Who Supplies Liquidity, and When? *

Sida Li[1]

University of Illinois, Urbana-Champaign

Xin Wang[2]

University of Illinois, Urbana-Champaign

Mao Ye[3]

University of Illinois, Urbana-Champaign and NBER

* We thank Hengjie Ai, Shmuel Baruch, Malcolm Baker, Hank Bessembinder, Eric Budish, Thierry Foucault, Katya Malinova, Maureen O'Hara, Monika Piazzesi, Veronika Pool, Neil Pearson, Shri Santosh, Andriy Shkilko, Brian Weller, Chen Yao, Bart Yueshen, Marius Zoican, and participants at the University of Rochester, UCLA, Texas A&M University, Carlson Junior Conference at the University of Minnesota, NYU Stern Market Microstructure Conference, Colorado Front Range Finance Seminar, Bank of Canada-Laurier Market Structure conference, Telfer Annual Conference on Accounting and Finance, Wabash River Conference at the 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 (jointed with the Office of Financial Research at U.S. Department of the Treasury).

[1] Department of Finance, University of Illinois at Urbana-Champaign. Email: sidali3@illinois.edu.

[2] Department of Economics, University of Illinois at Urbana-Champaign. Email: xinwang5@illinois.edu.

[3] Department of Finance, University of Illinois at Urbana-Champaign and NBER, 340 Wohlers Hall, 1206 S. 6th Street, Champaign, IL, 61820. Email: maoye@illinois.edu. Tel: 217-244-0474.

Abstract

We incorporate discrete tick size and allow non-high-frequency traders (non-HFTs) to supply liquidity in the framework of Budish, Cramton, and Shim (2015). When adverse selection risk is low or tick size is large, the bid-ask spread is typically below one tick, and HFTs dominate liquidity supply. In other situations, non-HFTs dominate liquidity supply by undercutting HFTs, because supplying liquidity to HFTs is always less costly than demanding liquidity from HFTs. A small tick size improves liquidity, but also leads to more mini-flash crashes. The cancellation-to-trade ratio, a popular proxy for HFTs, can have a negative correlation with HFTs' activity. Our model provides one explanation of flash crashes, and predicts when and where flash crashes are more likely to occur.

In decades past, specialists on the New York Stock Exchange and dealers in NASDAQ supply liquidity to other traders, that is, they buy when other traders sell and sell when other traders buy. The transition to electronic trading not only destroy these traditional liquidity suppliers, but also blurs the definition of liquidity supply. Everyone can supply liquidity, but no one is obligated to supply liquidity. Liquidity supply simply means to post a limit order, an offer to buy or sell at a certain price. A trade occurs when another trader (a liquidity demander) accepts the terms of a posted offer. Every trader has to decide whether to supply or demand liquidity in order to complete a trade. In this paper, we examine how the contemporary trading environment of voluntary liquidity supply and demand reaches its equilibrium. Who supplies liquidity and who demands liquidity? Can voluntary liquidity supply and demand lead to systemic risk such as a flash crash? And if this is possible, what conditions lead up to it?

In this paper, we show how the equilibria in liquidity supply and demand depend on the characteristics of securities, market structures, and market conditions. Our model extends Budish, Cramton, and Shim (2015; BCS hereafter) along two dimensions. BCS include two types of traders: high-frequency traders (HFTs) and non-HFTs. In the BCS model, non-HFTs can only demand liquidity, while in our model we allow non-HFTs to provide liquidity. In addition, BCS consider continuous price, whereas we consider discrete price to reflect the tick size (minimum price variation) imposed by the U.S. Security and Exchange Commission's (SEC's) Regulation National Market Systems (Reg NMS) Rule 612, and to reflect the recent policy debate to increase the tick size from one cent to five cents.

Our model includes one security, and its fundamental value is public information. However, liquidity suppliers in our model are subject to adverse selection risk, because they may fail to cancel stale quotes during value jumps. HFTs in our model have no private value to trade. They

consistently monitor the market for profit opportunities. For example, they supply liquidity when its expected profit is above 0, or snipe stale quotes after value jumps. Non-HFTs arrive at the market with private value to buy or sell one unit of a security. We allow a fraction of non-HFTs to choose between providing or demanding liquidity. We call these non-HFTs buy-side algorithmic traders (BATs) to represent algorithms used by buy-side institutions (e.g., mutual funds and pension funds) to minimize the cost of executing trades in portfolio transition (Hasbrouck and Saar, 2013; Frazzini, Israel, and Moskowitz, 2014). BATs are major players in modern financial markets (O'Hara, 2015). We build the first theoretical model to study their trading behavior. Our model captures two main features of BATs. First, BATs are slower than HFTs (O'Hara, 2015). Second, BATs supply liquidity to minimize the transaction costs of portfolio rebalancing (Hasbrouck and Saar, 2013), not to profit from the bid-ask spread. As both BATs and HFTs are algorithmic traders (Hasbrouck and Saar, 2013), we call the fraction of non-HFTs who are not BATs non-algorithmic traders (non-algos).

As in BCS, the adverse selection risk increases with the arrival rate of value jumps and decreases with the arrival rate of non-HFTs. Supplying liquidity to non-HFTs leads to revenue, but value jumps lead to sniping cost. With continuous price in BCS, the competitive bid-ask spread strictly increase with adverse selection risk. In our model, tick size constrains price competition in bid-ask spread. When the adverse selection risk is low or when tick size is large, the competitive bid-ask spread can be lower than one tick, which generate rents for liquidity supply. The rents are typically allocated to HFTs, because most U.S. stock exchanges use time to decide execution priority for orders quoted at identical prices. The market thus reaches equilibrium through queuing, not through price competition. In our first type of equilibrium, the queuing equilibrium, in which bid-ask spread is binding at one tick, HFTs dominate liquidity supply due to their speed advantage

over BATs.

When tick size is not binding, we find that BATs never demand liquidity from HFTs. Instead, they choose to provide liquidity at more aggressive price than HFTs. This result is surprising because Han, Khapko, and Kyle (2014), Hoffmann (2014), Bernales (2016), and Bongaerts and Van Achter (2016) maintain that HFTs cancel stale quotes faster, incur lower adverse selection cost, and quote more aggressive prices than other traders. Brogaard et al. (2015), however, show that non-HFTs quote a tighter bid-ask spread than HFTs. Our model reconciles the contraction between previous channels of speed competition and the empirical results by including the opportunity cost of liquidity supply. BATs have to trade in our model. The outside option for BATs is to demand liquidity and pay the bid-ask spread. We find that for BATs, supplying liquidity at a tighter bid-ask spread strictly dominate demanding liquidity from HFTs.

To show why BATs always choose to supply liquidity, we develop a new concept: the make-take spread. Without loss of generality, consider BATs' decision to buy and HFTs' decision to sell. HFTs quote an ask price higher than the fundamental value, and their difference, or the half bid-ask spread, reflects the compensation for adverse selection costs during value jumps. BATs pay the half bid-ask spread if they demand liquidity. BATs can reduce their transaction costs by supplying liquidity slightly above the fundamental value. We call this type of limit order a flash limit order, because it immediately triggers HFTs to demand liquidity. Flash limit orders execute immediately like market orders, but with a lower transaction cost. Flash limit orders exploit the make-take spread, the price difference between HFTs' willingness to make an offer and their willingness to accept one. HFTs accept a lower sell price when they demand liquidity, because when they immediately accept an order, they do not incur adverse selection costs during a value jump.

When tick size does not impose a constraint for BATs to quote more aggressive prices than HFTs, our model has two types of equilibria: flash and undercutting. In the flash equilibrium, BATs use flash limit orders to supply liquidity to HFTs. In the undercutting equilibrium, BATs quote a buy limit order price below the fundamental value or a sell limit order price above the fundamental value. These regular limit orders stay in the LOB to supply liquidity to non-algos or other BATs. We find that undercutting equilibrium are more likely to occur when the adverse selection risk is low, because flash limit orders incur no adverse selection cost, whereas the cost of regular limit orders increase in adverse selection risk.

We also examine mini-flash crashes, which are sharp price movements in one direction followed by quick reversion (Biais and Foucault, 2014), and predict their cross-sectional and time series patterns. In cross-section, mini-flash crashes are more likely to occur for stocks with a smaller tick size or higher adverse selection risk. Because BATs are able to undercut HFTs for these stocks, HFTs' limit orders face lower execution probability before value jumps. When the fraction of BATs is large enough, HFTs have to quote stub quotes, a bid-ask spread wider than the maximum value of the jump, to protect against sniping. Yet BATs do not always supply liquidity on both sides of the market. Thus, it is possible for incoming market orders to hit HFTs' stub quotes, which causes a mini-flash crash. In time series, a downward (upward) mini-flash crash is more likely to occur immediately after a downward (upward) price jump, because such jumps can snipe all BATs' limit orders on the bid (ask) side and increase the probability for market orders to hit stub quotes before BATs refill the limit order book (LOB).

Existing literature on HFTs focuses on the role of adverse selection. On the one hand, speed can allow HFTs to adversely select other traders, which has a detrimental effect on liquidity; on the other hand, speed can reduce adverse selection costs for liquidity suppliers and improve

liquidity [see Jones (2013), Biais and Foucault (2014), and Menkveld (2016) for surveys]. We contribute to the literature by identifying two new channels of speed competition, both of which are unrelated to adverse selection. For liquidity demand, we find that HFTs race to demand liquidity when BATs post flash limit orders, but HFTs impose no adverse selection cost to BATs. Instead, BATs prompt HFTs to demand liquidity to reduce their transaction costs. Thus, liquidity demand from HFTs may not necessarily be bad. Instead, the transaction costs are lower when HFTs demand liquidity than when they supply liquidity.

For liquidity supply, our queuing channel of speed competition rationalizes three contradictions between empirical evidence and channels focusing on adverse selection. If speed advantage predominantly helps HFTs to reduce adverse selection costs, HFTs should realize a comparative advantage in providing liquidity for stocks with higher adverse selection costs (Han, Khapko, and Kyle, 2014; Hoffmann, 2014; Bernales, 2016; Bongaerts and Van Achter, 2016). HFTs should also crowd out slow liquidity suppliers when tick size is smaller, because a smaller tick size reduces the constraints to offer better prices (Chordia et al., 2013). In addition, a higher cancellation-to-trade ratio likely indicates more liquidity supply from HFTs, because HFTs need to cancel lots of orders to avoid adverse selection risk [see Biais and Foucault (2014) and Menkveld (2016) for a survey]. Yet Jiang, Lo, and Valente (2014) and Yao and Ye (2017) show that non-HFTs dominate liquidity supply when adverse selection risk is high. O'Hara, Saar and Zhong and Yao and Ye (2017) show that a smaller tick size crowds out HFTs' liquidity supply. Yao and Ye (2017) show stocks with higher fractions of liquidity provided by HFTs have lower cancellation-to-trade ratios. The queuing channel of speed competition reconciles these three contradictions. Tick size is more likely to be binding when adverse selection risk is low or tick size is large. A binding tick size helps HFTs to establish time priority. HFTs dominate liquidity supply for stocks

with larger tick sizes, but they also have less incentive to cancel orders. A smaller tick size or higher adverse selection risk allows BATs to increase liquidity provision by establiShimg price priority, but smaller tick size or higher adverse selection risk also leads to more frequent order cancellations. This theoretical intuition, along with the empirical evidence in Yao and Ye (2017), suggests that the cancellation-to-trade ratio should not be used as a cross-sectional proxy for HFT activities.[4]

Our model casts doubt on the recent policy proposal in the U.S. to increase the tick size, initiated by the 2012 Jumpstart Our Business Startups Act (the JOBS Act). In October 2016, the SEC started a two-year pilot program to increase the tick size from one cent to five cents for 1,200 less liquid stocks. Proponents to increase the tick size assert that a larger tick size should control the growth of HFTs and increase liquidity (Weild, Kim, and Newport, 2012). We find that an increase in tick size would *encourage* HFTs. We also find that an increase in tick size constrains price competition and reduces liquidity. A larger tick size may reduce mini-flash crashes, or very high volatility in liquidity, but such a reduction decreases liquidity in normal times. We argue that a more effective way to reduce a mini-flash crash is a trading halt after value jumps so that liquidity supply from BATs can resume.

1. Model

In our model, the stock exchange operates as a continuous LOB. Each trade in the LOB requires a liquidity supplier and a liquidity demander. The liquidity supplier submits a limit order, which is an offer to buy or sell at a specified price and quality. The liquidity demander accepts the conditions of a limit order. Execution precedence for liquidity suppliers follows the price-time

[4] The cancellation-to-trade ratio can still be a good *time series* proxy for HFTs' activity (Hendershott, Jones, and Menkveld, 2011; Angel, Harris, and Spatt, 2015; Boehmer, Fong, and Wu, 2015).

priority rule. 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.

Our model has one security, x, whose fundamental value, v_t, evolves as a compound Poisson jump process with arrival rate λ_J. v_t starts from 0, and changes by a size of d or $-d$ in each jump with equal probability. As in BCS, v_t is common knowledge, but liquidity suppliers are subject to adverse selection risk when they fail to update stale quotes after value jumps. Traders start with a small latency to observe the common value jump, [5] but can reduce the latency to 0 by investing in a speed technology with cost c_{speed} per unit of time.

Our model includes HFTs and two types of non-HFTs: BATs and non-algo traders. HFTs place no private value on trading. They supply or demand liquidity as long as the expected profit is above 0. They submit a market order to buy (sell) x when its price is below (above) v_t. HFTs supply liquidity as long as the expected profit from the bid-ask spread is above 0. Non-HFTs, who arrive with a compound Poisson jump process with intensity λ_I, have to buy or sell one unit of x, each with probability $\frac{1}{2}$. Non-HFTs do not invest in speed technology because they only arrive at the market once.

Our model extends BCS along two dimensions. First, non-HFTs in the BCS model submit only market orders. In our model, we allow a proportion β of non-HFTs, BATs, to choose between limit and market orders to minimize transaction costs. The rest of the non-HFTs, non-algo traders, use only market orders. Second, BCS assume continuous pricing in their model, whereas we

[5] By small, we mean that no additional events, such as a trader arrival or a value jump, take place during the delay.

consider discrete pricing grids. The benchmark pricing grid in Section 2 $\left\{\ldots -\frac{3d}{2}, -\frac{d}{2}, \frac{d}{2}, \frac{3d}{2} \ldots\right\}$ has a tick size of $\Delta_0 = d$. This choice ensures that v_t is always at the midpoint of two price levels at any time. In Sections 3-6, we reduce the tick size to $\Delta_1 = \frac{d}{3}$, which creates additional price levels, such as $\frac{d}{6}$ and $-\frac{d}{6}$. Figure 1 shows the pricing grids with large and small tick sizes.

Following the dynamic LOB literature (e.g., Goettler, Parlour, and Rajan, 2005, 2009; Rosu, 2009; Colliard and Foucault, 2012), we examine the Markov perfect equilibrium, in which traders' actions condition only on state of the LOB and events at t. We assume that HFTs instantaneously build up the equilibrium LOB after any event. Under this simplification, six types of events trigger the transition of the LOB across states:

$$\begin{cases} \frac{1}{2}\beta\lambda_I & \text{BAT sells (BS)} \\ \frac{1}{2}\beta\lambda_I & \text{BAT buys (BB)} \\ \frac{1}{2}(1-\beta)\lambda_I & \text{Non-algo sells (NS)} \\ \frac{1}{2}(1-\beta)\lambda_I & \text{Non-algo buys (NB)} \\ \frac{1}{2}\lambda_J & \text{Price jumps up (UJ)} \\ \frac{1}{2}\lambda_J & \text{Price jumps down (DJ).} \end{cases} \quad (1)$$

BCS do not allow non-HFTs to supply liquidity. We extend their model by allowing BATs to submit limit orders. To convey the economic intuition in the most parsimonious way, we make a technical assumption that BATs can only submit limit orders when the price level contains no other limit orders. This assumption reduces the number of states of the LOB that we need to track. We can further relax the assumption in BCS by allowing BATs to queue for $n > 1$ shares, but such an extension only increases the number of LOB states without conveying new intuition. Non-HFTs in the BCS model never use limit orders, which can be justified by an infinitely large delay cost

(Menkveld and Zoican, 2017). Our extension effectively reduces the delay cost to allow BATs to submit limit orders.[6] The main intuition of our model stays the same as long as BATs do not queue for infinite length.

2. Benchmark: Binding at one tick under a large tick size

Our analysis starts from $\Delta_0 = d$. As in BCS, HFTs can choose to be *liquidity suppliers*, who profit from the bid-ask spread, or to be *stale-quote snipers*, who profit by demanding liquidity from stale quotes after a value jump. In BCS, the equilibrium bid-ask spread equalizes the HFTs' expected profits from these two strategies, which are both zero after speed investment. Lemma 1 shows that this break-even bid-ask spread is smaller than the tick size when adverse selection risk is low.

Lemma 1 (Binding Tick Size). When $\Delta_0 = d$ and $\frac{\lambda_I}{\lambda_J} > 1$, HFTs' profit from providing the first share at the ask price of $a_t^* = v_t + \frac{d}{2}$ and the bid price of $b_t^* = v_t - \frac{d}{2}$ is higher than HFTs' profit from stale-quote sniping.

Because non-HFTs trade for liquidity reasons and value jumps lead to sniping cost for stale quotes, $\frac{\lambda_I}{\lambda_J}$ measures adverse selection risk in our model. As in BCS and Menkveld and Zoican (2017), this adverse selection risk comes from the speed of the response to public information, not from exogenous information asymmetry (e.g., Glosten and Milgrom, 1985; Kyle, 1985). As the

[6] We can assume a finite delay cost so that BATs only queue for one share, and the results are available upon request. The value of the delay cost, however, conveys no intuition and only leads to a more complicated proof. In Section 4, we show that the exact size of the delay cost has little impact for BATs' choice between limit orders and market orders.

arrival rate of non-HFTs increases or the intensity of value jumps decreases, the adverse selection risk decreases and so does the break-even bid-ask spread. The break-even bid-ask spread drops below one tick when $\frac{\lambda_I}{\lambda_J} > 1$, making liquidity supply for the first share more profitable than stale-quote sniping.[7] The rents for liquidity supply then trigger the race to win time priority in the queue. As BATs do not have a speed advantage to win the race, they demand liquidity in the same manner as non-algo traders. As a result, Lemma 1 does not depend on β.[8]

Under a binding tick size, price competition cannot lead to economic equilibrium. It is the queue that restores the economic equilibrium. Next, we derive the equilibrium queue length for the ask side of the LOB, and the bid side follows symmetrically.

We evaluate HFTs' value of liquidity supply and stale-quote sniping for each queue position, though we allow an HFT to supply liquidity at multiple positions and to snipe shares in other positions where she is not a liquidity supplier. We denote the value of liquidity supply for the Q^{th} share as $LP(Q)$. A market sell order does not affect $LP(Q)$ on the ask side, because HFTs immediately restore the previous state of the LOB by refilling the bid side. A market buy order moves the queue forward by one unit, thereby changing the value to $LP(Q-1)$. A limit order execution leads to a profit of $\frac{d}{2}$ to the liquidity supplier, $LP(0) = \frac{d}{2}$. When v_t jumps upward, the liquidity providing HFT of the Q^{th} share races to cancel the stale quote, whereas the other $N-1$ HFTs (with N determined in equilibrium) race to snipe the stale quote. The loss from being sniped

[7] Throughout this paper, we consider $\frac{\lambda_I}{\lambda_J} > 1$ for expositional simplicity. When $\frac{\lambda_I}{\lambda_J} \leq 1$, Δ_0 is no longer binding, and the equilibrium structure is similar to that in Sections 3-6, where we reduce the tick size to $\Delta_1 = \frac{d}{3}$.

[8] An order with less time priority has lower probability of execution and higher probability of being sniped, both of which reduce BATs' incentives to queue. In addition, BATs have incentives to implement trades, and a positive delay cost would compel them to use market orders when the queue is long. We assume that BATs never queue after the first position to reflect these intuitions in a parsimonious way.

is $\frac{d}{2}$, while the probability of being sniped is $\frac{N-1}{N}$. When v_t jumps downward, the liquidity supplier cancels the order and joins the race to supply liquidity at a new BBO.[9] $LP(Q)$ then becomes 0. Equation (2) presents $LP(Q)$ in recursive form and Lemma 2 presents the solution for equation (2).

$$LP(Q) = \frac{\frac{1}{2}\lambda_I}{\lambda_I+\lambda_J} LP(Q) + \frac{\frac{1}{2}\lambda_I}{\lambda_I+\lambda_J} LP(Q-1) - \frac{N-1}{N}\frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J} \times \frac{d}{2} + \frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J} \times 0. \qquad (2)$$

Lemma 2 (Value of Liquidity Supply). The value of liquidity supply for the Q^{th} position is:

$$LP(Q) = \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q \frac{d}{2} - \frac{N-1}{N}\frac{1}{2}\left[1 - \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q\right]\frac{d}{2}. \qquad (3)$$

$LP(Q)$ decreases in Q.

Intuitively, Lemma 2 reflects the conditional probability of value-change events for $LP(Q)$ and their payoffs. Since $LP(Q)$ stays the same after a market sell order, the conditional probabilities of value-changing events are $\frac{\lambda_I}{\lambda_I+2\lambda_J}$ for a market buy, $\frac{\lambda_J}{\lambda_I+2\lambda_J}$ for an upward value jump, and $\frac{\lambda_J}{\lambda_I+2\lambda_J}$ for a downward value jump. The Q^{th} share executes when Q non-HFTs arrive in a row to buy, which has a probability of $\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q$, and the revenue conditional on execution is $\frac{d}{2}$. Their product, the first term in equation (3), reflects the expected revenue for liquidity suppliers. The Q^{th} share on the ask side fails to execute with non-HFTs when an upward or downward value jump occurs, each with probability $\frac{1}{2}[1 - \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q]$. After an upward value jump, the liquidity supplier has a probability of $\frac{1}{N}$ to cancel the stale quote, but failure to cancel the stale quote before

[9] We assume that the HFT liquidity supplier cancels the limit order to avoid the complexity of tracking infinite many price levels in the LOB.

sniping leads to a loss of $\frac{d}{2}$. The expected loss is $\frac{N-1}{N}\frac{1}{2}[1-\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q]\frac{d}{2}$, the second term in equation (3). A downward value jump before the order being snipped or executed leads to a zero payoff for the liquidity supplier. $LP(Q)$ decreases in Q, because an increase in a queue position reduces execution probability and increases the cost of being sniped.

The outside option for supplying liquidity for the Q^{th} share is to be the sniper of the share during the value jump. HFTs' liquidity supply decision for the Q^{th} share also needs to include this opportunity cost. With a probability of $\frac{1}{2}[1-\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q]$, the Q^{th} share becomes stale before it gets executed, and each sniper has a probability of $\frac{1}{N}$ to profit from the stale quote. The value for each sniper of the Q^{th} share is:

$$SN(Q) = \frac{1}{N}\frac{1}{2}[1-\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q]\frac{d}{2}. \quad (4)$$

$SN(Q)$ increases with Q, because shares in a later queue position offer more opportunities for snipers.

HFTs race to supply liquidity for the Q^{th} position as long as $LP(Q) > SN(Q)$, because the winner's payoff is higher than that of the losers. Equation (5) determines the equilibrium length:

$$\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q\frac{d}{2}-\frac{1}{2}[1-\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q]\frac{d}{2} > 0. \quad (5)$$

The solution for equation (5) is:

$$Q^* = \max\left\{Q \in \mathbb{N}^+ \text{ s.t.}\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q\frac{d}{2}-\frac{1}{2}\left[1-\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q\right]\frac{d}{2} > 0\right\}$$
$$= \max\left\{Q \in \mathbb{N}^+ \text{ s.t.}\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q > \frac{1}{3}\right\}$$

$$= \left\lfloor \log_{\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)} \frac{1}{3} \right\rfloor, \tag{6}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x.

Figure 2 shows the comparative statics for equilibrium queue length. The queue length at BBO decreases with $\frac{\lambda_I}{\lambda_J}$, which indicates that, for stocks with a bid-ask spread binding at one tick, the depth at the BBO may serve as a proxy for adverse selection risk. Traditionally, bid-ask spreads serve as a proxy for adverse selection risk (Glosten and Milgrom, 1985; Stoll, 2000). Yet Yao and Ye (2017) find that bid-ask spread is one-tick wide 41% of time for their stratified sample of Russell 3000 stocks in 2010. Depth at the BBO then serves as an ideal proxy to differentiate the level of adverse selection for these stocks.[10]

To derive N, note that HFTs' total rents come from the bid-ask spread paid by non-HFTs, because sniping only redistributes the rents among HFTs. Ex ante, each HFT obtains $\frac{1}{N}$ of the rents per unit of time. New HFTs continue to enter the market until:

$$\lambda_I \frac{d}{2} - N c_{speed} \leq 0. \tag{7}$$

In Proposition 1, we summarize the equilibrium under a large binding tick size.

Proposition 1. (**Large Binding Tick Size**): When $\Delta_0 = d$ and $\frac{\lambda_I}{\lambda_J} > 1$, N^* HFTs jointly supply Q^* units of sell limit orders at $a_t^* = v_t + \frac{d}{2}$ and Q^* units of buy limit orders at $b_t^* = v_t - \frac{d}{2}$, where:

$$Q^* = \left\lfloor \log_{\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)} \frac{1}{3} \right\rfloor, \text{ and}$$

[10] Certainly, the comparison also needs to control for price, because stocks with the same nominal bid-ask spread may have a different proportional bid-ask spread.

$$N^* = \max\left\{N \in \mathbb{N}^+ \text{ s.t. } \lambda_I \frac{d}{2} - N c_{speed} > 0\right\}. \quad (8)$$

BATs and non-algo traders demand liquidity when there is a large binding tick size.

In BCS, the depth at the BBO is one share, because the first share has a competitive price. The second share at that price, which faces lower execution probability and higher adverse selection costs, is not profitable. The discrete tick size in our model raises the profit of liquidity supply above the profit of stale-quote sniping for the first share, and generates a depth of multiple shares.

In BCS, the number of HFTs is determined by $\lambda_I \frac{s^*}{2} - N c_{speed} = 0$, where s^* is the break-even bid-ask spread. In our model, N is determined by $\lambda_I \frac{d}{2} - N c_{speed} > 0$. When tick size is binding, $d > s^*$, so tick size leads to more entries of HFTs. Taken together, our model contributes to the literature by identifying a queuing channel of speed competition, in which HFTs race for top queue positions to capture the rents created by tick size.

We assume that BATs do not queue after the first share to get the analytical solution of the queuing equilibrium. The intuition when BATs can queue more than one share, however, remains the same. As long as we do not allow BATs to queue for an infinitely long time, BATs will demand liquidity with positive probability. In Section 4, we show that BATs always supply liquidity when tick size is small.

3. Equilibrium types under a small tick size

Starting from this section, we reduce the tick size to $\frac{d}{3}$. BATs then always choose to supply liquidity by establiShimg price priority over HFTs, except when the adverse selection risk is very low.

Corollary 1 shows that a small tick size of $\frac{d}{3}$ is still binding when $\frac{\lambda_I}{\lambda_J} > 5$.

Corollary 1. (**Small Binding Tick Size**) If $\Delta_1 = \frac{d}{3}$ and $\frac{\lambda_I}{\lambda_J} > 5$, the bid-ask spread equals the tick size. N_s^* HFTs jointly post Q_s^* units of sell limit orders at $a_{s,t}^* = v_t + \frac{d}{6}$ and Q_s^* units of buy limit orders at $b_{s,t}^* = v_t - \frac{d}{6}$, where:

$$Q_s^* = \max\left\{Q \in \mathbb{N}^+ \text{ s.t.} \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q \frac{d}{6} - \frac{1}{2}\left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right] \frac{5d}{6} > 0\right\}$$

$$= \left\lfloor \log_{\left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)} \frac{5}{7} \right\rfloor < Q^*, \text{ and} \tag{9}$$

$$N_s^* = \max\left\{N \in \mathbb{N}^+ \text{ s.t.} \lambda_I \frac{d}{6} - N c_{speed} > 0\right\} < N^*. \tag{10}$$

Compared with Proposition 1, a small tick size reduces revenue from liquidity supply from $\frac{d}{2}$ to $\frac{d}{6}$, increases the cost of being sniped from $\frac{d}{2}$ to $\frac{5d}{6}$, and reduces the queue length from Q^* to Q_s^*. Figure 2 shows that Q_s^* is approximately $\frac{1}{3}$ of Q^*. A small tick size also discourages the entry of HFTs. N_s^* is approximately $\frac{1}{3}$ of N^*, because HFTs' expected profit per unit of time decreases from $\lambda_I \frac{d}{2}$ to $\lambda_I \frac{d}{6}$.

When $1 < \frac{\lambda_I}{\lambda_J} < 5$, the break-even bid-ask spread is larger than one tick. To profit from the bid-ask spread, HFTs have to quote the following bid-ask spread:[11]

[11] We defer the derivation of the boundary condition for HFTs' bid-ask spread to Sections 4-6. Another way to bypass tick size constraints is to randomize quotes immediately above and below the break-even bid–ask spread. In this paper, we consider only stationary HFT quotes.

$$\begin{cases} \frac{d}{2} & \frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5 \\ \frac{5d}{6} & \frac{1}{5(1-\beta)} < \frac{\lambda_I}{\lambda_J} < \frac{1}{1-\beta} \\ \frac{7d}{6} & 1 < \frac{\lambda_I}{\lambda_J} < \frac{1}{5(1-\beta)} \end{cases} \tag{11}$$

Figure 3 shows that the bid-ask spread quoted by HFTs weakly decreases with $\frac{\lambda_I}{\lambda_J}$, because an increase in $\frac{\lambda_I}{\lambda_J}$ decreases adverse selection risk. The bid-ask spread quoted by HFTs increases weakly with the fraction of BATs, because BATs' strategies for minimizing transaction costs reduce HFTs' expected profit from liquidity supply. Interestingly, when the adverse section risk or the fraction of BATs is high, HFTs effectively cease supplying liquidity by quoting a bid-ask spread that is wider than the size of a jump. In the following sections, we elaborate the equilibrium types when tick size is not binding.

Insert Figure 3 about Here

4. Make-take spread

In this section, we develop a new concept make-take spread, and we use the concept to explain why BATs never demand liquidity from HFTs when the tick size is not binding. Without loss of generality, we consider the decision for a BAT who wants to buy. We start from the case when $\frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5$, for which HFTs need to quote an ask price of $v_t + \frac{d}{2}$ and a bid price of $v_t - \frac{d}{2}$ to profit from the bid-ask spread.

A BAT can choose to accept the ask price of $v_t + \frac{d}{2}$, but submitting a limit order to buy at $v_t + \frac{d}{6}$ is always less costly, because a buy limit order above fundamental value immediately attracts HFTs to submit market orders to sell. This flash limit order immediately executes like a

market order, but with lower cost.

Why do HFTs quote a sell price of $v_t + \frac{d}{2}$, but are willing to sell at $v_t + \frac{d}{6}$ using market orders? It is because HFTs' limit price to sell includes the costs of adverse selection risk. An offer to sell is more likely to be executed when v_t jumps up. HFTs would accept a lower sell price when they demand liquidity, because immediate execution reduces adverse selection risk.

Flash limit orders exploit the make-take spread, which measures the price difference between the traders' willingness to list an offer and their willingness to accept an offer conditional on the trade direction (e.g., sell). We discover make-take spread because liquidity suppliers can demand liquidity. This new feature reflects reality in contemporary electronic platforms. In most exchanges, every trader can supply liquidity and encounter very limited, if any restrictions when demanding liquidity (Clark-Joseph, Ye, and Zi, Forthcoming)

BATs are able to quote more aggressive prices than HFTs because they have lower opportunity costs for supplying liquidity. BATs have to buy or sell, and they supply liquidity as long as its cost is less than demanding liquidity. BATs lose $\frac{d}{6}$ by using flash limit orders, but the cost of flash limit orders is lower than paying a half bid-ask spread $\frac{d}{2}$. O'Hara (2015) finds that sophisticated non-HFTs cross the spread only when it is absolutely necessary. The make-take spread provides one interpretation for why sophisticated non-HFTs seldom cross the bid-ask spread.

When $1 < \frac{\lambda_I}{\lambda_J} < \frac{1}{1-\beta}$, the half bid-ask spread quoted by HFTs are higher than $\frac{d}{2}$, leaving more price levels for BATs to use flash limit orders. Therefore, BATs never demand liquidity as long as HFTs quote a bid-ask spread that is wider than one tick.

5. Flash equilibrium versus undercutting equilibrium

In the previous section, we show that flash orders strictly dominate market orders. In this section, we show that, under some conditions, BATs can further reduce their transaction costs by submitting limit orders that do not cross the midpoint. These regular limit orders do not get immediate execution but stay in the LOB to wait for market orders.

We consider BATs' choice between flash and regular limit orders. In the flash equilibrium, BATs use flash limit orders to supply liquidity to HFTs, and HFTs supply liquidity to non-algos. In the undercutting equilibrium, BATs use regular limit orders to supply liquidity to non-algos and other BATs, whereas HFTs follow complex strategies with frequent order additions and cancellations. For simplicity, we focus on the case when $\frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5$, for which HFTs need to quote an ask price of $v_t + \frac{d}{2}$ and a bid price of $v_t - \frac{d}{2}$ to profit from the bid-ask spread. In this case, BATs only need to consider two price levels: a flash limit order (e.g., $v_t + \frac{d}{6}$ to buy) or a regular limit order (e.g., $v_t - \frac{d}{6}$ to buy).

5.1 Flash equilibrium

In Proposition 2, we characterize the flash equilibrium. Starting from now, we only characterize the equilibrium outcome. BATs' response to off-equilibrium paths are defined in the proofs.

Proposition 2. (**Flash Equilibrium**): When $\Delta_1 = \frac{d}{3}$ and $\frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < \frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta}$, the equilibrium is characterized as follows:

1. BAT buyers submit limit orders at $v_t + \frac{d}{6}$ and BAT sellers submit limit orders at price $v_t - \frac{d}{6}$.
2. N_f^* HFTs jointly supply Q_f^* units of sell limit orders at $v_t + \frac{d}{2}$ and Q_f^* units of buy limit

orders at $v_t - \frac{d}{2}$, where:

$$Q_f^* = \max\left\{Q \in \mathbb{N}^+ \text{ s.t.} \left(\frac{(1-\beta)\lambda_I}{(1-\beta)\lambda_I+2\lambda_J}\right)^Q \frac{d}{2} - \frac{1}{2}\left(1 - \left(\frac{(1-\beta)\lambda_I}{(1-\beta)\lambda_I+2\lambda_J}\right)^Q\right)\frac{d}{2} > 0\right\}$$

$$= \left\lfloor \log_{\left(\frac{(1-\beta)\lambda_I}{(1-\beta)\lambda_I+2\lambda_J}\right)} \frac{1}{3} \right\rfloor < Q^* \tag{12}$$

$$N_f^* = \max\left\{N \in \mathbb{N}^+ \text{ s.t.} \beta\lambda_I \frac{d}{6} + (1-\beta)\lambda_I \frac{d}{2} - Nc_{speed} > 0\right\} < N^*. \tag{13}$$

3. HFTs participate in three races: (1) HFTs race to fill the queue when the depth at $v_t + \frac{d}{2}$ or $v_t - \frac{d}{2}$ becomes less than Q_f^*. (2) HFTs race to take the liquidity offered by flash limit orders. (3) After a value jump, HFTs who supply liquidity race to cancel the stale quotes, whereas stale-quote snipers race to pick off the stale quotes.

In Proposition 2, we first derive the boundary between the flash equilibrium and the undercutting equilibrium. Figure 4 illustrates the boundary in. BATs choose flash limit orders over regular limit orders when adverse selection risk is high. Intuitively, flash limit orders execute immediately, but it costs $\frac{d}{6}$ relative to the midpoint; regular limit orders capture a half bid-ask spread of $\frac{d}{6}$ if executed against a non-HFT, but it is also subject to adverse selection risk. BATs tend to choose flash limit orders when the adverse selection risk is high. Figure 4 also shows BATs tend to choose regular limit orders when β decreases. Intuitively, because non-algo traders use only market orders, a regular limit order on the book would have higher execution probability before a value jump as the fraction of non-algo traders increases.

Insert Figure 4 about Here

Proposition 2 identifies a unique type of speed competition led by tick size: racing to be the first to take the liquidity offered by flash limit orders. If price is continuous, any buy limit order price above fundamental value would prompt HFTs to sell. In our model with discrete tick size, a BAT needs to place the buy limit order at $v_t + \frac{d}{6}$, which drives the speed race to capture the rent of $\frac{d}{6}$ through demanding liquidity.

In the literature, HFTs demand liquidity when they have advance information to adversely select other traders (BCS; Foucault, Kozhan, and Tham, Forthcoming; Menkveld and Zoican, 2017). Consequently, HFTs' liquidity demand often has negative connotations. Our model shows that HFTs can demand liquidity without adversely selecting other traders. Instead, the transaction cost is lower for BATs when HFTs demand liquidity than when HFTs supply liquidity. Therefore, researchers and policy makers should not evaluate the welfare impact of HFTs simply based on liquidity supply versus liquidity demand.

As BATs no longer demand liquidity from HFTs, HFTs respond to the reduced liquidity demand and higher adverse selection cost by decreasing their depth to Q_f^*. The profit to take liquidity from BATs, $\frac{d}{6}$, is less than the profit to supply liquidity to BATs at $\frac{d}{2}$ when the tick size is Δ_0. A smaller tick size, Δ_1, reduces the profit for HFTs, thereby reducing the number of HFTs.

5.2 Undercutting equilibrium

In flash equilibrium, the LOB only has one stable state. In the undercutting equilibrium, the LOB transits across different states. As indicated in Proposition 2, BATs choose regular limit orders over flash limit orders when adverse selection risk or β is low. In the undercutting equilibrium, their limit orders stay in the LOB, and their decisions, as well as those of HFTs, depend on the state of the LOB. Our technical assumption that BATs never queue at the second

position reduces the number of states. Still, the solution is complicated. We focus on deriving the equilibrium *strategies* of HFTs, as Proposition 2 and its proof in the Appendix demonstrate the strategy of BATs in undercutting equilibrium. BATs choose regular limit orders over flash limit orders when $\frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta} < \frac{\lambda_I}{\lambda_J} < 5$.

To show the equilibrium strategy of HFTs, we first define the state of the LOB as (i, j). Here i represents the number of BATs' limit orders on the same side of the LOB, and j denotes the number of BATs' limit orders on the opposite side of the LOB. For example, for a HFT who wants to buy, i represents the number of BATs' limit orders on the bid side, and j represents the number of BATs' limit orders on the ask side. The LOB then has four states:

(0,0)	No limit order from BATs
(1,0)	A BAT limit order on the same side
(0,1)	A BAT limit order on the opposite side
(1,1)	BAT limit orders on both sides

When $\frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta} < \frac{\lambda_I}{\lambda_J} < 5$, HFTs quote a half bid-ask spread of $\frac{d}{2}$, as a half bid-ask spread of $\frac{d}{6}$ loses money. Similar to the queuing equilibrium and the flash equilibrium, HFTs' decision to supply liquidity depends on the payoff of the liquidity supply relative to the outside option of sniping. The new feature of the undercutting equilibrium is that HFTs' decision also depends on the status of the LOB. We denote the payoff of the Q^{th} share to supply liquidity at half the bid-ask spread $\frac{d}{2}$ as $LP^{(i,j)}(Q)$, and the payoff to the snipers of the Q^{th} share as $SN^{(i,j)}(Q)$. The HFT's strategy depends on $D^{(i,j)}(Q) \equiv LP^{(i,j)}(Q) - SN^{(i,j)}(Q)$.

Figure 5 illustrates how $D^{(i,j)}(Q)$ changes with the six types of events defined in equation (1). For example, consider $D^{(0,0)}(Q)$ for an HFT on the ask side of the LOB.

1) A BAT buyer submits a limit order at $v_t - \frac{d}{6}$, which changes $D^{(0,0)}(Q)$ to $D^{(0,1)}(Q)$.

2) A BAT seller undercuts the ask side at $v_t + \frac{d}{6}$, which changes $D^{(0,0)}(Q)$ to $D^{(1,0)}(Q)$.

3) A non-algo buyer submits a market buy order, which moves the queue position forward by one unit. $D^{(0,0)}(Q)$ changes to $D^{(0,0)}(Q-1)$.

4) A non-algo seller submits a market sell order, which does not affect $D^{(0,0)}(Q)$ as the LOB on the bid side is refilled immediately by HFTs.

5) In an upward value jump, a liquidity providing HFT on the ask side gains $-\frac{d}{2}\frac{N-1}{N}$, a stale-quote sniper gains $\frac{d}{2}\frac{1}{N}$, and the difference between them is $-\frac{d}{2}$.

6) In a downward value jump, the liquidity supplier cancels the limit order, thereby changing the value of both the liquidity supply and stale-quote snipping to zero.

Insert Figure 5 about Here

These six types of events and the four states of the LOB are the key features of the undercutting equilibrium, which we summarize in Proposition 3. To simplify the notation, we use $p_1 \equiv \frac{1}{2}\cdot\frac{\lambda_I\beta}{\lambda_I+\lambda_J}$ to denote the arrival probability of a BAT buyer or seller, $p_2 \equiv \frac{1}{2}\cdot\frac{\lambda_I(1-\beta)}{\lambda_I+\lambda_J}$ to denote the arrival probability of a non-algo trader to buy or sell, and $p_3 \equiv \frac{1}{2}\cdot\frac{\lambda_J}{\lambda_I+\lambda_J}$ to denote the probability of an upward or downward value jump.

Proposition 3. (**Undercutting Equilibrium**): When $\Delta_1 = \frac{d}{3}$ and $\frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta} < \frac{\lambda_I}{\lambda_J} < 5$, the equilibrium is characterized as follows:

1. HFTs' strategy:
    a. Spread: HFTs quote ask price at $v_t + \frac{d}{2}$ and bid price at $v_t - \frac{d}{2}$.

b. Depth: The following system of equations determines the equilibrium depth in each state.

i. Difference in value between the liquidity supplier and the stale-queue sniper in each state:

$$\begin{cases} D^{(0,0)}(Q) = max\{0, p_1 D^{(0,1)}(Q) + p_1 D^{(1,0)}(Q) + p_2 D^{(0,0)}(Q-1) + p_2 D^{(0,0)}(Q) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0\} \\ D^{(1,0)}(Q) = max\{0, p_1 D^{(1,1)}(Q) + p_1 D^{(1,0)}(Q) + p_2 D^{0,0}(Q) + p_2 D^{(1,0)}(Q) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0\} \\ D^{(0,1)}(Q) = max\{0, p_1 D^{(0,1)}(Q) + p_1 D^{(1,1)}(Q) + p_2 D^{(0,1)}(Q-1) + p_2 D^{(0,0)}(Q) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0\} \\ D^{(1,1)}(Q) = max\{0, p_1 D^{(0,1)}(Q) + p_1 D^{(1,0)}(Q) + p_2 D^{(0,1)}(Q) + p_2 D^{(1,0)}(Q) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0\} \end{cases} . \quad (14)$$

ii. Difference in value for immediate execution: $D^{(0,0)}(0) = D^{(0,1)}(0) = \frac{d}{2}$.

iii. Equilibrium depth as a function of the difference in value:

$$Q^{(i,j)} = max\left\{Q \epsilon \mathbb{N}^+ \middle| D^{(i,j)}(Q) > 0\right\} \quad i = 0,1; j = 0,1.$$

c. In equilibrium there are $N_u^* < N^*$ HFTs.

2. BATs who intend to buy (sell) submit limit orders at price $v_t - \frac{d}{6}$ $(v_t + \frac{d}{6})$ if no existing limit orders sit at the price level, or buy (sell) limit orders at price $v_t + \frac{d}{6}$ $(v_t - \frac{d}{6})$ otherwise[12].

The depth from HFTs depends on $D^{(i,j)}(Q)$. $D^{(i,j)}(Q)$, is defined using the equation system in (14), because the value difference in each state also depends on the value differences in other states. The equations in (14) contain the $max\{0, .\}$ as HFTs do not queue at the Q^{th} position once the expected payoff is below 0.

[12] After an upward (downward) jump with size d, we assume BATs buy (sell) undercutting orders at $v_t - \frac{d}{6}$ $(v_t + \frac{d}{6})$ will be cancelled and resubmitted at price $v_t + \frac{5d}{6}$ $(v_t - \frac{5d}{6})$ to follow the value jump. Alternative BATs strategy does not change the equilibrium.

We present the solution for $D^{(i,j)}(Q)$ for any $i, j,$ and Q in the Appendix. Here we use a numerical example to present the main intuition of the undercutting equilibrium. Figure 6 shows that the value of the liquidity supply decreases in Q, while the value of stale-quote sniping increases in Q. HFTs supply liquidity as long as $LP^{(i,j)}(Q) > SN^{(i,j)}(Q)$. For example, in state$(0,0)$, the LOB has a depth of two shares.

Figure 6 also shows that $LP^{(i,j)}(Q)$ and $SN^{(i,j)}(Q)$ also depend on the state of the LOB. As the undercutting limit orders from BATs can change the states of the LOB, HFTs can add or cancel their limit orders even when the fundamental value stays the same. A comparison between Panel A and Panel B and between Panel C and Panel D of Figure 6 shows that an undercutting order reduces HFTs' depth on the same side of the LOB by approximately one share. Intuitively, when a BAT submits an undercutting order, the execution priority for all HFTs on the same side of the book decreases by one share.[13] An HFT who used to quote the last share at the half bid-ask spread $\frac{d}{2}$ has to cancel, because the share become unprofitable after the arrival of the undercutting order. For the same reason, once an undercutting order from a BAT executes, HFTs race to submit one more share at the half bid-ask spread $\frac{d}{2}$, because the execution priority in the LOB increases by one. One new feature of the undercutting equilibrium is the frequent order addition or cancellation of HFTs' limit orders in the absence of a change in fundamental value.

[13] An undercutting BAT order on the opposite side of the LOB has an indirect effect. For example, in state (1, 1), a BAT buyer takes liquidity at price $v_t + \frac{d}{6}$ and changes the state to (0, 1), which enables an HFT limit sell order at price $v_t + \frac{d}{2}$ to trade with the next buy market order from a non-algo trader. In state (1, 0), a BAT buyer chooses to submit a limit order at price $v_t - \frac{d}{6}$, which changes the state to (1, 1). An HFT limit sell order at price $v_t + \frac{d}{2}$ then needs to wait at least one more period for execution. More generally, an undercutting BAT limit buy (sell) order may attract future BAT sellers (buyers) to demand liquidity, making future BATs less likely to undercut HFTs. In turn, the value of liquidity supply increases relative to sniping, thereby incentivizing HFTs to supply larger depth. This indirect effect is so small that it does not affect depth in our numerical example, because the number of shares is an integer. It is possible for a depth of (1, 1) to be higher than (1, 0) for numerical values such as $\frac{\lambda_I}{\lambda_J} = 4.9$ and $\beta = 0.06$, and the results are available upon request.

One driver of HFTs' frequent additions and cancellations is small tick size. When tick size is binding, BATs cannot achieve execution priority over HFTs who are already in the queue. When tick size is small, BATs can achieve price priority over HFTs, which induces HFTs to cancel their earlier orders and to add new ones in response to the undercutting orders from BATs.

When $\frac{1}{5(1-\beta)} < \frac{\lambda_I}{\lambda_J} < \frac{1}{1-\beta}$, HFTs quote $\frac{5d}{6}$, and BATs' strategies follow the intuition outlined above, where they choose between flash limit orders and regular limit orders. The only main difference is that the four price levels between $v_t + \frac{5d}{6}$ and $v_t - \frac{5d}{6}$ increase the states to $2^4 = 16$. We do not report the results for brevity but they are available upon request. In Section 6, we discuss the case when the break-even spread equals $\frac{7d}{6}$.

6. Stub quotes and mini-flash

In Proposition 4, we show that HFTs quote a bid-ask spread wider than the size of the jump when adverse selection risk is high or the fraction of BATs is large. We call such quotes stub quotes. A mini-flash crash occurs when a market order hits a stub quote. In our model, the size of the mini-flash crash is $\frac{7d}{6}$, because the size of a value jump is d. An increase in the support of jump size can lead to stub quotes further away from the midpoint, thereby creating mini-flash crashes of larger size. Such an extension adds mathematical complexity without conveying new intuition.

Proposition 4 (**Stub Quotes and Mini-Flash Crash**). When $\Delta_1 = \frac{d}{3}$ and $1 < \frac{\lambda_I}{\lambda_J} < \frac{1}{5(1-\beta)}$, the equilibrium is characterized as follows.

1. HFTs quote a half bid-ask spread of $\frac{7d}{6}$.

2. A BAT buyer (seller) quotes $v_t - \frac{5d}{6}$ ($v_t + \frac{5d}{6}$) if the price level has no limit orders. Otherwise, the BAT buyer (seller) submits a flash limit order at price $v_t + \frac{d}{6}$ ($v_t - \frac{d}{6}$) to provide liquidity.
3. Compared with the case when $\Delta_0 = d$, the transaction cost for non-algo traders increases, but the average transaction cost for non-HFTs decreases.
4. The probability of mini-flash crashes decreases in $\frac{\lambda_I}{\lambda_J}$. The probability of mini-flash crashes first increases in β and then decreases in β.

Proposition 4 shows that HFTs are more likely to quote stub quotes when adverse selection risk is high. A higher adverse selection risk prompts HFTs to quote stub quotes through two channels. First, HFTs have to quote a wider bid-ask spread to reach the break even point. Second, when HFTs' quotes are wider than one tick, BATs are able to quote more aggressive prices than HFTs. HFTs then need to further widen the bid-ask spread due to reduced liquidity demand.

When HFTs quote stub quotes, BATs have six price levels to choose from. Fortunately, we are able to obtain analytical solutions for the BATs' strategy. Consider the decision for a BAT buyer. We find that the buyer chooses to queue at $v_t - \frac{5d}{6}$ if the price level contains no limit orders. The sniping cost is as low as $\frac{d}{6}$, and the BAT buyer can earn a half bid-ask spread of $\frac{5d}{6}$ if a non-algo trader arrives. When $v_t - \frac{5d}{6}$ contains a limit order, the BAT buyer will use a flash limit order at $v_t + \frac{d}{6}$ to obtain immediate execution with a transaction cost of $\frac{d}{6}$.[14] We show in the proof that

[14] This result is certainly a consequence of our simplifying assumption that BATS cannot queue for a second share. However, BATs should always have higher incentives to use flash limit orders when $v_t - \frac{5d}{6}$ contains a limit order,

BATs never quote at $v_t - \frac{d}{2}$ and $v_t - \frac{d}{6}$ as the execution cost is always higher than $\frac{d}{6}$. Flash buy limit orders at price $v_t + \frac{d}{6}$ also strictly dominate more aggressive flash limit orders of $v_t + \frac{d}{2}$ and $v_t + \frac{5d}{6}$, because a limit order price of $v_t + \frac{d}{6}$ is aggressive enough to trigger immediate execution.

In Section 3, we find that the transaction costs for both BATs and non-algo traders are $\frac{d}{2}$ when tick size is *d*. A decrease in tick size to $\frac{d}{3}$ increases the transaction cost for non-algo traders. A non-algo trader pays $\frac{5d}{6}$ when an order is she executed against a BAT and pays $\frac{7d}{6}$ if a stub quote is encountered. Meanwhile, a decrease in tick size to $\frac{d}{3}$ decreases the transaction cost for BATs. BATs' maximum transaction cost is $\frac{d}{6}$ if they use flash limit orders, although the cost is lower if they quote a half bid-ask spread of $\frac{5d}{6}$. Overall, we find that the average transaction cost decreases with tick size. Figure 3 shows that the proportion of BATs needs to be at least $\frac{4}{5}$ for stub quotes to occur. Non-algo traders' maximum transaction cost is $\frac{7d}{6}$ if they hit stub quotes. The average transaction cost for non-HFTs is then at most $\frac{11d}{30}$ $(\frac{4}{5} \times \frac{d}{6} + \frac{1}{5} \times \frac{7d}{6})$, which is lower than $\frac{d}{2}$. Therefore, a reduction in tick size reduces non-HFTs' average transaction costs, but increase the dispersion and volatility of their transaction costs.

An increase in adverse selection risk unambiguously increases the probability of mini-flash crashes. Figure 3 in Section 3 show that stub quotes are more likely to occur when there higher adverse selection risk. Conditional on stub quotes occurring, Figure 6 reveals another channel for adverse selection risk to increase the number of mini-flash crashes. An increase in adverse

because the second share has a lower probability of executing against a non-algo trader and a higher probability of executing against a sniper, whereas a flash limit order always incurs a constant cost of $\frac{d}{6}$.

selection risk implies more value jumps relative to the arrival rate of non-algo traders. During an upward (downward) value jump, BATs' limit orders on the bid (ask) side are all sniped and only stub quotes remain. If the limit orders from BATs fail to reconvene before a non-algo trader arrives, the market order from the non-algo trader hits the stub quote and causes a mini-flash crash.

The proportion of BATs, β, have an ambiguous effect on the probability of flash crashes because of two competing effects. On the one hand, Figure 3 in Section 3 shows that a larger β increases the probability for stub quotes as HFTs face less liquidity demand. On the other hand, a larger β decreases the probability of hitting stub quotes, because BATs never demand liquidity from HFTs. For example, mini-flash crashes never occur when $\beta = 0$ or $\beta = 1$. Therefore, mini-flash crashes need both BATs and non-algo traders. Figure 6 shows the simulated intensity of mini-flash crashes with respect to β. For each β, we first uniformly draw 100 $\frac{\lambda_I}{\lambda_J}$ from [1, 5], the support of the adverse selection risk in our paper. For each $\frac{\lambda_I}{\lambda_J}$, we simulate the first 100,000 trades. For all 10 million simulations, we count the number of trades that hit the stub quotes relative to the total number of trades.

Figure 7 shows that mini-flash crashes are most likely to occur when β is approximately 0.95, and we normalize this crash intensity to 1. The black square line shows that the intensity is hump-shaped with respect to β. The circle line shows that majority of mini-flash crashes occur after a value jump. An upward value jump removes BATs' limit orders from the ask side and a downward jump removes BATs' limit orders from the bid side. If BATs' limit orders do not reconvene in the LOB, a market buy (sell) order from non-algo trader would hit stub quotes. Therefore, most of the upward (downward) mini-flash crashes occur after an upward (downward) value jump. Only a small amount of crashes are due to BATs' liquidity being used up by non-algo traders.

An effective way to prevent a mini-flash crash is a trading halt to let the trading interest of BATs reconvene. The triangle line in Figure 7 shows the intensity of mini-flash crashes with trading halts. We impose the trading halt after a value jump, and the market reopens after 10 orders arrive at the market. We find that such a trading halt reduces mini-flash crashes by about 90%.

Insert Figure 7 About Here

6. Predictions and policy implications

Our model rationalizes a number of puzzles in the literature on HFTs and generates new empirical predictions that can be tested. In Subsection 6.1, we summarize the predictions on who supplies liquidity and when. In Subsection 6.2, we examine the predictions on liquidity demand. In Subsection 6.3, we evaluate the predictions on liquidity. In Subsection 6.4, we discuss the use of the cancellation ratio as the cross-sectional proxy for HFTs' activity.

6.1 Liquidity supply

Our model shows that who provides liquidity depends on the tick size, adverse selection risk, the motivation of the trade, and the speed of the trade. In Prediction 1, we posit that BATs dominate liquidity supply when tick size is not binding.

Prediction 1 (Price Priority): When tick size is not binding, Non-HFTs are more likely to establish price priority in liquidity supply.

Speed advantages in the LOB reduce HFTs' adverse selection costs (see Jones (2013) and Menkveld (2016) surveys), inventory costs (Brogaard et al., 2015), and operational costs (Carrion,

2013). These reduced costs of intermediation raise the concern 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" (Chordia et al., 2013, p. 644). However, Brogaard et al. (2015) find that non-HFTs quote a tighter bid-ask spread than HFTs, and Yao and Ye (2017) find that non-HFTs are more likely to establish price priority over HFTs as the tick size decreases. We find that the opportunity cost of supplying liquidity can reconcile the contradiction between the empirical results and the channels of speed competition. BATs incur lower opportunity costs when supplying liquidity. When they implement a trade, they supply liquidity as long as it is less costly to demand liquidity. The make-take spread that we introduce in Section 4 indicates that BATs never demand liquidity from HFTs when tick size is not binding.

Prediction 2 (Queuing): HFTs crowd out non-HFTs' liquidity supply when tick size is binding, that is, when the tick size is large or adverse selection risk is low.

When tick size is binding, HFTs' speed advantage allows them to establish time priority at the same price. Yao and Ye (2017) find that tick size is more likely to be binding when tick size increases. They also find that a large tick size crowds out non-HFTs' liquidity supply. Both results provide evidence to support Prediction 2.

Hoffmann (2014), Han, Khapko, and Kyle (2014), Bernales (2016), and Bongaerts and Van Achter (2016) find that HFTs have lower adverse selection costs than non-HFTs. Yao and Ye (2017), however, find that HFTs do not have a comparative advantage in providing liquidity for stocks with higher adverse selection risk. In Prediction 2, we provide the economic mechanism to reconcile this inconsistency. Comparing Corollary 1 with Proposition 2 and 3, we find that the tick

size is more likely to be binding when adverse selection risk is low. A binding tick size helps HFTs to supply liquidity through time priority. An increase in adverse selection risk raises the break-even bid-ask spread above one tick, allows non-HFTs to undercut HFTs, and decreases HFTs' liquidity supply.

In Prediction 3, we address who provides liquidity during a mini-flash crash.

Prediction 3. (**Stub Quotes and Mini-Flash Crashes**): A mini-flash crash is more likely to occur when the adverse selection risk is high or when the tick size is small. During a mini-flash crash, HFTs supply liquidity and non-HFTs demand liquidity. A downward (upward) mini-flash crash is more likely to follow a downward (upward) value jump.

A comparison of Propositions 1 and 4 shows that stub quotes are more likely to occur when the tick size is small. When the tick size is large, BATs cannot establish execution priority over HFTs. When the tick size is small, BATs can establish price priority over HFTs, which increases the adverse selection costs for HFTs through two channels. First, when BATs can undercut HFTs, they no longer demand liquidity from HFTs. HFTs then face reduced liquidity demand but the risk of value jump stay the same. Second, the undercutting orders by BATs reduce the execution priority of HFTs. In turn, HFTs' limit orders face lower execution probability and higher sniping cost. When the adverse selection cost is high enough, HFTs effectively quit liquidity supply by quoting stub quotes. HFTs are more likely to quote stub quotes when adverse selection risk is high as higher adverse selection risk widens the break-even bid-ask spread; a wider break-even bid-ask spread also allows BATs to undercut HFTs, which further increases the adverse selection costs for HFTs. Because BATs do not continuously supply liquidity in the market, non-algo traders' market

orders can hit stub quotes and cause mini-flash crashes. A high adverse selection risk also implies more value jumps relative to the arrival rate of non-HFTs. Non-algo traders' market orders are more likely to hit stub quotes after value jumps, because value jumps clear BATs' limit orders on the side of the jump.

In cross-section, our model predicts that stocks with smaller tick sizes or higher adverse selection risk are more likely to incur mini-flash crashes. This cross-sectional pattern has not been tested. In time series, our model predicts that an initial downward (upward) jump increases the probability of a downward (upward) mini-flash crash. The downward (upward) jump clears the LOB on the bid (ask) side, making the market orders from non-algo traders more likely to hit stub quotes.

Brogaard et al. (Forthcoming) analyze the time series pattern of mini-flash crashes. They show that, 20 seconds before a mini-flash crash, HFTs neither demand nor supply liquidity, whereas non-HFTs demand and supply the same amount of liquidity; 10 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 the liquidity supply from the mini-flash crash is profitable. This evidence is consistent with the theoretical mechanism for mini-flashes crash that we document. (1) In normal times, non-HFTs dominate both liquidity supply and liquidity demand; (2) slightly before a mini-flash crash, HFTs demand liquidity and remove limit orders from BATs; (3) a mini-flash crash occurs when a non-algo trader's market order hits HFTs' stub quotes, thus HFTs profit when a mini-flash crash occurs.

Our interpretations of mini-flash crashes are consistent with both negative and positive framing of the role of HFTs in a mini-flash crash. Brogaard et al. (2017) suggest that HFTs supply liquidity in extreme price movements, while Ait-Sahalia and Sağlam (2017) suggest that HFTs

withdraw liquidity supply when it is most needed. Both views, however, suggest that mini-flash crashes occur when the market orders of non-HFTs hit the stub quotes from HFTs.

Our interpretation of mini-flash crashes has two additional features that are consistent with economic reality. First, markets recover quite quickly from mini-flash crashes. In our model, mini-flash crashes disappear when the limit orders from BATs replenish the LOB. Second, Nanex, the firm that invented the concept of mini-flash crash, finds that mini-flash crashes are equally likely to be upward as downward. Indeed, even during the famous 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). In our model, upward and downward mini-flash crashes are equally likely, even though downward mini-flash crashes are more likely to occur conditional on an initial downward value jump.

6.2 Liquidity demanding

Our model discoveries a new channel of speed competition to demand liquidity. In Prediction 4, we summarize the empirical implications of this new channel.

Prediction 4. (**Speed Competition of Taking Liquidity**): Non-HFTs are more likely than HFTs to supply liquidity at price levels that cross the midpoint (flash limit orders). HFTs are also more likely to demand liquidity from flash limit orders, but they do not adversely select these orders.

Latza, Marsh, and Payne (2014) find evidence consistent with Prediction 4. They classify a market order as “fast” if it executes against a standing limit order that is less than 50 milliseconds old. Because of the speed of taking liquidity, it is natural to expect that fast market orders are from

HFTs. These authors also find that fast market orders often execute against limit orders that cross the midpoint, and they lead to virtually no permanent price impact.

In Prediction 4, we offer fresh perspectives on the liquidity demand from HFTs. Typically, HFTs demand liquidity when they employ a speed advantage to adversely select liquidity suppliers (BCS; Foucault, Kozhan, and Tham, 2017; Menkveld and Zoican, 2017). Therefore, liquidity demand from HFTs generally has negative connotations of reducing liquidity (Jones, 2013; Biais and Foucault, 2014). We find that HFTs' liquidity demand does not necessarily adversely select slow traders. Instead, the liquidity demand from HFTs can reduce the transaction costs of non-HFTs. In the flash equilibrium, BATs pay $\frac{d}{2}$ when HFTs supply liquidity, while BATs only pay $\frac{d}{6}$ when HFTs demand liquidity.

6.3 Liquidity

On April 5, 2012, President Barack Obama signed into law the Jumpstart Our Business Startups (JOBS) Act. Section 106 (b) of the Act requires the SEC to examine the effect 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 small- and mid-cap stocks. Proponents of the proposal argue that a larger tick size can improve liquidity (Weild, Kim, and Newport, 2012). In Prediction 5, however, we posit that an increase in tick size decreases liquidity.

Prediction 5. A larger tick size increases the depth at the BBO, but it also increases the effective bid-ask spread, the transaction costs paid by liquidity demanders.

Yao and Ye (2017) find evidence consistent with Prediction 5. Holding the BBO constant,

an increase in depth at the BBO implies an increase in liquidity. Yet these authors also find that the quoted bid-ask spread increases after an increase in tick size. When both quoted bid-ask spread and depth increase, the most relevant liquidity measure becomes the effective bid-ask spread, the transaction cost paid by liquidity demanders (Bessembinder, 2003). Our model shows that constrained price competition increases the effective bid-ask spread, which is consistent with Yao and Ye's (2017) findings. Our model prediction, along with the evidence in Yao and Ye (2017), shows that an increase in tick size would not improve liquidity.

Advocates for an increase in tick size also argue that a wider tick size increases market-making profits, supports sell-side equity research and, eventually, increases the number of IPOs (Weild, Kim, and Newport, 2012). We find that a wider tick size increases market-making profits, but the profit belongs to traders with higher transaction speeds. Therefore, a wider tick size is more likely to result in an arms race in latency reduction than in sell-side equity research.

We also find that an increase in tick size harms non-HFTs. An increase in tick size also does not benefit HFTs as the cost of the speed investment dissipates when larger tick size generates higher rents. In our model, non-HFTs trade no matter how large the bid-ask spread may be. In reality, a wider spread may prevent investors with low gains from trading, leading to a further reduction in welfare.

An increase in tick size reduces mini-flash crashes, but it also increases the transaction costs for average trades. A more effective solution to prevent mini-flash crashes would be to slow down the market, particularly during periods of market stress. In a standard Walrasian equilibrium, price is continuous and time is discrete. Modern financial markets exhibit exactly the opposite structure: price competition is constrained by the tick size, whereas time is divisible at the nanosecond level in electronic trading platforms (Gao, Yao, and Ye, 2013). Making price more

continuous and time more discrete would improve liquidity and also prevent mini-flash crashes at the same time.

6.4 Cancellation-to-trade ratio as a cross-sectional proxy for HFT activity

The cancellation-to-trade ratio is widely used as a proxy for HFTs' activities, particularly for HFTs' liquidity supplying activities (Biais and Foucault, 2014). Yet Yao and Ye (2017) find that stocks with a higher proportion of liquidity provided by HFTs have a lower cancellation-to-trade ratio. In Prediction 6, we offer one interpretation for this surprising negative correlation.

Prediction 6. (**Cancellation-to-trade Ratio**). Stocks with a smaller tick size and higher adverse selection risk have a lower proportion of liquidity provided by HFTs relative to non-HFTs but a higher cancellation-to-trade ratio.

A decrease in tick size decreases the proportion of liquidity provided by HFTs (Prediction 2), but it leads to more order cancellations. Under a large tick size in our model, HFTs do not need to cancel their orders when non-HFTs arrive, because non-HFTs cannot establish time priority over HFTs. A decrease in tick size increases the potential for non-HFTs to undercut HFTs. If non-HFTs submit flash limit orders, HFTs race to take liquidity, and the losers of the race cancel their orders. If non-HFTs submit regular limit orders, HFTs reduce their depth once non-HFTs undercut, and HFTs increase their depth once an undercutting order gets executed. These changes in depth lead to frequent order cancellations. We offer a new interpretation of flickering quotes. Yueshen (2014) shows that flickering quotes occur when new information causes the price to move to a new level. We show that HFTs can cancel orders in the absence of information. Periodic order additions

and cancellations also differ from Baruch and Glosten (2013), who rationalize flicking quotes using a mixed-strategy equilibrium. An increase in adverse selection risk, defined as the intensity of value jumps relative to the arrival rate of non-HFTs, also lead to more order cancellations, but HFTs also provide less liquidity for these stocks. Taken together, we suggest that the cancellation-to-trade ratio should not be used as a cross-sectional measure of HFTs' activity.

7. Conclusion

In this paper, we extend BCS by adding two unique characteristics in financial markets: discrete tick size and algorithmic traders who are not HFTs. We discover a queuing channel of speed competition for liquidity supply. BATs are more likely to supply liquidity when tick size is small, because supplying liquidity is less costly than demanding liquidity from HFTs. A large tick size constrains price competition, creates rents for liquidity supply, and encourages speed competition to capture such rents through the time priority rule. Higher adverse selection risk increases the break-even bid-ask spread relative to tick size, which allows BATs to establish price priority over HFTs and reduces the fraction of liquidity provided by HFTs.

We also discover a new channel of speed competition in liquidity demand. HFTs race to demand liquidity from BATs when BATs post flash limit orders to buy above the fundamental value or to sell below the fundamental value. BATs incur lower transaction cost when HFTs demand liquidity than when HFTs supply liquidity. Thus, an evaluation of the welfare impact of HFTs should not be based solely on demand versus supply liquidity. Our results also indicate that the definition of providing versus demanding liquidity blurs in model electronic markets.

Yao and Ye (2017) find that the cancellation ratio, a widely used empirical proxy for HFTs' activity, has a negative cross-sectional correlation with HFT liquidity supply. We provide a

theoretical foundation for their surprising negative correlation. A large tick sizes induces HFTs to race for the top queue position, and HFTs are less likely to cancel orders once they secure this spot. HFTs cancel orders more frequently for stocks with smaller tick sizes, but they also supply less liquidity. Both theoretical and empirical evidence suggests that researchers should not apply the cancellation ratio as a cross-sectional proxy for HFT activity.

We also provide new predictions to be tested. We predict that 1) non-HFTs are more likely than HFTs to supply liquidity at price levels that cross the midpoint, and these limit orders are more likely to be taken by HFTs; 2) a mini-flash crash is more likely to occur for stocks with smaller tick sizes and higher adverse selection risk; 3) an upward (downward) mini-flash crash is more likely to follow an initial price jump in the same direction.

Our model shows that a larger tick size increases transaction cost and negatively affects non-HFTs. Yet HFTs do not benefit from a larger tick size as an investment in high-speed technology dissipates the rents created by tick size. We challenge the rationale for increasing the tick size to five cents, and we encourage regulators to consider decreasing tick size, particularly for liquid stocks.

Our model is parsimonious. For example, BATs in our model do not have private information and they choose order types only upon arrival. It will be interesting to extending our model toward more realistic setups. Most studies in the finance literature ignore diversity among algorithms traders. We take the initial step to examine algorithmic traders who are not HFTs, and we believe that further examination on the relationship between HFTs and other algorithmic traders would prove to be fruitful.

References

Angel, J., L. Harris, and C. Spatt. 2015. Equity trading in the 21st century: An update. *The Quarterly Journal of Finance* 5:1550002-1-1550002-39.

Baruch, S., and L. R. Glosten. 2013. Fleeting orders. Columbia Business School Research Paper: 13-43.

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

Bessembinder, H. 2003. Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis* 38:747-777.

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

Boehmer, E., K. Fong, and J. Wu. 2015. International evidence on algorithmic trading. Working Paper, Singapore Management University, University of New South Wales, and University of Nebraska at Lincoln.

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* 28:3407-43.

Brogaard, J., A. Carrion, T. Moyaert, R. Riordan, A. Shkilko, and K. Sokolov. Forthcoming. High-frequency trading and extreme price movements. *Journal of Financial Economics*.

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

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., M. Ye, and C. Zi. Forthcoming. Designated market makers still matter: Evidence from two natural experiments. *Journal of Financial Economics*.

Colliard, J. E., and T. Foucault. 2012. Trading fees and efficiency in limit order markets. *Review*

of Financial Studies 25:3389-3421.

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

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.

————. 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.

Hendershott, T., C. M. Jones, and A. J. Menkveld. 2011. Does algorithmic trading improve liquidity?. *Journal of Finance* 66:1-33.

Hendershott, T. and A. J. Menkveld. 2014. Price pressures. *Journal of Financial Economics* 114:405-423.

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

Jiang, G., L. Ingrid, and V. Giorgio. 2014. High-Frequency Trading around Macroeconomic News Announcements: Evidence from the U.S. Treasury Market. Working Paper, Bank of Canada.

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

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315-1335.

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.

————, 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.

————., G. Saar, and Z. Zhong. 2015. 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.

Rosu, We. 2009. A dynamic model of the limit order book. *Review of Financial Studies* 22:4601-4641.

Stoll, H.R., 2000. Presidential address: friction. *The Journal of Finance* 55:1479-1514.

United States. Commodity Futures Trading Commission, and Securities and Exchange Commission. 2010. *Findings regarding the market events of May 6, 2010.*

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

Yao, C., and M. Ye. 2017. Conditionally accepted. Why trading speed matters: A tale of queue rationing under price controls. *Review of Financial Studies.*

Yueshen, B.Z. 2014. Queuing uncertainty in limit order market. Working Paper, INSEAD.

Appendix

Proof for Lemma 1

For the Q^{th} share in the queue at the half bid-ask spread $\frac{s}{2}$, we define its value for the liquidity supplier as $LP_{s/2}(Q)$ and its value for each sniper as $SP_{s/2}(Q)$. In all proofs, we drop the subscript if $\frac{s}{2} = \frac{d}{2}$. HFTs race to supply liquidity for the first share at $\pm\frac{d}{2}$ iff $LP(1) > SP(1)$.

We consider the first share on the ask side in the proof, and the race on the bid side follows symmetrically. When tick size in binding, both BATs and non-algo traders demand liquidity, so we use non-HFTs to refer to both in the proofs of Lemma 1 and Proposition 1. A non-HFT seller does not change the state of the LOB; an non-HFT buyer, who arrives with probability $\frac{\frac{1}{2}\lambda_I}{\lambda_I+\lambda_J}$ provides a profit of $\frac{d}{2}$ to HFT liquidity supplier; fundamental value jumps up with probability $\frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J}$ and costs an HFT firm $\frac{d}{2}\frac{N-1}{N}$; fundamental value jumps down with probability $\frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J}$, which reduces the value of the current queue position to 0. Therefore:

$$LP(1) = \frac{\frac{1}{2}\lambda_I}{\lambda_I+\lambda_J}\frac{d}{2} + \frac{\frac{1}{2}\lambda_I}{\lambda_I+\lambda_J}LP(1) - \frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J}\frac{d}{2}\frac{N-1}{N} + \frac{\frac{1}{2}\lambda_J}{\lambda_I+\lambda_J}\cdot 0$$

$$LP(1) = \frac{\lambda_I}{\lambda_I+2\lambda_J}\frac{d}{2} - \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{N-1}{N}.$$

Each sniper has a probability of $\frac{1}{N}$ to snipe the stale quote after an upward value jump. A successful sniping leads to a profit of $\frac{d}{2}$, so:

$$SP(1) = \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{1}{N}$$

$$LP(1) > SP(1) \Leftrightarrow \frac{\lambda_I}{\lambda_I+2\lambda_J}\frac{d}{2} - \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{N-1}{N} > \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{1}{N}$$

$$\frac{\lambda_I}{\lambda_J} > 1$$

Therefore, the tick size is binding at $\frac{d}{2}$ if $\frac{\lambda_I}{\lambda_J} > 1$. ■

Proof for Lemma 2

We prove Lemma 2 using mathematical induction.

1. From the proof for Lemma 1,

$$LP(1) = \frac{\lambda_I}{\lambda_I + 2\lambda_J}\frac{d}{2} - \frac{1}{2}\left[1 - \frac{\lambda_I}{\lambda_I + 2\lambda_J}\right]\frac{d}{2}\frac{N-1}{N},$$

   which satisfies equation (3).

2. Suppose that equation (3) holds for some $Q \in \mathbb{N}^+$. The following proof shows that it holds for $Q + 1 \in \mathbb{N}^+$ as well.

$$LP(Q+1) = \frac{\frac{1}{2}\lambda_I}{\lambda_I + \lambda_J}LP(\mathrm{Q}) + \frac{\frac{1}{2}\lambda_I}{\lambda_I + \lambda_J}LP(Q+1) - \frac{\frac{1}{2}\lambda_J}{\lambda_I + \lambda_J}\frac{d}{2}\frac{N-1}{N} + \frac{\frac{1}{2}\lambda_J}{\lambda_I + \lambda_J}\cdot 0$$

$$LP(\mathrm{Q}+1) = \frac{\lambda_I}{\lambda_I+2\lambda_J}LP(\mathrm{Q}) - \frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{N-1}{N} = \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^{Q+1}\frac{d}{2} - \frac{1}{2}\left[\frac{\lambda_I}{\lambda_I+2\lambda_J} - \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^{Q+1}\right]\frac{d}{2}\frac{N-1}{N} -$$

$$\frac{\lambda_J}{\lambda_I+2\lambda_J}\frac{d}{2}\frac{N-1}{N} = \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^{Q+1}\frac{d}{2} - \frac{1}{2}\left[1 - \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^{Q+1}\right]\frac{d}{2}\frac{N-1}{N}.$$

Thus, equation (3) holds with Q replaced by $Q + 1$. Hence equation (3) holds for all $\mathrm{Q} \in \mathbb{N}^+$. ■

Proof of Proposition 2

BATs use flash limit orders when regular limit orders are more costly. We start the proof by finding the boundary between the flash equilibrium and the undercutting equilibrium.

In an undercutting equilibrium, a BAT submits a limit order to an empty LOB $(0,0)$ and changes the state to $(1,0)$; a BAT submits a limit order to $(0,1)$ and changes the state to $(1,1)$. We denote the cost for the first case as $C(1,0)$ and the cost for the second case as $C(1,1)$. Then

$$\begin{cases} C(1,0) = p_1 \cdot C(1,1) + p_1 \cdot C(1,0) + p_2\left(-\frac{d}{6}\right) + p_2 \cdot C(1,0) + p_3\frac{5d}{6} + p_3 \cdot C(1,0) \\ C(1,1) = p_1(-\frac{d}{6}) + p_1 \cdot C(1,0) + p_2\left(-\frac{d}{6}\right) + p_2 \cdot C(1,0) + p_3\frac{5d}{6} + p_3 \cdot C(1,0) \end{cases} \quad \text{(A.1)}$$

Insert Figure A.1 about Here

In equation (A.1) 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 BAT buyer and a BAT seller each arrive each with probability p_1. A BAT buyer posts a limit order on the bid side and changes the state to $C(1,1)$; a BAT seller uses a flash limit order so the state remains at $C(1,0)$. A non-algo buyer and a non-algo seller arrive each with probability p_2. The BAT seller enjoys a negative transaction cost of $-\frac{d}{6}$ when the non-algo buyer takes his liquidity; the non-algo seller hits a HFT's quote on the bid side and does not change the state on the ask side. Upward and downward value jumps occur with probability p_3. An upward jump leads to a sniping cost of $\frac{5d}{6}$, whereas a downward jump does not change the state of the LOB.[15] $C(1,1)$ differs in two ways from $C(1,0)$. First, the arrival of a BAT buyer leads to execution of a sell limit order from a BAT.[16] Second, a downward jump under $C(1,1)$ leads to sniping on the opposite side of the LOB and changes the state to $C(1,0)$.

If an undercutting order gets immediate execution, the cost $-\frac{d}{6}$. $C(1,1)$ must be greater

[15] Here we assume that BATs position their order one tick above the new fundamental value. BATs are able to reposition their orders because they face no competition from other BATs in a short time period.

[16] The execution of this order results from our assumption that BATs do not queue after another limit order at the same price, but the intuition that a longer queue on the bid side increases the execution probability on the ask side holds true generally (Parlour, 1998).

than $-\frac{d}{6}$ because of the cost of being sniped. Therefore, $C(1,0) - C(1,1) = p_1\left(C(1,1) + \frac{d}{6}\right) > 0$. Intuitively, if a BAT chooses to post a sell limit order at $v_t + \frac{d}{6}$ on an empty LOB, he must post a sell limit order when the bid side has a limit order posed by a BAT, because the existence of a limit order on the bid side increases the execution probability for a limit order on the ask side. Note that our model starts with no limit orders from BATs, so $C(1,0) < \frac{d}{6}$ is needed to jumpstart the undercutting equilibrium.

The solution for equation (A.1) is:

$$C(1,1) = \frac{(-2+\beta)\lambda_I + 10\lambda_J}{(2-\beta)\lambda_I + 2\lambda_J}\frac{d}{6} = \frac{(-2+\beta)R + 10}{(2-\beta)R + 2}\frac{d}{6}$$

$$C(1,0) = \frac{d}{6}\left[\frac{\beta R}{R+1}\cdot\frac{(-2+\beta)R+10}{(2-\beta)R+2} + \frac{5-(1-\beta)R}{R+1}\right]$$

$C(1,0) < \frac{d}{6}$ iff $\frac{\beta R}{R+1}\cdot\frac{(-2+\beta)R+10}{(2-\beta)R+2} + \frac{5-(1-\beta)R}{R+1} < 1$, i.e.,

$(2-\beta)R^2 + (-2-4\beta)R - 4 > 0$.

Equation $(2-\beta)R^2 + (-2-4\beta)R - 4 = 0$ has two roots: $R_{1,2} = \frac{1+2\beta\pm\sqrt{4\beta^2+9}}{2-\beta}$,

$$R_2 < 0, R_1 = \frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta}.$$

So BATs choose to undercut when $R > R_1$, because $C(1,0) < \frac{d}{6}$; BATs choose to flash when $R < R_1$.

Above is the boundary between undercutting equilibrium and flash equilibrium. On both sides of the boundary, we let a BAT buyer (seller) use limit order to respond to the other side's limit order. Such a response is both rational and necessary. It is rational because $C(1,1) < C(1,0) = \frac{d}{6}$, thus a limit order response, which costs $C(1,1)$, is strictly better than flash order. It is necessary because otherwise all BATs buyers (sellers) will still use flash orders when off-

equilibrium sell (buy) order is present[17]. The off-equilibrium sell (buy) order will have an execution cost as follows:

$$C(1,0) = p_1\left(-\frac{d}{6}\right) + p_1 \cdot C(1,0) + p_2\left(-\frac{d}{6}\right) + p_2 \cdot C(1,0) + p_3\frac{5d}{6} + p_3 \cdot C(1,0)$$

$$C(1,0) = \frac{d}{6}\frac{5-R}{1+R}$$

$$C(1,0) < \frac{d}{6} \;\Leftrightarrow\; R > 2$$

Thus, undercutting is an optimal deviation when $\frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta} > R > 2$. The existence of deviation proves that, in the $R > 2$ region of flash equilibrium, BATs should use limit orders to respond to the other side's off-equilibrium undercutting order, otherwise the off-equilibrium undercutting order will become a profitable deviation.

However, in the $R < 2$ region of flash equilibrium, BATs should use flash orders to respond to the other side's off-equilibrium undercutting order, because the cost of limit order response, $C(1,1)$, is larger than $\frac{d}{6}$. On the other hand, even if other BATs use flash orders, the deviator is still not profiting.

In other words, regardless of whether $R > 2$ or $R < 2$, the equilibrium outcome is the same, but BATs need to use different rational strategies in off-equilibrium paths to eliminate profitable deviations, thus these deviations will never appear under equilibrium.

To sum up, the complete strategy (including the optimal response to off-equilibrium paths) of a BATs seller under flash equilibrium is:

1. If $2 < R < \frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta}$, use limit order under off-equilibrium path:

[17] In flash equilibrium, any BAT's undercutting limit order is off-equilibrium.

i> If there is no order at $-\frac{d}{6}$, submit a limit sell order at $-\frac{d}{6}$.

ii> Else, submit a limit sell order at $\frac{d}{6}$.

2. If $R < 2$, use flash order under off-equilibrium path:

i> Submit a limit sell order at $-\frac{d}{6}$ regardless of state of the book.

BATs buyer's strategy is symmetric. These strategies will generate the equilibrium outcome sketched in proposition 2.

Predictions on depth and HFT participation follow the proof of Proposition 1. ■

Proof of Proposition 3

1. In Proposition 2, we address the boundary between the flash equilibrium and the undercutting equilibrium.
2. The solution for HFT depth follows from Figure 5 and equation (14). The depth decreases because the revenue from liquidity supply for HFTs decreases. BATs never take HFTs' liquidity at $\frac{d}{2}$, and BATs can also supply liquidity to non-algo traders. The decreased revenue for HFTs also reduces their entry.
3. Equation (14) can be solved for any R and β. Here we give an example for $R = 4$ and $\beta = 0.1$.

First, we assume that all $D^{(i,j)}(1) > 0$. Thus we solve:

$$D^{(0,0)}(1) = p_1 D^{(0,1)}(1) + p_1 D^{(1,0)}(1) + p_2 \cdot \frac{d}{2} + p_2 D^{(0,0)}(1) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,0)}(1) = p_1 D^{(1,1)}(1) + p_1 D^{(1,0)}(1) + p_2 D^{(0,0)}(1) + p_2 D^{(1,0)}(1) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(0,1)}(1) = p_1 D^{(0,1)}(1) + p_1 D^{(1,1)}(1) + p_2 \cdot \frac{d}{2} + p_2 D^{(0,0)}(1) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,1)}(1) = p_1 D^{(0,1)}(1) + p_1 D^{(1,0)}(1) + p_2 D^{(0,1)}(1) + p_2 D^{(1,0)}(1) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

We then obtain:

$D^{(0,0)}(1)$

$$= \frac{8 + 12R + 12\beta R - 4R^2 + 24\beta R^2 + 2\beta^2 R^2 - 12R^3 + 21\beta R^3 - 2\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)}$$

$= 0.2202,$

$D^{(1,0)}(1)$

$$= \frac{8 + 24R + 20R^2 + 6\beta R^2 - 4\beta^2 R^2 + 12\beta R^3 - 5\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)}$$

$= 0.0527,$

$D^{(0,1)}(1) = \frac{8+12R+12\beta R-4R^2+24\beta R^2+2\beta^2 R^2-12R^3+21\beta R^3-5\beta^2 R^3+2\beta^3 R^3-4R^4+7\beta R^4-4\beta^2 R^4+\beta^3 R^4}{2(-16-48R-52R^2+12\beta R^2-4\beta^2 R^2-24R^3+18\beta R^3-8\beta^2 R^3+2\beta^3 R^3-4R^4+7\beta R^4-4\beta^2 R^4+\beta^3 R^4)} = 0.2205,$

$D^{(1,1)}(1)$

$$= \frac{8 + 24R + 20R^2 + 2\beta^2 R^2 + 6\beta R^3 + \beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)}$$

$= 0.0593.$

$D^{(i,j)}(1) > 0$ is satisfied. Therefore, the depth is at least one share in any state of the LOB.

Then we assume all $D^{(i,j)}(2) > 0$. Thus, we solve:

$$D^{(0,0)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,0)}(2) + p_2 D^{(0,0)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,0)}(2) = p_1 D^{(1,1)}(2) + p_1 D^{(1,0)}(2) + p_2 D^{0,0}(2) + p_2 D^{(1,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,1)}(2) + p_2 D^{(0,1)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,0)}(2) + p_2 D^{(0,1)}(2) + p_2 D^{(1,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

We get:

$$D^{(0,0)}(2) = 0.0448,$$

$$D^{(1,0)}(2) = -0.0602 < 0,$$

$$D^{(0,1)}(2) = 0.0451,$$

$$D^{(1,1)}(2) = -0.0561 < 0.$$[18]

We reject the assumption that all $D(2) > 0$. Therefore, under certain states of the LOB, HFTs would not supply the second share of liquidity. We start from the worst state for liquidity suppliers, $(1,0)$, in which a BAT undercuts HFTs on the same side of the LOB, but no BAT undercuts HFTs on the other side of LOB.[19] Therefore, $D^{(1,0)}(2) = 0$ and all other $D^{(i,j)}(2) > 0$. Thus we solve:

$$D^{(0,0)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,0)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_1 D^{(1,1)}(2) + p_2 D^{(0,1)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,1)}(2) = p_1 D^{(0,1)}(2) + p_2 D^{(0,1)}(2) + p_2 D^{(1,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

We obtain:

$$D^{(0,0)}(2) = 0.0475$$

$$D^{(0,1)}(2) = 0.0487$$

$$D^{(1,1)}(2) = -0.0310.$$

However, $D^{(1,1)}(2)$ is still smaller than 0. We further assume that $D^{(1,1)}(2)$ is also 0, i.e., HFTs cancel the second order when BATs submit limit orders on both sides. Therefore,

$$D^{(0,0)}(2) = p_1 D^{(0,1)}(2) + p_1 \cdot 0 + p_2 D^{(0,0)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(0,1)}(2) = p_1 D^{(0,1)}(2) + p_1 \cdot 0 + p_2 D^{(0,1)}(1) + p_2 D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

[18] For briefness, the closed-form solution is not presented, but it is available upon request.

[19] In this state, an HFT liquidity supplier on the ask side cannot trade with the next non-HFT buyer, because a BAT buyer chooses to supply liquidity and changes the state to $(1,1)$, and a non-algo buyer chooses to take the BAT seller's liquidity and changes the state to $(0,0)$.

We obtain:

$$D^{(0,0)}(2) = 0.0488$$

$$D^{(0,1)}(2) = 0.0489.$$

Further calculation shows $D^{(0,0)}(3) = 0, D^{(0,1)}(3) = 0$. We then conclude that $Q^{(0,0)} = Q^{(0,1)} = 2$ and $Q^{(1,0)} = Q^{(1,1)} = 1$ is the solution for equation (14) under R=4 and β=0.1. ■

Proof of Proposition 4

HFTs do not compete to supply liquidity at $\frac{5d}{6}$ when:

$$LP_{\frac{5d}{6}}(1) < SP_{\frac{5d}{6}}(1)$$

$$LP_{\frac{5d}{6}}(1) = p_1 \cdot LP_{\frac{5d}{6}}(1) + p_1 \cdot 0 + p_2 \cdot \frac{5d}{6} + p_2 \cdot LP_{\frac{5d}{6}}(1) - p_3 \frac{d}{6}\frac{N-1}{N} + p_3 \cdot 0$$

$$LP_{\frac{5d}{6}}(1) = \frac{(1-\beta)\lambda_I}{\lambda_I + 2\lambda_J}\frac{5d}{6} - \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{d}{6}\frac{N-1}{N}$$

$$SP_{\frac{5d}{6}}(1) = \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{d}{6}\frac{1}{N}$$

$$\therefore \frac{(1-\beta)\lambda_I}{\lambda_I + 2\lambda_J}\frac{5d}{6} - \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{d}{6}\frac{N-1}{N} < \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{d}{6}\frac{1}{N}$$

$$R < \frac{1}{5(1-\beta)}.$$

Thus, HFTs supply liquidity at $\frac{7d}{6}$. WOLOG, we consider a BATs seller's strategy. The complete strategy (including the optimal response to off-equilibrium paths, see proof of proposition 2) of a BAT seller is:

1. If there is no limit sell order on $\frac{d}{6}, \frac{d}{2}, and\ \frac{5d}{6}$, submit a limit sell order at $\frac{5d}{6}$.
2. Else, if there is no limit buy order on $-\frac{d}{6}$, submit a limit sell order at $-\frac{d}{6}$.

3. Else, there is a limit buy order on $-\frac{d}{6}$ (this is an off-equilibrium path, there are two possible responses, same intuition as the proof of proposition 2)

   i> If $R > 2$, submit a limit sell order at $\frac{d}{6}$, costs $C(1,1)$.

   ii> Else, submit a limit sell order at $-\frac{d}{6}$, costs $\frac{d}{6}$.

If all BATs follow this strategy, no limit sell (buy) order will be present at $\frac{d}{2}$ ($-\frac{d}{2}$) or $\frac{d}{6}$ ($-\frac{d}{6}$). We show that a deviator will suffer a higher execution cost.

Firstly, a BAT seller will not post a limit sell order at $\frac{d}{2}$, because only a non-algo buy order will trade with this seller. The seller's execution cost is:

$$C = p_1 \cdot C + p_1 \cdot C + p_2\left(-\frac{d}{2}\right) + p_2 \cdot C + p_3 \cdot \frac{d}{2} + p_3 \cdot C$$

$$C = \frac{d}{2} \cdot \frac{-(1-\beta)R+1}{(1-\beta)R+1}.$$

Since in flash crash equilibrium $R(1-\beta) < \frac{1}{5}$, the BAT's cost is at least $\frac{d}{2} \cdot \frac{4/5}{6/5} = \frac{d}{3} > \frac{d}{6} = Cost\ of\ flash\ order$. Thus, it is never optimal to submit a limit order at $\frac{d}{2}$.

Secondly, the BAT seller will not post a limit sell order at $\frac{d}{6}$. In this case, non-algo traders and other BAT buyers might trade with the seller: the non-algo trader will execute a buy order and a BAT will execute a flash buy order (when he cannot or finds not optimal to post a limit buy order at $-\frac{d}{6}$). The intuition is similar with formula (A.1) and Figure. A.1, but in the flash crash equilibrium, the BAT seller faces equal or higher costs than in an undercutting equilibrium: The BATs buyer does not have to post a limit buy order in a flash crash equilibrium. The solution of formula (A.1) is:

$$R_1 = \frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta}.$$

However, there is no combination of (R, β) in the flash crash equilibrium that satisfies $R > R_1$.

Finally, the BAT seller will post a sell limit order at $\frac{5d}{6}$. Her cost is:

$$C = p_1 \cdot C + p_1 \cdot C + p_2\left(-\frac{5d}{6}\right) + p_2 \cdot C + p_3 \cdot \frac{d}{6} + p_3 \cdot C$$

$$C = \frac{d}{6}\frac{-5(1-\beta)R+1}{5(1-\beta)R+1} < \frac{d}{6}. \blacksquare$$

Figure 1: Pricing Grid under Large vs. Small Tick Sizes

This figure demonstrates the pricing grids under a large tick size d and a small tick size $\frac{d}{3}$. The fundamental value of the asset is v_t.

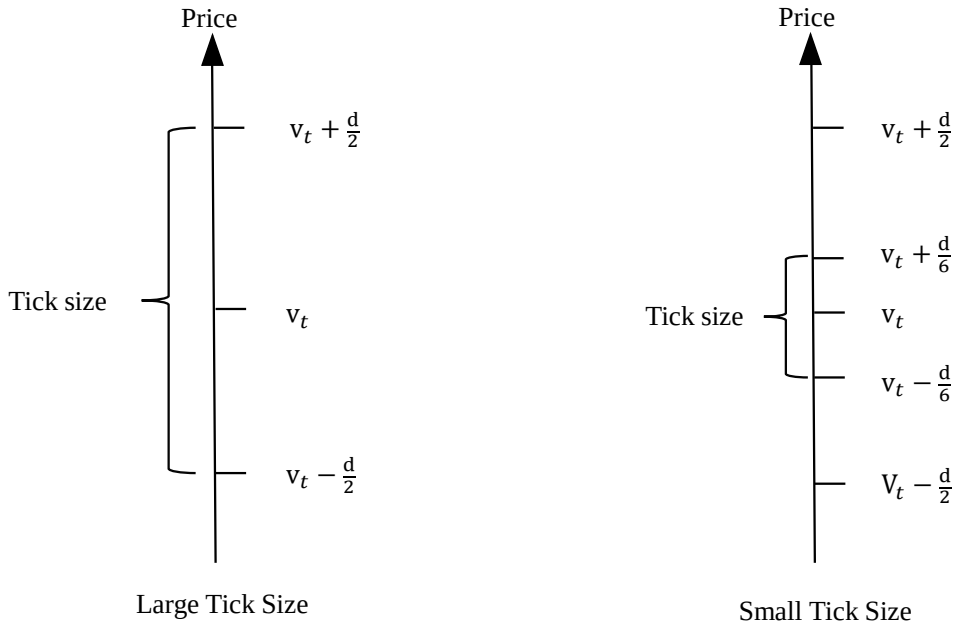

Figure 2: Depth and the Adverse Selection Risk under a Binding Tick Size

This figure demonstrates the relation between Q, the depth at the BBO, and $R = \frac{\lambda_I}{\lambda_J}$ under a binding tick size. An increase in the investor arrival rate (λ_I), or a decrease in intensity of jumps (λ_J), decreases the adverse selection risk and increases the depth. The solid line represents the depth under tick size d and the dashed line represents the depth under tick size $\frac{d}{3}$.

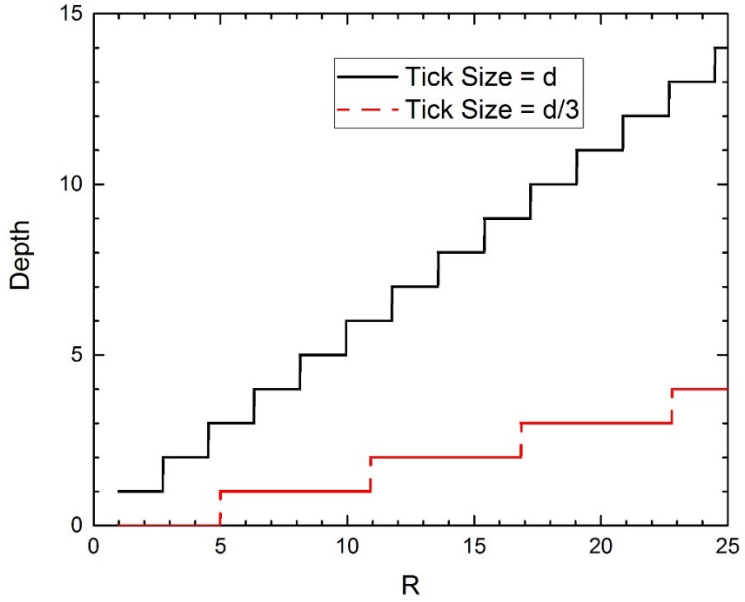

Figure 3: Bid-ask Spread Quoted by HFTs under a Small Tick Size

This figure demonstrates the half bid-ask spread quoted by HFTs as a function of β (the fraction of BATs) and $R \equiv \frac{\lambda_I}{\lambda_J}$ (the arrival intensity of non-HFTs relative to the value jump, a measure of adverse selection risk). When $R \geq 5$, adverse selection risk is low and the tick size is binding. HFTs quote a half bid-ask spread $\frac{d}{6}$ and the spread is independent of the fraction of BATs. When $R < 5$, HFTs' quoted bid-ask spreads weakly increase with the fraction of BATs and adverse selection risk.

Figure 4: The Undercutting and the Flash Trading Equilibrium

This figure demonstrates two types of equilibrium, undercutting equilibrium and flash equilibrium, when HFTs' ask price is at $v_t + \frac{d}{2}$ and their bid price is at $v_t - \frac{d}{2}$. In the undercutting equilibrium, BATs place limit buys at $v_t - \frac{d}{6}$ and limit sells at $v_t + \frac{d}{6}$. These limit orders undercut the BBO by one tick and establish price priority in the LOB. In the flash equilibrium, BATs place limit buys at $v_t + \frac{d}{6}$ and limit sells at $v_t - \frac{d}{6}$. These orders cross the midpoint and immediately attract market orders from HFTs. BATs are more likely to cross the midpoint when the fraction of BATs (β) is high or when the arrival intensity of non-HFTs relative to a value jump ($R \equiv \frac{\lambda_I}{\lambda_J}$) is low, because a high β and a low R reduce the potential for a limit order executing with non-HFTs before a value jump. To jumpstart an undercutting equilibrium, the expected transaction cost for a limit order that undercuts one tick must be lower than $\frac{d}{6}$. The short-dashed line, $C(1,0) = \frac{d}{6}$, illustrates the boundary for such a condition.

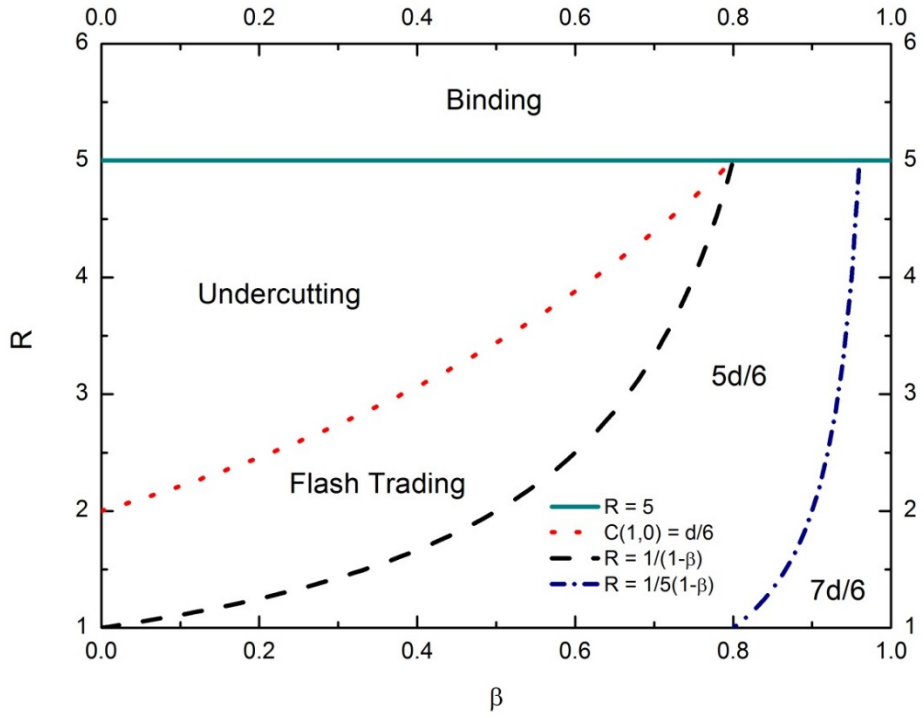

Figure 5: States and Profits for HFT Liquidity Suppliers with the Q^{th} Position on the Ask Side

This figure illustrates the dynamics of HFT queuing on $v_t + \frac{d}{2}$. In state (*i*, *j*), the number of undercutting BAT orders on the ask side is *i*, while the number on the bid side is *j*. BB and BS represent the arrival of BATs' buy and sell limit orders, NB and NS represent the arrival of non-algo traders' buy and sell market orders, and UJ and DJ denote the upward and downward value jumps. The number next to the event is the immediate payoff of the event.

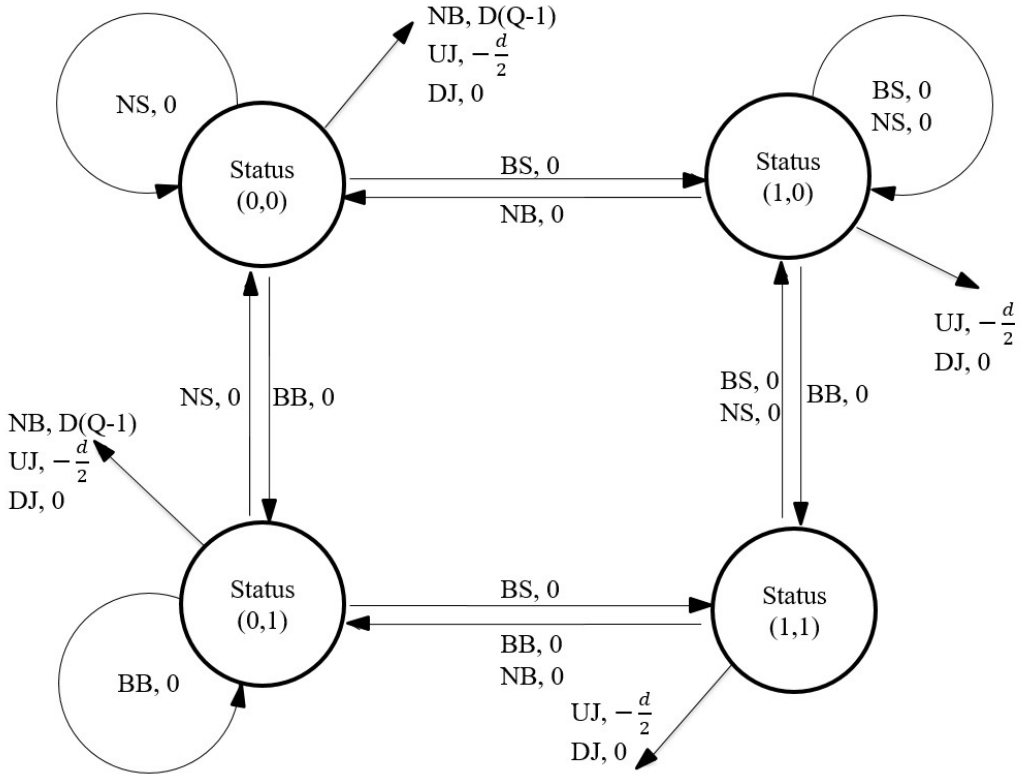

Figure 6: Value of Liquidity Supply and Stale-Queue Sniping and Queue Length
The x-axis is the value of HFT liquidity supply (*LP*) and stale-queue sniping (*SN*) for the four states of the LOB. In $Q(0,0)$, no BATs undercut HFTs in the LOB. In $Q(1,0)$, BATs undercut HFTs on the same side of the book. In $Q(0,1)$, BATs undercut HFTs on the opposite side of the book. In $Q(1,1)$, BATs undercut both sides of the book. LP decreases in the queue position, while SN increases in the queue position. HFTs supply liquidity as long as *LP* > *SN*.

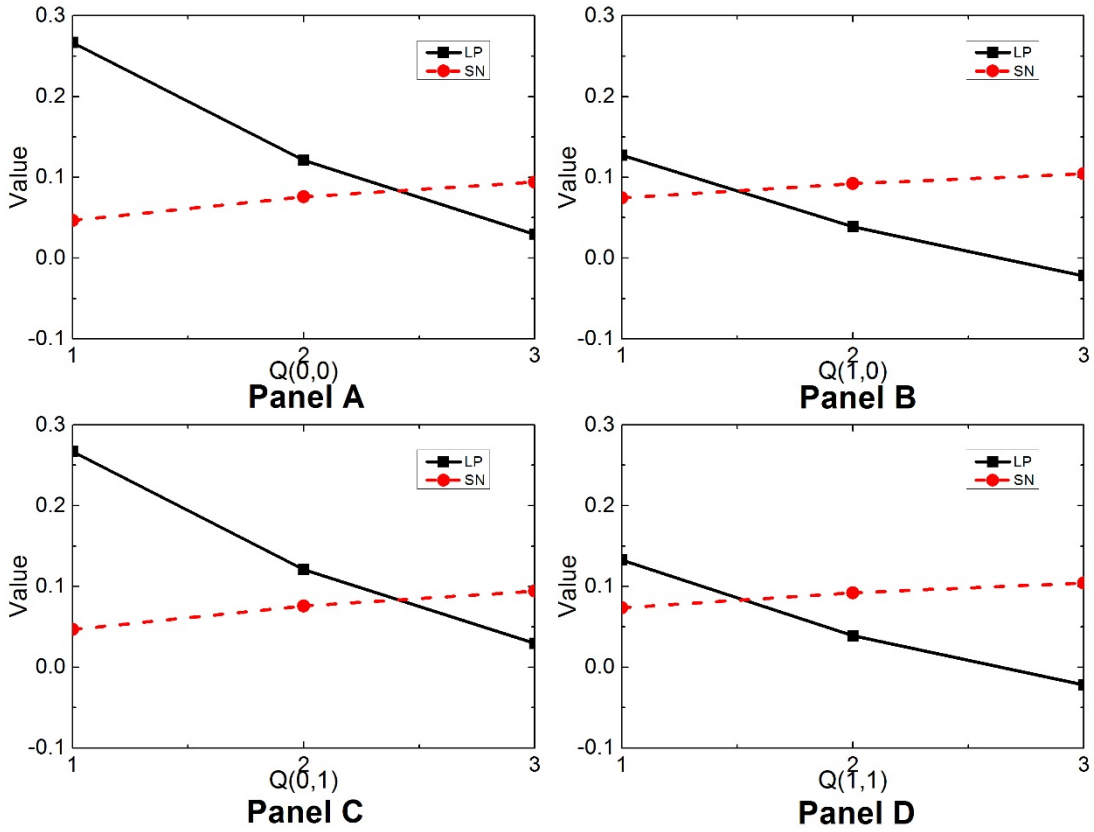

Figure 7. Flash Crash Intensity

This figure shows the intensity of mini-flash crashes with respect to the fraction of BATs. We normalize the highest intensity as 1. For each β, we uniformly draw 100 samples from [1,5] as $\frac{\lambda_I}{\lambda_J}$, which is the support of the adverse selection risk in our paper. For each $\frac{\lambda_I}{\lambda_J}$, we simulate 100,000 trades. For all these 10 million simulations, we count the number of trades hitting the stub quotes relative to the total number of trades. The line with squares shows the intensity for total crashes. The line with circles shows that the majority of mini-flash crashes occur after a value jump (and a small fraction of crashes occur after BATs' liquidity being consumed by non-algos). The line with triangles shows that trading halts reduce the number of mini-flash crashes. We impose trading halts after each value jump, and the market reopens when the market receives 10 orders.

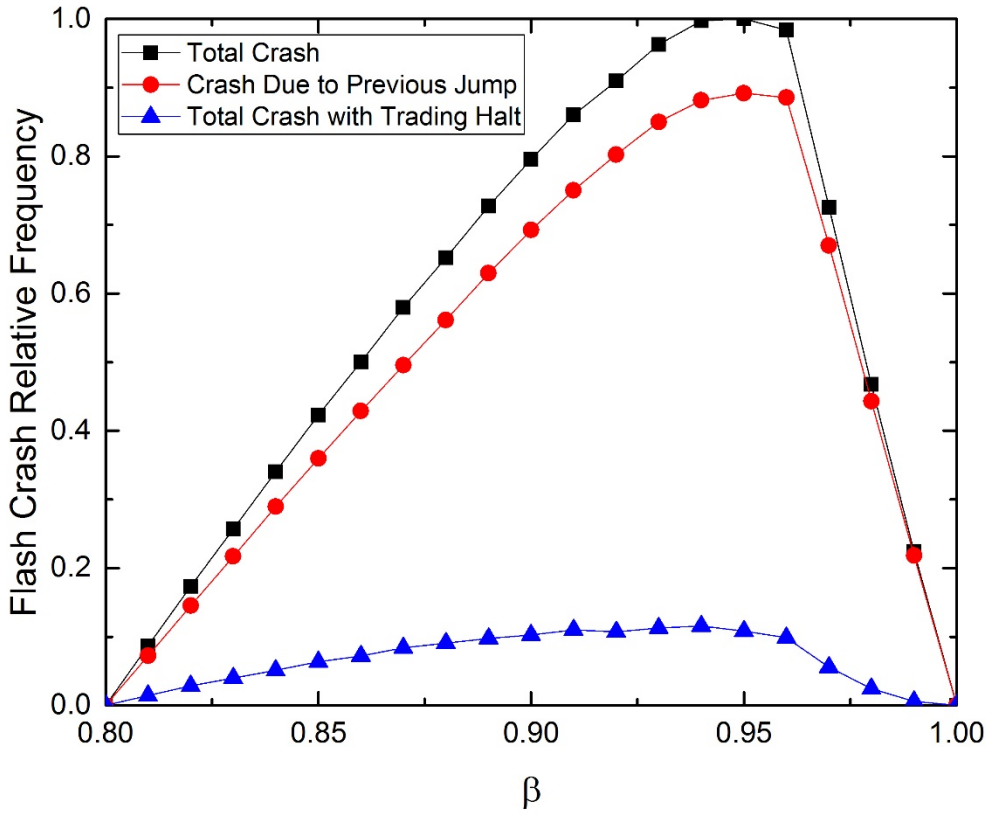

Figure A.1: States and Profits for BATs on the Ask Side

This figure illustrates the dynamics of the BAT seller who posts a limit order at $v_t + \frac{d}{6}$. State (i, j) implies the number of BAT orders on the ask and bid sides if the BAT seller add a regular limit order. BB and BS imply the arrival of BAT buy and sell orders, respectively. NB and NS are arrivals of non-algo buy and sell orders, respectively, while UJ and DJ are upward and downward jumps, respectively. For example, submitting a sell limit order to an empty LOB leads to state $(1,0)$, and the exepceted cost for the limit order is $C(1,0)$. If a BAT submits a limit order when a limit order already exists on the opposite side of the LOB, the state after submission is $(1,1)$ and the cost is $C(1,1)$.