs' quotes that are away from the market; thus, HFTs profit when a mini-flash crash occurs.

An increase in β creates two competing economic forces that might drive flash crashes. An increase in the fraction of B-Algos increases the probability that HFTs will provide quotes far away from the market, but sophisticated B-Algos never hit these quotes. When all non-HFTs are humans, or when all non-HFTs are B-Algos, flash crashes do not occur. Therefore, flash crashes result from interactions between the three types of traders.

5. Predictions and Policy Implications

By adding liquidity-providing non-HFTs and discrete pricing, our model not only rationalizes a number of puzzles in the literature, it also generates new testable predictions. In Subsection 5.1, we summarize the predictions that are driven mainly by liquidity-providing non-HFTs. In Subsection 5.2, we summarize the predictions that are driven by discrete pricing. In Subsection 5.3, we discuss the policy implications of our paper.

5.1 Predictions Driven by Liquidity-providing Non-HFTs

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

Prediction 1 (Price Priority): *Non-HFTs are more likely 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, as they incur lower adverse-selection costs (see Jones [2013] and Menkveld [2016]) surveys), lower inventory costs (Brogaard et al. 2015) and lower operational

costs (Carrion 2013). Our model shows that the opportunity cost of providing liquidity can reconcile this contradiction. B-Algos incur lower opportunity costs when providing liquidity, and they can afford more aggressive limit orders as long as they cost less to execute than market orders.

Prediction 2 (Negative Correlation between the Bid–Ask Spread and Liquidity): *Technology shocks that increase the fraction of B-Algos widen the bid–ask spread but reduce the overall transaction cost.*

Black (1971) describes a liquid market intuitively in the following manner:

"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 first 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) may be internally inconsistent. In Proposition 2, an increase in β will widen the bid–ask spread because HFT market makers receive less non-HFT order flows. On the other hand, the average transaction cost for non-HFTs drops. In the extreme case in which $\beta = 1$, a market is infinitely liquid when all non-HFTs are B-Algos, because every trader pays zero transaction costs. At the same time, the bid–ask spread is at its widest. Proposition 2 and Corollary 1 suggest that we should

update the definition of liquidity and the measure of liquidity for modern electronic markets. Corollary 1 also leads to Prediction 2, which can be tested by examining whether technology shocks that support B-Algos reduce transaction costs for institutional traders (such as implementation shortfalls measured by Acerno data) while increasing the bid–ask spread.

5.2 Predictions Driven by a Discrete Tick Size

When pricing is continuous, B-Algos always provide liquidity to HFTs, and HFTs always provide liquidity to humans for any parameter value. When pricing is discrete, who provides liquidity to whom depends on the parameter value, and such dependence generates cross-sectional and time-series predictions regarding liquidity provision and liquidity demand.

Prediction 3 (Time Priority versus Price Priority): *HFTs crowd out non-HFTs' liquidity provision when the tick size is large.*

Prediction 3 works against the grain of arguments advanced in Chordia et al. (2013), who 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." Their concern would be valid if HFTs quote more aggressive prices than non-HFTs when pricing is more continuous. Yet Brogaard et al. (2015) and Yao and Ye (2018) find that non-HFTs are more likely than HFTs to establish price priority, and our paper provides the theoretical foundation for this finding. In our model, B-Algos can quote tighter bid–ask spreads than HFTs because B-Algos face worse outside options. HFTs place no private value in trade. B-Algos have an internal need to trade, and they use limit orders as long as their costs are lower than the costs of using market orders. A large tick size, however, imposes a constraint that 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 one-cent uniform tick size

leads to larger relative tick size. They also find that HFTs provide a larger share of liquidity for low-priced securities. These results are consistent with Prediction 3.

Prediction 4 (Adverse Selection and Liquidity Provision): *An increase in adverse-selection risk decreases the share of liquidity provided by HFTs.*

Prediction 4 differs significantly from findings reported in the existing literature on HFTs. Prior studies typically model HFTs as traders who can access information more rapidly than other traders. In this framework, speed competition should be more active when there is more information. In particular, Hoffmann (2014), Han, Khapko, and Kyle (2014), Bernales (2016), and Bongaerts and Van Achter (2016) find that HFTs incur lower adverse-selection costs than non-HFTs. Therefore, an increase in the level of information should give HFTs a comparative advantage in liquidity provision.

Prediction 4, however, implies that less information drives speed competition. Compare Proposition 3 with Propositions 4, 5, and 6: when the level of sniping risk is low, the binding bid–ask spread drives speed competition at a constrained spread because liquidity provision is highly profitable. If the incidence of sniping rises high enough, the spread is wider than one tick, allowing non-HFTs to undercut HFTs and decreasing liquidity provision on the part of HFTs. One limitation of our model is that we model only adverse selection led by sniping, but other types of adverse selection should provide the same economic mechanism. Generally, the break-even bid–ask spread should be lower when adverse-selection risk is low. Once the break-even spread falls below one tick, speed competition to achieve time priority should be more critical. Yao and Ye (2018) provide cross-sectional evidence consistent with Prediction 4: stocks with higher adverse-selection risk have a lower fraction of liquidity provided by HFTs. It would be interesting to test whether Prediction 4 holds in time series, that is whether, for a given security, HFTs provide less fraction

of liquidity when adverse-selection risk is high.

Prediction 5 addresses the question of who provides liquidity during flash crashes.

Prediction 5. (**Flash Crashes**): *A flash crash is more likely to occur when the sniping risk is high. Limit orders from non-HFTs, however, incur much lower transaction costs.*

Again, here we do not distinguish market wide flash crashes from mini-flash crashes at individual security level. HFTs' limit orders are less likely to be executed and the sniping cost is likely to be higher when κ is high. Moreover, higher sniping risk widens the break-even bid–ask spread; a wider break-even bid–ask spread also allows B-Algos to undercut HFTs, further increasing the adverse-selection costs for HFTs. When κ is high enough, HFTs effectively stop providing liquidity by placing quotes far away from the market. Because B-Algos do not continuously provide liquidity in the market, humans' market orders can hit HFTs' quotes and cause flash crashes.

Our interpretations of flash crashes are consistent with both negative and positive evidence of the role of HFTs in flash crashes. Brogaard et al. (2018) suggest that HFTs provide liquidity in extreme price movements, while Ait-Sahalia and Sağlam (2017) suggest that HFTs withdraw liquidity when it is most needed. Both views suggest, however, that flash crashes occur when the market orders of non-HFTs hit quotes from HFTs that are placed away from the market.

If mini-flash crashes are preceded or signaled by high sniping risk, our model predicts that transaction costs for non-HFTs are much lower if they use limit orders. At least non-HFTs can use stimulating limit orders to encourage HFTs to demand liquidity. This unique prediction can be tested to see whether our model captures the main driver of flash crashes.

Prediction 6. (**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.

Latza, Marsh, and Payne (2014) find evidence consistent with Prediction 6. 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 they lead to virtually no permanent price impacts. It will be interesting to test Prediction 6 more directly using data that include account information of traders.

5.3 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 price. Particularly, we show that, when all non-HFTs are B-Algos, transaction costs are zero and there is no incentive for HFTs to engage in speed competition even if time is continuous. In this sense, our paper supports Kyle and Lee's (2017) vision of a fully continuous market.

On April 5, 2012, President Barack Obama signed the Jumpstart Our Business Startups (JOBS) Act. Section 106 (b) of the Act requires the SEC to examine the effects of tick size on initial public offerings (IPOs). On October 3, 2016, the SEC implemented a pilot program to increase the tick size from one cent to five cents for 1,200 common stocks that have a market capitalization of $3 billion or less, a closing price of at least $2.00, and a consolidated average daily volume of one million shares or fewer. Proponents of the proposal argue that a larger tick size can improve liquidity (Weild, Kim, and Newport, 2012). In Prediction 7, however, we posit that an increase in the tick size decreases liquidity.

Prediction 7. *A larger tick size increases transaction costs.*

Discrete pricing can create rents for HFTs and push up non-HFTs' execution costs in two ways. First, when the tick size is binding and non-HFTs have to rely more on market orders, they pay higher than the break-even spread to HFTs. Second, when non-HFTs use limit orders to trigger HFTs, they have to pay beyond the marginal valuation of HFTs.[17] Yao and Ye (2018) and Albuquerque, Song, and Yao (2018) find evidence consistent with Prediction 7. Our model's prediction along with their empirical evidence shows that an increase in the tick size would not improve liquidity.

6. Conclusion

This paper contributes to the literature by including two salient features that are found in financial markets: algorithmic traders who are not HFTs and discrete pricing. B-Algos incur lower opportunity costs than HFTs when providing liquidity, because providing liquidity is always less costly than demanding liquidity from HFTs. When prices are continuous enough, B-Algos can establish price priority over HFTs. 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 break-even bid–ask spread relative to the tick size, which allows B-Algos 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) Non-HFTs are more likely than HFTs to provide liquidity at price levels that cross the midpoint, and these limit orders are

[17] Certainly, when two non-HFTs trade with each other, one side may benefit from discrete pricing while the other side may lose, but such gains and losses zero out between them.

more likely to be taken by HFTs almost immediately. 2) A flash crash is more likely to occur for stocks subject to higher sniping risk, but transaction costs for B-Algos do not change much during flash crashes. 3) The bid–ask spread widens when technological shocks increase the proportion of B-Algos, but overall transaction costs decrease. Therefore, the bid–ask spread can move in the opposite direction to the true liquidity when all traders can use limit orders.

Our model also shows 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 five cents.

By adding trading algorithms designed by sophisticated non-HFTs, our model adds significant new insight to the understanding of how HFTs affect financial markets. For example, we find that B-Algos can prompt HFTs to demand liquidity using stimulating limit orders to reduce transaction costs. Therefore, we should not evaluate the impact of HFTs on liquidity and social welfare based on whether they demand or provide liquidity.

References

Albuquerque, R., Song, S., and Yao, C. (2017). The Price Effects of Liquidity Shocks: A Study of SEC's Tick-Size Experiment.

Ait-Sahalia, Y., and Sağlam, M. (2017). High frequency market making: Implications for liquidity.

Biais, B., and T. Foucault. 2014. HFT and market quality. *Bankers, Markets & Investors* 128:5-19.

Bernales, A. 2016. Algorithmic and High Frequency Trading in Dynamic Limit Order Markets. Working Paper, Universidad de Chile.

Black, F. (1971). Toward a fully automated stock exchange, part I. *Financial Analysts Journal*, 27(4), 28-35.

Bongaerts, D., and M. V. Achter. 2016. High-Frequency Trading and Market Stability. Working Paper, Erasmus University Rotterdam.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan. 2015. Trading fast and slow: Colocation and liquidity. *Review of Financial Studies* 28:3407-3443.

Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., and Sokolov, K. (2018). High frequency trading and extreme price movements. *Journal of Financial Economics*, *128*(2), 253-265.

Budish, E., P. Cramton, and J. Shim. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics* 130:1547-1621.

Carrion, A. 2013. Very fast money: High-frequency trading on the NASDAQ. *Journal of Financial Markets* 16:680-711.

Chordia, T., A. Goyal, B. N. Lehmann, and G. Saar. 2013. High-frequency trading. *Journal of Financial Markets* 16:637-645.

Clark-Joseph, A. D., Ye, M., and Zi, C. (2017). Designated market makers still matter: Evidence from two natural experiments. *Journal of Financial Economics*, *126*(3), 652-667.

Foucault, T., R. Kozhan, and W.W. Tham. 2017. Toxic arbitrage. *Review of Financial Studies* 30:1053-1094.

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

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies*, 18(4), 1171-1217.

Frazzini, A., R. Israel, and T. J. Moskowitz. 2014. Trading costs of asset pricing anomalies. Working paper, AQR Capital Management, and University of Chicago.

Glosten, L. R., and P. R. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14:71-100.

Goettler, R. L., C. A. Parlour, and U. Rajan. 2005. Equilibrium in a dynamic limit order market. *Journal of Finance* 60:2149-2192.

Goettler, R. L., C. A. Parlour, and U. Rajan. 2009. Informed traders and limit order markets. *Journal of Financial Economics* 93:67-87.

Han. J., M. Khapko, and A. S. Kyle. 2014. Liquidity with High-Frequency Market Making. Working Paper, Swedish House of Finance, University of Toronto, and University of Maryland.

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

Hoffmann, P. 2014. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics* 113:156-169.

Huang, J., and Wang, J. (2009). Liquidity and market crashes. *Review of Financial Studies*, 22(7), 2607-2643.

Johnson, N., Zhao, G., Hunsader, E., Qi, H., Johnson, N., Meng, J., and Tivnan, B. (2013). Abrupt rise of new machine ecology beyond human response time. *Scientific reports*, 3, 2627.

Jones, C. 2013. What do we know about high-frequency trading? Working paper, Columbia University.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3), 967-998.

Kyle, A. S., and Lee, J. (2017). Toward a fully continuous exchange. *Oxford Review of Economic Policy*, 33(4), 650-675.

Kyle, A. S., and Obizhaeva, A. A. (2016). Market microstructure invariance: Empirical hypotheses. *Econometrica*, 84(4), 1345-1404.

Kyle, A. S., and Obizhaeva, A. A. (2016). Large bets and stock market crashes. Working paper, University of Maryland, and New Economic School (NES)

Latza, T., We. W. Marsh, and R. Payne. 2014. Fast aggressive trading. Working paper, Blackrock, and City University London.

Menkveld, A. J. 2016. The economics of high-frequency trading: Taking stock. *Annual Review of*

Financial Economics 8:1-24.

Menkveld, A. J., and M. A. Zoican. 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.

O'Hara, M., G. Saar, and Z. Zhong. 2018. Relative tick size and the trading environment. Working Paper, Cornell University, and University of Melbourne.

Parlour, C.A. 1998. Price dynamics in limit order markets. *Review of Financial Studies* 11:789-816.

Weild, D., E. Kim, and L. Newport. 2012. The trouble with small tick sizes. Grant Thornton.

Yao, C., and Ye, M. (2018). Why trading speed matters: A tale of queue rationing under price controls. *The Review of Financial Studies*, *31*(6), 2157-2183.

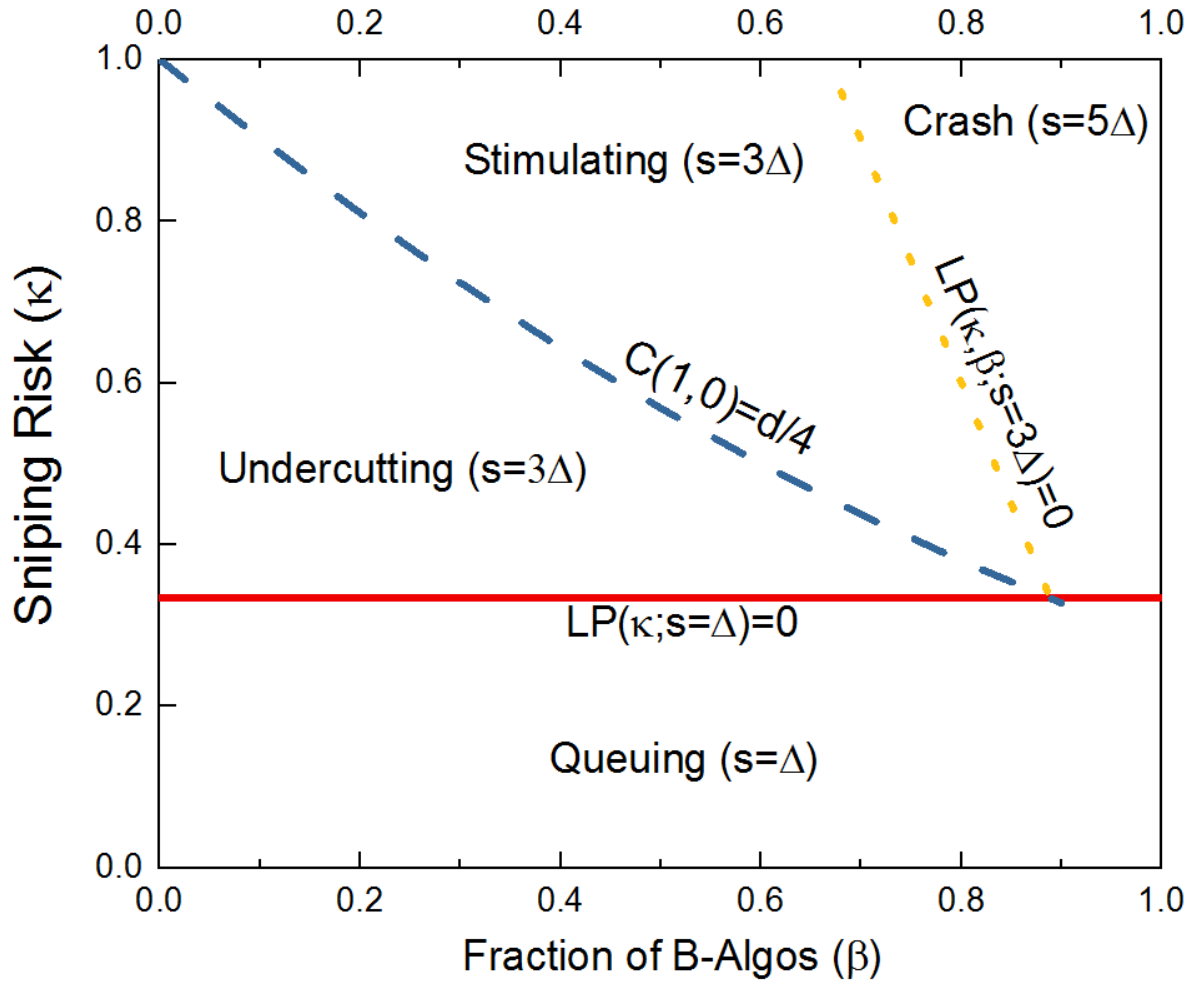

Figure 1: Four Types of Equilibrium

This figure demonstrates four types of equilibrium depending on $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ and β, given $d = 2\Delta$. When $\kappa \leq \frac{1}{3}$, HFTs queue at the one-tick bid–ask spread at $v_t \pm \frac{d}{4}$ (Proposition 3). When $\frac{1}{3} < \kappa \leq 3(1-\beta)$, HFTs quote the three-tick bid–ask spread at $v_t \pm \frac{3d}{4}$, and B-Algo buyers can choose to submit limit orders at either $v_t + \frac{d}{4}$ or $v_t - \frac{d}{4}$. B-Algos use stimulating buy (sell) limit orders at $v_t + \frac{d}{4}$ ($v_t - \frac{d}{4}$) when the sniping risk is relatively high ($\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa \leq 3(1-\beta)$), and HFTs immediately take liquidity from B-Algos (Proposition 4). B-Algos choose to undercut HFTs when the sniping risk is relatively low ($\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$), and they wait to provide liquidity to humans (Proposition 5). When the sniping risk is very high ($\kappa > 3(1-\beta)$), the liquidity provision profit from the three-tick spread $LP(\kappa, \beta; s = 3\Delta)$ is negative, and HFTs quote at $v_t \pm \frac{5d}{4}$ (Proposition 6). Boundary conditions are defined in the propositions.

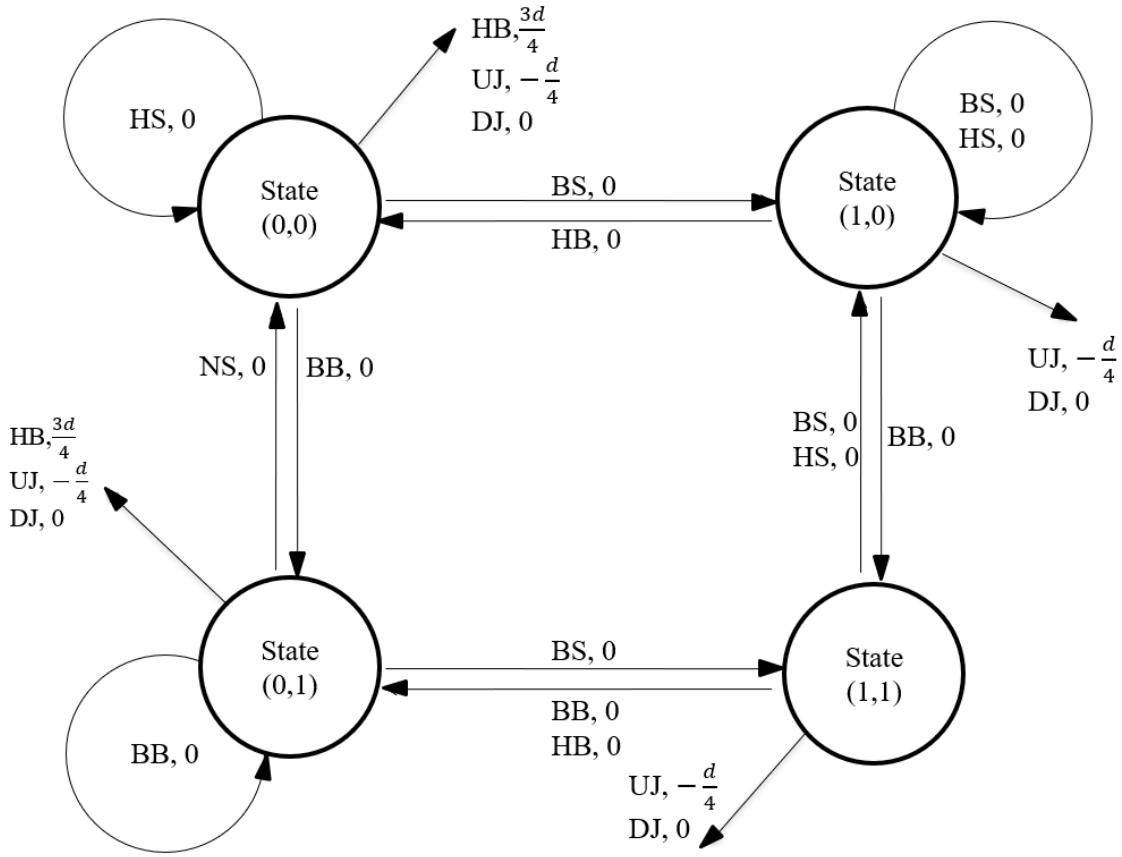

Figure 2. States and Payoffs for the HFT Liquidity Supplier on the Ask-Side

This figure illustrates the transition between LOB states and payoffs from the point of view of HFT liquidity providers on the ask side. In undercutting equilibrium, HFTs quote $v_t \pm \frac{3d}{4}$ and B-Algos can submit undercutting orders at $v_t \pm \frac{d}{4}$. In state (i, j), the number of undercutting B-Algo sell orders at $v_t + \frac{d}{4}$ is i, and the number of buy orders at $v_t - \frac{d}{4}$ is j. BB and BS represent the arrival of B-Algos' buy and sell limit orders, HB and HS represent the arrival of human traders' buy and sell market orders, and UJ and DJ denote upward and downward value jumps. The arrows between states represent state transitions, while arrows pointing toward the outside represent either order executions or cancellations. The number next to each event is the payoff of the event.

Appendix. Proofs

Proof of Proposition 1

We verify that the strategies described in Proposition 1 for non-HFTs and HFTs are their best responses:

First, it is optimal for non-HFTs to trade immediately upon arrival. Although we do not impose a delay cost on non-HFTs, there is no benefit for non-HFTs who delay trades because the bid–ask spread s_1^* is a constant and v_t is martingale.

Second, no HFT would deviate from the quoted bid–ask spread at $v_t \pm \frac{s_1^*}{2}$:

1. Any HFT who crosses the midpoint (sells below v_t or buys above v_t) always loses money instantly.
2. Liquidity provider(s) and snipers earn the same expected profit for each share in the LOB. Any HFT who narrows the bid–ask spread will (1) earn less than the original liquidity provider when she executed with a non-HFT, and (2) lose more than the original liquidity provider when being sniped during a value jump. Thus, there is no profitable deviation strategy for HFTs to narrow the spread.
3. Any HFT who quotes at $v_t \pm \frac{s_1^*}{2}$ after an existing limit order will be less likely to trade with a non-HFT because the second share has less execution priority. She has to wait longer in expectation and is more likely to be sniped. Thus, the liquidity provision revenue from the second share is lower than the sniping profits from the second share. All HFTs prefer to be snipers for the second share; no HFT is willing to submit the second share at $v_t \pm \frac{s_1^*}{2}$.
4. No HFT who quotes a spread wider than $v_t \pm \frac{s_1^*}{2}$ but within $v_t \pm d$ can trade with a non-

HFT, because each non-HFT trades only one share, and other HFTs will refill the liquidity-provision share after it has been consumed by a non-HFT. Thus, liquidity-provision revenue is negative if one quotes a half-spread that is between $\left(\frac{s_1^*}{2}, d\right)$.

5. Quoting outside $v_t \pm d$, though, is possible because we have restricted our value jump size to d. It is also possible, in the analysis of BCS, that HFTs can submit "orders that trade with probability zero." To simplify the state space of our model, we assumed in the main text that no traders can submit limit orders far away from the book. Even if we were to allow HFTs to quote far away from the market, the equilibrium bid–ask spread and transaction costs for B-Algos both remain the same. ■

Proof of Proposition 2

The difference between Proposition 2 and Proposition 1 is that a fraction β of non-HFT, buy-side algorithmic traders (B-Algos) can use limit orders to minimize their transaction costs.

First, the equilibrium bid–ask spread s_2^* is given by:

$$\frac{(1-\beta)\lambda_I}{(1-\beta)\lambda_I+2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{(1-\beta)\lambda_I+2\lambda_J}\frac{N-1}{N}\left(d-\frac{s}{2}\right) = \frac{\lambda_J}{(1-\beta)\lambda_I+2\lambda_J}\frac{1}{N}\left(d-\frac{s}{2}\right) \qquad \text{(A.1)}$$

The left-hand side of A.1 is the HFT's liquidity-provision profit, and the right-hand side is a sniper's profit. Compared with the equations (2) and (3), the only difference is the factor $(1-\beta)$, reflecting that now only humans, with an arrival rate of $(1-\beta)\lambda_I$, take liquidity from HFTs.

Secondly, submitting limit orders at v_t (stimulating orders) and facing zero transaction costs is the best outcome for B-Algos. All other execution strategies would lead to positive transaction costs. B-Algos who cross the midpoint always incur an instant positive transaction cost. B-Algos who

narrow the $v_t \pm \frac{s_2^*}{2}$ bid–ask spread by posting sell limit orders at $v_t + \frac{s}{2}$ or buy limit orders at $v_t - \frac{s}{2}$, where $s < s_2^*$, have expected transaction costs:

$$C(\kappa, \tfrac{s}{2}) = -LP_{B-Algos}(\kappa, \tfrac{s}{2}) = -[\tfrac{(1-\beta)}{(1-\beta)+2\kappa}\tfrac{s}{2} - \tfrac{\kappa}{(1-\beta)+2\kappa}\left(d - \tfrac{s}{2}\right)] \qquad \text{(A.2)}$$

where $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ and we denote a B-Algo's liquidity-provision revenue at spread s as $LP_{B-Algos}\left(\kappa, \frac{s}{2}\right)$. When B-Algos provide liquidity and earn positive liquidity-provision revenue, they incur negative transaction costs to execute their trades; thus, we have $C\left(\kappa, \frac{s}{2}\right) = -LP_{B-Algos}(\kappa, \frac{s}{2})$. Comparing a B-Algo's liquidity-provision revenue with an HFT's liquidity-provision revue (the left-hand side of A.1) demonstrates the following difference: during value jumps B-Algos are sniped at a probability of one because they are slower than HFTs. $C(\kappa, \frac{s}{2})$ is monotonically decreasing in s, and $C\left(\kappa, \frac{s}{2}\right) = 0$ when $s = s_2^*$ (from A.1). Thus, we have $C > 0$ when $s < s_2^*$. Moreover, $C\left(\kappa, \frac{s_2^*}{2}\right) = 0$ means B-Algos have zero transaction costs if they provide the first unit at $v_t \pm \frac{s_2^*}{2}$. Therefore, they will incur positive transaction costs if they add orders at $v_t \pm \frac{s_2^*}{2}$ after the existing limit order because they do not have execution priority and face a higher sniping risk.[18] B-Algos who quote limit orders wider than $v_t \pm \frac{s_2^*}{2}$ can trade only with snipers, because each non-HFT trades only one share, and other HFTs will refill the liquidity-provision share after the share has been consumed by a non-HFT. A quote outside $v_t \pm d$ is ruled out by assumption.[19]

[18] When pricing is discrete, our Assumption 1 requires all limit orders to be price-improving, which is not a binding constraint here.

[19] Without the assumption, B-Algos still cannot quote outside $v_t \pm d$ because HFTs will undercut B-Algos' quotes

Third, HFTs who accept a B-Algo's order at v_t receive zero payoffs. No HFT can receive a payoff from a B-Algo that is greater than zero by deviating to the strategy in virtue of which she does not accept the B-Algo's order, because other HFTs immediately accept the B-Algo's order. Thus, the deviator cannot extract a sniping profit from the B-Algo's order if she does not attempt to take the B-Algo's order immediately. Also, for the same reason in Proposition 1, no HFT can earn a greater payoff than $LP\left(\kappa, \frac{s_2^*}{2}\right) = SN\left(\kappa, \frac{s_2^*}{2}\right)$ on the shares quoted by HFT(s) at $v_t \pm \frac{s_2^*}{2}$.

To summarize, no market participant can receive a higher payoff by deviating from the strategy defined in Proposition 2. Thus, Proposition 2 is an equilibrium. ■

Proof of Corollary 1

All the results follow directly by taking the derivative of s_2^* and $\bar{C}(\beta)$ with respect to β:

$$\frac{ds_2^*}{d\beta} = \frac{2\lambda_I\lambda_J}{((1-\beta)\lambda_I+\lambda_J)^2} d > 0 \tag{A.3}$$

$$\bar{C}(\beta) = \beta \cdot 0 + (1-\beta) \cdot \frac{s_2^*}{2} = \frac{(1-\beta)\lambda_J}{(1-\beta)\lambda_I+\lambda_J} d \tag{A.4}$$

$$\frac{d\bar{C}(\beta)}{d\beta} = \frac{-\lambda_J^2}{((1-\beta)\lambda_I+\lambda_J)^2} d < 0 \tag{A.5}$$

Thus, the quoted spread s_2^* increases in β and the average transaction cost $\bar{C}(\beta)$ decreases in β.■

when the market moves to B-Algos' limit orders. In particular, HFTs undercut B-Algos if they observe $-C(\lambda_I, \lambda_J, \frac{s}{2}) = LP_{B-Algos}\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = \frac{(1-\beta)\lambda_I}{(1-\beta)\lambda_I+2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{(1-\beta)\lambda_I+2\lambda_J}\left(d - \frac{s}{2}\right) > 0$ at any time t. In other words, HFTs allow B-Algos to be at the top of the LOB only when $-C(\lambda_I, \lambda_J, \frac{s}{2}) \leq 0 \Leftrightarrow C(\lambda_I, \lambda_J, \frac{s}{2}) \geq 0$. Therefore, there is no way that B-Algos can attain negative transaction costs.

Proof of Proposition 3

HFTs provide liquidity at $v_t \pm \frac{d}{4}$ if:

$$LP\left(\kappa, \frac{d}{4}\right) = \frac{1}{2\kappa+1}\frac{d}{4} - \frac{\kappa}{2\kappa+1}\frac{3d}{4} \geq 0, \tag{A.6}$$

which is equivalent, as $\kappa \leq \frac{1}{3}$. No HFT wants to cancel her order and give up $LP\left(\kappa, \frac{d}{4}\right) \geq 0 = SN\left(\kappa, \frac{d}{4}\right)$, except when a fundamental value jump occurs. No HFT can quote a spread narrower than $\frac{d}{2}$ because of the tick-size constraint. No HFT wants to quote a spread wider than $\frac{d}{2}$ because she can never trade with non-HFTs while she still faces sniping risks. All HFTs want to snipe a stale quote during value jumps; otherwise, the quote would be immediately sniped by other HFTs. Thus, Proposition 3 is an equilibrium. ■

Proof of Proposition 4

We present the proof in the following three parts. First, we calculate the bid–ask spread quoted by HFTs. Second, we calculate the boundary at which B-Algos are indifferent between using stimulating orders (Proposition 4) and undercutting limit orders (Proposition 5). Third, we check for the off-equilibrium path and formally pin down the subgame perfect equilibrium.

First, we show that HFTs quote at $v_t \pm \frac{3d}{4}$ when $\frac{1}{3} < \kappa \leq 3(1-\beta)$. From A.6, when $\kappa > \frac{1}{3}$, we have $LP\left(\kappa, \beta, \frac{d}{4}\right) < 0$ and the liquidity-provision profit also depends on β. This implies that when $\kappa > \frac{1}{3}$, HFTs lose money if they provide liquidity at $v_t \pm \frac{d}{4}$ even if all non-HFTs take liquidity from HFTs. Therefore, HFTs will widen the spread to the next available prices that are $v_t \pm \frac{3d}{4}$. We show that when $\kappa \leq 3(1-\beta)$, HFTs can make non-negative liquidity-provision

profits at $v_t \pm \frac{3d}{4}$.

Note that when HFTs provide liquidity at $v_t \pm \frac{3d}{4}$, B-Algos will never take liquidity from HFTs, because B-Algos always have a better option: buy at $v_t + \frac{d}{4}$ or sell at $v_t - \frac{d}{4}$. Those orders will be immediately taken by HFTs, as in Proposition 2, and B-Algos will incur transaction cost of $\frac{d}{4}$, which is lower than the transaction cost of taking liquidity from HFTs ($\frac{3d}{4}$). Therefore, if HFTs provide liquidity at $v_t \pm \frac{3d}{4}$, only humans will take HFTs' orders. When quoting at $v_t \pm \frac{3d}{4}$ and there is no B-Algo undercutting the order, the HFT seller who provides the first share of liquidity at $v_t + \frac{3d}{4}$ reaps the following expected profit:

$$LP\left(\kappa,\beta,\tfrac{3d}{4}\right) = \tfrac{1-\beta}{2\kappa+2}\tfrac{3d}{4} + \tfrac{1-\beta}{2\kappa+2}LP\left(\kappa,\beta,\tfrac{3d}{4}\right) + \tfrac{\beta}{2\kappa+2}\cdot LP\left(\kappa,\beta,\tfrac{3d}{4}\right) + \tfrac{\beta}{2\kappa+2}\cdot LP\left(\kappa,\beta,\tfrac{3d}{4}\right) - \tfrac{\kappa}{2\kappa+2}\tfrac{d}{4} + \tfrac{\kappa}{2\kappa+2}\cdot 0 \quad \text{(A.7)}$$

The right-hand side terms are: A human buyer arrives and trades with the HFT seller; a human seller arrives on the contra side and the LOB does not change; a B-Algo buyer arrives and uses a stimulating order,[20] which does not change the LOB state; a B-Algo seller arrives and uses a stimulating order, which does not change the LOB state; an upward jump occurs and loses $\frac{d}{4}$; a downward jump occurs and cancels the order, respectively. The solution to $LP\left(\kappa,\beta,\frac{3d}{4}\right) \geq 0$ is:

$$\kappa \leq 3(1-\beta) \quad \text{(A.8)}$$

Secondly, when HFTs quote at $v_t \pm \frac{3d}{4}$, B-Algos can undercut HFTs to sell at $v_t + \frac{d}{4}$ or buy at $v_t - \frac{d}{4}$ (undercutting equilibrium) or cross the midpoint to buy at $v_t + \frac{d}{4}$ or sell at $v_t - \frac{d}{4}$ (stimulating equilibrium). B-Algos choose orders that minimize their transaction costs.

Now we determine the boundary between the stimulating equilibrium and the undercutting

[20] Later we show that, at the boundary where HFTs quote $v_t \pm \frac{3d}{4}$ and $v_t \pm \frac{5d}{4}$, i.e., $LP\left(\kappa,\beta,\frac{3d}{4}\right) = 0$ and $\kappa = 3(1-\beta)$, the sniping risk is too high for B-Algos to use undercutting orders (Short-dashed line in Figure 1).

equilibrium. In an undercutting equilibrium, a B-Algo submits a limit order to an empty LOB $(0,0)$ and changes the state to $(1,0)$; a B-Algo submits a limit order to $(0,1)$ and changes the state to $(1,1)$. We denote the B-Algo's transaction cost for the first case as $C(1,0)$ and for the second case as $C(1,1)$.[21] Then

$$\begin{cases} C(1,0) = \frac{1-\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{1-\beta}{2\kappa+2}\cdot C(1,0) + \frac{\beta}{2\kappa+2}\cdot C(1,1) + \frac{\beta}{2\kappa+2}\cdot C(1,0) + \frac{\kappa}{2\kappa+2}\cdot\frac{3d}{4} + \frac{\kappa}{2\kappa+2}\cdot C(1,0) \\ C(1,1) = \frac{1-\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{1-\beta}{2\kappa+2}\cdot C(1,0) + \frac{\beta}{2\kappa+2}\cdot\left(-\frac{d}{4}\right) + \frac{\beta}{2\kappa+2}\cdot C(1,0) + \frac{\kappa}{2\kappa+2}\cdot\frac{3d}{4} + \frac{\kappa}{2\kappa+2}\cdot C(1,0) \end{cases} \quad (A.9)$$

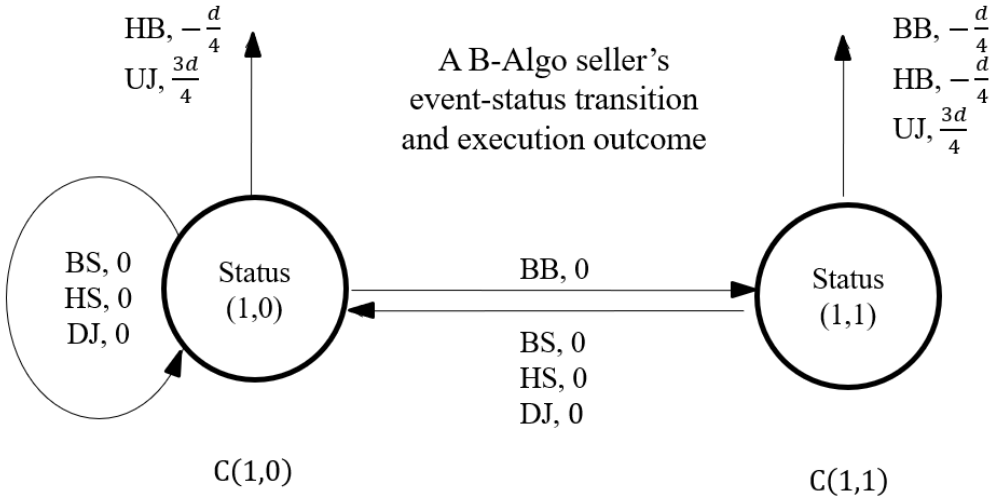

Figure A.1

In equation (A.9) and Figure A.1, we describe six event types that can change the LOB in an undercutting equilibrium. Consider $C(1,0)$ on the ask side. A human buyer and a human seller each arrive at probability $\frac{1-\beta}{2\kappa+2}$. The B-Algo seller enjoys a negative transaction cost of $-\frac{d}{4}$ when the human buyer takes his liquidity; the human seller hits an HFT's quote on the bid side and does not change the state on the ask side. A B-Algo buyer and a B-Algo seller arrive, each at

[21] Note that $C(1,j)$ is the B-Algo's cost of execution using regular limit orders at $v_t \pm \frac{d}{4}$. Formally, it is $C^{(1,j)}\left(\kappa,\beta,\frac{d}{4}\right)$. There is no $C(0,j)$ because the undercutting B-Algo herself becomes the "1." $C^{(i,j)}\left(\kappa,\beta,\frac{d}{4}\right)$ is not the same with $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)$, which is an HFT's liquidity provision profit at $v_t \pm \frac{3d}{4}$ under state (i,j).

probability $\frac{\beta}{2\kappa+2}$. A B-Algo buyer posts a limit order on the bid side and changes the state to $(1,1)$; a B-Algo seller uses a stimulating limit order, so the state remains at $(1,0)$. Upward and downward value jumps occur at probability $\frac{\kappa}{2\kappa+2}$. An upward jump leads to a sniping cost of $\frac{3d}{4}$, whereas a downward jump does not change the state of the LOB because the undercutting B-Algo updates her order accordingly. $C(1,1)$ differs from $C(1,0)$ in the sense that the arrival of a B-Algo buyer leads to the execution of a sell limit order from a B-Algo.[22]

From Proposition 3 we know that, when $\kappa > \frac{1}{3}$, HFTs suffer negative liquidity-provision profits at $v_t \pm \frac{d}{4}$. Thus, when B-Algos provide liquidity at $v_t \pm \frac{d}{4}$, they will suffer negative liquidity-provision profits as well. Therefore, $C(1,0) > 0$ and $C(1,1) > 0$, because the transaction costs in A.9 is the negative of B-Algos' liquidity-provision profits.

It is easy to see that $C(1,0) - C(1,1) = \frac{\beta}{2\kappa+2}\left(C(1,1) + \frac{d}{4}\right) > 0$, i.e., a B-Algo's undercutting order-execution cost will be lower if the contra side has another undercutting order. The solution for equation (A.9) is:

$$C(1,0) = \frac{\kappa(2\kappa + 2 + \beta)d}{(\kappa + 1)(2\kappa + 2 - \beta)} - \frac{d}{4}$$

$$C(1,1) = \frac{\kappa(2\kappa + 2)d}{(\kappa + 1)(2\kappa + 2 - \beta)} - \frac{d}{4}$$

Thus, $C(1,0) < \frac{d}{4} \Leftrightarrow \frac{\kappa(2\kappa+2+\beta)}{(\kappa+1)(2\kappa+2-\beta)} < \frac{1}{2} \Leftrightarrow 2\kappa^2 + 3\kappa\beta + \beta - 2 < 0$, where $C(1,0) < \frac{d}{4}$ is the condition for B-Algos to use regular limit orders at $v_t \pm \frac{d}{4}$ when the price level is availiable.

Equation $2\kappa^2 + 3\kappa\beta + \beta - 2 = 0$ has two roots: $\kappa_{1,2} = \frac{-3\beta \pm \sqrt{9\beta^2 - 8\beta + 16}}{4}$:

[22] The execution of this order results from Assumption 1, but the intuition that a longer queue on the bid side increases the execution probability on the ask side holds true generally (Parlour 1998).

$$\kappa_2 < 0, \kappa_1 = \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}.$$

Thus, B-Algos choose to undercut when $\kappa < \kappa_1$, because $C(1,1) < C(1,0) < \frac{\mathrm{d}}{4}$; B-Algos choose to stimulate when $\kappa \geq \kappa_1$ and trade immediately at a transaction cost of $\frac{d}{4}$.

Third, we check whether a B-Algo has an incentive to deviate when all other B-Algos use stimulating limit orders, and we construct off-equilibrium path strategies. We consider the deviator on the ask side of the LOB. When a B-Algo seller arrives, if she uses a stimulating limit order to sell at $v_t - \frac{d}{4}$, her transaction cost is $\frac{d}{4}$. Now suppose she wants to deviate and sell at $v_t + \frac{d}{4}$; in that case we denote $\tilde{C}(1,0)$ as her transaction cost. Then we have:

$$\tilde{C}(1,0) = \frac{1}{2+2\kappa}\left(-\frac{d}{4}\right) + \frac{1}{2+2\kappa}\tilde{C}(1,0) + \frac{\kappa}{2+2\kappa}\left(\frac{3d}{4}\right) + \frac{\kappa}{2+2\kappa}\tilde{C}(1,0) \qquad \text{(A.10)}$$

Similarly, the four terms on the right-hand side of A.10 indicate the transaction cost for our B-Algo seller with the arrival of a human or B-Algo buyer, a human or B-Algo seller, an upward jump, and a downward jump. Note that the difference between A.10 and $C(1,0)$ in A.9 is that a B-Algo buyer will take liquidity at $v_t + \frac{d}{4}$ in A.10, because all late-arriving B-Algos are supposed to use stimulating orders. Therefore if $\tilde{C}(1,0) \geq \frac{d}{4}$, no B-Algos will deviate from the stimulating equilibrium, because they can always use stimulating orders at transaction cost $\frac{d}{4}$. From A.10, it is easy to see that $\tilde{C}(1,0) = \frac{3\kappa-1}{\kappa+1}\frac{d}{4}$. Thus,

$$\tilde{C}(1,0) \geq \frac{d}{4} \Leftrightarrow \kappa \geq 1.$$

Therefore, when $1 \leq \kappa \leq 3(1-\beta)$, all B-Algos will use stimulating limit orders, and no one has an incentive to deviate from her current strategy. Before moving forward, we summarize the results thus far:

1. When $\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, B-Algos will buy at $v_t - \frac{d}{4}$ and sell at $v_t + \frac{d}{4}$ when the price level is available (reflecting the need to satisfy the price-improving assumption). Thus, the equilibrium outcome is the undercutting equilibrium.
2. When $1 \leq \kappa \leq 3(1-\beta)$ all B-Algos will use stimulating limit orders to buy at $v_t + \frac{d}{4}$ and sell at $v_t - \frac{d}{4}$. Thus, the equilibrium outcome is the stimulating equilibrium.

Now we analyze the last case, where $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa < 1$. In this case, B-Algos' order-placement strategies cannot be as simple as they are in the two abovementioned cases. If all B-Algos use stimulating limit orders regardless of the status of the LOB, then a B-Algo has an incentive to deviate and she undercuts HFTs' quotes because when $\kappa < 1$, then $\tilde{C}(1,0) < \frac{d}{4}$, an outcome that is better than stimulating. Similarly, if all B-Algos regularly use limit orders as in the undercutting equilibrium, then when a B-Algo arrives, and if there are no other B-Algos' orders in the LOB, the arriving B-Algo has an incentive to use a stimulating order because $C(1,0) \geq \frac{d}{4}$ when $\kappa \geq \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$. Thus, it is better to use a stimulating limit order at cost $\frac{d}{4}$.

Therefore, when $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa < 1$, a B-Algo's order-placement strategy depends on the status of the LOB to be able to punish potential deviators. Specifically, the following order-placement strategy for B-Algos constructs an equilibrium:

- A B-Algo buyer will buy at $v_t - \frac{d}{4}$ only when there is an order at $v_t + \frac{d}{4}$ and no order at $v_t - \frac{d}{4}$. Otherwise, she will use a stimulating order to buy at $v_t + \frac{d}{4}$;
- A B-Algo seller will sell at $v_t + \frac{d}{4}$ only when there is an order at $v_t - \frac{d}{4}$ and no order at $v_t + \frac{d}{4}$. Otherwise, she will use a stimulating order to sell at $v_t - \frac{d}{4}$.

Intuitively, under this strategy, all B-Algos use stimulating orders to trade. But if they find a B-Algo's limit order on the contra of the LOB (deviator), they will switch to a regular limit order. For instance, when a B-Algo seller arrives and there is a buy limit order at $v_t - \frac{d}{4}$, then the B-Algo seller will submit a limit order to sell at $v_t + \frac{d}{4}$. Now we show why this is optimal for the B-Algo seller. Denote $\tilde{C}(1,1)$ as the B-Algo seller's transaction cost; then:

$$\tilde{C}(1,1) = \frac{1-\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{1-\beta}{2\kappa+2}\cdot\frac{d}{4} + \frac{\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{\beta}{2\kappa+2}\cdot\frac{d}{4} + \frac{\kappa}{2\kappa+2}\cdot\frac{3d}{4} + \frac{\kappa}{2\kappa+2}\cdot\frac{d}{4} \quad \text{(A.11)}$$

Note that the difference between A.11 and $C(1,1)$ in A.9 is that whenever a human or B-Algo seller takes an order at $v_t - \frac{d}{4}$, according to the above strategy, our B-Algo seller (punisher) will immediately cancel her sell order at $v_t - \frac{d}{4}$ and use a stimulating order to complete her trade. In A.9, though, the B-Algo seller keeps her order at $v_t + \frac{d}{4}$ and, thus, the state transits to $C(1,0)$. From A.11 we have:

$$\tilde{C}(1,1) = \frac{\kappa}{2\kappa+2}d$$

All B-Algos have incentives to follow the above strategy only when $\tilde{C}(1,1) < \frac{d}{4} \Leftrightarrow \kappa < 1$, because B-Algos use only regular limit orders when the expected transaction cost is below $\frac{d}{4}$, the cost of using stimulating orders. Therefore, when $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa < 1$, the above strategy for B-Algos defines an equilibrium, because all B-Algos have an incentive to follow the strategy. In other words, B-Algos use regular limit orders only when the state is $(1,1)$, i.e., when $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa < 1$. Because the first-arriving B-Algo uses stimulating orders, all late-arriving B-Algos use stimulating orders to trade. As a result, the equilibrium outcome is still stimulating equilibrium when $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} \leq \kappa < 1$. This completes our proof. ■

Proof of Proposition 5

In the proof of Proposition 4, we have shown in (A.7) – (A.8) that, when the undercutting order is absent, $LP^{(0,j)}\left(\kappa,\beta,\frac{3d}{4}\right)\geq 0$ when $\kappa\leq 3(1-\beta)$. Also, we solved the undercutting equilibrium regime $\frac{1}{3}<\kappa<\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$ in virtue of which a B-Algo's order-placement strategy is to use regular limit orders whenever possible. Possible deviations by B-Algos, as well as the deviation-punishers' strategy, have also been discussed in the proof of Proposition 4. We need only to determine $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)$ and whether an HFT wants to supply liquidity at price $v_t\pm\frac{3d}{4}$, given the existence of an undercutting order, i.e. $LP^{(1,j)}\left(\kappa,\beta,\frac{3d}{4}\right)$.

Here we give an example for $\kappa=0.5,\beta=0.6$. We analytically solve the four linear formulas assuming $LP^{(i,j)}\left(\kappa,\beta,\frac{3d}{4}\right)>0$ without truncation, and we insert $\kappa=0.5,\beta=0.6$:[23]

$$LP^{(0,0)}\left(\kappa,\beta,\frac{3d}{4}\right)=0.0711$$

$$LP^{(1,0)}\left(\kappa,\beta,\frac{3d}{4}\right)=-0.0591$$

$$LP^{(0,1)}\left(\kappa,\beta,\frac{3d}{4}\right)=0.0757$$

$$LP^{(1,1)}\left(\kappa,\beta,\frac{3d}{4}\right)=-0.0361$$

Now, $LP^{(1,j)}\left(\kappa,\beta,\frac{3d}{4}\right)<0$ and liquidity provision is profitable only when there is no undercutting order, i.e., the state $(0,j)$. The HFT supplying liquidity in state $(0,j)$ will cancel her order when a

[23] The solution has tens of terms in its denominator, because the denominator is the determinant of a 4 by 4 matrix with p_1,p_2,p_3 as its elements. We have solved it analytically, and we believe both HFTs and B-Algos have the ability to solve it numerically at a minimum.

B-Algo arrives and undercuts her. In other words, the truncation is in effect, and we solve the following equations instead,[24] where $\mathbf{0} = \overline{LP}^{(1,j)}\left(\kappa, \beta, \frac{3d}{4}\right)$:

$$\begin{cases} LP^{(0,0)}\left(\kappa, \beta, \frac{3d}{4}\right) = p_1 LP^{(0,1)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_1 \cdot \mathbf{0} + p_2 LP^{(0,0)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_2 LP^{(0,0)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0 \\ LP^{(0,1)}\left(\kappa, \beta, \frac{3d}{4}\right) = p_1 LP^{(0,1)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_1 \cdot \mathbf{0} + p_2 LP^{(0,1)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_2 LP^{(0,0)}\left(\kappa, \beta, \frac{3d}{4}\right) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0 \end{cases} \quad \text{(A.12)}$$

We have:

$$LP^{(0,0)}\left(\kappa, \beta, \frac{3d}{4}\right) = LP^{(0,1)}\left(\kappa, \beta, \frac{3d}{4}\right) = 0.0875 > 0$$

Neither supplying liquidity when $LP^{(i,j)}\left(\kappa, \beta, \frac{3d}{4}\right) < 0$ nor cancelling the limit order when $LP^{(i,j)}\left(\kappa, \beta, \frac{3d}{4}\right) > 0$ is a profitable deviation. ■

Proof of Proposition 6

We have shown in the proof of Proposition 4 that HFTs quote $v_t \pm \frac{5d}{4}$ when $\kappa > 3(1-\beta)$, because $v_t \pm \frac{3d}{4}$ is no longer profitable for liquidity provision. We will construct the equilibrium as well as the off-equilibrium path strategies in the spirit of Proposition 4.

Without loss of generality, consider the B-Algo seller's problem. A B-Algo seller can submit an order at $v_t + \frac{3d}{4}, v_t + \frac{d}{4}, v_t - \frac{d}{4}, v_t - \frac{3d}{4}$. It is easy to see that a B-Algo seller never sells at $v_t - \frac{3d}{4}$, because selling at $v_t - \frac{d}{4}$ (a stimulating order) strictly dominates it. Also, a B-Algo

[24] In some cases, we might have $LP^{(1,0)}\left(\kappa, \beta, \frac{3d}{4}\right) < 0$ and $LP^{(1,1)}\left(\kappa, \beta, \frac{3d}{4}\right) > 0$, which means the HFT supplying liquidity in state $(1,1)$ will cancel her order when the B-Algo on the contra side trades with a non-HFT. This is because, in the absence of the B-Algo on the contra side, the undercutting order on the same side is less likely to be consumed, and the HFT has lower expected profit in state $(1,0)$ than in state $(1,1)$.

seller would not use a stimulating order at $v_t - \frac{d}{4}$ when $v_t + \frac{3d}{4}$ is empty, because the first pays $\frac{d}{4}$, and the second pays less than $\frac{d}{4}$.[25] When $v_t + \frac{3d}{4}$ is occupied, the B-Algo should choose between $v_t + \frac{d}{4}$ and $v_t - \frac{d}{4}$, and we show a perfect equilibrium in which B-Algo sellers never sell at $v_t + \frac{d}{4}$. A B-Algo's equilibrium strategy depends on κ. Specifically:

When $\kappa \geq 1$, the order-placement strategy is as follows:

1. A B-Algo will sell at $v_t + \frac{3d}{4}$ or buy at $v_t - \frac{3d}{4}$ whenever possible (reflecting the need to satisfy the price-improving assumption).
2. A B-Algo seller will sell at $v_t - \frac{d}{4}$ when there is already a sell limit order at $v_t + \frac{3d}{4}$. Similarly, a B-Algo buyer will buy at $v_t + \frac{d}{4}$ when there is already a buy limit order at $v_t - \frac{3d}{4}$.

When $3(1-\beta) < \kappa < 1$, the strategy is as follows:

1. A B-Algo will sell at $v_t + \frac{3d}{4}$ or buy at $v_t - \frac{3d}{4}$ whenever the price level is available.
2. For B-Algo sellers: When there is already an order at $v_t + \frac{3d}{4}$, a B-Algo seller will sell at $v_t - \frac{d}{4}$ when there is no buy limit order at $v_t - \frac{d}{4}$. Otherwise (the off-equilibrium path), the B-Algo seller will sell at $v_t + \frac{d}{4}$. Symmetrically for B-Algo buyers: When there is already an order at $v_t - \frac{3d}{4}$, a B-Algo buyer will buy at $v_t + \frac{d}{4}$ when there is no

[25] If the order at $v_t + \frac{3d}{4}$ has been sniped, the B-Algo seller pays $\frac{d}{4}$; it is also possible, however, that the B-Algo seller trades with a non-HFT and incurs transaction cost $-\frac{3d}{4}$. Thus, the expected cost should be between 0 and $\frac{d}{4}$, strictly less than $\frac{d}{4}$, the cost of using stimulating orders.

sell limit order at $v_t + \frac{d}{4}$. Otherwise (the off-equilibrium path), the B-Algo buyer will buy at $v_t - \frac{d}{4}$.

We illustrate the above strategies when the B-Algo is a seller. The B-Algo seller prefers to sell at $v_t + \frac{3d}{4}$. When there is already an order at $v_t + \frac{3d}{4}$, the B-Algo seller needs to choose between a regular limit order at $v_t + \frac{d}{4}$ or a stimulating limit order at $v_t - \frac{d}{4}$. When the sniping risk is high enough ($\kappa \geq 1$), the B-Algo seller always uses a stimulating limit order at $v_t - \frac{d}{4}$. When the sniping risk is moderate ($3(1-\beta) < \kappa < 1$), the B-Algo seller will still use a stimulating order at $v_t - \frac{d}{4}$ in equilibrium. If there are buy limit orders at $v_t - \frac{d}{4}$, the B-Algo seller (a deviation punisher) will post a limit order at $v_t + \frac{d}{4}$.

Now we verify that there is no profitable deviation for B-Algos. In the first case, when $\kappa \geq 1$, and there is already a sell limit order at $v_t + \frac{3d}{4}$, if the B-Algo seller wants to sell at $v_t + \frac{d}{4}$ instead of using a stimulating limit order, we denote her transaction cost as C_1. C_1 is smaller when there is a B-Algo buyer order at $v_t - \frac{3d}{4}$, because the next arriving B-Algo buyer will use a stimulating limit order. Thus,

$$C_1 > \frac{1-\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{1-\beta}{2\kappa+2} \cdot C_1 + \frac{\beta}{2\kappa+2}\left(-\frac{d}{4}\right) + \frac{\beta}{2\kappa+2} \cdot C_1 + \frac{\kappa}{2\kappa+2} \cdot \frac{3d}{4} + \frac{\kappa}{2\kappa+2} \cdot C_1 \qquad \text{(A.13)}$$

We have:

$$C_1 > \frac{1}{\kappa+1}\frac{d}{2} \geq \frac{d}{4} \text{ when } \kappa \geq 1$$

Therefore, when using a regular limit order at $v_t + \frac{d}{4}$, the B-Algo seller incurs higher transaction costs than when using a stimulating limit order at $v_t - \frac{d}{4}$. A.13 offers similar explanations to those

offered by the first equation in A.9. There is an inequality in A.13 because, whenever a human seller arrives, a B-Algo seller arrives, or the asset's value jumps downward (the second, fourth, and sixth terms on the right-hand side of A.13), which clears the order at $v_t - \frac{3d}{4}$. The B-Algo seller who has an order at $v_t + \frac{d}{4}$ incurs higher transaction costs than C_1, because the next-arriving B-Algo buyer will submit a buy order at $v_t - \frac{3d}{4}$ but will not take an order at $v_t + \frac{d}{4}$.

We then check that the deviating B-Algo seller at $v_t + \frac{d}{4}$ is indeed losing no less than $\frac{d}{4}$ when $3(1-\beta) < \kappa < 1$. In this case, all B-Algo buyers who observe the deviator will use undercutting limit orders; thus, the deviator will incur an execution cost of $C_1 > C(1,0) > \frac{\mathrm{d}}{4}$. In this case, the counter-deviating B-Algo still realizes $\tilde{C}(1,1)$, as in Proposition 4. Also, the regular undercutting B-Algo sell order at $v_t + \frac{3d}{4}$ would not cancel, because cancelling and using a stimulating order would incur a cost of $\frac{\mathrm{d}}{4}$, while waiting always incurs a lower cost. Therefore, B-Algo sellers never jumpstart the off-equilibrium path by quoting $v_t + \frac{d}{4}$.

Finally, we check whether the deviation-punisher's strategy is subgame perfect. As in the proof of Proposition 4, the punisher receives all non-HFT stimulating sell order flows when the selling deviator is present; she can therefore pay an execution cost lower than $\frac{d}{4}$ when $3(1-\beta) < \kappa < 1$. When an upward fundamental value jump occurs, the deviator has been sniped and the punisher switches her limit buy order at $v_t - \frac{d}{4}$ to a stimulating order at $v_t + \frac{5d}{4}$ (a half-tick higher than the new fundamental value $v_t + d$). The update enables the punisher to avoid becoming a deviator herself, because stimulating is less costly than deviating, i.e., she keeps posting a limit order at $v_t - \frac{d}{4}$. Thus, the punisher's strategy is subgame perfect, as in Proposition 4, preventing

deviators from realizing lower execution costs.

HFTs also have no profitable deviation strategy, for the same reason in the previous propositions. They will lose money if they narrow the spread or cross the midpoint, and they can never profit from not satisfying the trading need of any stimulating order because other HFTs immediately trade with the stimulating order.

The off-equilibrium-path strategies we describe above, together with the equilibrium path, will generate the equilibrium outcome sketched in Proposition 6. ■

Proof of Corollary 2

When $m = 1$, the problem degenerates to Proposition 6. For $J = md$ where $m \in N^+$, the HFT liquidity provider at $v_t + \left(m - \frac{1}{4}\right) d$ will receive a profit of $\left(m - \frac{1}{4}\right) d$ from human buys at arrival intensity $\frac{1}{2}(1-\beta)\lambda_I$, and incur a loss of $\frac{d}{4}$ from upward jumps at intensity $\frac{1}{2}\lambda_J$. Then $LP(\kappa, \beta, \left(m - \frac{1}{4}\right) d) = 0$ if and only if $\frac{\kappa}{1-\beta} = 4m - 1$.

For any symmetric jump size distribution $J(\cdot)$ without a tick size constraint, HFTs set the equilibrium bid–ask spread as the smallest spread at which they break even, i.e. $LP(\kappa, \beta, \frac{s}{2}) = 0$.

$$LP(\kappa, \beta, \tfrac{s}{2}) = \tfrac{(1-\beta)}{2\kappa+1}\tfrac{s}{2} - \tfrac{\kappa \cdot Pr(|J| > \frac{s}{2})}{2\kappa+1} E\left(|J| - \tfrac{s}{2} \middle| |J| > \tfrac{s}{2}\right) \quad \text{(A.14)}$$

$\frac{(1-\beta)}{2\kappa+1}$ is the probability of a limit-order trade with a human, and $\frac{\kappa \cdot Pr(|J| > \frac{s}{2})}{2\kappa+1}$ is the probability that the order is sniped. Thus, the spread s^* at which the HFT breaks even satisfies:

$$\frac{\kappa}{1-\beta} = \frac{s/2}{Pr(|J| > \frac{s}{2}) E\left(|J| - \frac{s}{2} \middle| |J| > \frac{s}{2}\right)}$$

Right-hand side (RHS) is a continuous function and monotonically increases in s. We have $RHS(s=0)=0$ and $RHS(s=\infty)=\infty$. If $J(\cdot)$ is massless, there exist a unique s^* that makes $LP(\kappa,\beta,\frac{s^*}{2})=0$. Otherwise, $s^*=min\left\{s\middle|LP(\kappa,\beta,\frac{s}{2})\geq 0\right\}$. When $\beta\to 1$, we have $\frac{\kappa}{1-\beta}\to\infty$ and $s^*\to max|J|$.

When pricing is discrete, competitive HFTs simply choose the narrowest possible spread that guarantees $LP(\kappa,\beta,\frac{s}{2})\geq 0$. Formally speaking,

$$s_\Delta^*=\min_{\Phi\in N^+}\left\{s=(2\Phi-1)\Delta\left|\frac{\kappa}{1-\beta}\leq\frac{s/2}{Pr(|J|>\frac{s}{2})E\left(|J|-\frac{s}{2}\middle||J|>\frac{s}{2}\right)}\right.\right\}$$

where Δ is the tick size. ■