Who Provides Liquidity and When:

An Analysis of Price vs. Speed Competition on Liquidity and Welfare

Mao Ye[1]

Abstract

I model the interaction between buy-side algorithmic traders (BATs) and high-frequency traders (HFTs). When the minimum price variation (tick size) is small, BATs dominate liquidity provision by establishing price priority over HFTs in the limit order book (LOB), because providing liquidity is less costly than demanding liquidity from HFTs. A large tick size, however, constrains price competition and encourages HFTs to provide liquidity by establishing time priority. An increase in adverse selection risk raises the unconstrained bid-ask spread, reduces tick size constraints, and discourages HFTs' liquidity provision. An increase in tick size increases transaction costs and harms liquidity demanders, but it does not benefit liquidity providers because the costs of speed investments dissipate the rents resulting from the tick size. I predict that mini-flash crashes are more likely to occur for stocks with a smaller tick size and higher adverse selection risk. I suggest that the literature should not use the message-to-trade ratio as a cross-sectional proxy for HFTs' liquidity provision because stocks with more liquidity provided by HFTs have a lower message-to-trade ratio.

[1] I thank Hengjie Ai, Malcolm Baker, Hank Bessembinder, Thierry Foucault, Maureen O'Hara, Neil Pearson, Brian Weller, Chen Yao, Bart Yueshen, Marius Zoican, and participants at the Carlson Junior Conference at the University of Minnesota 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). Send correspondence to Mao Ye, 340 Wohlers Hall, 1206 S. 6th Street, Champaign, IL, 61820. Email: maoye@illinois.edu. Telephone: 217-244-0474. I thank Xin Wang and Sida Li for excellent research assistance.

To minimize their transaction costs, buy-side institutions, such as mutual funds and pension funds, extensively use computer algorithms to execute their trades (Frazzini, Israel, and Moskowitz 2014; O'Hara 2015). These buy-side algorithmic traders (BATs) differ from high-frequency traders (HFTs) in two fundamental ways (Hasbrouck and Saar 2013; Jones 2013; O'Hara 2015). First, BATs may provide liquidity, but their goal is to minimize transaction costs of portfolio rebalancing rather than to profit from the bid-ask spread; second, BATs are faster than humans, but are slower than HFTs (O'Hara 2015). Although buy-side institutions are major players in financial markets in the United States, their trading algorithms do not have an independent identity in existing models. The model I build in this paper bridges the gap between the economic reality and the theoretical literature by considering three types of traders: HFTs, BATs, and non-algorithmic (non-algo) traders. I use the model to address three questions: 1) Who provides liquidity and who demands liquidity, and when? 2) What drives speed competition? and 3) Does speed competition in the limit order book (LOB) improve liquidity and social welfare?

In my model, HFTs and two types of non-HFTs (BATs and non-algo traders) trade a security in a dynamic LOB. A liquidity provider in the LOB submits limit orders (offers to buy or sell a stock at a specified price and quantity), and a liquidity demander accepts a limit order using a market order. Limit order execution follows the price-time priority rule. Limit buy orders with higher price or limit sell orders with lower price are executed before those at less aggressive prices; for limit orders queuing at the same price, the time priority rule gives precedence for the order arriving first. HFTs have no private value to trade, but simply provide or demand liquidity when its expected profit is above 0. Non-HFTs, who arrive at the market through a compound Poisson process, have inelastic demand to buy or sell one unit of a security. Some of the non-HFTs are BATs, who can choose to provide or demand liquidity to minimize transaction costs, and the rest

are non-algo traders, who only demand liquidity.

In the model, two exogenous variables, adverse selection risk and tick size, determine who provides liquidity. The fundamental value of a security is public information in the model, but continuous time trading generates adverse selection risk for liquidity providers (Budish, Cramton, and Shim (2015; BCS hereafter). Even if liquidity providers cancel stale quotes immediately after the value jump, orders to snipe the stale quotes may arrive before their cancellation. Because non-HFTs trade for liquidity reasons and value jump leads to adverse selection of stale quotes, I use the arrival rate of non-HFTs relative to the intensity of value jumps to measure adverse selection risk. [2] If price is continuous, the adverse selection risk dictates the break-even bid-ask spread. The U.S. Securities and Exchange Commission's (SEC's) Rule 612, however, impose discrete tick size (minimum price variation), which prevents the bid-ask spread from reaching its competitive level.[3]

Tick size and the time priority rule then drive a queuing channel of speed competition in liquidity supply. Tick size creates rents for liquidity provision, the rents generate the queue of liquidity providers, and the rents in the queue are allocated following the time priority rule. I predict that HFTs are the dominate liquidity providers when tick size is large, because a large tick size constrains price competition. In addition, a decrease in adverse selection risk reduces the break-even spread relative to the tick size, which also constrains price competition and incentivizes speed competition.

As a small tick size or a high adverse selection risk drives HFTs' break-even bid-ask spread above one tick, BATs no longer demand liquidity from HFTs. One way to reduce transaction costs

[2] In this paper, adverse selection risk refers to the degree of adverse selection for the whole market. Each trader's adverse selection cost also depends on her execution priority and strategies of other traders.

[3] As tick size is one cent for all stocks valued at $1.00 or above, the relative tick sizes for low-priced stocks are larger than those for high-priced stocks. Consequently, the comparative statics with respect to tick size explains the differences in the trading environments for low-priced and high-priced stocks.

is to provide liquidity to HFTs. Without loss of generality, consider BATs' decision to buy and HFTs' decision to sell. HFTs incur adverse selection costs when they use sell limit orders, but not when they accept buy limit orders from BATs. Therefore, BATs can submit a limit buy order with a price slightly below HFTs' ask price (limit price to sell) and immediately prompts HFTs to submit market orders to sell. This type of limit order, which I call a "flash" limit order, strictly dominates market orders, because flash limit order is also immediately executed but at a lower cost.

In the equilibrium where BATs use flash limit orders, tick size creates rents for demanding liquidity, because it prevents BATs from submitting limit orders with the exact price that prompts HFTs to demand liquidity. BATs have to offer limit orders with more aggressive prices, and the difference between BATs' quoted prices and HFTs' valuation generates the race for HFTs to demand liquidity. Under certain parameters, BATs can further reduce transaction costs by providing liquidity to non-HFTs. This undercutting strategy for limit orders works to establish price priority over HFTs, yet is not so aggressive as to prompt HFTs to take liquidity.

Existing literature on speed competition focuses on the role of information. On the one hand, speed can reduce adverse selection costs for liquidity providers and improve liquidity; on the other hand, speed can allow HFTs to adversely select other traders, which has a detrimental effect on liquidity [see Jones (2013), Biais and Foucault (2014), and Menkveld (2016) for surveys]. I incorporate these two traditional speed competition in my model, but the main drivers of the model are two types of speed competitions that relate not to information but to tick size. By identifying liquidity supply and demand unrelated to information, I can reconcile a number of contradictions between existing channels of speed competition and empirical results.

Carrion (2013), Hoffmann (2014) and Brogaard et al. (2015) show that speed reduces HFTs'

intermediation costs, particularly adverse selection costs. The reduced costs imply that HFTs should quote a tighter bid-ask spread than non-HFTs, should have a competitive advantage in providing liquidity for stocks with higher adverse selection risk, and should dominate liquidity provision when tick size is small, because the constraints to offer better prices is less binding. Yet Brogaard et al. (2015) find that slow traders quote a tighter bid-ask spread than fast traders, and Yao and Ye (2016) find that a reduction in tick size and an increase in adverse selection risk reduces HFTs' fraction of liquidity provision. My model helps to reconcile these contradictions. Slow traders have higher incentives to quote a tighter spread because they are less likely to establish time priority over HFTs; when tick size is small or adverse selection risk is high, non-HFTs are able to establish price priority over HFTs, because the break-even bid-ask spread is large relative to tick size; a large tick size or low adverse selection risk constrains price competition and increases HFTs' liquidity provision through time priority.

Yao and Ye (2016) find that the message-to-trade ratio, a widely-used proxy for HFTs' liquidity provision (Biais and Foucault 2014), is negatively correlated with the true measure in cross-section. My model rationalizes this negative correlation. HFTs dominate liquidity provision for stocks with larger tick sizes, but they also have less incentive to cancel orders, which results in a loss of their queue positions. A smaller tick size allows BATs to establish price priority and reduces HFTs' liquidity provision, but cancellations increase because price competition occurs at a finer grid. This theoretical intuition, along with the empirical evidence in Yao and Ye (2016), suggests that the message-to-trade ratio should not be used as a cross-sectional proxy for HFT activities.[4]

My model provides an interpretation for mini-flash crashes, defined as sharp price

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

movements in one direction followed by quick reversion (Biais and Foucault 2014), and is predictive of their cross-sectional and time series variations. In cross-section, mini-flash crashes are more likely to occur for stocks with smaller tick size or higher adverse selection risk. HFTs have fewer liquidity demanders for these stocks, because BATs no longer demand liquidity from HFTs, and because non-algo traders' market orders may execute first against BATs' limit orders. HFTs' limit orders then face lower execution probability and higher adverse selection costs, forcing HFTs to quote wide bid-ask spreads (stub quotes) to protect themselves from sniping. Yet BATs provide liquidity only as the need arises to trade. It is then possible for incoming market orders to hit HFTs' stub quotes, which causes a mini-flash crash. In time series, I find that a downward (upward) 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 LOB.

My model extends BCS along two dimensions. BCS considers *continuous* prices, while I consider *discrete* prices to reflect the tick size regulation. I find that an increase in tick size raises transaction costs for liquidity *demanders*, but does not benefit liquidity *providers* as speed investment dissipates all the rents created by tick size. Along with Chao, Yao, and Ye (2016) and Yao and Ye (2016), I question the rationale to increase the tick size to five cents, proposed by the 2012 U.S. Jumpstart Our Business Startups Act (the JOBS Act). Proponents of increasing the tick size argue that a larger tick size increases liquidity, discourages HFTs, increases market-making profits, supports sell-side equity research and, eventually, increases the number of initial public offerings (IPOs) (Weild, Kim, and Newport 2012). My results show that an increase in the tick size reduces liquidity, encourages HFTs, and allocates resources to latency reduction.

In BCS, non-HFTs only demand liquidity, but in my model I allow a fraction of non-HFTs

to choose between demanding and supplying liquidity. By taking the initial step to model sophisticated non-HFTs, I develop new predictions and perceptions. For example, liquidity demanding from HFTs generally has a negative connotation, because liquidity demand from HFTs usually adverse selects liquidity suppliers (BCS; Foucault, Kozhan, and Tham Forthcoming; Menkveld and Zoican Forthcoming). In my model, BATs can use aggressive limit orders to prompt HFTs to demand liquidity, which involves no adverse selection and reduces BATs' transaction costs. This may be one reason for why Latza, Marsh, and Payne (2014) find that limit orders executed within 50 milliseconds after their submission incur no adverse selection costs.

This article is organized as follows. In Section 1, I describe the model. In Section 2, I present the benchmark model with a large and binding tick size. In Section 3, I provide an overview of equilibrium types under a small tick size. In Section 4, I analyze the flash equilibrium and the undercutting equilibrium. In Section 5, I offer a theoretical interpretation of flash crashes and predict their occurrence in cross-section and time series. In Section 6, I summarize the empirical predictions and policy implications of this paper. I conclude the paper in Section 7. All proofs are presented in the Appendix.

1. Model

In my model, the stock exchange operates as a continuous LOB. Traders can choose to be liquidity providers by submitting limit orders that specify a price, a quantity, and the direction of trade (buy or sell), or they can choose to be liquidity demanders and accept the conditions of the existing limit orders. Execution precedence for liquidity suppliers follows the price-time 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 orders arriving

later. 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.

My 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. To simply the model, I assume that v_t is common knowledge. Even so, liquidity providers are still subjected to adverse selection risk when they fail to update stale quotes after value jumps. All traders observe the common value jump with a small latency, [5] but can choose to reduce the latency to 0 by investing in a speed technology with cost c_{speed} per unit of time.

My model includes HFTs and two types of non-HFTs: BATs and non-algo traders. HFTs place no private value on trading. They buy x when its price is below v_t, or sell x when its price is above v_t. One such profit opportunity occurs after the value jump, when HFTs can snipe the stale quotes. HFTs can also choose to be liquidity providers to profit from the bid-ask spread. Each non-HFT has to buy or sell one unit of x, each with probability $\frac{1}{2}$. The speed choices of HFTs and non-HFTs follow directly with my assumption on their trading motivations. HFTs need to invest in speed technology because they constantly monitor the market for opportunities to be the first to provide or demand liquidity, and non-HFTs do not invest in speed technology because they only arrive at the market once.

My model has two major extensions on BCS. First, non-HFTs in the BCS model submit only market orders. In my model, I 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,

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

use only market orders. Second, BCS assumes continuous pricing in their model, whereas I 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 3-5, I 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 coarse and fine pricing grids.

Following the dynamic LOB literature (e.g., Goettler, Parlour, and Rajan 2005, 2009; Rosu 2009; Colliard and Foucault 2012), I examine the Markov perfect equilibrium, in which traders' actions condition only on state variables and events at t. LOB is a natural state variable and represents the history of the play (Goettler, Parlour, and Rajan 2005). I 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:

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

To reduce the number of states, I make a technical assumption that BATs never queue after existing limit orders at the same price. I can relax the one-share restriction by assuming that BATs can queue until the n^{th} position, but such an extension only increases the number of LOB states to $(1+n)^2$ without conveying new intuition.[6] This assumption can be justified by a delay cost for

[6] Each side of the LOB can have zero to n shares. There are $(1+n)^2$ possible states for each side of the LOB.

BATs. Non-HFTs in the BCS model never use limit orders, which can be justified by an infinitely large delay cost (Menkveld and Zoican Forthcoming). My extension effectively reduce the delay cost to allow BATs to submit limit orders. I do not include exogenous delay cost as a parameter because it would dramatically complicate the proof of the model. In addition, in Section 4 I show that the exact size of the delay cost hardly plays any role for BATs' choice between limit order and market order. The other justification for assumption is BATs' trading motivation and speed disadvantage. BATs do not consistently monitor the market for opportunities to provide liquidity; even if they do, they are slower than HFTs to submit limit orders. Therefore, BATs cannot establish time priority at the same price as HFTs. An order with less time priority has lower probability of execution and higher probability of being sniped, both of which reduces BATs' incentives to queue. I assume that BATs never queue to reflect this intuition in a parsimonious way.

2. Benchmark: Binding at One Tick under a Large Tick Size

My analysis starts from $\Delta_0 = d$. HFTs can choose to be *liquidity providers*, who profit from the bid-ask spread. The outside option for liquidity providers is to be *stale-quote snipers*, who profit by taking liquidity from stale quotes after a value jump. In BCS, the equilibrium bid-ask spread equalizes the 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 $\frac{\lambda_I}{\lambda_J} > 1$, the 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 the profit from stale-quote sniping.

Because non-HFTs trade for liquidity reasons but value jumps lead to adverse selection risk for stale quotes, $\frac{\lambda_I}{\lambda_J}$ measures adverse selection risk in my model. As in BCS and Menkveld and Zoican (Forthcoming), this adverse selection risk comes from the speed to respond to public information but not from exogenous information asymmetry (e.g., Glosten and Milgrom 1985; Kyle 1985). As the 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 provision for the first share strictly more profitable than stale-quote sniping.[7] The rents for liquidity provision then trigger the race to win time priority in the queue. BATs do not have a speed advantage to win the race to provide liquidity, they demand liquidity as non-algo traders do. As a result, Lemma 1 does not depend on β.

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

I evaluate HFTs' value of liquidity provision and stale-quote sniping for each queue position, though I allow an HFT to provide liquidity at multiple queue positions and to snipe shares in any other positions where the HFT is not a liquidity provider. I denote the value of liquidity provision 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.[8] A

[7] Throughout this paper, I consider the case in which $\frac{\lambda_I}{\lambda_J} > 1$ 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 3-5, where I reduce the tick size to $\Delta_1 = \frac{d}{3}$.

[8] This result no longer holds in Section 4, when the tick size is not binding at one tick.

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 provider, $LP(0)=\frac{d}{2}$. 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 is determined in equilibrium) race to snipe the stale quote. The loss from being sniped is $\frac{d}{2}$, and the probability of being sniped is $\frac{N-1}{N}$. When v_t jumps downward, the liquidity provider cancels his order and joins the race to provide 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 Provision). The value of liquidity provision 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

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

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 providers. 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 a value jump, the liquidity provider has a probability of $\frac{1}{N}$ to cancel the stale quote, but failing to cancel the stale quote before 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 provider. $LP(Q)$ decreases in Q, because an increase in a queue position reduces execution probability and increases the cost of being sniped.

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 to act.

HFTs race to provide 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 of the race, who can only be the snipers during value jumps. Equation (5) determines the equilibrium length:

$$\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q \frac{d}{2} - \frac{1}{2}\left[1 - \left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)^Q\right]\frac{d}{2} > 0.^{10} \quad (5)$$

The solution for equation (5) is:

$$\begin{aligned} Q^* &= \max\left\{Q \in \mathbb{N}^+ \text{ s.t.} \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q \frac{d}{2} - \frac{1}{2}\left[1 - \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q\right]\frac{d}{2} > 0\right\} \\ &= \max\left\{Q \in \mathbb{N}^+ \text{ s.t.} \left(\frac{\lambda_I}{\lambda_I + 2\lambda_J}\right)^Q > \frac{1}{3}\right\} \\ &= \left\lfloor \log_{\left(\frac{\lambda_I}{\lambda_I+2\lambda_J}\right)} \frac{1}{3} \right\rfloor. \end{aligned} \quad (6)$$

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 the proxy for adverse selection risk (Glosten and Milgrom 1985; Stoll 2000). Yet Yao and Ye (2016) find that bid-ask spread is exactly 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.[11]

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

[10] Liquidity providers in traditional limit order models continue to add limit orders until their marginal profits become zero (Seppi 1997; Parlour and Seppi 2003, 2008). HFTs in my model stop increasing the depth as long as $SN(Q^*+1) > LP(Q^*+1)$, even if the marginal profit for the $(Q^*+1)^{th}$ unit is greater than zero. This is a consequence of HFTs' option to be a sniper. Because sniping only reallocates rents among HFTs, the total rents for HFTs come only from non-HFTs. Because the liquidity provider for the $(Q^*+1)^{th}$ position earns below average rents, HFTs find it optimal to leave the $(Q^*+1)^{th}$ position empty until a market order moves the queue forward. Each HFT expects average rents in the race for the Q^{*th} position, and the winner obtains above-average rents $\left(LP(Q^*) > SN(Q^*)\right)$. In summary, the depth in my model is different from that in traditional LOB models because I allow liquidity providers to demand liquidity.

[11] 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.

per unit of time. New HFTs continue to enter the market until:

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

In Proposition 1, I 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 provide 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:

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

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

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 my model raises the profit of liquidity provision above the profit of stale-quote sniping for the first share, and generates the queue for liquidity provision.

In BCS, the number of HFTs is determined by $\lambda_I \frac{s^*}{2} - Nc_{speed} = 0$, where s^* is the break-even bid-ask spread. In my model, N is determined by $\lambda_I \frac{d}{2} - Nc_{speed} > 0$. When tick size is binding, $d > s^*$, so tick size leads to more entries of HFTs. Taken together, my 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.

3. Equilibrium Types under a Small Tick Size

Starting from this section, I show that a reduction in tick size to $\frac{d}{3}$ prompts BATs to become liquidity providers by establishing 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 provision 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:[12]

$$\begin{cases} \frac{d}{2} & \frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5 \\ \frac{5d}{6} & \frac{1}{5(1-\beta)} < \frac{\lambda_I}{\lambda_J} < \frac{1}{1-\beta} \\ \frac{7d}{6} & \frac{\lambda_I}{\lambda_J} < \frac{1}{5(1-\beta)} \end{cases}^{13} \tag{11}$$

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 and the break-even bid-ask spread. 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 provision. Interestingly, when the adverse section risk or the fraction of BATs is high, HFTs effectively cease liquidity provision by quoting a bid-ask spread that is wider than the size of a jump. The following sections elaborate the equilibrium types when tick size is not binding.

Insert Figure 3 about Here

4. Make-take Spread, Flash Equilibrium, and Undercutting Equilibrium

In this section, I study the case that $\frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5$, 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 spread. In Section 4.1, I explain why BATs always choose to provide liquidity when tick size is not binding. Section 4.2 shows how adverse selection risk affects BATs' limit order prices.

[12] I defer the derivation of the boundary condition for HFTs' bid-ask spread to Sections 4-5.

[13] Another way to bypass tick size constraints is to randomize quotes immediately above and below the break-even bid–ask spread. In this paper, I consider only stationary HFT quotes.

4.1 Make-take spread

In this subsection, I show that BATs never take liquidity from HFTs when $\frac{1}{1-\beta} < \frac{\lambda_I}{\lambda_J} < 5$. Without loss of generality, I consider the decision for a BAT who wants to buy, and the intuition is the same for a BAT who wants to sell. 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 prompts HFTs to submit market orders to sell. This flash limit order gets immediate execution like market orders, but with lower cost.

Flash limit orders exploit the make-take spread, a new concept I develop in this paper. A HFT's limit price to sell includes the adverse selection risk. The HFT would accept a lower price for a market sell order, because immediate execution reduces adverse selection risk. The make-take spread 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).

As HFTs would take liquidity for any order that crosses the midpoint, make-take spread happens to be half of the bid-ask spread in my model. Two intuitions, however, should hold more generally. First, as the limit order price of a BAT approaches that of the HFTs, it prompts HFTs to demand liquidity. Second, the make-take spread should nest in the bid-ask spread, which implies that BATs can no longer find a price level to take advantage of the make-take spread if the bid-ask spread is exactly one tick.

In most market microstructure models, traders cannot take advantage of the make-take spread, because liquidity suppliers cannot demand liquidity. This assumption reflects the economic reality at the time, when some exchanges even prohibited market makers from demanding liquidity (Clark-Joseph, Ye, and Zi Forthcoming). In modern electronic platforms, every trader can provide liquidity, and liquidity providers' face very limited, if any restriction to demand liquidity (Clark-

Joseph, Ye, and Zi Forthcoming). O'Hara (2015) points out that demand or supply liquidity now simply implies "cross the spread" or "do not cross the spread." She 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.

4.2 Flash versus regular limit orders

Although flash orders strictly dominate market orders, BATs can choose to submit 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. In this subsection, I consider BATs' choice between flash and regular limit orders. A flash limit order (e.g., $v_t + \frac{d}{6}$ to buy) executes immediately, but it costs $\frac{d}{6}$ relative to the midpoint. A regular limit order (e.g., $v_t - \frac{d}{6}$ to buy) captures a half bid-ask spread of $\frac{d}{6}$ if executed against a non-HFT, but it is also subject to adverse selection risk. A higher adverse selection risk, therefore, increases the costs of regular limit orders and prompts BATs to submit flash limit orders. BATs choose flash limit orders when β increases, because a large β reduces the probability of execution before a value jump. Figure 4 shows the boundary between the flash equilibrium, in which BATs choose flash limit orders, and the undercutting equilibrium, in which BATs choose regular limit orders.

Insert Figure 4 about Here

In both equilibria, BATs quote a tighter spread than HFTs. This theoretical prediction is in the opposite direction to existing channels of speed competition, but is supported by empirical evidence. Conventional wisdom maintains that HFTs should quote tighter bid-ask spreads than

non-HFTs, because speed reduces their adverse selection cost (Jones 2013; Menkveld 2016). Yet Brogaard et al. (2015) and Yao and Ye (2016) find that non-HFTs quote tighter spreads than HFTs. I provide possible explanations based on their motivations and trading speed. BATs' motivation to complete a trade allows them to submit more aggressive limit orders, as long as the limit orders are less costly than market orders; BATs' speed disadvantage prevents them from achieving time priority in the queue and incentivizes them to undercut HFTs.[14]

4.2.1 Flash equilibrium

Proposition 2 characterizes the flash equilibrium in which BATs use flash limit orders.

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 provide 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:

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

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

[14] I model time priority parsimoniously by assuming that BATs do not queue after existing limit orders at the same price, but the intuition similar if I assume that they only queue until the Q^{th} as long as Q is less than the maximum depth offered by HFTs.

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

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 provide liquidity race to cancel the stale quotes, whereas stale-quote snipers race to pick off the stale quotes.

Proposition 2 identifies a second 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 limit buy price above fundamental value would prompt HFTs to sell. 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 taking liquidity.

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

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 provide 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.

In flash equilibrium, the LOB only has one stable state. Next, I discuss the undercutting equilibrium, in which LOB transits across different states.

4.2.2 Undercutting equilibrium

In the undercutting equilibrium, BATs submit limit orders that remain in the LOB. HFTs' and BATs' decisions now depend on the state of the book (i, j). Here i represents the number of BATs' limit orders on the same of the LOB, and j stands for the number of BATs' limit orders on the opposite side of the LOB. For example, for a BAT or 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.

(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

HFTs' and BATs' strategies depends on the states of LOB and the probability of future events. Their actions also lead to state transitions of the LOB, which are shown in Figure 5. To simplify the notation, I denote the probability of events as follows. $p_1 \equiv \frac{1}{2} \cdot \frac{\lambda_I \beta}{\lambda_I + \lambda_J}$ denotes the arrival probability of a BAT buyer or seller, $p_2 \equiv \frac{1}{2} \cdot \frac{\lambda_I (1-\beta)}{\lambda_I + \lambda_J}$ denotes the arrival probability of a non-algo trader to buy or sell, and $p_3 \equiv \frac{1}{2} \cdot \frac{\lambda_J}{\lambda_I + \lambda_J}$ denotes the probability of an upward or downward value jump. In Proposition 3, I summarize the undercutting equilibrium.

Insert Figure 5 about Here

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: Define $D^{(i,j)}(Q) \equiv LP^{(i,j)}(Q) - SN^{(i,j)}(Q)$. The following system of equations determines the equilibrium depth in each state.
        i. Difference in value between the liquidity provider and the stale-queue sniper in each state:

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

        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 difference in value:

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

    c. Entry: 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.

Even under the simplifying assumption that BATs do not queue behind existing limit orders at the same price, the solution for the equilibrium depth offered by HFTs is rather complex. The depth for each state (i,j) depends on the value difference between liquidity provision and stale-quote snipping in this state, $D(i,j)$, which then depends recursively on the value difference in other states of the LOB and the probability of transition. For example, consider HFTs' decision on the ask side under state $(0,0)$. The six types of events defined in equation (1) change $D^{(0,0)}(Q)$ in the following way. BAT buyers (sellers) arrive with probability p_1; a BAT buyer chooses to undercut HFTs on the bid side and changes the value difference to $D^{(0,1)}(Q)$; a BAT seller chooses to undercut the bid side and changes the value difference to $D^{(1,0)}(Q)$. Non-algo buyers (sellers) arrive with probability p_2; a non-algo buyer submits a market buy order, moves the queue position forward by one unit, and changes the value difference to $D^{(0,0)}(Q-1)$; a non-algo seller submits a market sell order and does not affect $D^{(0,0)}(Q)$, because the LOB on the bid side is refilled immediately by HFTs. Value jumps occur with probability p_3. 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 their difference is $-\frac{d}{2}$. In a downward value jump, the liquidity provider cancels the limit order, thereby changing the value of both the liquidity provision and stale-quote snipping to zero. Equation (14) contains $max\{0,.\}$ operator, because HFTs do not queue at the Q^{th} position once the expected payoff is below 0.

I present the solution for $D^{(i,j)}(Q)$ for any i,j, and Q in the Appendix, and Figure 6 provides a numerical example. Figure 6 shows that the value of liquidity provision decreases in Q, while the value of stale-quote sniping increases in Q. HFTs provide 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.

$LP^{(i,j)}(Q)$ and $SN^{(i,j)}(Q)$ also depend on the state of the LOB. A comparison between the left and right panels of Figure 6 shows that an undercutting order reduces HFTs' depth on the same side of the LOB by approximately one share. Intuitively, an undercutting order on the same side of HFTs' limit orders directly reduces the execution priority of HFTs.

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 $v_t - \frac{d}{6}$ and 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 to get execution. More generally, an undercutting BAT limit buy (sell) order may attract future BAT sellers (buyers) to take liquidity, making future BATs less likely to undercut HFTs. In turn, the value of liquidity provision increases relative to sniping, thereby incentivizing HFTs to provide larger depth. This indirect effect, however, is rather small.[15]

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 in Section 4, 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 $2^4 = 16$. I do not report the results for brevity but they are available upon request. In Section 5, I discuss the case when the break-even spread equals $\frac{7d}{6}$.

[15] This indirect effect is so small that it does not affect depth in my 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.

5. Stub Quotes and Mini-Flash Crashes

In Proposition 4, I 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. I call such quotes stub quotes. A mini-flash crash occurs when a market order hits a stub quote.[16]

Proposition 4 (**Stub Quotes and Mini-Flash Crash**). HFTs quote a half bid-ask spread of $\frac{7d}{6}$ when $\frac{\lambda_I}{\lambda_J} < \frac{1}{5(1-\beta)}$. Although HFTs quote wider bid-ask spreads, the average transaction cost for non-HFTs is lower than the case under which the tick size is d.

A reduction in tick size increases the bid-ask spread quoted by HFTs, because BATs no longer demand liquidity from them. Surprisingly, liquidity improves on average despite the increase in the bid-ask spread quoted by HFTs. Figure 3 shows that the fraction of BATs needs to be at least $\frac{4}{5}$ for stub quotes to occur. BATs' maximum transaction cost is $\frac{d}{6}$ if they use flashing limit orders. Non-algo traders' maximum transaction cost is $\frac{7d}{6}$ if they always 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 $\frac{d}{2}$, the half bid-ask spread under larger tick size d. The average transaction cost for non-HFTs must be even lower, because 1) the fraction of BATs is higher than $\frac{4}{5}$ in the stub quote area; 2) BATs may further reduce transaction costs by submitting regular limit orders; and 3) non-algo traders may take liquidity from BATs.

[16] In my model, the size of the mini flash crash can only be as large as $\frac{7d}{6}$, because the size of a value jump is d. An increase in the support of jump size leads 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.

As BATs do not constantly stay in the market to provide liquidity, non-algo traders may hit HFTs' stub quotes. A small tick size then increases the volatility of the bid-ask spread and harms non-algo traders that encounter the stub quotes. I discuss policy implication for mini-flash crashes in the next section.

6. Predictions and Policy Implications

By incorporating a new type of algorithmic trader and discrete pricing in my model, I offer a number of predictions on traders' behaviors, liquidity, and the impact of speed competition on social welfare. Some predictions rationalize existing empirical findings, whereas others have not been tested. In Subsection 6.1, I summarize the predictions on who provides liquidity and when. In Subsection 6.2, I summarize the predictions on the welfare implications on speed versus price competition. In Subsection 6.2, I discuss the use of message-to-trade ratio as the cross-sectional proxy for HFTs' activity.

6.1 Liquidity provision

My model shows that who provides liquidity depends on the tick size, adverse selection risk, motivation of the trade, and the speed of the trade. In Prediction 1, the queuing hypothesis, I posit that HFTs provide a larger fraction of liquidity when tick size is large.

Prediction 1. (**Queening Channel of Liquidity Provision**): An increase in tick size increases the fraction of liquidity provided by HFTs. Non-HFTs are more likely to undercut HFTs when the tick size is small. An increase in the tick size increases the revenue from liquidity provision, but forces non-HFTs to use more market orders.

Researchers in existing literature find that speed advantages reduce HFTs' adverse selection cost (see Jones (2013) and Menkveld (2016) for the survey) and inventory cost (Brogaard et al. 2015). These reduced costs of intermediation raise the concern that "HFTs use their speed advantage to crowd out liquidity provision when the tick size is small and stepping in front of standing limit orders is inexpensive" (Chordia et al. 2013). Yao and Ye (2016) and O'Hara, Saar and Zhong (2015), however, find that it is a large tick size that crowds out non-HFT liquidity provision.[17] My model provides two economic mechanisms to bridge the gap: (1) non-HFTs have incentives to undercut HFTs because they are less likely to establish time priory at the same price as HFTs; and (2) non-HFTs are able to undercut HFTs if aggressive limit orders reduce their transaction costs relative to marker orders. In Proposition 1, when the tick size is large, HFTs dominate liquidity provision due to their top position in the queue. When tick size becomes smaller, as posited in Propositions 2-4, BATs provide liquidity by establishing price priority over HFTs.

Brogaard et al. (2015) find that non-HFTs quote a tighter bid-ask spread than HFTs, and Yao and Ye (2016) find that non-HFTs are more likely to establish price priority over HFTs as the tick size decreases, both of which provide additional empirical support for the queuing channel. Yao and Ye (2016) also find that an increase in tick size increases the revenue generated from using limit orders, but decreases non-HFTs' use of limit orders. My model provides an economic mechanism to rationalize this puzzle. A large tick size increases the length of the queue, which forces traders who do not have a speed advantage to demand liquidity.

Prediction 2 discusses the speed competition of taking liquidity.

[17] Yao and Ye (2016) consider relative tick size (1 cent uniform tick size divided by price) the economically meaningful tick size for empirical work.

Prediction 2. (**Speed Competition of Taking Liquidity**): Non-HFTs are more likely to provide liquidity at price levels that cross the midpoint (flash limit orders) than HFTs do. HFTs are more likely to take liquidity from flash limit orders, but they do not adversely select flash limit orders.

In the flash equilibrium, BATs offer aggressive limit orders to prompt HFTs to take liquidity. To my knowledge, no existing empirical work tests Prediction 2 using data with traders' identities. Yet Latza, Marsh, and Payne (2014) provide empirically evidence consistent with Prediction 2. 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. Latza, Marsh, and Payne (2014) find that fast market orders often execute against limit orders that cross the midpoint, and they lead to virtually no permanent price impact, whereas market orders from slow trades have positive long run price impacts. In my model, liquidity providers use flash limit orders in a fast trade, which suffer less from the adverse selection; liquidity providers use regular limit orders in a slow trade, which suffer more from adverse selection costs.

Prediction 3. (**Liquidity Provision and Adverse Selection Risk**): The fraction of liquidity provided by HFTs decreases with the level of adverse selection.

Yao and Ye (2016) find that HFTs provide less liquidity for stocks with higher adverse selection risk, a result inconsistent with HFTs' reduced adverse selection cost (Hoffmann 2014). Queuing serves as a channel to reconcile this contradiction. When adverse selection risk is low, the tick size is binding and HFTs provide more liquidity through winning time priority in the queue.

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 provision.

Prediction 4 addresses who provides liquidity during a mini-flash crash, an extreme price movement.

Prediction 4. (**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 provide liquidity and non-HFTs demand liquidity. A downward mini-flash crash is more likely to follow a downward value jump, while an upward mini-flash crash is more likely to follow an 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 cost for HFTs through two channels. First, it reduces the liquidity demand for HFTs. Second, BATs' undercutting orders reduce the execution probability of HFTs' limit orders. When the adverse selection cost is higher enough, HFTs back away from liquidity provision by quoting stub quotes.

HFTs are more likely to retreat to stub quotes when adverse selection risk is high, because (1) higher adverse selection risk widens the break-even bid-ask spread; and (2) a wider break-even bid-ask spread also allows BATs to undercut HFTs, which further increases the adverse selection cost for HFTs.

HFTs' retreat to stub quotes is only a necessary for a mini-flash crash, because BATs are

able to provide liquidity if HFTs retreat. Yet BATs do not continuously provide liquidity in the market, making it possible for non-algo traders' market orders to hit stub quotes and to cause a mini-flash crash. In cross-section, stocks with smaller tick sizes or higher adverse selection risk are more likely to incur mini-flash crashes.

In time series, an initial downward (upward) jump increases the probability of a downward (upward) mini-flash crash, because the initial downward (upward) jump clears all limit orders from BATs on the bid (ask) side of the LOB. A market order becomes more likely to hit stub quotes before BATs refill the book.

To the best of my knowledge, no existing paper tests the cross-sectional predictions of the mini-flash crashes. Brogaard et al. (2016) analyze the time series pattern of mini-flash crashes. They show that, 20 seconds before the mini-flash crash, HFTs neither demand nor supply liquidity, whereas non-HFTs demand and supply the same amount of liquidity; 10 seconds before the mini-flash crash, HFTs demand liquidity from non-HFTs; at the time of the mini-flash crash, HFTs provide liquidity to non-HFTs, but at much wider spread. The authors also find that liquidity provision from the mini-flash crash is profitable. This evidence is consistent with the theoretical mechanism for mini-flashes crash I document. I find that: (1) in normal times, non-HFTs dominate both liquidity provision and liquidity demanding; (2) slightly before a mini-flash crash, HFTs take liquidity and remove limit orders from BATs; (3) a mini-flash crash occurs when a non-algo trader's market order hit HFTs' stub quotes, which leaves positive profit to HFTs.

6.2 Liquidity, social welfare, and policy

On April 5, 2012, the U.S. Congress passed the Jumpstart Our Business Startups (JOBS) Act. Section 106 (b) of the Act requires the Securities and Exchange Commission (SEC) to examine

the effect of the tick size on IPOs. On October 3, 2016, the SEC started to implement a pilot program to increase the tick size 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). Prediction 5, however, posits that an increase in tick size increases transaction costs.

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 (2016) find empirical evidence consistent with Prediction 5. Holding the BBO constant, an increase in depth implies an increase in liquidity. Yet these authors also find that the quoted spread increases after an increase in tick size. When both quoted spread and depth increase, the most relevant liquidity measure becomes the effective bid-ask spread, the transaction cost paid by liquidity demanders (Bessembinder 2003). My model shows that constrained price competition increases effective spread, which is consistent with Yao and Ye (2016). My model prediction, along with the empirical evidence in Yao and Ye (2016), shows that an increase in tick size would not improve liquidity.

Proponents to increase the 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). I find that a wider tick size increases market-making profits, but the profit belongs to traders with higher speed. Therefore, a wider tick size is more likely to result in an arms race in latency reduction than in sell-side equity research.

Prediction 6 (Welfare): An increase in tick size harms liquidity demanders and does not benefit

liquidity suppliers after accounting for the cost of the investment in speed.

An increase in tick size leads to an increase in the effective spread, which harms liquidity demanders. An increase in tick size also does not benefit liquidity providers as the cost of the speed investment dissipates the higher rents created by tick size.

Yao and Ye (2016) show that tick size is currently surprisingly binding. In their stratified sample of 117 Russell 300 stocks, the tick size is binding 41% of the time. This finding, along with Predictions 5 and 6, suggest that the SEC should consider reducing the tick size, particularly for large liquid stocks.

One possible side effect of decreasing the tick size is more frequent mini-flash crashes, as posited in Prediction 4. An increase in tick size prevents mini-flash crashes, but it also increases the transaction costs for average trades. A more effective solution to prevent mini-flash crashes is 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). I argue that the best solution to mini-flash crashes is to slow down the market to wait for natural trading interests to reconvene.

6.3 Message-to-trade as a cross-sectional proxy for HFT activity

Because HFTs are known for their high order cancellation rates, the message-to-trade ratio is widely used as a proxy for HFTs' activity, particularly for HFTs' liquidity provision (Biais and Foucault 2014). Yet Yao and Ye (2016) find that stocks with a higher fraction of liquidity provided

by HFTs have a lower message-to-trade ratio. I provide one interpretation for this surprising negative correlation in Prediction 7.

Prediction 7. (**Message-to-Trade Ratio**). Stocks with a smaller tick size have a lower fraction of liquidity provided by HFTs but a higher message-to-trade ratio.

A decrease in tick size decreases the fraction of liquidity provided by HFTs (Prediction 4), but it leads to more cancellations. Under a large tick size in my model, HFTs with liquidity provision positions in the queue 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 possibility 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 cancellation.

My model also provide a new interpretation of flickering quotes. Yueshen (2014) shows that flickering quotes occur when new information causes the price to move a new level. I show in this paper that HFTs can cancel orders in the absence of information. These periodic additions and cancellations of orders also differ from Baruch and Glosten (2013), who interpret the existence of flicking quotes using a mixed-strategy equilibrium.

7. Conclusions

This paper contributes to the literature on HFTs by including two salient features in current financial markets: discrete tick size and algorithmic traders who are not HFTs. BATs are more

likely to provide liquidity when tick size is small, because providing liquidity is less costly than demanding liquidity from HFTs. A large tick size constrains price competition, creates rents for liquidity provision, and encourages speed competition to capture such rents through the time priority rule. 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. All these predictions are consistent with the empirical findings by Yao and Ye (2016).

Yao and Ye (2016) find that the message-to-trade ratio, a widely used empirical proxy for HFTs' activity, has a negative cross-sectional correlation with HFT liquidity provision. This paper provides a theoretical foundation for their surprising negative correlation. A large tick sizes induces HFTs to race for the top queue position, but HFTs are less likely to cancel orders once they secure the queue position. HFTs cancel orders more frequently for stocks with smaller tick sizes, but they also provide a lower fraction of liquidity. Both theoretical and empirical evidences suggest that researchers should not apply the message-to-trade ratio as a cross-sectional proxy for HFT activity.

My model also provides several new testable predictions. I predict that 1) non-HFTs are more likely to provide liquidity at price levels that cross the midpoint than HFTs do, 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.

My model shows that a larger tick size increases transaction cost and harms liquidity demanders. Yet liquidity suppliers do not benefit from a larger tick size, because the investment in speed dissipates the rents created by tick size. I challenge the rationale for the recent movement

to increase the tick size to five cents, and I encourage regulators to consider decreasing tick size, particularly for liquid stocks.

The inclusion of trading algorithms designed by sophisticated non-HFTs adds significant new insight. For example, I find that BATs can prompt HFTs to demand liquidity using flash limit orders to reduce transaction costs. This finding cautions the evaluation of the welfare impact of HFTs based on demand versus supply liquidity. I take an initial step to examine the interaction between high-frequency and non-high-frequency algorithms, but my model is parsimonious. For example, BATs in my model do not have private information and they choose order types only upon arrival. Extending my model toward more realistic setups 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.

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.

Brogaard, J., A. Carrion, T. Moyaert, R. Riordan, A. Shkilko, and K. Sokolov. 2016. High-frequency trading and extreme price movements. Working paper, University of Washington, University of Utah, Louvain School of Management, Queen's School of Business, Wilfrid Laurier University.

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.

Chao, Y., C. Yao, and M. Ye. 2016. What Drives Price Dispersion and Market Fragmentation across US Stock Exchanges?. Working paper, University of Louisville, University of Warwick, and University of Illinois at Urbana-Champaign.

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. Forthcoming. Toxic arbitrage. *Review of Financial Studies.*

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.

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.

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

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

Menkveld, A. J., and M. A. Zoican. Forthcoming. Need for speed? Exchange latency and liquidity. *Review of Financial Studies*.

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.

————, and D. J. Seppi. 2008. Limit order markets: A survey. *Handbook of financial intermediation and banking* 5: 63-95.

————. 2003. Liquidity-based competition for order flow. *Review of Financial Studies* 16:301-

343.

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

Seppi, D. J. 1997. Liquidity provision with limit orders and a strategic specialist. *Review of Financial Studies* 10:103-150.

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

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

Yao, C., and M. Ye. 2016. Why trading speed matters: A tale of queue rationing under price controls. Working paper, University of Warwick, and University of Illinois at Urbana-Champaign.

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

I consider the first share on 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 I 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 provider; 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, and 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

I 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. I 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 $(1,1)$. Denote the cost for the first case $C(1,0)$ and the cost for the second case $C(1,1)$. 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, I 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 arrives 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 limit order book.[18] $C(1,1)$ 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.[19] Second, a downward jump under $C(1,1)$ leads to sniping on the opposite side of the LOB and changes the state to $C(1,0)$.

[18] Here I 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.

[19] The execution of this order results from my 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).

If an undercutting order gets immediate execution, the cost is $-\frac{d}{6}$. $C(1,1)$ must be greater than the $-\frac{d}{6}$ 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 book, 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 my 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 $R < R_1$.

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

Proof of Proposition 3

1. Proposition 2 shows the boundary between the flash equilibrium and the undercutting equilibrium.
2. The solution for HFT depth follows Figure 5 and equation (14). The depth decreases because the revenue of liquidity provision for HFTs decreases. BATs never take HFTs' liquidity at $\frac{d}{2}$, and BATs can also provide 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 I give an example for $R = 4$ and $\beta = 0.1$.

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

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

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

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

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

   I then obtain:

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

$$= 0.2202,$$

$$D^{(1,0)}(1) = \frac{8 + 24R + 20R^2 + 6\beta R^2 - 4\beta^2 R^2 + 12\beta R^3 - 5\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)}$$

$$= 0.0527,$$

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

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

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

$= 0.0593.$

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

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

$$D^{(0,0)}(2) = p_1D^{(0,1)}(2) + p_1D^{(1,0)}(2) + p_2D^{(0,0)}(1) + p_2D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,0)}(2) = p_1D^{(1,1)}(2) + p_1D^{(1,0)}(2) + p_2D^{0,0}(2) + p_2D^{(1,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(0,1)}(2) = p_1D^{(0,1)}(2) + p_1D^{(1,1)}(2) + p_2D^{(0,1)}(1) + p_2D^{(0,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

$$D^{(1,1)}(2) = p_1D^{(0,1)}(2) + p_1D^{(1,0)}(2) + p_2D^{(0,1)}(2) + p_2D^{(1,0)}(2) + p_3\left(-\frac{d}{2}\right) + p_3 \cdot 0$$

I get:

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

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

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

$$D^{(1,1)}(2) = -0.0561 < 0.$$[20]

I reject the assumption that all $D(2) > 0$. Therefore, under certain states, HFTs would not provide the second share of liquidity in the LOB. I start from the worst state for liquidity providers, $(1,0)$, in which a BAT undercuts HFTs on the same side of the LOB, but no BAT undercuts HFTs on the other side of LOB.[21] Therefore, $D^{(1,0)}(2) = 0$ and all other $D^{(i,j)}(2) > 0$. Thus I solve:

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

[21] In this state, an HFT liquidity provider on the ask side cannot trade with the next non-HFT buyer, because a BAT buyer chooses to provide 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 $(0,0)$.

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

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

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

I 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. I 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,

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

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

I obtain:

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

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

Further calculation shows $D^{(0,0)}(3) = 0, D^{(0,1)}(3) = 0$. I 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 provide liquidity at $\frac{5d}{6}$ when:

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

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

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

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

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

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

Figure 1: 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.

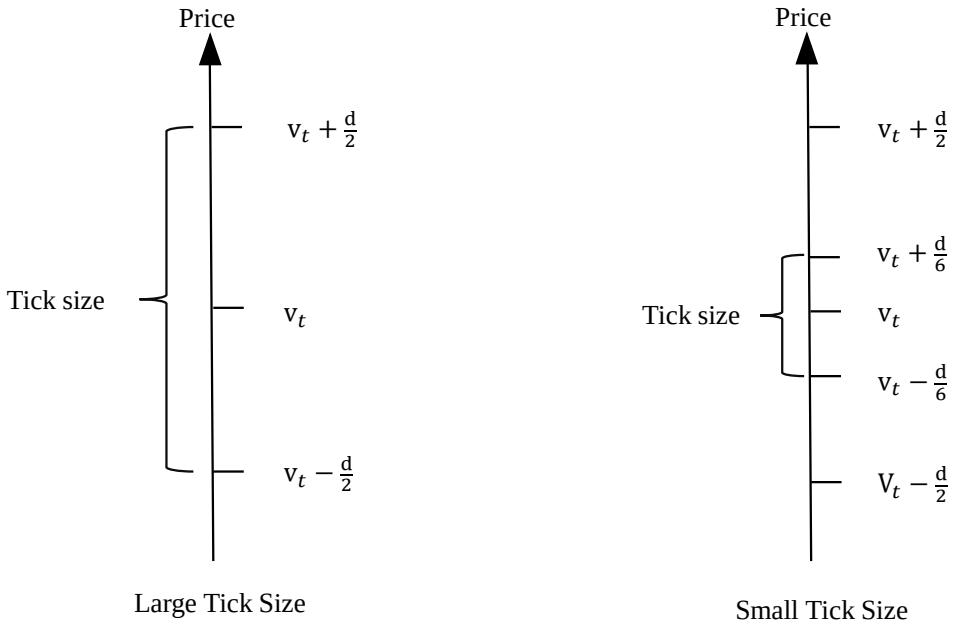

Figure 2: Depth and the Adverse Selection Risk under a Binding Tick Size

This figure demonstrates the relationship 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 cost and increases the depth. The solid line represents the depth under tick size d and the dash line represents the depth under tick size $\frac{d}{3}$.

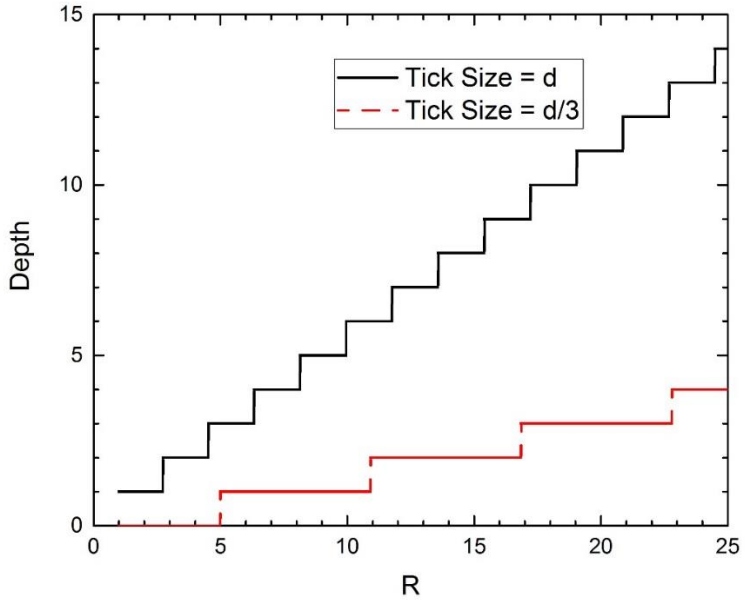

Figure 3: Bid-ask Spread Quoted by HFTs under a Small Tick Size

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 spreads weakly increase with the fraction of BATs and adverse selection risk.

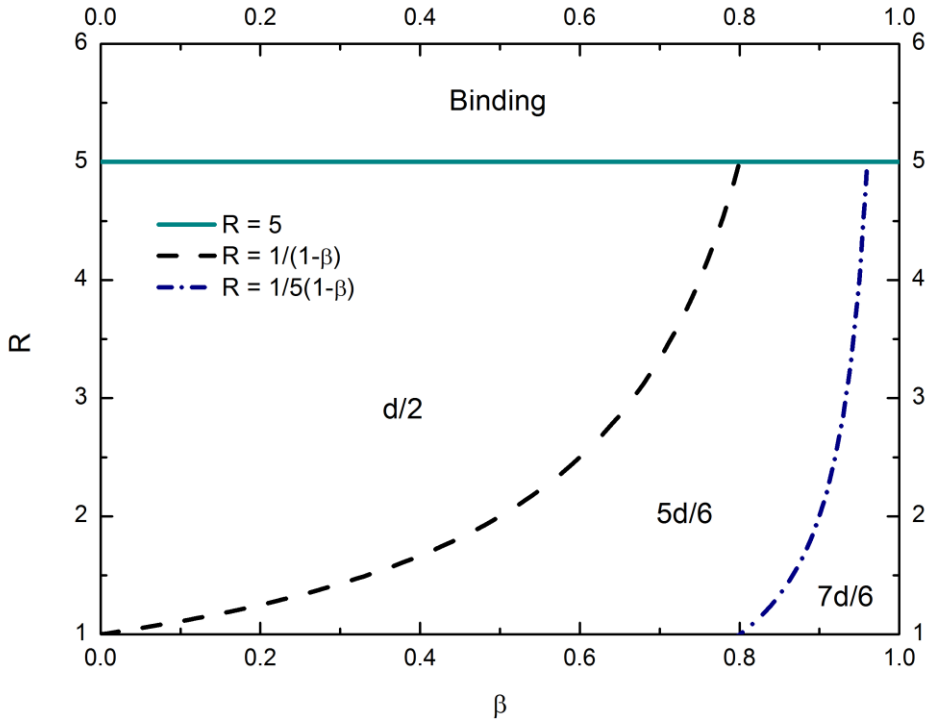

Figure 4: The Undercutting and the Flash Trading Equilibrium

This figure demonstrates two types of equilibrium 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 probability that a limit order executes with non-HFTs before a value jump. To jumpstart an undercutting equilibrium, the expected transaction cost for a limit order undercutting 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.

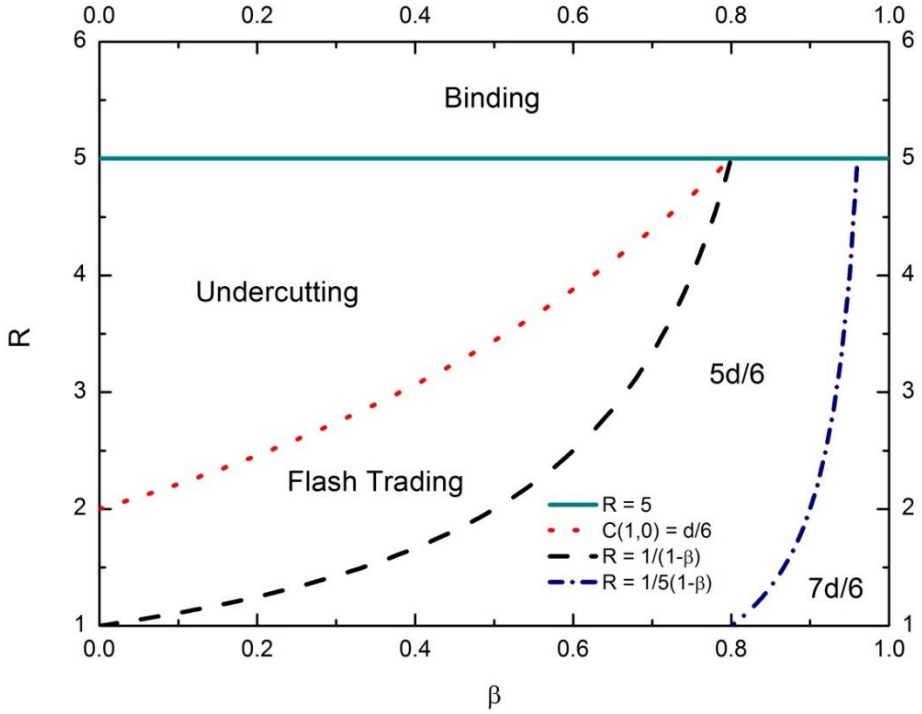

Figure 5: States and Profits for HFT Liquidity Providers with the Q^{th} Position on the Ask Side

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 and number of that on the bid side is j. BB and BS represent the arrival of BATs' buy and sell limit orders, NB and NS represents 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.

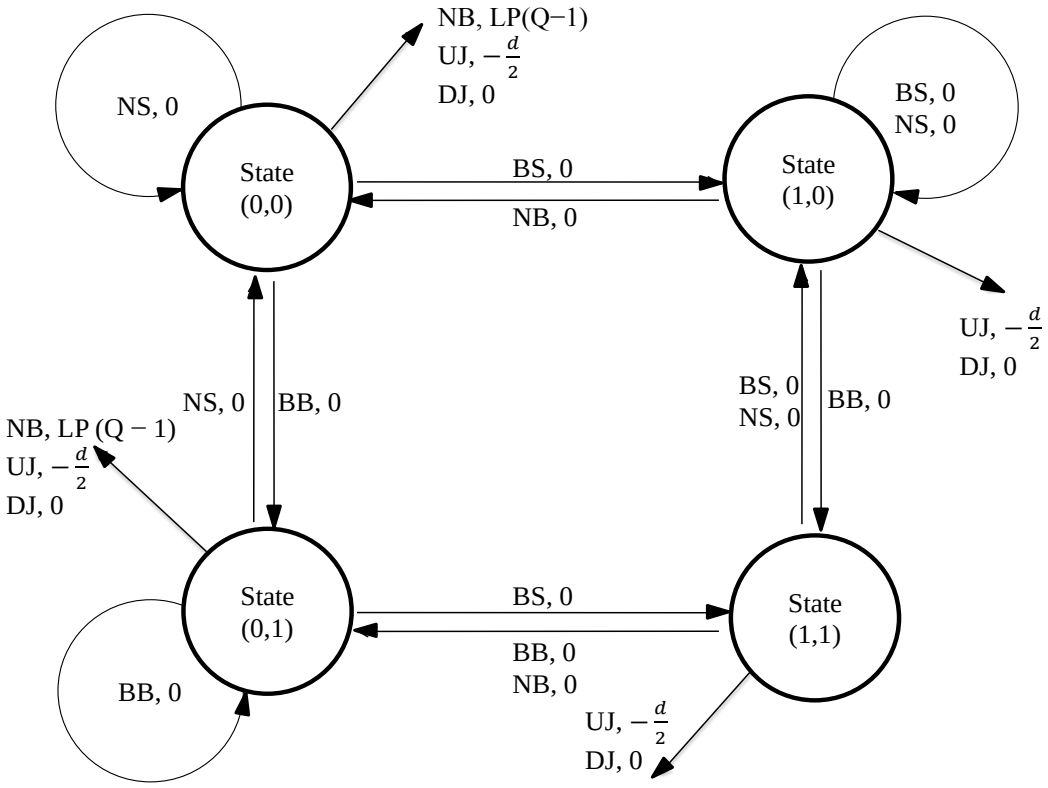

Figure 6: Value of Liquidity Provision and Stale-Queue Sniping and Queue Length

The x-axis is the value of HFT liquidity provision (*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 and SN increases in the queue position. HFTs provide liquidity as long as *LP>SN*.

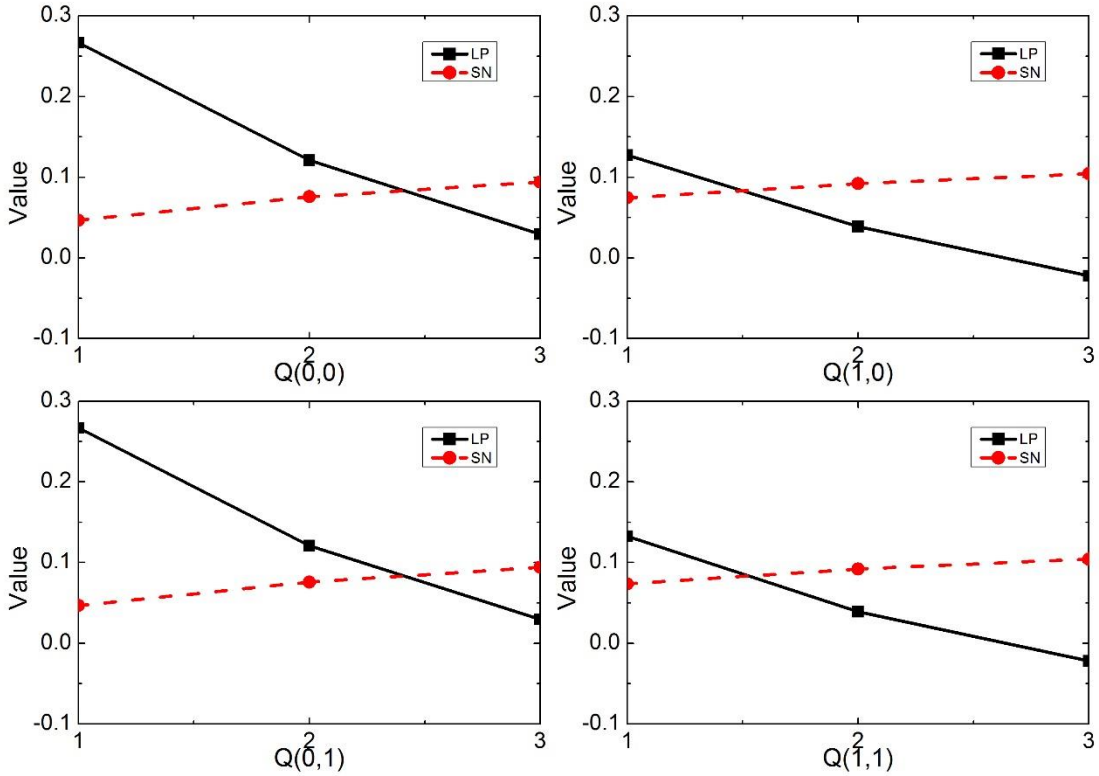

Figure A.1: States and Profits for BATs on the Ask Side

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. BB and BS implies the arrival of BAT buy and sell orders. NB and NS are arrivals of non-algo buy and sell orders, while UJ and DJ are upward and downward jumps. The states are defined after BATs submit the limit orders. For example, submitting a sell limit order to an empty book 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 book, the state after submission is $(1,1)$ and the cost is $C(1,1)$.