Who Provides Liquidity, and When?

Abstract

We model the competition in liquidity provision between high-frequency traders (HFTs) and the relatively slow execution algorithms initiated by buy-side institutions (BATs). As BATs have to trade, their opportunity cost of supplying liquidity is lower, and we show that they always provide liquidity at better prices than HFTs when price is continuous. When tick size (minimum price variation) is large or when the probability of being adversely selected is low, the break-even bid-ask spread is lower than one tick. The bind tick size constrains price competition and encourages HFTs to provide liquidity through time priority. We show that transaction costs can be perfectly negatively correlated with the bid-ask spread when all traders are able to provide liquidity. Our model shows that a large tick size increases transaction costs, which brings into question the rationale of the policy initiative to increase the tick size from one cent to five cents. Flash crashes arise at an equilibrium under certain parameter values, which not only helps to predict when and where flash crashes are more likely to happen, but also helps to inform policy design to prevent these crashes.

1 Introduction

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 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, their trading algorithms do not have an independent identity in existing models. In one view, the financial markets include HFTs and everyone else, where the latter includes both sophisticated institutions and unsophisticated retail traders [see the survey by O'Hara (2015)]. In the other view, algorithmic traders and HFTs are interchangeable [see the survey by Biais and Foucault (2014)]. In this paper, we offer the first theoretical study of BATs, and we examine how they interact with HFTs and humans.

In an environment of HFTs, BATs, and humans, who provide liquidity, who demands liquidity, and when? These questions are important because traditional liquidity providers, such as the NYSE specialist and NASDAQ dealers, almost disappear in modern electronic markets (Clark-Joseph, Ye, and Zi 2017). Everyone can supply liquidity, but no one is obligated to supply liquidity. We examine how this new environment of voluntary liquidity supply and demand reaches equilibrium.

In our model, HFTs and two types of non-HFTs (BATs and humans) trade a security in a dynamic limit order book (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. HFTs have no private value to trade, but simply provide or demand liquidity when its expected profit is above zero. Non-HFTs, who arrive at the market following a Poisson process, have the 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 humans, who only demand liquidity.

Our model includes one security whose fundamental value is public information. However, liquidity providers are subject to sniping risk (Budish, Cramton, and Shim 2015; BCS hereafter), because they may fail to cancel stale quotes during value jumps. BATs are always sniped during value jumps, and HFTs can reduce their probability of being sniped by $\frac{1}{N}$, where N is the number of equally fast HFTs. As in BCS, HFTs in our model quote a positive bid-ask spread because of the sniping risk. Surprisingly, we find that BATs always quote better prices than HFTs as long as the price is continuous enough, even though BATs incur a higher cost of being sniped. Two economic mechanisms, the opportunity cost of liquidity provision and the make-take spread, drive this counterintuitive result.

The opportunity cost stems from outside options for BATs and HFTs. BATs have inelastic need to trade in our model. They can both demand liquidity by paying the bid-ask spread, or they can provide liquidity. As a consequence, their opportunity cost to provide liquidity is negative, and they choose to provide liquidity as long as it is less costly than paying the bid-ask spread. HFTs, on the other hand, do not have to trade, and their speed advantage leads to a positive opportunity cost to provide liquidity. If an HFT provides liquidity for one share, she loses the opportunity to snipe the share if its quote becomes stale. Therefore, although speed advantage decreases HFTs sniping cost by a factor of $\frac{1}{N}$, speed advantage increases HFTs' opportunity cost of liquidity provision by the same amount. Taking the sniping cost and opportunity cost together, BATs have lower overall costs of liquidity provision, and they always choose to provide liquidity when the price is continuous enough.

The make-take spread, a new concept introduced by our model, stems from the flexibility to choose between providing liquidity and demanding liquidity for both HFTs and BATs. For any potential trades, HFTs have two price levels: one price they are willing to offer and another they are willing to accept. For example, an HFT offers an ask price to sell above the fundamental value, because the ask price is subject to the sniping risk. The same HFT accepts an offer to buy at fundamental value, because demanding liquidity entails no sniping risk. Therefore, a BAT can immediately trigger HFTs to demand liquidity if BATs submit a buy limit order at the fundamental value. Consequently, BATs pay zero transaction costs despite the positive bid-ask spread. Our model then offers a dramatic contrast to BCS. In BCS, non-HFTs can only demand liquidity, and the sniping risk leads to a positive bid-ask spread, which motivates BCS to recommend frequent batch auction as the alternative market design. Our model shows that when all non-HFTs can choose between demanding and providing liquidity, transaction cost becomes zero despite the sniping risk. Our model has a unique equilibrium under continuous price, in which BATs provide liquidity at the fundamental value and incur zero transaction cost, while HFTs demand liquidity from BATs.

Next, we add discrete price into our model to reflect the tick size (minimum price variation) of one cent imposed by the U.S. Security and Exchange Commission's (SEC's) Regulation National Market Systems (Reg NMS) Rule 612, and to evaluate the recent policy initiative to increase the tick size from one cent to five cents. Discrete price generates four types of equilibria depending on the sniping risk, the fraction of BATs, and the tick size. We find a model with BATs and discrete price can rationalize a number of existing puzzles in the literature on HFTs, as well as generate a number of new testable predictions.

We find that discrete price leads to speed competition in liquidity provision. Such competition is most apparent when sniping risk is small relative to the tick size. In our model, the breakeven spread increases with sniping risk. When sniping risk is low, the break-even bid-ask spread becomes smaller than one tick. The difference between one tick mandated bid-ask spread and the breakeven bid-ask spread generate rents for liquidity provision. Speed then allocate these rents to HFTs, because U.S. stock exchanges use order arrival time to decide execution priority for orders quoted at identical prices. In our first type of equilibrium, queuing equilibrium, HFTs dominate liquidity provision either when sniping risk is too low or when tick size is too large. Yao and Ye (2018) find empirical support for these two predictions.

Interestingly, discrete price also leads to speed competition in liquidity demand. As sniping risk increases, the break-even spread for HFTs becomes wider than one tick, and BATs are able to quote a more aggressive bid-ask spread than HFTs. We show that “flash” limit orders, or limit orders that cross the midpoint of the fundamental value, strictly dominate market orders. Flash limit orders follow the intuition of the make-take spread, with a new feature created by discrete price. When price is continuous, BATs can use flash limit orders priced at the fundamental value to attract HFTs. When price is discrete, BATs may have to use more aggressive limit orders to attract HFTs, because the fundamental value may not be at a discrete price level. The difference between the flash order price and the fundamental value creates rents and generates speed races to demand liquidity. In the second type of equilibrium, flash equilibrium, BATs provide liquidity to HFTs; HFTs provide liquidity to humans because limit orders from BATs do not stay in the LOB.

BATs provide liquidity to humans in our third types of equilibrium: undercutting equilibrium. This equilibrium occurs when sniping risk is higher than the queuing equilibrium, but lower than the flash equilibrium. If sniping risk is too low, HFTs will dominate liquidity provision at binding one tick, which is the queuing equilibrium. The comparison with the flash equilibrium comes from the following trade-off. A flash limit order that just crosses the fundamental value executes immediately and its payoff is certain. A limit order that does not cross the midpoint leads to a profit when humans arrive, and it loses money when HFTs snipe the order. Therefore, flash limit orders are more attractive when sniping risk is high. Also, flash limit orders become more attractive when the fraction of BATs is high, because the probability to execute with a human is lower.

In the final type of equilibrium, the crash equilibrium, neither HFTs nor BATs provide liquidity at a reasonable price, and liquidity demand from humans[1] leads to flash crashes, which are sharp price movements in one direction followed by a quick reversion (Biais and Foucault, 2014). HFTs retreat from liquidity provision when the sniping risk or the fraction of BATs is too high. The probability of flash crashes first increases in the fraction of BATs and then decreases with it. On the one hand, BATs do not demand liquidity from HFTs as long as tick size is not binding. As the fraction of BATs approach to one, HFTs face almost no non-HFT order flows and they have to quote a spread wider than the maximum possible size of value jumps to avoid being sniped. We call such spread “stub quotes” and flash crashes occur when humans hit stub quotes. An increase in the fraction of BATs creates more stub quotes. On the other hand, an increase in the fraction of BATs reduces the probability to hit the stub quotes, because BATs never choose to hit stub quotes. Taking together, flash crashes never occur when all non-HFTs are humans or when all non-HFTs are BATs. Our model predicts that flash crashes are more likely to occur with an

[1] By humans we mean those traders who take liquidity regardless of bid and ask prices. They demand liquidities even if they are stub quotes. Therefore, naïve algorithms that always demand liquidity behave like humans in our model.

intermediate level of BATs. Also, flash crashes are 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. If BATs do not refill the LOB before a human arrives, flash crashes occur. These predictions can be tested using the data from market wide Flash Crashes or individual security mini-flash crashes.

In the existing literature, information drives the arms race in speed.[2] Our paper discovers new channels of speed competition that work in the opposite direction. In our model, speed competition is most profitable when there is no information. In the absence of information, the breakeven spread is zero, which generate maximum rents for racing to the top of the queue of liquidity provision. Speed competition in the absence of information addresses three puzzles in the literature. 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 (Hoffmann 2014)), should have a competitive advantage in providing liquidity for stocks with higher adverse selection risk (Han, Khapko and Kyle 2014), and should dominate liquidity provision when tick size is small, because the constraints to offer better prices is less binding (Chordia et al. 2013). Yet Brogaard et al. (2015) find that slow traders quote a tighter bid-ask spread than fast traders. Yao and Ye (2018) find that an increase in adverse selection risk reduces HFTs' fraction of liquidity provision. Yao and Ye (2018) and O'Hara, Saar and Zhong (2018) find that a reduction in tick size decreases HFTs' fraction of liquidity provision.

Our model helps to reconcile these three contradictions. First, slow traders have higher

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

incentives to quote a tighter bid-ask spread because they are less likely to establish time priority over HFTs, and they are able to quote a tighter spread because the need to complete a trade reduces their opportunity costs of providing liquidity. Second, low sniping risk reduces the break-even spread below one tick and drives speed competition at constrained price. Third, a large tick size drives speed competition, because it raises the spread above the break-even level.

The closest paper to ours is BCS. We differ from BCS in two dimensions. BCS considers *continuous* prices, while we consider *discrete* prices. Our results 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). Our results show that an increase in tick size would reduce liquidity, encourage HFTs, and allocate resources to latency reduction.

In BCS, non-HFTs only demand liquidity. We allow non-HFTs to choose between demanding and providing liquidity. By taking the initial step to model sophisticated non-HFTs, we develop new predictions and perceptions. Liquidity demand from HFTs used to have a negative connotation, because in existing models, HFTs usually adversely select liquidity providers when they demand liquidity (BCS; Foucault, Kozhan, and Tham, 2017; Menkveld and Zoican, 2017). In our model, BATs can use aggressive limit orders to prompt HFTs to demand liquidity, which involves no adverse selection 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.

2 Model Setup

We consider a continuous time model with an infinite horizon. All the random variables in our model are mutually independent. Our model has one security, whose fundamental value, v_t, evolves as a compound Poisson jump process with arrival rate λ_J, where t runs continuously on $[0, \infty)$. $v_0 = 0$ and jumps by $J = d$ or $-d$ occur with equal probability, where $d > 0$. v_t is common knowledge, but liquidity providers are still subject to adverse selection risk when they fail to update stale quotes after value jumps.

Limit Order Book. The stock exchange operates as a continuous limit order book (LOB). Each trade in the LOB requires a liquidity provider and a liquidity demander. The liquidity provider submits a limit order, which is an offer to buy or sell at a specified price and quantity. The liquidity demander accepts the price and quantity of a limit order. The minimum price variation for a trade and quote, i.e., the tick size, is Δ. Execution precedence for liquidity providers follows the price-time priority. Limit orders with higher buy or lower sell prices execute before less aggressive limit orders. For limit orders queuing at the same price, orders arriving earlier execute before later orders. The LOB contains all outstanding limit orders. Outstanding orders to buy are called bids and outstanding orders to sell are called asks. The highest bid and lowest ask are called the best bid and ask (offer) (BBO), and the difference between them is the bid-ask spread. Because our paper focuses on the competition in liquidity provision, we examine the BBO provided by each type of traders so that we do not need to track infinite price levels in the LOB. For that reason, we assume that HFTs and BATs only place orders at their BBO and do not engage in laying, that is, placing limit orders far away from their BBO and wait for execution after future value jumps. The appendix provides a more complex model with laying, but the intuition is the same.

Traders. Our model includes HFTs and two types of non-HFTs: BATs with fraction β and humans with fraction $1 - \beta$. Each non-HFT wants to buy or sell one share of the security with equal probability.

HFTs. N $(2 \leq N \leq \infty)$ HFTs always present at the market. HFTs place no private value on trading, and they provide or demand liquidity as long as the expected payoff is above zero. HFTs are equally fast. If multiple HFT order messages (limit orders, market orders or cancelations) reach the exchange at the same time, the exchange processes them in a random order. HFTs' strategy space involves the following choices.

The first strategy is to provide liquidity. In the absence of value jumps, liquidity provision profits from the bid-ask spread paid by the liquidity demanders. During value jumps, liquidity providers always cancel their stale quotes, but their orders may be sniped before cancelation. Liquidity providers need to balance the profit from liquidity traders with the cost of being sniped. As HFTs are equally fast, liquidity provision also incurs opportunity cost. A HFT who provides liquidity for a particular share loses the profit opportunity to snipe the share when value jumps. Our paper considers HFTs strategy for each share in the book. That is, a HFT can provide liquidity for one share but not for other shares in the LOB. This share-by-share analysis not only simplifies the model, but also generates the main intuition as to why flash crashes occur (Subsection 5.4).

The second strategy is to demand liquidity. The optimal strategy to demand liquidity is straightforward: HFTs demand liquidity as long as the ask price is below the fundamental value or the bid price is above the fundamental value. These opportunities occur during value jumps, where each HFT sends orders to snipe all stale quotes in the LOB. Providing liquidity for some shares does not affect an HFT's ability to snipe other shares. For simplicity in notation, we assume that HFTs can snipe their own stale quotes, which is economically equivalent to canceling orders. In

BCS, HFTs demand liquidity only when they snipe stale quotes. Section 4 and 5 of our paper will show that HFTs can demand liquidity without adversely selecting stale quotes. Instead, liquidity demand from HFTs can reduce the transaction cost of BATs.

Non-HFTs. Non-HFTs arrive at the market with Poisson intensity λ_I. Each of them has an inelastic need to buy or sell one unit of the risky asset with equal probability. Among these Non-HFTs, a fraction of β of them is BATs and the rest $1-\beta$ are humans. Humans only demand liquidity, and our model degenerates into BCS if $\beta = 0$. BATs choose to demand or supply liquidity to minimize the transaction cost. We consider that BATs are the execution desk of mutual funds or hedge funds, or the brokers who represent their order flow. As we are building the first model for these type of traders, we make the following assumptions to simplify the strategy of BATs.

First, we assume that BATs make a one-time choice between limit and market orders upon arrival, and they leave the market after the decision. Second, if BATs' limit orders are unexecuted during value jump, we assume their limit order would be automatically adjusted so that the bid-ask spread remains the same. Therefore, we assume that BATs use limit orders that pegged to the fundamental value. Without this assumption, we would need to track infinitely many price levels for BATs if fundamental values continue to jump in the opposite direction.

Equilibrium Concept: We aims to find the Markov perfect equilibrium, in which traders' actions condition only on the state of the LOB and events at time t. As BATs make decisions only upon arrival, they do not respond to the future events. They simply choose between market and (pegged) limit orders conditional on the state of the LOB when they arrive, and their object is to minimize the transaction costs conditional on the probability of future events. HFTs consistently monitor the market, and their strategies depend not only the state of LOB and the probability of future events, but also depend on the realization of the event.

Six types of events can trigger the reaction of HFTs.

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

When v_t jumps up, all HFTs have a strictly dominant strategy, that is, sniping all stale quotes on the ask side. Again, sniping their own quotes is economically equivalent to cancelation. Because we do not allow laying, HFTs reestablish equilibrium BBO around the new fundamental value after each value jump. Because we assume HFTs have no latency, they instantaneously build up the equilibrium LOB after any event. HFTs condition their strategy on an event and the state of the LOB immediately before an event, and their collective actions change the LOB to the new equilibrium state. Finally, the Poisson arrival processes are memoryless, i.e., the outstanding limit orders' expected payoffs do not change over time until an event arrives, because the probability distribution of next event does not change. Thus, HFTs would not submit and cancel their orders between two events, so the LOB state will remain unchanged.

Our analysis evolves as follows. We first establish a benchmark model in which price is continuous ($\Delta= 0$) and non-HFTs only demand liquidity ($\beta = 0$). Next, we allow non-HFTs to provide liquidity ($\beta > 0$) under continuous price. After considering continuous price in Section 3 and 4, we examine the impact of discrete price in Section 5.

3 Benchmark: Continuous Price and No BATs

Our benchmark is similar to BCS, in which price is continuous and non-HFTs only demand liquidity. In this benchmark, we only need to consider the strategy of HFTs because non-HFTs are passive players. As in BCS, in equilibrium, the best bid and best ask only contain one share each, because the second one has lower probability of execution but suffers higher sniping risk. The equilibrium is characterized by the bid-ask spread.

The equilibrium bid-ask spread should equalizes the payoff of providing liquidity to the payoff of sniping the stale quotes. Without loss of generality, we consider the expected payoff for the HFT's limit sell order at $v_t + \frac{s}{2}$, $LP\left(\lambda_I, \lambda_J, \frac{s}{2}\right)$.

$$LP\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = \frac{\lambda_I/2}{\lambda_I+\lambda_J} \times \frac{s}{2} + \frac{\lambda_I/2}{\lambda_I+\lambda_J} \times LP\left(\lambda_I, \lambda_J, \frac{s}{2}\right) + \frac{N-1}{N}\frac{\lambda_J/2}{\lambda_I+\lambda_J} \times \left(\frac{s}{2} - d\right) + \frac{\lambda_J/2}{\lambda_I+\lambda_J} \times 0. \qquad (2)$$

As there are no BATs in the benchmark model, the six types of events in equation (1) degenerate to four types. In equation (2), we combine the payoff in each type of event with its probability. With probability $\frac{\lambda_I/2}{\lambda_I+\lambda_J}$, the next event is a human buy order, which leads to a profit of $\frac{s}{2}$ to the liquidity provider. With probability $\frac{\lambda_I/2}{\lambda_I+\lambda_J}$, a human sell order arrives, which does not affect $LP\left(\lambda_I, \lambda_J, \frac{s}{2}\right)$ on the ask-side, because HFTs immediately restore the previous state of the LOB by refilling the bid-side. With probability $\frac{\lambda_J/2}{\lambda_I+\lambda_J}$, v_t jumps upward by d, all HFTs race to snipe stale quotes on the ask side. The conditional probability of being sniped by other HFTs is $\frac{N-1}{N}$. The payoff of being sniped by other traders is $(v_t + \frac{s}{2}) - (v_t + d) = \frac{s}{2} - d$. When v_t jumps downward, the liquidity provider cancels the order and the payoff is zero.

The solution for equation (2) is:

$$LP\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = \frac{\lambda_I}{\lambda_I + 2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{N-1}{N}\left(d - \frac{s}{2}\right). \tag{3}$$

Equation (3) reveals additional intuition for the payoff for a liquidity provider. With probability $\frac{\lambda_I}{\lambda_I + 2\lambda_J}$, a human takes the limit order, and the payoff is $\frac{s}{2}$; with probability $\frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{N-1}{N}$, the limit order is sniped by other HFTs, and the payoff is $\left(\frac{s}{2} - d\right)$; with the remaining probability of $\frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{N+1}{N}$, the limit order is cancelled and the payoff is zero.

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

$$SN\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = \frac{\lambda_I/2}{\lambda_I + \lambda_J} \times 0 + \frac{\lambda_I/2}{\lambda_I + \lambda_J} \times SN\left(\lambda_I, \lambda_J, \frac{s}{2}\right) + \frac{1}{N}\frac{\lambda_J/2}{\lambda_I + \lambda_J} \times \left(d - \frac{s}{2}\right) + \frac{\lambda_J/2}{\lambda_I + \lambda_J} \times 0. \tag{4}$$

The solution for equation (4) is:

$$SN\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = \frac{\lambda_J}{\lambda_I + 2\lambda_J}\frac{1}{N}\left(d - \frac{s}{2}\right). \tag{5}$$

The equilibrium bid-ask spread s_1^* should ensure that HFTs are indifferent between liquidity provision and stale-quote sniping. Thus, equating (3) and (5) solve s_1^*. We summarize the equilibrium as follows:

Proposition 1 (BCS 2015). With zero tick size ($\Delta = 0$) and with non-HFTs only demanding liquidity ($\beta = 0$), the equilibrium bid-ask spread is $s_1^* = \frac{2\lambda_J}{\lambda_I + \lambda_J} d$.

(1) At almost all times t, HFTs always maintain one unit in the LOB at the ask price $v_t + \frac{s_1^*}{2}$ and one unit at the bid price $v_t - \frac{s_1^*}{2}$. The bid and ask prices may belong to different HFTs.

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

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

[Insert Figure 1 about here]

Figure 1 illustrates the dynamics of fundamental value v_t and the corresponding BBOs. Value jumps arrive with Poisson intensity λ_J and change v_t by size d or $-d$. Investors arrive with Poisson intensity λ_I and do not change v_t. HFTs quote $v_t - \frac{s_1^*}{2}$ as the best bid price and $v_t + \frac{s_1^*}{2}$ as the best ask price. As $s_1^* = \frac{2\lambda_J}{\lambda_I + \lambda_J} d < 2d$, quotes at best bid (ask) are sniped when the fundamental value jump upward (downward).

4 Continuous Price Model with BATs

In this section, we allow a fraction of $\beta > 0$ non-HFTs to provide liquidity. In the unique equilibrium of the model, BATs provide liquidity to HFTs, and HFTs provide liquidity to humans.

The first step to establish the equilibrium is to show that BATs never demand liquidity from HFTs. We show this result by contradiction.

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

The previous proof uncovers two new economic mechanisms. The first mechanism is the make-take spread. This spread measures the difference in prices between a trader's willingness to post and accept an offer. These two prices are different because limit orders are subject to sniping risk, whereas market orders do not. As HFTs and BATs are both able to provide and demand liquidity, each group has two price levels. For the decision to sell, the price to provide liquidity is higher than the price to demand liquidity. The make-take spread does not exist in most of the market microstructure models, because the existing literature often assign the role for traders. Some of them have to provide liquidity, while others have to demand liquidity. Therefore, the literature focuses more on the bid-ask spread, that is, conditional on some traders having to provide liquidity, which is their price to buy or sell.

The make-take spread for HFTs happens to be half of the bid-ask spread in our model.[3] An HFT quoting an ask price of $v_t + \frac{s}{2}$ would accept a limit buy price of v_t. Therefore, a BAT can use an aggressive limit order at price $v_t + \varepsilon$ ($\varepsilon \rightarrow 0$) to save the half spread. The limit order at $v_t + \varepsilon$ executes like a market order because of immediate execution.

The second mechanism, the opportunity cost of liquidity provision, offers one interpretation to Yao and Ye (2018), who find that non-HFTs quote more aggressive prices than HFTs as long as the price is continuous enough. The Yao and Ye finding is surprising because the literature suggests the opposite. HFTs, as market makers, have lower adverse selection costs [see the Jones (2013) and Menkveld (2016) surveys], inventory costs (Brogaard et al., 2015), and operational costs (Carrion, 2013). All of these cost advantages indicate that HFTs should be able to quote better prices as market makers. Our model covers an important cost overlooked by the literature: the opportunity cost. BATs face negative opportunity cost of liquidity provision because they have to execute their trade. The outside option to provide liquidity is to demand liquidity by paying $\frac{s}{2}$. HFTs' do not have to trade, and they are not willing to buy at price $v_t + \varepsilon$. In fact, HFTs have positive opportunity costs of providing liquidity. An HFT liquidity provider for a share cannot profit from sniping the share during value jumps, and the probability of sniping conditional on a value jump is $\frac{1}{N}$. This increased cost of $\frac{1}{N}$ exactly offset the reduced sniping cost, because HFT is sniped with a probability of $\frac{N-1}{N}$, whereas a BAT is sniped with a probability of 1. Taking the sniping cost and the opportunity cost together, BATs always have lower cost of liquidity provision.

In Proposition 2, we characterize the equilibrium. In the equilibrium, BATs always use

[3] This is because sniping is the only cost in our model. If there exists: (1) a permanent price impact of BATs order; (2) inventory cost for HFTs; (3) tick constraints forbid BATs from quoting v_t, HFTs would only take more aggressive limit orders, and the make-take spread would be less than half of the bid-ask spread.

limit orders, and we show that BATs choose a limit order price at v_t. The LOB contains no limit order from BATs, as HFTs always immediately demand liquidity from BATs. Therefore, the LOB effectively contains only one state: HFTs quote one share at $v_t + \frac{s_2^*}{2}$ and one share at $v_t - \frac{s_2^*}{2}$. In summary, BATs provide liquidity to HFTs and HFTs provides liquidity to humans. s_2^* equalizes the payoff of liquidity provision and stale quote sniping.

Proposition 2. With zero tick size ($\Delta = 0$) and the portion of BATs that are positive ($\beta > 0$), the equilibrium bid-ask spread $s_2^* = \frac{2\lambda_J}{(1-\beta)\lambda_I + \lambda_J} d$.

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

(2) BATs submit limit orders at v_t when they arrive, and all HFTs immediately take liquidity from BATs.

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

The equilibrium spread s_2^* has two interesting features. First, just as for s_1^*, s_2^* is independent of the number of HFTs. This result is a consequence of the opportunity cost of providing liquidity. During the value jump, the probability of being sniped is $\frac{N-1}{N}$. An increase in the number of HFTs increases this possibility, which reduces the value to be a liquidity provider. An increase in the number of HFTs, however, also reduces the probability to successfully snipe the stale quotes for each HFT. As N affects the value of liquidity provision and sniping by the same amount, N never affects the bid-ask spread. For this reason, we assume $N = \infty$ for the rest of the paper to simplify the notation.

Second, $s_2^* > s_1^* > 0$, which means that each human pays more when BATs can use limit orders. When more non-HFTs use limit orders, HFTs have to quote a wider bid-ask spread for the rest of the market orders. In this sense, BATs reduce their transaction costs at the expense of humans. Corollary 1 shows that the total transaction costs for non-HFTs reduces as β increases. That is, the decrease in the transaction costs for BATs is more than the increase in transaction costs for humans. Therefore, an increase in β increases the overall market liquidity and benefits BATs, but it reduces liquidity for humans. We denote $\bar{C}(\beta)$ as the weighted average transaction cost for BATs and humans.

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

Corollary 1 shows that the quoted bid-ask spread, a common measure of liquidity, may be a misleading of true liquidity when every trader can provide liquidity. As β increases, the quoted bid-ask spread widens, but such an increase is associated with a decrease in the transaction cost. At an extreme, when all non-HFTs are BATs, HFTs' bid-ask spread reaches $2d$ but transaction cost is zero. BCS show that continuous trading leads to sniping risk and positive transaction cost for non-HFTs. Corollary 1 shows that their results also require that some traders cannot supply liquidity. Once all traders can supply liquidity, the transaction cost is again zero even if there is sniping risk.

When $\beta = 1$, the equilibrium with continuous price leads to two predictions that are inconsistent with empirical results. First, the model predicts that HFTs never provide liquidity, but Yao and Ye (2018) show that they provide liquidity when tick size is large. Second, HFTs make zero profit in equilibrium, and they have no economic incentive to invest for arms race in speed. In next section, we show that discrete price generates the rents for arms race in speed and allows

HFTs to provide liquidity.

5. Discrete Price Model

In this section, we compare the differences in market outcome for a discrete tick size. For illustration purposes, we set the pricing grid as $\left\{\dots, -\frac{\boldsymbol{3d}}{\boldsymbol{4}}, -\frac{\boldsymbol{d}}{\boldsymbol{4}}, \frac{\boldsymbol{d}}{\boldsymbol{4}}, \frac{\boldsymbol{3d}}{\boldsymbol{4}}, \dots\right\}$**.** Therefore, tick size is $\Delta = \frac{d}{2}$ and v_t is always at the midpoint of the two nearest ticks. We find four types of equilibrium depending on the sniping risk and fraction of BATs. Figure 2 shows that $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ and β define the boundaries of each type of equilibrium, and we derivate the boundaries in Proposition 3 – 6. BATs have unique strategy in each type of equilibrium. The four subsections each contain one proposition that characterize one type of equilibrium, in which we describes BATs' strategy and HFTs' best response to BATs' strategy.

[Insert Figure 2 about here]

5.1 Queuing Equilibrium

We first examine low sniping risk. When $\kappa \equiv \frac{\lambda_J}{\lambda_I} < \frac{1}{3}$, the break-even spread in Proposition 1, $s_1^* = \frac{2\kappa}{\kappa+1} d$, is lower than one tick. The difference between the break-even spread and the mandated one tick minimum spread generates rents for liquidity provision, which drives the speed race to win time priority in liquidity provision.

When BATs can use limit orders, the state the LOB can explode as more BATs arrive. To reduce the number of states, we make the following assumption.

Assumption 1: The limit orders of BATs must be price improving, that is, they do not queue after existing orders at the same price.

Assumption 1 does not offer a binding constraint for equilibrium in continuous price, in which BATs aggressively undercut HFTs by submitting limit orders at the midpoint. We introduce Assumption 1 here to reduce the state of the LOB to 2^k, where k is the number of price level for BATs to improve the quotes from HFTs. If we relax the assumption that BATs can queue up to n shares, we need to track $(n+1)^k$ states of the LOB. The case for $n > 1$ only increases mathematical complexity without offering any additional intuition.[4] Assumption 1 is standard in the LOB literature. Foucault et al. (2005) make a similar assumption to reduce the states of the LOB. They require all limit order submitters to be price improving, while we only require BATs to do so. We can replace Assumption 1 by another standard assumption in the literature. If BATs are not infinitely patient, a discount factor would imply that they must improve the existing quotes because of price-time priority. Again, the construction of the discount factor is ad hoc, and it does not offer additional intuition.

When $\kappa < \frac{1}{3}$, providing liquidity at the first share is more profitable than sniping the first share, and the bid-ask spread is binding at $\frac{d}{2}$. To see that, consider the expected profit for the first share of liquidity at $v_t + \frac{d}{4}$, $LP(\kappa)$.

$$LP(\kappa) = \frac{1}{2\kappa+2} \times \frac{d}{4} + \frac{1}{2\kappa+2} \times LP(\kappa) + \frac{\kappa}{2\kappa+2} \times \left(\frac{d}{4} - d\right) + \frac{\kappa}{2\kappa+2} \times 0. \qquad (6)$$

As BATs no longer provide liquidity under Assumption 1, four types of events can change the status of the LOB. 1) With probability $\frac{1}{2\kappa+2}$ $(= \frac{\lambda_I/2}{\lambda_I+\lambda_J})$, the next event is a non-HFT buy order,

[4] The queue from BATs is finite, because the execution probability of later queue positions is so low, and the sniping risk is so high that BATs would rather use market orders or limit orders with better prices. Tracking the finite queue, however, can be complex. As BATs do not consistently monitor the LOB, the state of the LOB depends on the random arrival of previous BATs.

and it leads to a profit of $(v_t + \frac{d}{4}) - v_t = \frac{d}{4}$. 2) With probability $\frac{1}{2\kappa+2}$, a non-HFT sell order arrives and it does not affect $LP(\kappa)$ on the ask-side, because HFTs immediately refill the bid-side and restore the previous state of the LOB. 3) With probability $\frac{\kappa}{2\kappa+2}$, the fundamental value v_t jumps upward to $v_t + d$, and the HFT sell order is sniped. The payoff from being sniped is $(v_t + \frac{d}{4}) - (v_t + d) = -\frac{3d}{4}$. 4) When v_t jumps downward, the liquidity supplier cancels the order and joins the race to supply liquidity at a new BBO, and the payoff is 0. The solution for equation (6) is:

$$LP(\kappa) = \frac{1}{2\kappa+1}\frac{d}{4} - \frac{\kappa}{2\kappa+1}\frac{3d}{4}. \qquad (7)$$

Equation (7) also has an intuitive interpretation. With probability[5] $\frac{1}{2\kappa+1}$, a non-HFT takes the limit order, and the payoff is $\frac{d}{4}$; with probability $\frac{\kappa}{2\kappa+1}$, the limit order is sniped, and the payoff is $\frac{3d}{4}$; With probability $\frac{\kappa}{2\kappa+1}$, HFTs will cancel the first share of the order, and the payoff is zero.

When $\kappa \equiv \frac{\lambda_J}{\lambda_I} < \frac{1}{3}$, equation (7) is greater than 0, which means that HFTs would race to provide the first share of liquidity to capture the rent created by the tick size. HFTs would compete in price if it is continuous, but tick size imposes a constraint for such competition. Furthermore, a constrained price cannot clear the market. Next, we show how queuing, or speed competition for the front queue position, clears the market.

Denote the liquidity provision profit of the Q^{th} share as $LP(\kappa; Q)$:

$$LP(\kappa; Q) = \frac{1}{2\kappa+2} \times LP(\kappa; Q-1) + \frac{1}{2\kappa+2} \times LP(\kappa; Q) + \frac{\kappa}{2\kappa+2} \times \left(\frac{d}{4} - d\right) + \frac{\kappa}{2\kappa+2} \times 0. \qquad (8)$$

[5] The probability is not $\frac{1}{2\kappa+2}$ because the arrival of humans to the contra side (second term in the right-hand side of equation (6)) does not change the LOB state as well as the payoff of liquidity provision, $LP(\kappa)$.

The main difference between equation (6) and equation (8) is that shares with position $Q > 1$ cannot get immediate execution when a non-HFT arrives. Instead, it moves to position $Q - 1$. Therefore, instead of achieving an instantaneous payoff of $\frac{d}{4}$, the liquidity providers simply update the value of queuing from $LP(\kappa; Q)$ to $LP(\kappa; Q-1)$. The boundary condition $LP(\kappa; 0) = \frac{d}{4}$ allows us to solve equation (8) recursively:

$$LP(\kappa; Q) = \left(\frac{1}{2\kappa+1}\right)^{Q} \frac{d}{4} - \frac{1}{2}\left[1 - \left(\frac{1}{2\kappa+1}\right)^{Q}\right] \frac{3d}{4} \qquad (9)$$

Equation (9) has a similar intuition as equation (7). The Qth share executes only when Q humans arrive in a row, which has a probability of $\left(\frac{1}{2\kappa+1}\right)^{Q}$ and a payoff of $\frac{d}{4}$. Other than that, the Qth share is either canceled before execution or is sniped, each with probability $\frac{1}{2}\left[1 - \left(\frac{1}{2\kappa+1}\right)^{Q}\right]$, and with payoff 0 and $-\frac{3d}{4}$, respectively. $LP(\kappa; Q)$ is a decreasing function of Q. A share at a later queue position has a lower probability of execution and a higher probability of being sniped.

In Proposition 3, we define and solve the queuing equilibrium. HFTs are the only active players in the queuing equilibrium and their bid-ask spread is always $\frac{d}{2}$. HFTs' key choice is whether to add additional shares to the queue. Their breakeven condition, $LP(\kappa; Q) > 0$, determines the equilibrium queue length.

Proposition 3. (**Binding Tick Size** $\boldsymbol{\Delta = \frac{d}{2} > 0}$): When $\kappa \equiv \frac{\lambda_J}{\lambda_I} < \frac{1}{3}$,.

(1) At almost all times t, HFTs maintain Q^* units at the ask price $v_t + \frac{d}{4}$ and Q^* units at the bid price $v_t - \frac{d}{4}$, where $Q^* = \lfloor \log_{(2\kappa+1)}(5/3) \rfloor$.

(2) When v_t jumps up (down), all HFTs race to snipe Q^* shares at the ask (bid) prices.

(3) HFTs race to fill the queue when the depth at $v_t \pm \frac{d}{4}$ becomes less than Q^*.

Under the queuing equilibrium, HFTs' strategies the state of the LOB and evolve as follows. As before, HFTs' strategy during value jumps is to snipe all stale quotes. The new feature of the queuing equilibrium is to speed race for the top queue position. When the market opens or the security's fundamental value jumps, each HFT sends two orders of size Q^*: one sell order to the price grid $\frac{d}{4}$ above the fundamental value and one buy order to the price grid $\frac{d}{4}$below the fundamental value. HFTs who fail to achieve the first Q^* position cancel their orders. When a non-HFT arrives to reduce the queue length by 1, HFTs race to fill in the Q^{*th} position, and HFTs who fail to fill in the Q^{*th} position cancel their orders. Therefore, a queue of Q^* may contain limit orders from different HFTs. The LOB has only one state: a queue length of Q^* from HFTs on the tick immediately above or below the fundamental value.

Figure 3 shows the comparative statics for the equilibrium queue length with respect to sniping risk. Q^* decreases with $\kappa \equiv \frac{\lambda_J}{\lambda_I}$, which indicates that, for stocks with a bid-ask spread binding at one tick, the depth at the BBO may serve as a proxy for adverse selection risk. Traditionally, bid-ask spreads serve as a proxy for adverse selection risk (Glosten and Milgrom 1985; Stoll 2000). Yet Yao and Ye (2018) find that the bid-ask spread is one-tick wide 41% of the 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.[6]

[Insert Figure 3 about here]

5.2 Flash Equilibrium

[6] 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.

When $\kappa > \frac{1}{3}$, the value of supplying liquidity at $v_t \pm \frac{d}{4}$ is less than the value of sniping the share. Therefore, HFTs no longer quote a bid-ask spread at one binding tick, and BATs are able to quote more aggressive prices than HFTs.

For a BAT who wants to buy, the strategy of providing liquidity at $v_t + \frac{d}{4}$ strictly dominate demanding liquidity at $v_t + \frac{3d}{4}$. A limit buy order priced at $v_t + \frac{d}{4}$ immediately executes like a market order and pays a transaction cost of $\frac{d}{4}$. Due to the tick size constraint, BATs can no longer place limit order at v_t and execute with zero transaction costs. Instead, the lowest price to trigger HFTs is $v_t + \frac{d}{4}$, which has a transaction cost of $\frac{d}{4}$. We call this type of limit orders flash limit orders. Next, BATs need to choose between flash limit orders and regular limit orders, which are limit orders that do not cross the midpoint. Regular limit orders make profits when executed with humans, but lose money when value jumps.

In this section, we characterize the flash equilibrium, in which BATs choose to provide liquidity to HFTs using flash limit orders and HFTs choose to provide liquidity to humans at $v_t \pm \frac{3d}{4}$. Proposition 4 summarizes the flash equilibrium.

Proposition 4. (The Flash Equilibrium) The flash equilibrium occurs when $\Delta = \frac{d}{2}$ and $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} < \kappa < 3(1-\beta)$.[7]

(1) At almost all times t, HFTs maintain Q_F^* units at the ask price $v_t + \frac{3d}{4}$ and Q_F^* units at the bid price $v_t - \frac{3d}{4}$, where $Q_F^* = \lfloor \log_{(2\kappa+1-\beta)} 7 \rfloor$.

[7] $\kappa = \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, i.e. the blue dash line in Figure 2, is the condition that BATs are indifferent in using flash orders and undercutting limit orders. BATs face similar LOB state transition problem in the undercutting equilibrium as in Figure 4, which we demonstrate in Figure A.1. We obtain the boundary between flash and undercutting equilibrium in the proof of Proposition 4.

(2) BAT buyers (sellers) provide liquidity at $v_t + \frac{d}{4}$ $(v_t - \frac{d}{4})$ to HFTs.

(3) HFTs participate in three speed races: (i) race to fill the queue when the depth at $v_t \pm \frac{3d}{4}$ becomes less than Q_F^*; (ii) race to pick off all stale quotes after value jumps; and (iii) race to take the liquidity offered by BATs.

In the flash equilibrium, BATs' strategy is to use flash limit orders to trigger HFTs to demand liquidity. Proposition 5 in the next section pins down the parameter values under which BATs use regular limit orders to provide liquidity to humans.

Similar to the queuing equilibrium, HFTs still race for the top queue position in liquidity provision, because tick size provides rents for liquidity provision even if it is not binding. HFTs always race to snipe all stale quote after values jumps. Part 3 of Proposition 4 reveals a new type of speed competition: racing to be the first to take the liquidity offered by flash limit orders. This race is created by tick size. When price is continuous, Proposition 2 shows that BATs can place limit orders at v_t, which leaves no rents for HFTs who demand liquidity. When price is discrete, a BAT needs to place the buy limit order at $v_t + \frac{d}{4}$, which drives the speed race to capture the rent of $\frac{d}{4}$ through demanding liquidity.

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

Because the limit orders from BATs do not stay in the book, HFTs immediately restore the unique equilibrium state of the LOB after each event. For example, HFTs race to fill in the last position when a human order moves the queue forward, and they race to snipe stale quotes and reestablish their quotes $\frac{3d}{4}$ above and below the fundamental value after each value jump.

5.3 Undercutting Equilibrium

Next, we consider the undercutting equilibrium, in which HFTs quote $v_t + \frac{3d}{4}$ to sell and $v_t - \frac{3d}{4}$ to buy, and BATs quote $v_t + \frac{d}{4}$ to sell and $v_t - \frac{d}{4}$ to buy when these price levels contain no other limit orders. The undercutting equilibrium occurs when the sniping risk is higher than the queuing equilibrium but lower than the flash equilibrium. If the sniping risk is $\kappa \leq \frac{1}{3}$, queuing equilibrium occurs because HFTs will provide liquidity at binding one tick. As κ increases slightly above $\frac{1}{3}$, HFTs would lose money by quoting at $v_t \pm \frac{d}{4}$. However, BATs are able to provide liquidity at $v_t \pm \frac{d}{4}$ as long as the loss is less than $\frac{d}{4}$, because the best outside option for BATs is to use flash limit orders and pay a cost of $\frac{d}{4}$. When $\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, BATs lose more than zero but less than $\frac{d}{4}$ by quoting at $v_t \pm \frac{d}{4}$. So, BATs will use regular limit order to sell at $v_t + \frac{d}{4}$ and buy at $v_t - \frac{d}{4}$.

The undercutting equilibrium have more complex structure than both the queuing equilibrium and the flash equilibrium, because LOB contains more than one state. We define the state of the LOB as (i, j), where i represents the number of BATs' limit orders on the same side of the LOB, and j denotes the number of BATs' limit orders on the opposite side of the LOB. For example, for a trader 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. With Assumption 1, there are four possible states:

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

The characterization of the undercutting equilibrium involves three elements: the strategies of BATs conditional on the states; the strategies of HFTs conditional on states and events; and the transition of the LOB depending on the strategies of HFTs and BATs.

Proposition 5 and Figure 4 characterize the undercutting equilibrium. BATs choose to improve HFTs' quotes by one tick if there is no limit order at the price. When there is a limit order from another BAT, incoming BATs improve the quote by two ticks. Their use of flash limit orders certainly results from our simplifying Assumption 1, but the economic force behind such choice should be the same if we relax Assumption 1. The second share at $v_t \pm \frac{d}{4}$ should have lower execution probability and higher sniping risk, both of which offer higher incentives to choose flash limit orders.

The core of Proposition 5 is to characterize the strategy of HFTs in each state and for each event. Despite the complexity of states and events, we are able to characterize HFTs' strategy as the strategy space of HFTs is rather limited. Proposition 5 shows when $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} < \kappa < 3(1-\beta)$, HFTs always quote at $v_t \pm \frac{3d}{4}$. An HFT loses money by quoting a tighter spread, and she loses price competition by quoting a wider spread. HFTs add more shares to the queue at $v_t \pm \frac{3d}{4}$ if $LP^{(i,j)}(\kappa, \beta; Q) > 0$, and their choice depend on the state of the LOB (i, j).

[Insert Figure 4 about here]

To understand the dynamics of Figure 4, consider $LP^{(0,0)}(\kappa,\beta;Q)$ for an HFT on the ask-side of the LOB.

1) A BAT buyer submits a limit order at $v_t - \frac{d}{4}$, which changes $LP^{(0,0)}(\kappa,\beta;Q)$ to $LP^{(0,1)}(\kappa,\beta;Q)$.

2) A BAT seller undercuts the ask-side price at $v_t + \frac{d}{4}$, which changes $LP^{(0,0)}(\kappa,\beta;Q)$ to $LP^{(1,0)}(\kappa,\beta;Q)$.

3) A human buyer submits a market buy order, which moves the queue position forward by one unit: $LP^{(0,0)}(\kappa,\beta;Q)$ to $LP^{(0,0)}(\kappa,\beta;Q-1)$.

4) A human seller submits a market sell order, which does not affect $LP^{(0,0)}(\kappa,\beta;Q)$ as the LOB on the bid-side is refilled immediately by HFT orders.

5) In an upward value jump, a liquidity-providing HFT on the ask-side gains $-\frac{d}{4}$.

6) In a downward value jump, the liquidity supplier cancels the limit order, thereby changing the value of the liquidity supply to zero.

Across all these six events, 5 and 6 lead to immediate payoffs, and other events only lead to a transition to another state, expect when an HFT is at the top of the queue in state 3. In this case, the limit order executes and the payoff is $\frac{3d}{4}$. Therefore, we obtain the following boundary conditions:

$$LP^{(0,0)}(\kappa,\beta;0) = LP^{(0,1)}(\kappa,\beta;0) = \frac{3d}{4}.$$

The transition across the four states under the six types of events defines the undercutting equilibrium. Proposition 5 summarizes the undercutting equilibrium.

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

1. BATs who intend to buy (sell) submit limit orders at price $v_t - \frac{d}{4}$ $(v_t + \frac{d}{4})$ if no existing limit orders sit at the price level, or buy (sell) limit orders at price $v_t + \frac{d}{4}$ $(v_t - \frac{d}{4})$ otherwise.
2. The transition across states is shown in Figure 4.
3. HFTs' strategy:
   a. HFTs compete for the first Q^{th} position when $LP^{(i,j)}(\kappa, \beta; Q) > 0$, where $LP^{(i,j)}(\kappa, \beta; Q)$ is defined by the following four equations:

$$\begin{cases} LP^{(0,0)}(\kappa,\beta;Q) = max\{0, p_1 LP^{(0,1)}(\kappa,\beta;Q) + p_1 LP^{(1,0)}(\kappa,\beta;Q) + p_2 LP^{(0,0)}(\kappa,\beta;Q-1) + p_2 LP^{(0,0)}(\kappa,\beta;Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,0)}(\kappa,\beta;Q) = max\{0, p_1 LP^{(1,1)}(\kappa,\beta;Q) + p_1 LP^{(1,0)}(\kappa,\beta;Q) + p_2 LP^{(0,0)}(\kappa,\beta;Q) + p_2 LP^{(0,1)}(\kappa,\beta;Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(0,1)}(\kappa,\beta;Q) = max\{0, p_1 LP^{(0,1)}(\kappa,\beta;Q) + p_1 LP^{1,1}(\kappa,\beta;Q) + p_2 LP^{(0,1)}(\kappa,\beta;Q-1) + p_2 LP^{(0,0)}(\kappa,\beta;Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,1)}(\kappa,\beta;Q) = max\{0, p_1 LP^{(0,1)}(\kappa,\beta;Q) + p_1 LP^{(1,0)}(\kappa,\beta;Q) + p_2 LP^{(0,1)}(\kappa,\beta;Q) + p_2 LP^{(1,0)}(\kappa,\beta;Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \end{cases} \quad (10)$$

   and

   $LP^{(0,0)}(\kappa, \beta; 0) = LP^{(0,1)}(\kappa, \beta; 0) = \frac{3d}{4}$.

   b. HFTs race to snipe the stale quotes from HFTs and BATs during value jumps.
   c. HFTs race to take the flash limit orders from BATs.
4. The LOB in each state (i, j) : the depth from HFTs is defined as $Q^{(i,j)} = max\{Q \epsilon \mathbb{N} \big| LP^{(i,j)}(\kappa, \beta; Q) > 0\} \quad i = 0,1; j = 0,1.$

Equation system (10) has four equations. Each equation shows the payoff for liquidity provision for each state (i, j). The payoff for state (i, j) not only depends on the queue position, but also on the payoff of other states. Because the six types of events can change the LOB from one state to the other, each equation in (10) is the sum of six parts. Each part is the product of the probability of the transition and the payoff in the new state. Each equation contains $max\{0, .\}$, because HFTs can simply choose not to submit a limit order or cancel an existing limit order once the payoff is negative.

In Proposition 5, we also characterize the equilibrium LOB for each state. As state (i,j)has already defined the depth from BATs, we focus on the characterization of the depth from HFTs. We present the solution for $LP^{(i,j)}(\kappa,\beta;Q)$ for any $i,j,$ and Q in the Appendix.

[Insert Figure 5 about here]

Figure 5 shows $LP^{(i,j)}(\kappa,\beta;Q)$ decreases in κ and β. Figure 5 also shows the ranking of $LP^{(i,j)}(\kappa,\beta;Q)$ based on states. The first order effect comes from the BATs who undercutting the same side, because the undercutting order from BATs reduces the execution priority of the HFTs by one share. Therefore $LP^{(0,j)}(\kappa,\beta;Q)$ is greater than $LP^{(1,j)}(\kappa,\beta;Q)$ and $LP^{(1,j)}(\kappa,\beta;Q) \approx LP^{(0,j)}(\kappa,\beta;Q+1)$. Figure 5 shows the BATs who undercut on the opposite side have second order effect. $LP^{(i,1)}(\kappa,\beta;Q)$ is higher than $LP^{(i,0)}(\kappa,\beta;Q)$, but not by much. For example, in state (1, 1), a BAT buyer takes liquidity at price $v_t+\frac{d}{4}$ and changes the state to (0, 1), which enables an HFT limit sell order at price $v_t+\frac{3d}{4}$ to trade with the next buy market order from a human. In state (1, 0), a BAT buyer chooses to submit a limit order at price $v_t-\frac{d}{4}$, which changes the state to (1, 1). An HFT limit sell order at price $v_t+\frac{3d}{4}$ then needs to wait at least one more period for execution. More generally, an undercutting BAT limit buy (sell) order may attract future BAT sellers (buyers) to demand liquidity, making future BATs less likely to undercut HFTs. In turn, the value of liquidity provision increases, thereby incentivizing HFTs to supply larger depth. This indirect effect is so small that it does not affect depth.

5.4 Stub Quotes and Mini-flash Crashes

In Proposition 6, we show that HFTs quote a bid-ask spread wider than the size of the jump when sniping risk is high or the fraction of BATs is large. We call such quotes stub quotes. A mini-flash crash occurs when a market order hits a stub quote. In our model, the size of the mini-flash crash is only $\frac{5d}{4} > d$, because the size of a value jump is fixed at d. An increase in the support of jump size can lead to stub quotes further away from the midpoint, thereby creating larger mini-flash crashes. Such an extension is presented in the online Appendix, which shows that the crash size can be as large as the maximum possible jump size.

Proposition 6 (**The Crash Equilibrium**). When $\Delta = \frac{d}{2} > 0$ and $\kappa > 3(1-\beta)$,

(1) HFTs quote a bid price of $v_t - \frac{5d}{4}$ and an ask price of $v_t + \frac{5d}{4}$.

(2) A BAT buyer quote a bid price $v_t - \frac{3d}{4}$ and a BAT seller quote an ask price $v_t + \frac{3d}{4}$ if the price levels has no limit orders. Otherwise, BAT buyers submit flash limit orders at price $v_t + \frac{d}{4}$ and BAT sellers submit flash limit orders at price $v_t - \frac{d}{4}$

(3) HFTs participate in two speed races: (i) race to pick off all stale quotes after value jumps; and (ii) race to take the flash liquidity offered by BATs at $v_t \pm \frac{d}{4}$.

Proposition 6 shows when sniping risk is high, HFTs quote outside the maximum possible fundamental value jump range. We call these quotes stub quotes because HFTs effectively quit liquidity provision. The depth at a half spread of $\frac{5d}{4}$ is infinite because sniping cost is always lower than the profit of liquidity provision. Therefore, we do not directly model the speed race on top of the queue.

High sniping risk is one driver for stub quotes. Also, when HFTs' quotes are wider than one tick, BATs are able to quote more aggressive prices than HFTs. HFTs, who face a reduced liquidity demand, have to further widen their bid-ask spread. Therefore, HFTs initiate stub quotes when sniping risk or the fraction of BATs is high.

The strategy for HFTs is simple. HFTs quote infinite depth at $v_t \pm \frac{5d}{4}$. These wider quotes leave four price levels for BATs. For example, for a BAT to sell, they can quote $v_t + \frac{3d}{4}$, $v_t + \frac{d}{4}$, $v_t - \frac{d}{4}$, or $v_t - \frac{3d}{4}$. We find BATs only use two price levels $v_t + \frac{3d}{4}$ and $v_t - \frac{d}{4}$. We prove this results in the appendix, and we offer the intuition for this result below.

Sell at $v_t - \frac{d}{4}$ strictly dominates $v_t - \frac{3d}{4}$ because both price levels immediately attract HFTs to demand liquidity, and $v_t - \frac{d}{4}$ has a lower cost of $\frac{d}{4}$. $v_t - \frac{d}{4}$ also strictly dominates$v_t + \frac{d}{4}$. Proposition 4 shows that HFTs choose flash limit orders at $v_t - \frac{d}{4}$ over regular limit orders at $v_t + \frac{d}{4}$ when sniping risk is higher than $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$. As $\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} < 3(1-\beta) < \kappa$, the sniping risk in this section is even higher, giving higher incentive to choose $v_t - \frac{d}{4}$ over $v_t + \frac{d}{4}$. In summary, a flash limit order at price $v_t - \frac{d}{4}$ strictly dominates both a more aggressive flash limit order and a regular limit order at $v_t + \frac{d}{4}$.

Finally, limit orders at $v_t + \frac{3d}{4}$ dominate flash limit orders at $v_t - \frac{d}{4}$. Quotes at $v_t + \frac{3d}{4}$ lose $\frac{d}{4}$ in value jumps but have strictly positive profits when a human takes the quote. Therefore, BAT sellers start with a quote at $v_t + \frac{3d}{4}$ when the price level does not contain an order and use flash limit order otherwise.[8]

As BATs do not consistently provide liquidity, humans will hit the stub quote when BATs

[8] Once again, the use of a flash limit order is a consequence of Assumption 1.

are not present. When humans hit a stub quote, a mini-flash crash occurs. We predict the probability of mini-flash crashes in Corollary 2.

Corollary 2. A mini-flash crash is more likely to occur when sniping risk is high. The probability of mini-flash crashes first increases with the fraction of BATs and then decreases with the fraction of BATs. A downward (upward) mini-flash crash is more likely to follow a downward (upward) value jump.

An increase in sniping risk increases the probability of mini-flash crashes by increase the probability for stub quotes. An increase in the fraction of BATs also creates more stub quotes, but at the same time reduce the probability to hit the stub quotes. Therefore, mini-flash crashes never occur when all non-HFTs are humans ($\beta = 0$) and when all non-HFTs are BATs ($\beta = 1$), and probability of flash crashes first increase and then decrease in β. Figure 6 shows the simulated probability of mini-flash crashes are the highest at medium level of β.[9]

[Insert Figure 6 about here]

Figure 6 shows that the majority of flash crashes occur after a value jump. Without a value jump, limit orders from BATs would accumulate in the limit order book, preventing humans from hitting stub quotes. After an upward (downward) jump, HFTs would remove BATs' limit orders from the ask (bid) side. If BATs' limit orders do not reconvene in the LOB, a market buy (sell) order from a human trader would hit stub quotes. Therefore, most of the upward (downward) mini-flash crashes occur after an upward (downward) value jump. Only a few crashes are due to BATs' liquidity being used up by human traders.

Figure 6 shows that an effective way to prevent a mini-flash crash is a trading halt to let

[9] For each β, we first uniformly draw 100 κ from [0, 1], the support of the sniping risk in our paper. For each κ, we simulate 100,000 trades. For all 10 million simulations, we count the number of trades that hit the stub quotes relative to the total number of trades.

the trading interests of BATs reconvene. The blue circle line in Figure 6 shows the intensity of mini-flash crashes with trading halts. We impose the trading halt after a value jump, and the market reopens after five orders arrive. We find that such a trading halt reduces mini-flash crashes by about 90%.

6. Predictions and policy implications

Our model rationalizes a number of puzzles in the literature on HFTs and generates new empirical predictions that can be tested. In Subsection 6.1, we summarize the predictions on who supplies liquidity and when. In Subsection 6.2, we examine the predictions on liquidity demand. In Subsection 6.3, we evaluate the predictions on liquidity.

6.1 Liquidity supply

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

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

Speed advantages in the LOB reduce HFTs' adverse selection costs (see Jones (2013) and Menkveld (2016) surveys), inventory costs (Brogaard et al., 2015), and operational costs (Carrion, 2013). These reduced costs of intermediation raise the concern that "HFTs use their speed advantage to crowd out liquidity supply when the tick size is small and stepping in front of standing limit orders is inexpensive" (Chordia et al., 2013, p. 644). However, Brogaard et al. (2015) find that non-HFTs quote a tighter bid-ask spread than HFTs, and Yao and Ye (2018) find that non-HFTs are more likely to establish price priority over HFTs as the tick size decreases. We find that the opportunity cost of supplying liquidity can reconcile the contradiction between the empirical results and the channels of speed competition. BATs incur lower opportunity costs when supplying liquidity. When they implement a trade, they supply liquidity as long as it is less costly to demand liquidity. The make-take spread that we introduce in Section 4 indicates that BATs never demand liquidity from HFTs when tick size is not binding.

Prediction 2 (Queuing and Time Priority): HFTs crowd out non-HFTs' liquidity supply when tick size is large.

When tick size is binding, the speed advantage of HFTs allows them to establish time priority at the same price. Yao and Ye (2018) find that tick size is more likely to be binding when tick size increases. They also find that a large tick size crowds out non-HFTs' liquidity supply. Both results provide evidence to support Prediction 2.

Prediction 3 (Adverse Selection and Liquidity Provision): An increase in adverse selection risk decreases the fraction of liquidity provided by HFTs.

Prediction 3 differs significantly from the existing literature on HFTs. Existing literature usually model HFTs as traders who have faster access to information. In this framework, speed competition should be more active when there is more information. Particularly, Hoffmann (2014), Han, Khapko, and Kyle (2014), Bernales (2016), and Bongaerts and Van Achter (2016) find that HFTs have lower adverse selection costs than non-HFTs. Therefore, an increase in the level of information should give HFTs a comparative advantage in liquidity provision.

Prediction 3, however, argues that less information drives speed competition. Compare Proposition 3 with Proposition 4, 5 and 6; when the level of sniping risk is low, the binding bid-ask spread drives the speed competition at constrained spread. Once sniping risk is high, the spread is wider than one tick, which allows non-HFTs to undercut HFTs, and decreases HFTs' liquidity supply. One limitation of our model is that we only model adverse selection led by sniping, but other types of adverse selection should provide the same economic mechanism. Generally, the breakeven bid-ask spread should be lower when adverse selection risk is low. Once the breakeven spread is below one tick, speed competition to achieve time priority should be more critical. Empirically, Yao and Ye (2018) show that the fraction of liquidity provided by HFTs is low when adverse selection risk is high, consistent with prediction 3.

Prediction 4 addresses who provide liquidity during a mini-flash crash.

Prediction 4. (**Stub Quotes and Mini-Flash Crashes**): A mini-flash crash is more likely to occur when the sniping risk is high. The probability of flash crashes first increases in the fraction of BATs and then decreases in the fraction of BATs. During a mini-flash crash, HFTs supply liquidity and non-HFTs demand liquidity. A downward (upward) mini-flash crash is more likely to follow a downward (upward) value jump.

HFTs' limit orders face lower execution probability and higher sniping cost when sniping risk is high. When the sniping cost is high enough, HFTs effectively quit liquidity supply by quoting stub quotes. HFTs are more likely to quote stub quotes when sniping risk is high as higher sniping risk widens the break-even bid-ask spread; a wider break-even bid-ask spread also allows BATs to undercut HFTs, which further increases the adverse selection costs for HFTs. Because BATs do not continuously supply liquidity in the market, human's market orders can hit stub quotes and cause mini-flash crashes. A high sniping risk also implies more value jumps relative to the arrival rate of non-HFTs. Human's market orders are more likely to hit stub quotes after value jumps because value jumps clear BATs' limit orders on the side of the jump.

An increase in β creates two competing economic forces for mini-flash crashes. An increase in the fraction of BATs increases the probability for HFTs to provide stub quotes, but BATs are less likely to hit stub quotes. When all non-HFTs are humans or when all non-HFTs are BATs, mini-flash crashes do not happen. Therefore, mini-flash crashes result from the interaction of three types of traders. The probability of min-flash crashes increases in β when it is low, and the probability of min-flash crashes decreases in β when it is high. To the best of our knowledge, these implications have not yet been tested.

Our model predicts that an initial downward (upward) jump increases the probability of a downward (upward) mini-flash crash. The downward (upward) jump clears the LOB on the bid (ask) side, making the market orders from humans more likely to hit stub quotes.

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

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

Our interpretation of mini-flash crashes has two additional features that are consistent with economic reality. First, markets recover quite quickly from mini-flash crashes. In our model, mini-flash crashes disappear when the limit orders from BATs replenish the LOB. Second, Nanex, the firm that invented the concept of mini-flash crash, finds that mini-flash crashes are equally likely to be upward as downward. Indeed, even during the famous Flash Crash on May 6, 2010, in which the Dow Jones plunged 998.5 points, some stocks, including Sotheby's, Apple Inc., and Hewlett-Packard, increased in value to over $100,000 in price (SEC, 2010). In our model, upward and downward mini-flash crashes are equally likely.

6.2 Liquidity demanding

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

Prediction 5. (**Speed Competition of Taking Liquidity**): Non-HFTs are more likely to supply liquidity at price levels that cross the midpoint (flash limit orders) than HFTs. HFTs are also more likely to demand liquidity from flash limit orders, but they do not adversely select these orders.

Latza, Marsh, and Payne (2014) find evidence consistent with Prediction 3. They classify a market order as “fast” if it executes against a standing limit order that is less than 50 milliseconds old. These fast market orders should come from HFTs. These authors also find that fast market orders often execute against limit orders that cross the midpoint, and they lead to virtually no permanent price impact.

Prediction 5 offers fresh perspectives on the liquidity demand from HFTs. Typically, HFTs demand liquidity when they employ a speed advantage to adversely select liquidity suppliers (BCS; Foucault, Kozhan, and Tham, 2017; Menkveld and Zoican, 2017). Therefore, liquidity demand from HFTs generally has negative connotations of reducing liquidity (Jones, 2013; Biais and Foucault, 2014). We find that HFTs’ liquidity demand does not necessarily adversely select slow traders. Instead, the liquidity demand from HFTs can reduce the transaction costs of non-HFTs. In the flash equilibrium, BATs pay $\frac{3d}{4}$ when HFTs supply liquidity, while BATs only pay $\frac{d}{4}$ when HFTs demand liquidity.

6.3 Liquidity

On April 5, 2012, President Barack Obama signed the Jumpstart Our Business Startups (JOBS) Act. Section 106 (b) of the Act requires the SEC to examine the effect of tick size on initial public offerings (IPOs). On October 3, 2016, the SEC implemented a pilot program to increase the tick size from one cent to five cents for 1,200 small- and mid-cap stocks. Proponents of the proposal argue that a larger tick size can improve liquidity (Weild, Kim, and Newport, 2012). In Prediction 6, however, we posit that an increase in tick size decreases liquidity.

Prediction 6. A larger tick size increases the transaction costs.

Yao and Ye (2018) and Albuquerque, Song, and Yao (2018) find evidence consistent with Prediction 6. Our model prediction, along with their empirical evidence shows that an increase in tick size would not improve liquidity.

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

7. Conclusion

This paper contributes to the literature on HFTs by including two salient features in financial markets: discrete tick size and algorithmic traders who are not HFTs. BATs are more likely to provide liquidity when prices are continuous enough, because providing liquidity is always 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 sniping 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 Yao and Ye's (2018) empirical findings.

Our model also provides several new testable predictions. We predict that 1) non-HFTs are more likely than HFTs to provide liquidity at price levels that cross the midpoint, and these limit orders are more likely to be taken by HFTs; 2) a mini-flash crash is more likely to occur for stocks with higher sniping risk; 3) the probability of a mini-flash crash first increases with the fraction of

BATs, and then decreases with it; and 4) an upward (downward) mini-flash crash is more likely to follow an initial price jump in the same direction.

Our model shows that a larger tick size increases transaction costs and drives arms race in speed. These results challenge the rationale for the recent policy initiative to increase the tick size to five cents, and we encourage regulators to consider decreasing the tick size, particularly for liquid stocks.

The inclusion of trading algorithms designed by sophisticated non-HFTs adds significant new insight. For example, we find that BATs can prompt HFTs to demand liquidity using flash limit orders to reduce transaction costs. Therefore, we should not evaluate the impact of HFTs on liquidity and social welfare based on whether they demand or provide liquidity.

We take an initial step to examine the interaction between high-frequency and non-high-frequency algorithms, but our model is parsimonious. For example, BATs in our model do not have private information and they choose order types only upon arrival. Extending our model toward more realistic setups would prove to be fruitful.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

———. 2009. Informed traders and limit order markets. *Journal of Financial Economics* 93:67-87.

Han. J., M. Khapko, and A. S. Kyle. 2014. Liquidity with High-Frequency Market Making.

Working Paper, Swedish House of Finance, University of Toronto, and University of Maryland.

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

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

Menkveld, A. J. 2016. The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics* 8:1-24.

————, and M. A. Zoican. 2017. Need for speed? Exchange latency and liquidity. *Review of Financial Studies* 30:1188-1228.

O'Hara, M. 2015. High frequency market microstructure. *Journal of Financial Economics* 116:257-270.

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

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

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

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

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

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

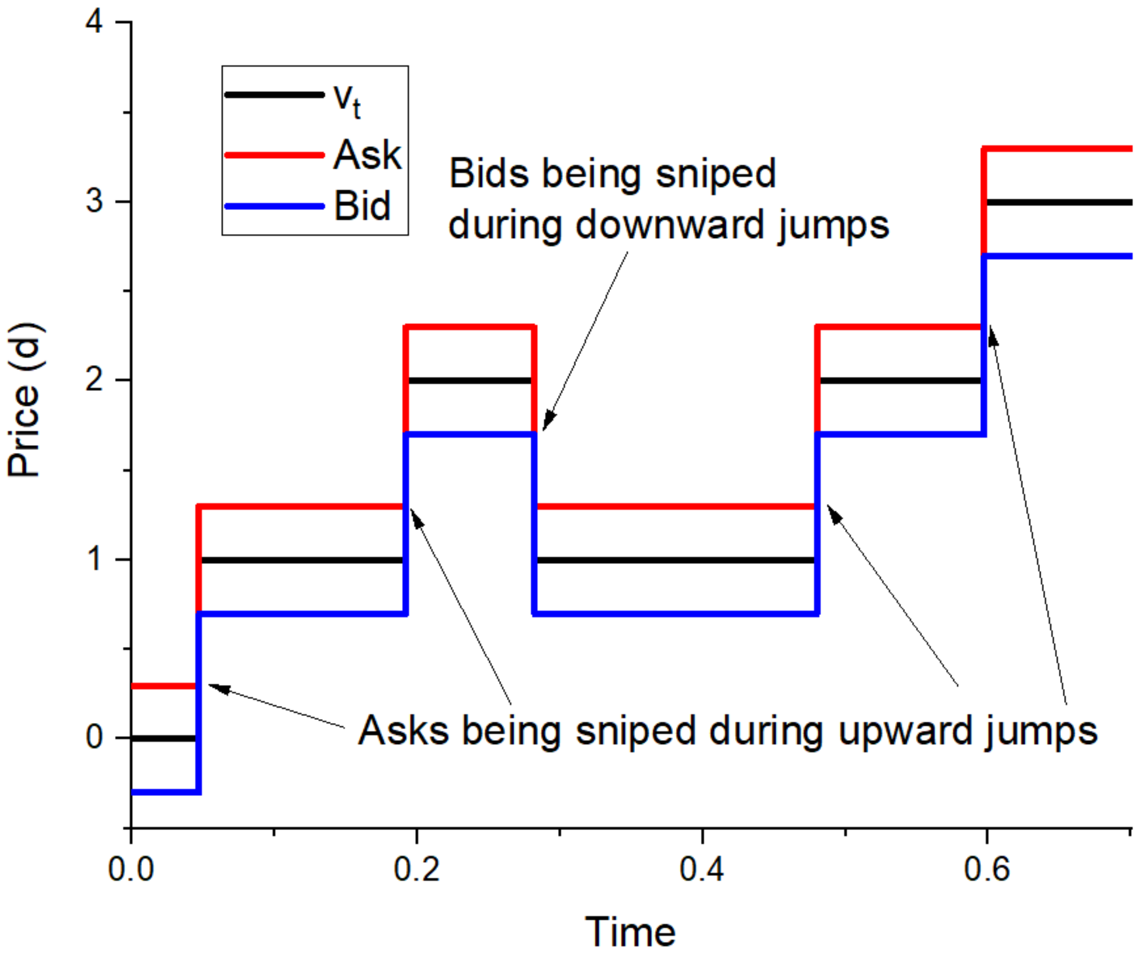

Figure 1. Evolution of fundamental value and bid-ask spread

This figure illustrates the dynamics of fundamental value v_t and the corresponding BBOs. Value jumps arrive with Poisson intensity λ_J and change v_t by size d or $-d$. Investors arrive with Poisson intensity λ_I and do not change v_t. HFTs quote $v_t - \frac{s_1^*}{2}$ as the best bid price and $v_t + \frac{s_1^*}{2}$ as the best ask price. As $s_1^* = \frac{2\lambda_J}{\lambda_I+\lambda_J} d < 2d$, quotes at best bid (ask) are sniped when the fundamental value jump upward (downward). Parameter: $\kappa \equiv \frac{\lambda_J}{\lambda_I} = \frac{1}{2}$.

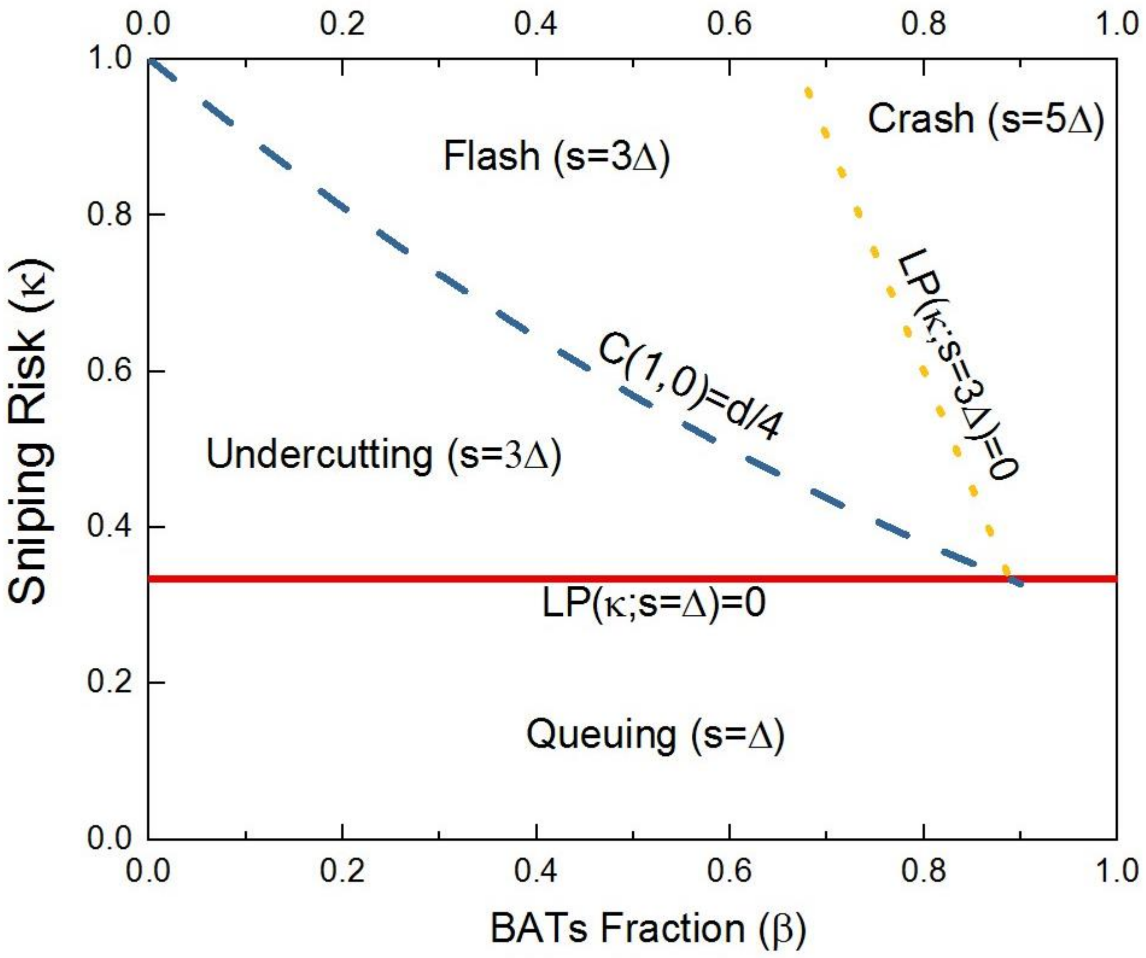

Figure 2: Four Types of Equilibrium

This figure demonstrates four types of equilibrium depending on $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ and β. When $\kappa < \frac{1}{3}$, HFTs queue at one-tick bid-ask spread at $v_t \pm \frac{d}{4}$ (Proposition 3). When $\frac{1}{3} < \kappa < 3(1-\beta)$, HFTs quote three-tick bid-ask spread at $v_t \pm \frac{3d}{4}$, and BAT buyers can either choose to submit limit orders at $v_t + \frac{d}{4}$ and $v_t - \frac{d}{4}$. BATs use flash buy (sell) limit orders at $v_t + \frac{d}{4}$ ($v_t - \frac{d}{4}$) when the sniping risk is relatively high ($\frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4} < \kappa < 3(1-\beta)$), and HFTs immediately take liquidity from BATs (Proposition 4). BATs choose to undercut HFTs when the sniping risk is relatively low ($\frac{1}{3} < \kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$), and they wait to provide liquidity to humans (Proposition 5). When the sniping risk is very high ($\kappa > 3(1-\beta)$), the liquidity provision profit for three-tick spread $LP(\kappa; s = 3\Delta) < 0$, and HFTs quote stub quotes at $v_t \pm \frac{5d}{4}$. Mini-flash crash happens when a human order hit stub quotes (Proposition 6). Boundary conditions are defined in the propositions.

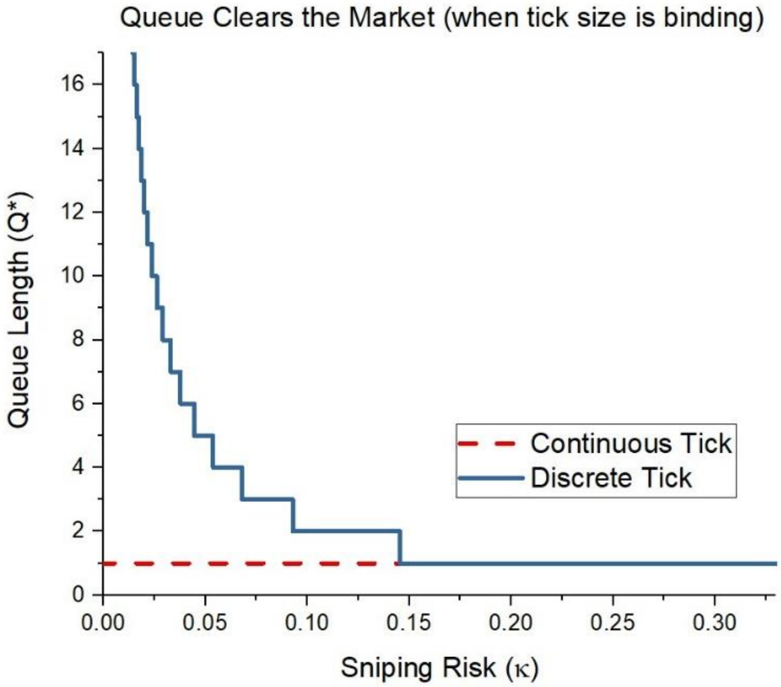

Figure 3. Queue Length Monotonically Decreases with Sniping Risk

This figure demonstrates the relation between HFTs' jointly quoted queue length and sniping risk $\kappa \equiv \frac{\lambda_J}{\lambda_I}$ when tick size is binding (preventing spread from going below one tick). HFTs only quote one share when sniping risk is large; they form longer queues at the BBO when sniping risk is lower. $Q^* \to \infty$ when $\kappa \to 0$. Parameters: $J = d,\ \Delta = \frac{d}{2}$.

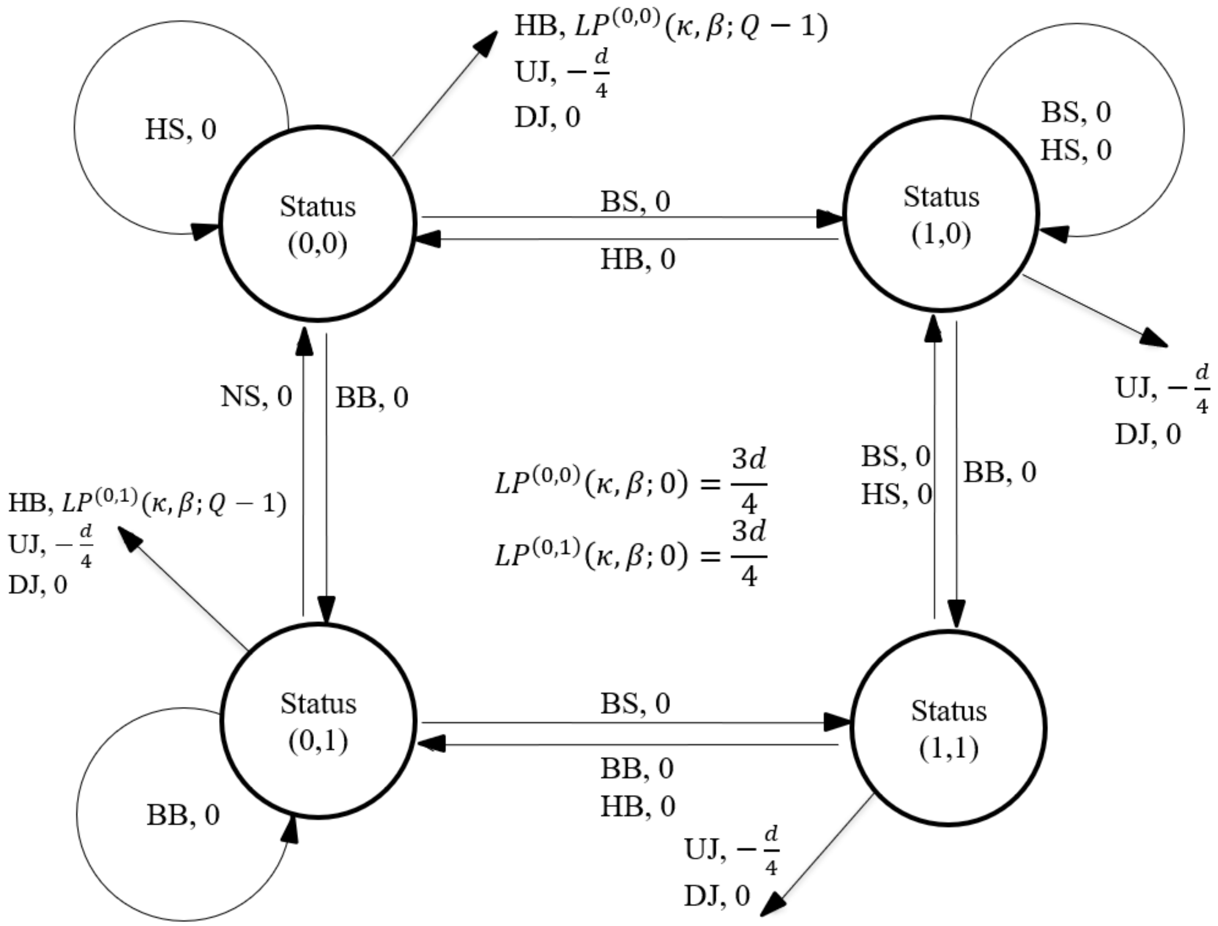

Figure 4. States and Payoffs for the Q^{th} HFT Liquidity Supplier on the Ask-Side

This figure illustrates the Markov transition between LOB states and payoffs from ask side HFT liquidity providers' view. In undercutting equilibrium, HFTs quote $v_t \pm \frac{3d}{4}$ and BATs can submit undercutting orders at $v_t \pm \frac{d}{4}$. In state (i,j), the number of undercutting BAT sell orders at $v_t + \frac{d}{4}$ is i, while the number of buy order at $v_t - \frac{d}{4}$ is j. BB and BS represent the arrival of BATs' buy and sell limit orders, NB and NS represent the arrival of human traders' buy and sell market orders, and UJ and DJ denote the upward and downward value jumps. The arrows between states represent state transitions, and while arrows toward the outside represent either order execution, cancellation, or HFT sell queue moves. The number next to the event is the payoff of the event.

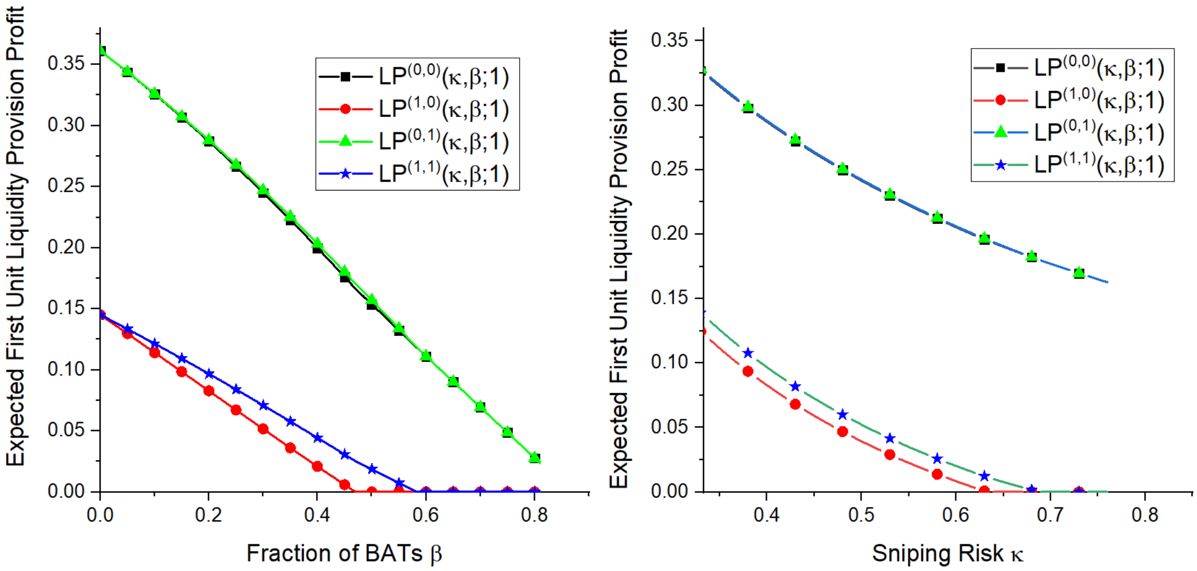

Figure 5. Liquidity Provision Profit for Different LOB States

Panel A shows that HFTs' liquidity provision profit decreases in the fraction of BATs β, as the larger the β, the less order flow that HFTs can receive. Panel B shows the profit decreases in sniping risk κ. The square and triangle lines are much higher than the star and circle lines, i.e. $LP^{(0,j)}(\kappa,\beta;1) > LP^{(1,j)}(\kappa,\beta;1)$. It is the first order effect because the undercutting BATs order directly lowers the execution priority of HFT limit orders. Star lines are also higher than circle lines, i.e. $(LP^{(1,1)}(\kappa,\beta;1) > LP^{(1,0)}(\kappa,\beta;1)$, because the BAT undercut order on the contra side can help accelerate the execution of BATs undercutting order on the same side, and help recover the state to $(0,j)$. The triangle line is only slightly higher than the square line, because $LP^{(0,1)}(\kappa,\beta;1)$ only faces uncertain future threats of being undercut, and the acceleration benefit from the contra side BAT is only marginal. These relative relationships also hold for any queue position of HFTs. The graph cuts off when β and κ is large, and flash equilibrium kicks in.

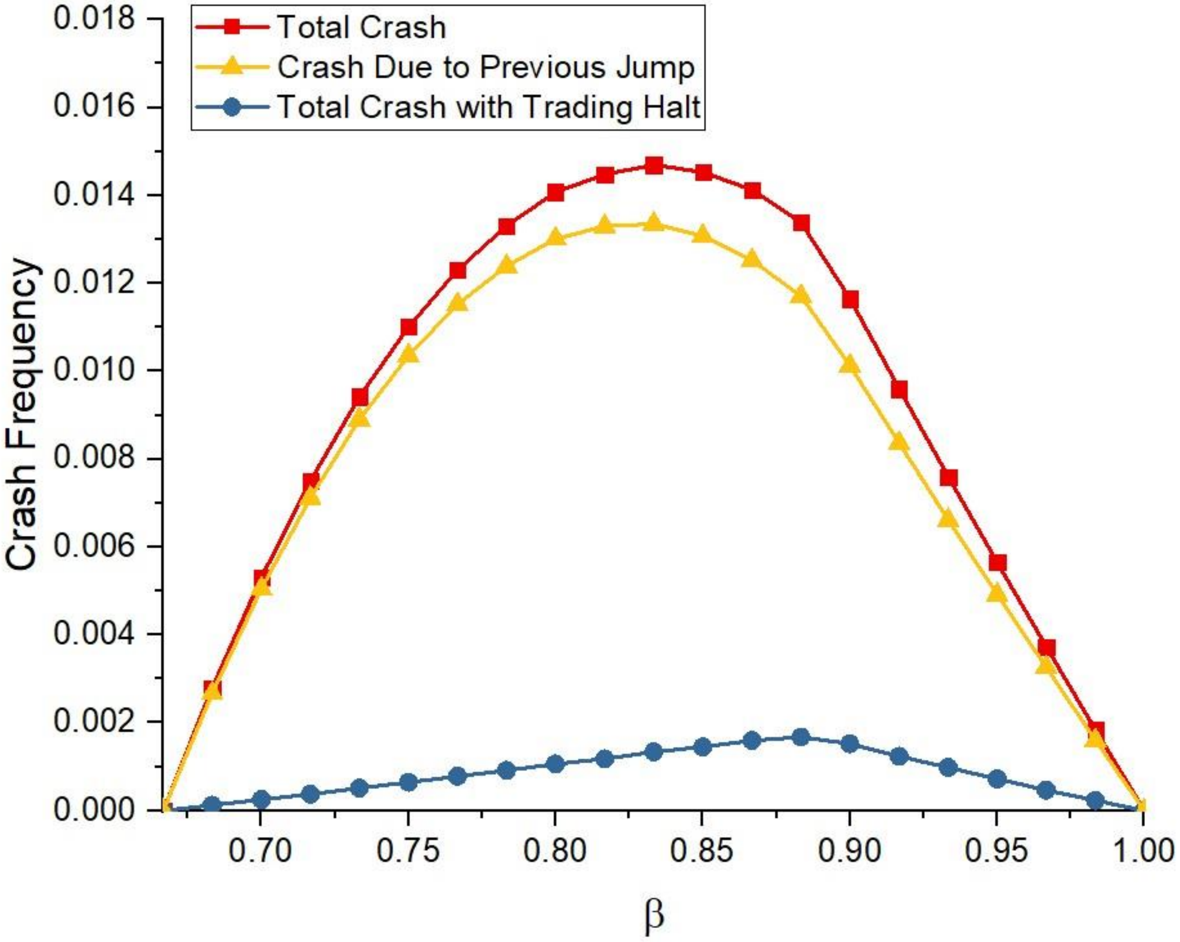

Figure 6: Mini-flash Crash Frequency and $\boldsymbol{\beta}$

This figure shows the mini-flash crash frequency (normalized by number of trades) with respect to the fraction of BATs. For each β, we uniformly draw 100 samples from [0,1] as κ, which is the support of the sniping risk in our paper. For each κ, we initialize the system with 10,000 trades and simulate over 90,000 later trades. For all 9 million simulations, we count the number of trades that hit the stub quotes relative to the total number of trades. The square line shows the intensity for total crashes. The triangle line shows that the majority of mini-flash crashes occur after a value jump (and a small fraction of crashes occur after BATs' liquidity is consumed by humans). The circle line shows that trading halts reduce the frequency of mini-flash crashes. We impose trading halts after each value jump, and the market reopens when the market receives five orders.

Appendix A. Proofs

Proof of Proposition 1

We exclude all off-equilibrium paths in the proof as in BCS.

Firstly, we stated in the model setup that all non-HFTs trade immediately when they arrive. Though we do not impose a delay cost on non-HFTs, there is no benefit for non-HFTs who delay trades because the bid-ask spread s_1^* is a constant and v_t is martingale.

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

1. Any HFT who crosses the midpoint always loses money instantly.
2. Liquidity provider(s) and snipers earn the same expected profit for each share on the LOB. Any HFT who narrows the bid-ask spread will (1) earn less than the original liquidity provider when executed with a non-HFT, and (2) lose more than the original when being sniped during a value jump. Thus, there's no profitable deviation strategy in narrowing the spread for HFTs.
3. Any HFT who quotes at $v_t \pm \frac{s_1^*}{2}$ after an existing limit order will be less likely to trade with a non-HFT because the second share does not have execution priority. She has to wait longer in expectation and is more likely to be sniped.
4. Any HFT who quotes wider than $v_t \pm \frac{s_1^*}{2}$ but within $v_t \pm d$ can never trade with a non-HFT, because each non-HFT only trades one share, and other HFTs will refill the liquidity provision share after it has been consumed by a non-HFT.
5. Quoting stub quotes outside $v_t \pm d$, though, is possible because we have restricted our value jump size to d. It is also possible in BCS that HFTs can submit "orders that trade

with probability zero.” To simplify the state space of our model, we assumed in the main text that no traders can submit limit orders far away from the book. ■

Proof of Proposition 2

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

Firstly, submitting limit orders at v_t (flash orders) and getting zero transaction cost is the best outcome for BATs. All other execution strategies would lead to a positive transaction cost. BATs who cross the midpoint always realize an instant positive transaction cost. BATs who narrow the $v_t \pm \frac{s_2^*}{2}$ bid-ask spread, post limit sell orders at $v_t + \frac{s}{2}$ or limit buy order at $v_t - \frac{s}{2}$, have expected transaction cost $C\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = -LP\left(\lambda_I, \lambda_J; s\right) = -[\frac{\lambda_I}{\lambda_I+2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{\lambda_I+2\lambda_J}\left(d - \frac{s}{2}\right)$ [10] . The cost is monotonically decreasing in s, and $C\left(\lambda_I, \lambda_J; s\right) = 0$ when $s = s_2^*$. Thus, we have $C > 0$ when $s < s_2^*$. BATs who add orders at $v_t \pm \frac{s_2^*}{2}$ after the existing limit order have positive transaction cost because they do not have execution priority. BATs who quote wider than $v_t \pm \frac{s_2^*}{2}$ can only trade with snipers, because each non-HFT only trades one share, and other HFTs will refill the liquidity provision share after the share have been consumed by a non-HFT. Our assumption in Section 3 forbids any trader from quoting outside $v_t \pm d.$ [11]

Secondly, HFTs who accept the BAT order at v_t get zero payoffs. No HFT can get a payoff larger

[10] when BATs provide liquidity and earn positive liquidity provision revenue, it essentially means that BATs have negative transaction cost to execute their trades.

[11] BATs would still not quote outside $v_t \pm d$ without this assumption, because HFTs would undercut BATs if they observe $-C\left(\lambda_I, \lambda_J, \frac{s}{2}\right) = LP\left(\lambda_I, \lambda_J; s\right) = \frac{\lambda_I}{\lambda_I+2\lambda_J}\frac{s}{2} - \frac{\lambda_J}{\lambda_I+2\lambda_J}\left(d - \frac{s}{2}\right) > 0$ at any time t. In other words, HFTs only allow BATs to be on the top of the LOB when $-C \leq 0 \Leftrightarrow C\left(\lambda_I, \lambda_J, \frac{s}{2}\right) \geq 0$. Therefore, there is no way that BATs can attain negative transaction cost.

than zero by deviating to the strategy that she does not accept the BAT's order, because other HFTs immediately accepted the BAT order. Thus, the deviator cannot extract a sniping profit from the BAT's order. Also, for the same reason in Proposition 1, no HFT can get a higher payoff than $LP(\lambda_I, \lambda_J, s) = SN(\lambda_I, \lambda_J, s)$ on the shares quoted by HFT(s) at $v_t \pm \frac{s_2^*}{2}$.

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

Proof of Corollary 1

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

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

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

Proof of Proposition 3

We prove $LP(\kappa; Q) = \left(\frac{1}{2\kappa+1}\right)^Q \frac{d}{4} - \frac{1}{2}\left[1 - \left(\frac{1}{2\kappa+1}\right)^Q\right]\frac{3d}{4}$ using mathematical induction.

1. First, we have $LP(\kappa; 1) = \frac{1}{2\kappa+1}\frac{d}{4} - \frac{\kappa}{2\kappa+1}\frac{3d}{4}$, 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.

$$LP(\kappa; Q+1) = \frac{1}{2\kappa+2} LP(\kappa; Q) + \frac{1}{2\kappa+2} LP(\kappa; Q+1) - \frac{\kappa}{2\kappa+2}\frac{3d}{4} + \frac{\kappa}{2\kappa+2} \cdot 0$$

$$
\begin{aligned}
LP(\kappa; Q+1) \; &= \; \frac{1}{2\kappa+1} LP(\kappa; Q) - \frac{\kappa}{2\kappa+1}\frac{3d}{4} \\
&= \left(\frac{1}{2\kappa+1}\right)^{Q+1}\frac{d}{4} - \frac{1}{2}\left[\frac{1}{2\kappa+1} - \left(\frac{1}{2\kappa+1}\right)^{Q+1}\right]\frac{3d}{4} - \frac{\kappa}{2\kappa+1}\frac{3d}{4} \\
&= \left(\frac{1}{2\kappa+1}\right)^{Q+1}\frac{d}{4} - \frac{1}{2}[1 - \left(\frac{1}{2\kappa+1}\right)^{Q+1}]\;\frac{3d}{4}
\end{aligned}
$$

Thus, $LP(\kappa; Q) = \left(\frac{1}{2\kappa+1}\right)^{Q}\frac{d}{4} - \frac{1}{2}\left[1 - \left(\frac{1}{2\kappa+1}\right)^{Q}\right]\frac{3d}{4}$ holds with Q replaced by $Q+1$. Hence it holds for all $Q \in \mathbb{N}^{+}$. The biggest Q that makes $LP(\kappa; Q) > 0$ is $Q^{*} = \lfloor \log_{(2\kappa+1)}(5/3) \rfloor$.

No HFT can deviate from Proposition 3 by adding liquidity beyond Q^{*} because $LP(\kappa; Q) < 0$ when $Q > Q^{*}$. No HFT wants to cancel orders within the first Q^{*} shares, giving up $LP(\kappa; Q) > 0 = SN(\kappa; Q)$. No HFT can quote a spread narrower than $\frac{d}{2}$ because of the tick size constraint. No HFT wants to quote a spread wider than $\frac{d}{2}$ because they can never trade with non-HFTs. All HFTs want to snipe a stale quote during value jumps, otherwise it would be sniped by other HFTs immediately. Thus, Proposition 3 is an equilibrium. ■

Proof of Propositions 4

First, HFTs quote $v_t \pm \frac{3d}{4}$ when $\frac{1}{3} < \kappa < 3(1-\beta)$. When quoting $v_t \pm \frac{3d}{4}$ and there is no BAT undercutting the order, the HFT seller who provides the first share of liquidity has expected profit:

$$LP(\kappa; 1) = \frac{1-\beta}{2\kappa+2}\frac{3d}{4} + \frac{1-\beta}{2\kappa+2}LP(\kappa; 1) + \frac{\beta}{2\kappa+2}\cdot LP(\kappa; 1) + \frac{\beta}{2\kappa+2}\cdot LP(\kappa; 1) - \frac{\kappa}{2\kappa+2}\frac{d}{4} + \frac{\kappa}{2\kappa+2}\cdot 0.$$

The RHS terms are: Human buyer arrives and trades with the HFT seller; Human seller arrives on the contra side and the LOB does not change afterwards; BAT buyer arrives and uses flash order,[12] which does not change the LOB state; BAT seller arrives and uses flash order, which does not change the LOB state; Upward jump arrives and loses $\frac{d}{4}$; Downward jump arrives and cancels the order, respectively. The solution of $LP(\kappa;1) > 0$ is $\kappa < 3(1-\beta)$.

Then we calculate the equilibrium queue length. In flash equilibrium, there is no long-lasting order at $v_t \pm \frac{d}{4}$, and HFTs only receive human order flows at $v_t \pm \frac{3d}{4}$. Similar to Proposition 3, we have:

$$LP(\kappa;Q_F) = \left(\frac{1}{2\kappa+1-\beta}\right)^{Q_F}\frac{3d}{4} - \frac{1}{2}\left[1-\left(\frac{1}{2\kappa+1-\beta}\right)^{Q_F}\right]\frac{d}{4}$$

$\left(\frac{1}{2\kappa+1-\beta}\right)^{Q_F}$ is the probability that the order eventually executes with a human order and gains $\frac{3d}{4}$, and $\frac{1}{2}\left[1-\left(\frac{1}{2\kappa+1-\beta}\right)^{Q_F}\right]$ is the probability that the order eventually executes with a snipe order and loses $\frac{d}{4}$. The biggest Q_F that makes $LP(\kappa;Q_F) > 0$ is $Q_F^* = \lfloor \log_{(2\kappa+1-\beta)} 7 \rfloor$.

Then we find the boundary between the flash equilibrium and the undercutting equilibrium. In an undercutting equilibrium, a BAT submits a limit order to an empty LOB $(0,0)$ and changes the state to $(1,0)$; a BAT submits a limit order to $(0,1)$ and changes the state to $(1,1)$. We denote the cost for the first case as $C(1,0)$[13] and the cost for the second case as $C(1,1)$. Then

[12] At the boundary where HFTs quote $v_t \pm \frac{3d}{4}$ and $v_t \pm \frac{5d}{4}$, the sniping risk is too large for BATs to use undercutting orders (Figure 2).

[13] Note that $C(1,j)$ is BAT's cost for execution using limit orders at $v_t \pm \frac{d}{4}$. There is no $C(0,j)$ because the undercutting BAT itself becomes the "1." It is not the same with $LP^{(i,j)}(\kappa,\beta;Q)$, which is HFT's liquidity provision profit at $v_t \pm \frac{3d}{4}$.

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

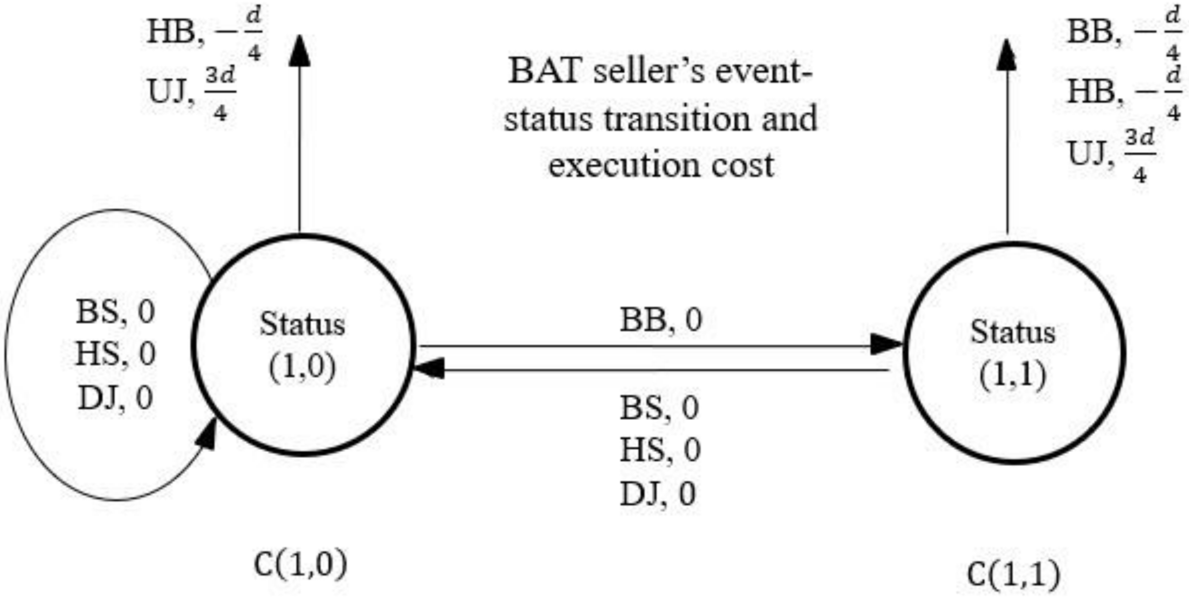

Figure A.1

In equation (A.1) and Figure A.1, we describe six event types that can change the LOB in an undercutting equilibrium. Consider $C(1,0)$ on the ask-side. A human buyer and a human seller each arrive with probability $\frac{1-\beta}{2\kappa+2}$. The BAT seller enjoys a negative transaction cost of $-\frac{d}{4}$ when the human buyer takes his liquidity; the human seller hits a HFT's quote on the bid-side and does not change the state on the ask-side. A BAT buyer and a BAT seller each arrive each with probability $\frac{\beta}{2\kappa+2}$. A BAT buyer posts a limit order on the bid-side and changes the state to $(1,1)$; a BAT seller uses a flash limit order, so the state remains at $(1,0)$. Upward and downward value jumps occur with probability $\frac{\kappa}{2\kappa+2}$. An upward jump leads to a sniping cost of $\frac{3d}{4}$, whereas a downward jump does not change the state of the LOB because the undercutting BAT order pegs to the fundamental value. $C(1,1)$ differs in two ways from $C(1,0)$. First, the arrival of a BAT buyer leads to the execution of a sell limit order from a BAT.[14] Second, a downward jump under $C(1,1)$ leads to sniping on the opposite side of the LOB and changes the state to $C(1,0)$.

The boundary between the flash equilibrium and the undercutting equilibrium is $C(1,0) = \frac{d}{4}$, where BATs are indifferent to using flash orders with a transaction cost of $\frac{d}{4}$ and using undercutting orders, which turns the LOB to state $(1,0)$. Our model starts with no limit orders from BATs, and $C(1,0) < \frac{d}{4}$ is needed to attract the first BAT to use an undercutting order.[15] If $C(1,0) > \frac{d}{4}$, all BATs will face a LOB with state (0,0) and use flash orders (which cost $\frac{d}{4}$).

It is easy to see that $C(1,0) - C(1,1) = p_1\left(C(1,1) + \frac{d}{4}\right) > 0$, i.e., a BAT's undercutting order execution cost will be lower if the contra side has another undercutting order.

The solution for equation (A.1) is:

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

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

$C(1,0) < \frac{d}{4}$ iff $\frac{\kappa(2\kappa+2+\beta)}{(\kappa+1)(2\kappa+2-\beta)} < \frac{1}{2}$, i.e., $2\kappa^2 + 3\kappa\beta + \beta - 2 < 0$.

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

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

Therefore, BATs choose to undercut when $\kappa < \kappa_1$, because $C(1,0) < \frac{\mathrm{d}}{4}$; BATs choose to flash when $\kappa > \kappa_1$.

Finally, we formally describe the perfect equilibrium that the BATs follow:

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

[15] The off-equilibrium path, that there is a BAT who mistakenly jumpstarted the undercutting equilibrium, is discussed beneath this proof.

1. If observing state (0,0) when they arrive at the market, they flash if $\kappa > \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$ and undercut if $\kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$.
2. If observing state (1,0) when they arrive at the market, use flash orders.
3. If observing state (0,1), calculate $C(1,1)$ as the cost of using undercutting limit orders.
   i. If $\kappa < \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, submit a limit order.
   ii. If $\kappa \geq \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, this is an off-equilibrium path[16] because the contra order should have $C(1,0) > \frac{\mathrm{d}}{4}$. The BATs deploy the following strategy to guarantee the off-equilibrium path is not optimal for the deviator, and also guarantees the strategy is perfect: (a) submit a limit order and turn the book status to (1,1) if $1 \geq \kappa \geq \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$; (b) cross midpoint and transact with the off-equilibrium order if $\kappa > 1$.

We check the deviator is indeed losing no less than $\frac{d}{4}$. When $1 \geq \kappa \geq \frac{-3\beta+\sqrt{9\beta^2-8\beta+16}}{4}$, all contra side BATs who observe the deviator will use undercutting limit orders, and the deviator will have an execution cost of $C(1,0) > \frac{\mathrm{d}}{4}$. When $\kappa > 1$, all contra side BATs will use flash orders, which will give the deviator more order flow than in the undercutting equilibrium, i.e. deviator's execution cost $\tilde{C}(1,0)$ is lower than $C(1,0)$. The first term in the right-hand side is different with $C(1,0)$:

$$\tilde{C}(1,0) = p_1\left(-\frac{d}{4}\right) + p_1 \cdot \tilde{C}(1,0) + p_2\left(-\frac{d}{4}\right) + p_2 \cdot \tilde{C}(1,0) + p_3\frac{3d}{4} + p_3 \cdot \tilde{C}(1,0)$$

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

To have $\tilde{C}(1,0) \leq \frac{d}{4}$ the deviator needs $\kappa \leq 1$, which contradicts the condition. Thus, deviation is also suboptimal when $\kappa > 1$.

Similar calculations will show the strategy is subgame perfect. The off-equilibrium path strategies we described above, together with the equilibrium path, will generate the equilibrium outcome sketched in Propositions 4 and 5.

Proof of Propositions 5

In this proof, we solve the HFT liquidity provision depth of the undercutting equilibrium, which follows the four equations as well as the boundary condition $LP^{(0,0)}(\kappa, \beta; 0) = LP^{(0,1)}(\kappa, \beta; 0) = \frac{3d}{4}$.

$$\begin{cases} LP^{(0,0)}(\kappa, \beta; Q) = max\{0, p_1 LP^{(0,1)}(\kappa, \beta; Q) + p_1 LP^{(1,0)}(\kappa, \beta; Q) + p_2 LP^{(0,0)}(\kappa, \beta; Q-1) + p_2 LP^{(0,0)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,0)}(\kappa, \beta; Q) = max\{0, p_1 LP^{(1,1)}(\kappa, \beta; Q) + p_1 LP^{(1,0)}(\kappa, \beta; Q) + p_2 LP^{(0,0)}(\kappa, \beta; Q) + p_2 LP^{(0,1)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(0,1)}(\kappa, \beta; Q) = max\{0, p_1 LP^{(0,1)}(\kappa, \beta; Q) + p_1 LP^{(1,1)}(\kappa, \beta; Q) + p_2 LP^{(0,1)}(\kappa, \beta; Q-1) + p_2 LP^{(0,0)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,1)}(\kappa, \beta; Q) = max\{0, p_1 LP^{(0,1)}(\kappa, \beta; Q) + p_1 LP^{(1,0)}(\kappa, \beta; Q) + p_2 LP^{(0,1)}(\kappa, \beta; Q) + p_2 LP^{(1,0)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \end{cases}$$

The equilibrium depth under state (i, j) is the largest Q that makes $LP^{(0,0)}(\kappa, \beta; Q) > 0$. Sometimes $LP^{(1,\mathrm{j})}(\kappa, \beta; Q) < 0$, which means HFTs find $v_t \pm \frac{3d}{4}$ not profitable because an undercutting order from BATs is present. $LP^{(0,\mathrm{j})}(\kappa, \beta; Q)$ never drops to zero, otherwise HFTs will widen the spread by one more tick to $v_t \pm \frac{5d}{4}$.

This is a linear equation system with $\max\{\cdot\,,0\}$ function. Thus, we can solve the system iteratively, until all $\max\{\cdot\,,0\}$ functions are not binding. Here we give an example for $\kappa = 0.4, \beta = 0.2$. First, we assume that all $LP^{(i,j)}(\kappa = 0.4, \beta = 0.2; 1) > 0$. We analytically solve the four formulas with $Q = 1$ and without the $max\{\cdot\,,0\}$ function, and we insert $\kappa = 0.4, \beta = 0.2$:

$$\begin{aligned}
&LP^{(0,0)}(\kappa = 0.4, \beta = 0.2; 1)\\
&= -\frac{12 - 21\beta + 12\beta^2 - 3\beta^3 + 44\kappa - 61\beta\kappa + 10\beta^2\kappa + \beta^3\kappa + 44\kappa^2 - 64\beta\kappa^2 - 2\beta^2\kappa^2 + 4\kappa^3 - 28\beta\kappa^3 - 8\kappa^4}{4(-4 + 7\beta - 4\beta^2 + \beta^3 - 24\kappa + 18\beta\kappa - 8\beta^2\kappa + 2\beta^3\kappa - 52\kappa^2 + 12\beta\kappa^2 - 4\beta^2\kappa^2 - 48\kappa^3 - 16\kappa^4)}\\
&= 0.2871
\end{aligned}$$

$$\begin{aligned}
&LP^{(1,0)}(\kappa = 0.4, \beta = 0.2; 1)\\
&= -\frac{12 - 21\beta + 12\beta^2 - 3\beta^3 + 16\kappa - 40\beta\kappa + 17\beta^2\kappa + \beta^3\kappa - 12\kappa^2 - 22\beta\kappa^2 + 12\beta^2\kappa^2 - 24\kappa^3 - 8\kappa^4}{4(-4 + 7\beta - 4\beta^2 + \beta^3 - 24\kappa + 18\beta\kappa - 8\beta^2\kappa + 2\beta^3\kappa - 52\kappa^2 + 12\beta\kappa^2 - 4\beta^2\kappa^2 - 48\kappa^3 - 16\kappa^4)}\\
&= 0.0828
\end{aligned}$$

$$\begin{aligned}
&LP^{(0,1)}(\kappa = 0.4, \beta = 0.2; 1)\\
&= -\frac{12 - 21\beta + 12\beta^2 - 3\beta^3 + 44\kappa - 61\beta\kappa + 17\beta^2\kappa - 6\beta^3\kappa + 44\kappa^2 - 64\beta\kappa^2 - 2\beta^2\kappa^2 + 4\kappa^3 - 28\beta\kappa^3 - 8\kappa^4}{4(-4 + 7\beta - 4\beta^2 + \beta^3 - 24\kappa + 18\beta\kappa - 8\beta^2\kappa + 2\beta^3\kappa - 52\kappa^2 + 12\beta\kappa^2 - 4\beta^2\kappa^2 - 48\kappa^3 - 16\kappa^4)}\\
&= 0.2881
\end{aligned}$$

$$\begin{aligned}
&LP^{(1,1)}(\kappa = 0.4, \beta = 0.2; 1)\\
&= -\frac{12 - 21\beta + 12\beta^2 - 3\beta^3 + 44\kappa - 26\beta\kappa + 3\beta^2\kappa + \beta^3\kappa - 12\kappa^2 - 8\beta\kappa^2 - 2\beta^2\kappa^2 - 24\kappa^3 - 8\kappa^4}{4(-4 + 7\beta - 4\beta^2 + \beta^3 - 24\kappa + 18\beta\kappa - 8\beta^2\kappa + 2\beta^3\kappa - 52\kappa^2 + 12\beta\kappa^2 - 4\beta^2\kappa^2 - 48\kappa^3 - 16\kappa^4)}\\
&= 0.0967
\end{aligned}$$

$LP^{(i,j)}(\kappa = 0.4, \beta = 0.2; 1) > 0$ is satisfied. Therefore, the depth is at least one share in any state of the LOB.

Then we assume all $LP^{(i,j)}(\kappa = 0.4, \beta = 0.2; 2) > 0$. Thus, we solve the four formulas with $Q = 2$ and without the $\max\{\cdot, 0\}$ function, and we insert $\kappa = 0.4, \beta = 0.2$ as well as the solution of $LP^{(i,j)}(\kappa = 0.4, \beta = 0.2; 1)$:[17]

$$LP^{(0,0)}(\kappa = 0.4, \beta = 0.2; 2) = 0.0691$$

$$LP^{(1,0)}(\kappa = 0.4, \beta = 0.2; 2) = -0.0271$$

$$LP^{(0,1)}(\kappa = 0.4, \beta = 0.2; 2) = 0.0699$$

$$LP^{(1,1)}(\kappa = 0.4, \beta = 0.2; 2) = -0.0204$$

Now, $LP^{(1,j)}(\kappa = 0.4, \beta = 0.2; 2) < 0$ and liquidity provision at the second place is only profitable when there is no undercutting order, i.e. state $(0, j)$. The HFT occupying the second liquidity provision share in state $(0, j)$ will cancel their orders when a BAT arrives and undercuts them. In other words, the $\max\{\cdot, 0\}$ function is in effect, and we solve the following equations instead:

$$\begin{cases} LP^{(0,0)}(\kappa, \beta; 2) = max\{0, p_1 LP^{(0,1)}(\kappa, \beta; Q) + p_1 \cdot \mathbf{0} + p_2 LP^{(0,0)}(\kappa, \beta; 1) + p_2 LP^{(0,0)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,0)}(\kappa, \beta; 2) = \mathbf{0} \\ LP^{(0,1)}(\kappa, \beta; 2) = max\{0, p_1 LP^{(0,1)}(\kappa, \beta; Q) + p_1 \cdot \mathbf{0} + p_2 LP^{(0,1)}(\kappa, \beta; 1) + p_2 LP^{(0,0)}(\kappa, \beta; Q) + p_3\left(-\frac{d}{4}\right) + p_3 \cdot 0\} \\ LP^{(1,1)}(\kappa, \beta; 2) = \mathbf{0} \end{cases}$$

We have:

$$LP^{(0,0)}(\kappa = 0.4, \beta = 0.2; 2) = 0.0720 > 0$$

$$LP^{(0,1)}(\kappa = 0.4, \beta = 0.2; 2) = 0.0723 > 0$$

[17] The analytical solution is upon request because it is too large to fit the page. Its numerator has 41 terms from constant to $8\kappa^4\beta^4$ and $128\kappa^8$.

Further calculation shows $LP^{(i,j)}(\kappa = 0.4, \beta = 0.2; 3) < 0$. Thus, we conclude that $Q^{(0,0)} = Q^{(0,1)} = 2$ and $Q^{(1,0)} = Q^{(1,1)} = 1$ is the solution for equation (14) when $\kappa = 0.4, \beta = 0.2$.

■

Proof of Proposition 6

We have shown in the proofs of Propositions 4 and 5 that HFTs quote stub quotes at $v_t \pm \frac{5d}{4}$ when $\kappa > 3(1-\beta)$, because $v_t \pm \frac{3d}{4}$ is no longer profitable for liquidity provision.

Without loss of generality, consider the BAT seller's problem, and denote (i, j, k, l) as the outstanding limit orders at $v_t + \frac{3d}{4}, v_t + \frac{d}{4}, v_t - \frac{d}{4}, v_t - \frac{3d}{4}$, respectively. It is easy for BATs seller to exclude the strategy selling at $v_t - \frac{3d}{4}$, because selling at $v_t - \frac{d}{4}$ (flash order) strictly dominates it. Also, BATs seller would not use flash order at $v_t - \frac{d}{4}$ when $v_t + \frac{3d}{4}$ is empty, because the first pays $\frac{d}{4}$ and the second pays less than $\frac{d}{4}$.[18] When $v_t + \frac{3d}{4}$ is occupied, the BAT should choose between $v_t + \frac{d}{4}$ and $v_t - \frac{d}{4}$, and we show a perfect equilibrium that BATs never sell at $v_t + \frac{d}{4}$ as follows:

1. If observing state $(0,0,0,l)$ when they arrive at the market, submit a limit sell order at $v_t + \frac{3d}{4}$ and change the state to $(1,0,0,l)$.
2. If observing state $(1,0,0,l)$ when they arrive at the market, submit a limit sell order at

[18] If the order at $v_t + \frac{3d}{4}$ has been sniped, the BATs seller pays $\frac{d}{4}$, but it is also possible that the BATs seller trade with a non-HFT and realized profit. Of course, the weighted average payoff should be negative, otherwise HFTs would provide liquidity at $v_t + \frac{3d}{4}$. Thus, the expected payoff should be between $-\frac{d}{4}$ and 0, strictly higher than $-\frac{d}{4}$, the payoff of using flash orders.

$v_t - \frac{d}{4}$ (flash order).

3. If observing state $(i, 1, k, l)$, it is an off-equilibrium path and submits a limit sell order at $v_t - \frac{d}{4}$.
4. If observing state $(i, 0{,}1, l)$, it is an off-equilibrium path:
    i. If $3(1-\beta) < \kappa < 1$, submit a limit sell order at $v_t + \frac{d}{4}$ and change the state to $(i, 1{,}1, l)$. This would cost less than submitting a limit sell order at $v_t - \frac{d}{4}$ and trading with the off-equilibrium $k = 1$. Furthermore, this prevents BATs from initial deviation.

    ii. If $\kappa > 1$, submit a limit sell order at $v_t - \frac{d}{4}$ and change the state to $(i, 0{,}0, l)$.

We check the deviating BAT seller at $v_t + \frac{d}{4}$ is indeed losing no less than $\frac{d}{4}$. When $3(1-\beta) < \kappa < 1$, all BAT buyers who observe the deviator will use undercutting limit orders, and the deviator will have an execution cost of $C(1{,}0) > \frac{\mathrm{d}}{4}$. When $\kappa > 1$, all BAT buyers will use flash orders, giving the deviating BAT seller a cost of $\tilde{C}(1{,}0)$:

$$\tilde{C}(1{,}0) = p_1\left(-\frac{d}{4}\right) + p_1 \cdot \tilde{C}(1{,}0) + p_2\left(-\frac{d}{4}\right) + p_2 \cdot \tilde{C}(1{,}0) + p_3\frac{3d}{4} + p_3 \cdot \tilde{C}(1{,}0)$$

To have $\tilde{C}(1{,}0) \leq \frac{d}{4}$ the deviator needs $\kappa \leq 1$ as in the flash equilibrium, which contradicts the condition. Thus, deviation is also suboptimal when $\kappa > 1$. Therefore, BAT sellers never jumpstart the off-equilibrium path by quoting $v_t + \frac{d}{4}$.

HFTs also have no profitable deviation strategy for the same reason in the previous equilibriums. They'll lose money if they narrow the spread or cross the midpoint, and they can never snipe a flash order because other HFTs immediately trade with the flash order.

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