

# Who Supplies Liquidity, and When? \*

Xin Wang<sup>1</sup>

University of Illinois, Urbana-Champaign

Mao Ye<sup>2</sup>

University of Illinois, Urbana-Champaign and NBER

## Abstract

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

---

\* We thank Hengjie Ai, Malcolm Baker, Hank Bessembinder, Eric Budish, Thierry Foucault, Maureen O'Hara, Neil Pearson, Brian Weller, Chen Yao, Bart Yueshen, Marius Zoican, and participants at the Carlson Junior Conference at the University of Minnesota, Wabash River Conference at the Indiana University, and the Smokey Mountain Conference at the University of Tennessee for their helpful suggestions. This research is supported by National Science Foundation grant 1352936 (jointed with the Office of Financial Research at U.S. Department of the Treasury). We thank Sida Li for excellent research assistance.

<sup>1</sup> Department of Economics, University of Illinois at Urbana-Champaign. Email: xinwang5@illinois.edu. Tel: 217-419-9000.

<sup>2</sup> College of Business, University of Illinois at Urbana-Champaign and NBER, 340 Wohlers Hall, 1206 S. 6th Street, Champaign, IL, 61820. Email: maoye@illinois.edu. Tel: 217-244-0474.

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

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

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

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

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

their speed advantage over BATs.

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

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

When the tick size does not impose a constraint for BATs to quote more aggressive prices

than HFTs, our model has two types of equilibria: flash and undercutting. In the flash equilibrium, BATs use flash limit orders to supply liquidity to HFTs. In the undercutting equilibrium, BATs quote a buy limit order price below the fundamental value or a sell limit order price above the fundamental value. These regular limit orders stay in the LOB to supply liquidity to non-algos or other BATs. We find that undercutting equilibrium are more likely to occur when the adverse selection risk is low, because flash limit orders incur no adverse selection cost, whereas the cost of regular limit orders increases with the adverse selection risk.

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

Existing literature on HFTs focuses on the role of adverse selection. On the one hand, speed can allow HFTs to adversely select other traders, which harms liquidity; on the other hand, speed can reduce adverse selection costs for liquidity suppliers and improve liquidity [see Jones (2013), Biais and Foucault (2014), and Menkveld (2016) for surveys]. We contribute to the literature by

identifying two new channels of speed competition, both of which are unrelated to adverse selection. For liquidity demand, we find that HFTs race to demand liquidity when BATs post flash limit orders, but HFTs impose no adverse selection cost on BATs. Instead, BATs prompt HFTs to demand liquidity to reduce their transaction costs. Thus, liquidity demand from HFTs need not be bad. Indeed, transactions costs are lower when HFTs demand liquidity than when they supply liquidity.

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

cancel orders. A smaller tick size or higher adverse selection risk allows BATs to increase liquidity provision by establishing price priority, but smaller tick size or higher adverse selection risk also leads to more frequent order cancellations. This theoretical intuition, along with the empirical evidence in Yao and Ye (2017), suggests that the cancellation-to-trade ratio should not be used as a cross-sectional proxy for HFT activities.<sup>3</sup>

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

## 1. Model

In our model, the stock exchange operates as a continuous limit order book (LOB). Each trade in the LOB requires a liquidity supplier and a liquidity demander. The liquidity supplier submits a limit order, which is an offer to buy or sell at a specified price and quantity. The liquidity demander accepts the conditions of a limit order. Execution precedence for liquidity suppliers follows the price-time priority rule. Limit orders with higher buy or lower sell prices execute before

---

<sup>3</sup> The cancellation-to-trade ratio can still be a good *time series* proxy for HFTs' activity (Hendershott, Jones, and Menkveld, 2011; Angel, Harris, and Spatt, 2015; Boehmer, Fong, and Wu, 2015).

less aggressive limit orders. For limit orders queuing at the same price, orders arriving earlier execute before later orders. The LOB contains all outstanding limit orders. Outstanding orders to buy are called “bids” and outstanding orders to sell are called “asks.” The highest bid and lowest ask are called the “best bid and ask (offer)” (BBO), and the difference between them is the bid-ask spread.

Our model has one security,  $x$ , whose fundamental value,  $v_t$ , evolves as a compound Poisson jump process with arrival rate  $\lambda_J$ .  $v_t$  starts from 0, and changes by a size of  $d$  or  $-d$  in each jump with equal probability. As in BCS,  $v_t$  is common knowledge, but liquidity suppliers are subject to adverse selection risk when they fail to update stale quotes after value jumps. Traders start with a small latency to observe the common value jump,<sup>4</sup> but can reduce the latency to 0 by investing in a speed technology with cost  $c_{speed}$  per unit of time.

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

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

---

<sup>4</sup> By small, we mean that no additional events, such as a trader arrival or a value jump, take place during the delay.



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

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

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

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

(Menkveld and Zoican, 2017). Our extension effectively reduces the delay cost to allow BATs to submit limit orders.<sup>5</sup> The main intuition of our model stays the same as long as BATs do not queue for infinite length.

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

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

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

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

---

<sup>5</sup> We can assume a finite delay cost so that BATs only queue for one share, and the results are available upon request. The value of the delay cost, however, conveys no intuition and only leads to a more complicated proof. In Section 4, we show that the exact size of the delay cost has little impact for BATs' choice between limit orders and market orders.

arrival rate of non-HFTs increases or the intensity of value jumps decreases, the adverse selection risk decreases and so does the break-even bid-ask spread. The break-even bid-ask spread drops below one tick when  $\frac{\lambda_I}{\lambda_J} > 1$ , making liquidity supply for the first share more profitable than stale-quote sniping.<sup>6</sup> The rents for liquidity supply then trigger the race to win time priority in the queue. As BATs do not have a speed advantage to win the race, they demand liquidity in the same manner as non-algo traders. As a result, Lemma 1 does not depend on  $\beta$ .<sup>7</sup>

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

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

---

<sup>6</sup> Throughout this paper, we consider  $\frac{\lambda_I}{\lambda_J} > 1$  for expositional simplicity. When  $\frac{\lambda_I}{\lambda_J} \leq 1$ ,  $\Delta_0$  is no longer binding, and the equilibrium structure is similar to that in Sections 3-6, where we reduce the tick size to  $\Delta_1 = \frac{d}{3}$ .

<sup>7</sup> An order with less time priority has lower probability of execution and higher probability of being sniped, both of which reduce BATs' incentives to queue. In addition, BATs have incentives to implement trades, and a positive delay cost would compel them to use market orders when the queue is long. We assume that BATs never queue after the first position to reflect these intuitions in a parsimonious way.

is  $\frac{d}{2}$ , while the probability of being sniped is  $\frac{N-1}{N}$ . When  $v_t$  jumps downward, the liquidity supplier cancels the order and joins the race to supply liquidity at a new BBO.<sup>8</sup>  $LP(Q)$  then becomes 0. Equation (2) presents  $LP(Q)$  in recursive form and Lemma 2 presents the solution for equation (2).

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

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

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

$LP(Q)$  decreases in  $Q$ .

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

---

<sup>8</sup> We assume that the HFT liquidity supplier cancels the limit order to avoid the complexity of tracking infinite many price levels in the LOB.

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

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

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

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

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

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

The solution for equation (5) is:

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

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

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

Figure 2 shows the comparative statics for equilibrium queue length. The queue length at BBO decreases with  $\frac{\lambda_I}{\lambda_J}$ , which indicates that, for stocks with a bid-ask spread binding at one tick, the depth at the BBO may serve as a proxy for adverse selection risk. Traditionally, bid-ask spreads serve as a proxy for adverse selection risk (Glosten and Milgrom, 1985; Stoll, 2000). Yet Yao and Ye (2017) find that bid-ask spread is one-tick wide 41% of time for their stratified sample of Russell 3000 stocks in 2010. Depth at the BBO then serves as an ideal proxy to differentiate the level of adverse selection for these stocks.<sup>9</sup>

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

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

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

**Proposition 1. (Large Binding Tick Size):** When  $\Delta_0 = d$  and  $\frac{\lambda_I}{\lambda_J} > 1$ ,  $N^*$  HFTs jointly supply  $Q^*$

units of sell limit orders at  $a_t^* = v_t + \frac{d}{2}$  and  $Q^*$  units of buy limit orders at  $b_t^* = v_t - \frac{d}{2}$ , where:

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

---

<sup>9</sup> Certainly, the comparison also needs to control for price, because stocks with the same nominal bid-ask spread may have a different proportional bid-ask spread.

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

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

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

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

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

### 3. Equilibrium types under a small tick size

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

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

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

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

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

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

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

When  $1 < \frac{\lambda_I}{\lambda_J} < 5$ , the break-even bid-ask spread is larger than one tick. To profit from the bid-ask spread, HFTs have to quote the following bid-ask spread:<sup>10</sup>

---

<sup>10</sup> We defer the derivation of the boundary condition for HFTs' bid-ask spread to Sections 4-6. Another way to bypass tick size constraints is to randomize quotes immediately above and below the break-even bid-ask spread. In this paper, we consider only stationary HFT quotes.



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

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

**Insert Figure 3 about Here**

#### 4. Make-take spread

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

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

market order, but with lower cost.

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

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

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

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

## 5. Flash equilibrium versus undercutting equilibrium

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

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

### 5.1 Flash equilibrium

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

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

1. BAT buyers submit limit orders at  $v_t + \frac{d}{6}$  and BAT sellers submit limit orders at price

$$v_t - \frac{d}{6}.$$

2.  $N_f^*$  HFTs jointly supply  $Q_f^*$  units of sell limit orders at  $v_t + \frac{d}{2}$  and  $Q_f^*$  units of buy limit orders at  $v_t - \frac{d}{2}$ , where:

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

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

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

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

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

**Insert Figure 4 about Here**

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

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

As BATs no longer demand liquidity from HFTs, HFTs respond to the reduced liquidity demand and higher adverse selection cost by decreasing their depth to  $Q_f^*$ . The profit to take liquidity from BATs,  $\frac{d}{6}$ , is less than the profit to supply liquidity to BATs at  $\frac{d}{2}$  when the tick size is  $\Delta_0$ . A smaller tick size,  $\Delta_1$ , reduces the profit for HFTs, thereby reducing the number of HFTs.

## 5.2 Undercutting equilibrium

In flash equilibrium, the LOB only has one stable state. In the undercutting equilibrium, the LOB transits across different states. As indicated in Proposition 2, BATs choose regular limit orders over flash limit orders when adverse selection risk or  $\beta$  is low. In the undercutting equilibrium, their limit orders stay in the LOB, and their decisions, as well as those of HFTs,

depend on the state of the LOB. Our technical assumption that BATs never queue at the second position reduces the number of states. Still, the solution is complicated. We focus on deriving the equilibrium *strategies* of HFTs, as Proposition 2 and its proof in the Appendix demonstrate the strategy of BATs in undercutting equilibrium. BATs choose regular limit orders over flash limit orders when  $\frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta} < \frac{\lambda_I}{\lambda_J} < 5$ .

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

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

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

Figure 5 illustrates how  $D^{(i,j)}(Q)$  changes with the six types of events defined in equation

(1). For example, consider  $D^{(0,0)}(Q)$  for an HFT on the ask side of the LOB.

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

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

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

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

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

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

### Insert Figure 5 about Here

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

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

1. HFTs' strategy:

- a. Spread: HFTs quote ask price at  $v_t + \frac{d}{2}$  and bid price at  $v_t - \frac{d}{2}$ .
- b. Depth: The following system of equations determines the equilibrium depth in each state.
  - i. Difference in value between the liquidity supplier and the stale-queue sniper in each state:

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

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

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

$$Q^{(i,j)} = \max\{Q \in \mathbb{N}^+ \mid D^{(i,j)}(Q) > 0\} \quad i = 0,1; j = 0,1.$$

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

2. BATs who intend to buy (sell) submit limit orders at price  $v_t - \frac{d}{6}$  ( $v_t + \frac{d}{6}$ ) if no existing limit orders sit at the price level, or buy (sell) limit orders at price  $v_t + \frac{d}{6}$  ( $v_t - \frac{d}{6}$ ) otherwise<sup>11</sup>.

The depth from HFTs depends on  $D^{(i,j)}(Q)$ .  $D^{(i,j)}(Q)$ , is defined using the equation system in (14), because the value difference in each state also depends on the value differences in other

---

<sup>11</sup> After an upward (downward) jump with size  $d$ , we assume BATs buy (sell) undercutting orders at  $v_t - \frac{d}{6}$  ( $v_t + \frac{d}{6}$ ) will be cancelled and resubmitted at price  $v_t + \frac{5d}{6}$  ( $v_t - \frac{5d}{6}$ ) to follow the value jump. Alternative BATs strategy does not change the equilibrium.



states. The equations in (14) contain the  $\max\{0, \cdot\}$  as HFTs do not queue at the  $Q^{th}$  position once the expected payoff is below 0.

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

Figure 6 also shows that  $LP^{(i,j)}(Q)$  and  $SN^{(i,j)}(Q)$  also depend on the state of the LOB. As the undercutting limit orders from BATs can change the states of the LOB, HFTs can add or cancel their limit orders even when the fundamental value stays the same. A comparison between Panel A and Panel B and between Panel C and Panel D of Figure 6 shows that an undercutting order reduces HFTs' depth on the same side of the LOB by approximately one share. Intuitively, when a BAT submits an undercutting order, the execution priority for all HFTs on the same side of the book decreases by one share.<sup>12</sup> An HFT who used to quote the last share at the half bid-ask spread  $\frac{d}{2}$  has to cancel, because the share become unprofitable after the arrival of the undercutting order. For the same reason, once an undercutting order from a BAT executes, HFTs race to submit one more share at the half bid-ask spread  $\frac{d}{2}$ , because the execution priority in the LOB increases by

---

<sup>12</sup> An undercutting BAT order on the opposite side of the LOB has an indirect effect. For example, in state (1, 1), a BAT buyer takes liquidity at price  $v_t + \frac{d}{6}$  and changes the state to (0, 1), which enables an HFT limit sell order at price  $v_t + \frac{d}{2}$  to trade with the next buy market order from a non-algo trader. In state (1, 0), a BAT buyer chooses to submit a limit order at price  $v_t - \frac{d}{6}$ , which changes the state to (1, 1). An HFT limit sell order at price  $v_t + \frac{d}{2}$  then needs to wait at least one more period for execution. More generally, an undercutting BAT limit buy (sell) order may attract future BAT sellers (buyers) to demand liquidity, making future BATs less likely to undercut HFTs. In turn, the value of liquidity supply increases relative to sniping, thereby incentivizing HFTs to supply larger depth. This indirect effect is so small that it does not affect depth in our numerical example, because the number of shares is an integer. It is possible for a depth of (1, 1) to be higher than (1, 0) for numerical values such as  $\frac{\lambda_I}{\lambda_J} = 4.9$  and  $\beta = 0.06$ , and the results are available upon request.

one. One new feature of the undercutting equilibrium is the frequent order addition or cancellation of HFTs' limit orders in the absence of a change in fundamental value.

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

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

## 6. Stub quotes and mini-flash

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

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

1. HFTs quote a half bid-ask spread of  $\frac{7d}{6}$ .
2. A BAT buyer (seller) quotes  $v_t - \frac{5d}{6}$  ( $v_t + \frac{5d}{6}$ ) if the price level has no limit orders. Otherwise, the BAT buyer (seller) submits a flash limit order at price  $v_t + \frac{d}{6}$  ( $v_t - \frac{d}{6}$ ) to provide liquidity.
3. Compared with the case when  $\Delta_0 = d$ , the transaction cost for non-algo traders increases, but the average transaction cost for non-HFTs decreases.
4. The probability of mini-flash crashes decreases in  $\frac{\lambda_I}{\lambda_J}$ . The probability of mini-flash crashes first increases in  $\beta$  and then decreases in  $\beta$ .

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

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

at  $v_t + \frac{d}{6}$  to obtain immediate execution with a transaction cost of  $\frac{d}{6}$ .<sup>13</sup> We show in the proof that BATs never quote at  $v_t - \frac{d}{2}$  and  $v_t - \frac{d}{6}$  as the execution cost is always higher than  $\frac{d}{6}$ . Flash buy limit orders at price  $v_t + \frac{d}{6}$  also strictly dominate more aggressive flash limit orders of  $v_t + \frac{d}{2}$  and  $v_t + \frac{5d}{6}$ , because a limit order price of  $v_t + \frac{d}{6}$  is aggressive enough to trigger immediate execution.

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

An increase in adverse selection risk unambiguously increases the probability of mini-flash crashes. Figure 3 in Section 3 show that stub quotes are more likely to occur when there higher

---

<sup>13</sup> This result is certainly a consequence of our simplifying assumption that BATS cannot queue for a second share. However, BATs should always have higher incentives to use flash limit orders when  $v_t - \frac{5d}{6}$  contains a limit order, because the second share has a lower probability of executing against a non-algo trader and a higher probability of executing against a sniper, whereas a flash limit order always incurs a constant cost of  $\frac{d}{6}$ .

adverse selection risk. Conditional on stub quotes occurring, Figure 6 reveals another channel for adverse selection risk to increase the number of mini-flash crashes. An increase in adverse selection risk implies more value jumps relative to the arrival rate of non-algo traders. During an upward (downward) value jump, BATs' limit orders on the bid (ask) side are all sniped and only stub quotes remain. If the limit orders from BATs fail to reconvene before a non-algo trader arrives, the market order from the non-algo trader hits the stub quote and causes a mini-flash crash.

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

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

value jump. Only a small amount of crashes are due to BATs' liquidity being used up by non-algo traders.

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

**Insert Figure 7 About Here**

## **6. Predictions and policy implications**

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

### **6.1 Liquidity supply**

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

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

Speed advantages in the LOB reduce HFTs' adverse selection costs (see Jones (2013) and Menkveld (2016) surveys), inventory costs (Brogaard et al., 2015), and operational costs (Carrion, 2013). These reduced costs of intermediation raise the concern that "HFTs use their speed advantage to crowd out liquidity supply when the tick size is small and stepping in front of standing limit orders is inexpensive" (Chordia et al., 2013, p. 644). However, Brogaard et al. (2015) find that non-HFTs quote a tighter bid-ask spread than HFTs, and Yao and Ye (2017) find that non-HFTs are more likely to establish price priority over HFTs as the tick size decreases. We find that the opportunity cost of supplying liquidity can reconcile the contradiction between the empirical results and the channels of speed competition. BATs incur lower opportunity costs when supplying liquidity. When they implement a trade, they supply liquidity as long as it is less costly to demand liquidity. The make-take spread that we introduce in Section 4 indicates that BATs never demand liquidity from HFTs when tick size is not binding.

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

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

Hoffmann (2014), Han, Khapko, and Kyle (2014), Bernales (2016), and Bongaerts and Van Achter (2016) find that HFTs have lower adverse selection costs than non-HFTs. Yao and Ye (2017), however, find that HFTs do not have a comparative advantage in providing liquidity for

stocks with higher adverse selection risk. In Prediction 2, we provide the economic mechanism to reconcile this inconsistency. Comparing Corollary 1 with Proposition 2 and 3, we find that the tick size is more likely to be binding when adverse selection risk is low. A binding tick size helps HFTs to supply liquidity through time priority. An increase in adverse selection risk raises the break-even bid-ask spread above one tick, allows non-HFTs to undercut HFTs, and decreases HFTs' liquidity supply.

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

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

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



spread also allows BATs to undercut HFTs, which further increases the adverse selection costs for HFTs. Because BATs do not continuously supply liquidity in the market, non-algo traders' market orders can hit stub quotes and cause mini-flash crashes. A high adverse selection risk also implies more value jumps relative to the arrival rate of non-HFTs. Non-algo traders' market orders are more likely to hit stub quotes after value jumps, because value jumps clear BATs' limit orders on the side of the jump.

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

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

Our interpretations of mini-flash crashes are consistent with both negative and positive

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

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

## 6.2 Liquidity demanding

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

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

Latza, Marsh, and Payne (2014) find evidence consistent with Prediction 4. They classify

a market order as “fast” if it executes against a standing limit order that is less than 50 milliseconds old. Because of the speed of taking liquidity, it is natural to expect that fast market orders are from HFTs. These authors also find that fast market orders often execute against limit orders that cross the midpoint, and they lead to virtually no permanent price impact.

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

### **6.3 Liquidity**

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

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

Yao and Ye (2017) find evidence consistent with Prediction 5. Holding the BBO constant, an increase in depth at the BBO implies an increase in liquidity. Yet these authors also find that the quoted bid-ask spread increases after an increase in tick size. When both quoted bid-ask spread and depth increase, the most relevant liquidity measure becomes the effective bid-ask spread, the transaction cost paid by liquidity demanders (Bessembinder, 2003). Our model shows that constrained price competition increases the effective bid-ask spread, which is consistent with Yao and Ye's (2017) findings. Our model prediction, along with the evidence in Yao and Ye (2017), shows that an increase in tick size would not improve liquidity.

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

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

An increase in tick size reduces mini-flash crashes, but it also increases the transaction costs for average trades. A more effective solution to prevent mini-flash crashes would be to slow down the market, particularly during periods of market stress. In a standard Walrasian equilibrium, price is continuous and time is discrete. Modern financial markets exhibit exactly the opposite

structure: price competition is constrained by the tick size, whereas time is divisible at the nanosecond level in electronic trading platforms (Gao, Yao, and Ye, 2013). Making price more continuous and time more discrete would improve liquidity and also prevent mini-flash crashes at the same time.

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

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

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

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

(2014) shows that flickering quotes occur when new information causes the price to move to a new level. We show that HFTs can cancel orders in the absence of information. Periodic order additions and cancellations also differ from Baruch and Glosten (2013), who rationalize flicking quotes using a mixed-strategy equilibrium. An increase in adverse selection risk, defined as the intensity of value jumps relative to the arrival rate of non-HFTs, also lead to more order cancellations, but HFTs also provide less liquidity for these stocks. Taken together, we suggest that the cancellation-to-trade ratio should not be used as a cross-sectional measure of HFTs' activity.

## **7. Conclusion**

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

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

Yao and Ye (2017) find that the cancellation ratio, a widely used empirical proxy for HFTs' activity, has a negative cross-sectional correlation with HFT liquidity supply. We provide a theoretical foundation for their surprising negative correlation. A large tick sizes induces HFTs to race for the top queue position, and HFTs are less likely to cancel orders once they secure this spot. HFTs cancel orders more frequently for stocks with smaller tick sizes, but they also supply less liquidity. Both theoretical and empirical evidence suggests that researchers should not apply the cancellation ratio as a cross-sectional proxy for HFT activity.

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

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

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

## References

- Angel, J., L. Harris, and C. Spatt. 2015. Equity trading in the 21st century: An update. *The Quarterly Journal of Finance* 5:1550002-1-1550002-39.
- Baruch, S., and L. R. Glosten. 2013. Fleeting orders. Columbia Business School Research Paper: 13-43.
- Bernales, A. 2016. Algorithmic and High Frequency Trading in Dynamic Limit Order Markets. Working Paper, Universidad de Chile.
- Bessembinder, H. 2003. Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis* 38:747-777.
- Biais, B., and T. Foucault. 2014. HFT and market quality. *Bankers, Markets & Investors* 128:5-19.
- Boehmer, E., K. Fong, and J. Wu. 2015. International evidence on algorithmic trading. Working Paper, Singapore Management University, University of New South Wales, and University of Nebraska at Lincoln.
- Bongaerts, D., and M. V. Achter. 2016. High-Frequency Trading and Market Stability. Working Paper, Erasmus University Rotterdam.
- Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan. 2015. Trading Fast and Slow: Colocation and Liquidity. *Review of Financial Studies* 28:3407-43.
- Brogaard, J., A. Carrion, T. Moyaert, R. Riordan, A. Shkilko, and K. Sokolov. Forthcoming. High-frequency trading and extreme price movements. *Journal of Financial Economics*.
- Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan. 2015. Trading fast and slow: Colocation and liquidity. *Review of Financial Studies* 28:3407-3443.
- Budish, E., P. Cramton, and J. Shim. 2015. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics* 130:1547-1621.
- Carrion, A. 2013. Very fast money: High-frequency trading on the NASDAQ. *Journal of*



- Financial Markets* 16:680-711.
- Chordia, T., A. Goyal, B. N. Lehmann, and G. Saar. 2013. High-frequency trading. *Journal of Financial Markets* 16:637-645.
- Clark-Joseph, A.D., M. Ye, and C. Zi. Forthcoming. Designated market makers still matter: Evidence from two natural experiments. *Journal of Financial Economics*.
- Colliard, J. E., and T. Foucault. 2012. Trading fees and efficiency in limit order markets. *Review of Financial Studies* 25:3389-3421.
- Foucault, T., R. Kozhan, and W.W. Tham. 2017. Toxic arbitrage. *Review of Financial Studies* 30:1053-1094.
- Frazzini, A., R. Israel, and T. J. Moskowitz. 2014. Trading costs of asset pricing anomalies. Working paper, AQR Capital Management, and University of Chicago.
- Glosten, L. R., and P. R. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics* 14:71-100.
- Goettler, R. L., C. A. Parlour, and U. Rajan. 2005. Equilibrium in a dynamic limit order market. *Journal of Finance* 60:2149-2192.
- . 2009. Informed traders and limit order markets. *Journal of Financial Economics* 93:67-87.
- Han, J., M. Khapko, and A. S. Kyle. 2014. Liquidity with High-Frequency Market Making. Working Paper, Swedish House of Finance, University of Toronto, and University of Maryland.
- Hasbrouck, J., and G. Saar. 2013. Low-latency trading. *Journal of Financial Markets* 16:646-679.
- Hendershott, T., C. M. Jones, and A. J. Menkveld. 2011. Does algorithmic trading improve liquidity?. *Journal of Finance* 66:1-33.
- Hendershott, T. and A. J. Menkveld. 2014. Price pressures. *Journal of Financial Economics* 114:405-423.
- Hoffmann, P. 2014. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics* 113:156-169.
- Jiang, G., L. Ingrid, and V. Giorgio. 2014. High-Frequency Trading around Macroeconomic News Announcements: Evidence from the U.S. Treasury Market. Working Paper, Bank of Canada.
- Jones, C. 2013. What do we know about high-frequency trading? Working paper, Columbia

- University.
- Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica* 53:1315-1335.
- Latza, T., We. W. Marsh, and R. Payne. 2014. Fast aggressive trading. Working paper, Blackrock, and City University London.
- Menkveld, A. J. 2016. The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics* 8:1-24.
- , and M. A. Zoican. 2017. Need for speed? Exchange latency and liquidity. *Review of Financial Studies* 30:1188-1228.
- O'Hara, M. 2015. High frequency market microstructure. *Journal of Financial Economics* 116:257-270.
- ., G. Saar, and Z. Zhong. 2015. Relative tick size and the trading environment. Working Paper, Cornell University, and University of Melbourne.
- Parlour, C.A. 1998. Price dynamics in limit order markets. *Review of Financial Studies* 11:789-816.
- Rosu, We. 2009. A dynamic model of the limit order book. *Review of Financial Studies* 22:4601-4641.
- Stoll, H.R., 2000. Presidential address: friction. *The Journal of Finance* 55:1479-1514.
- United States. Commodity Futures Trading Commission, and Securities and Exchange Commission. 2010. *Findings regarding the market events of May 6, 2010*.
- Weild, D., E. Kim, and L. Newport. 2012. The trouble with small tick sizes. Grant Thornton.
- Yao, C., and M. Ye. 2017. Conditionally accepted. Why trading speed matters: A tale of queue rationing under price controls. *Review of Financial Studies*.
- Yueshen, B.Z. 2014. Queuing uncertainty in limit order market. Working Paper, INSEAD.

## Appendix

### Proof for Lemma 1

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

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

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

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

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

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

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

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

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

### Proof for Lemma 2

We prove Lemma 2 using mathematical induction.

1. From the proof for Lemma 1,

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

which satisfies equation (3).

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

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

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

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

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

### Proof of Proposition 2

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

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

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

**Insert Figure A.1 about Here**

In equation (A.1) and Figure A.1, we describe six event types that can change the LOB in an undercutting equilibrium. Consider  $C(1,0)$  on the ask side. A BAT buyer and a BAT seller each arrive each with probability  $p_1$ . A BAT buyer posts a limit order on the bid side and changes the state to  $C(1,1)$ ; a BAT seller uses a flash limit order so the state remains at  $C(1,0)$ . A non-algo buyer and a non-algo seller arrive each with probability  $p_2$ . The BAT seller enjoys a negative transaction cost of  $-\frac{d}{6}$  when the non-algo buyer takes his liquidity; the non-algo seller hits a HFT's quote on the bid side and does not change the state on the ask side. Upward and downward value jumps occur with probability  $p_3$ . An upward jump leads to a sniping cost of  $\frac{5d}{6}$ , whereas a downward jump does not change the state of the LOB.<sup>14</sup>  $C(1,1)$  differs in two ways from  $C(1,0)$ . First, the arrival of a BAT buyer leads to execution of a sell limit order from a BAT.<sup>15</sup> Second, a downward jump under  $C(1,1)$  leads to sniping on the opposite side of the LOB and changes the state to  $C(1,0)$ .

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

---

<sup>14</sup> Here we assume that BATs position their order one tick above the new fundamental value. BATs are able to reposition their orders because they face no competition from other BATs in a short time period.

<sup>15</sup> The execution of this order results from our assumption that BATs do not queue after another limit order at the same price, but the intuition that a longer queue on the bid side increases the execution probability on the ask side holds true generally (Parlour, 1998).

than  $-\frac{d}{6}$  because of the cost of being sniped. Therefore,  $C(1,0) - C(1,1) = p_1 \left( C(1,1) + \frac{d}{6} \right) > 0$ .

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

The solution for equation (A.1) is:

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

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

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

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

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

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

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

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

equilibrium sell (buy) order is present<sup>16</sup>. The off-equilibrium sell (buy) order will have an execution cost as follows:

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

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

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

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

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

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

---

<sup>16</sup>In flash equilibrium, any BAT's undercutting limit order is off-equilibrium.

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

1. If  $2 < R < \frac{1+2\beta+\sqrt{4\beta^2+9}}{2-\beta}$ , use limit order under off-equilibrium path:
  - i> If there is no order at  $-\frac{d}{6}$ , submit a limit sell order at  $-\frac{d}{6}$ .
  - ii> Else, submit a limit sell order at  $\frac{d}{6}$ .
2. If  $R < 2$ , use flash order under off-equilibrium path:
  - i> Submit a limit sell order at  $-\frac{d}{6}$  regardless of state of the book.

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

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

### **Proof of Proposition 3**

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

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



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

We then obtain:

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

$$\begin{aligned}
&= \frac{8 + 12R + 12\beta R - 4R^2 + 24\beta R^2 + 2\beta^2 R^2 - 12R^3 + 21\beta R^3 - 2\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)} \\
&= 0.2202,
\end{aligned}$$

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

$$\begin{aligned}
&= \frac{8 + 24R + 20R^2 + 6\beta R^2 - 4\beta^2 R^2 + 12\beta R^3 - 5\beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)} \\
&= 0.0527,
\end{aligned}$$

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

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

$$\begin{aligned}
&= \frac{8 + 24R + 20R^2 + 2\beta^2 R^2 + 6\beta R^3 + \beta^2 R^3 - \beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4}{2(-16 - 48R - 52R^2 + 12\beta R^2 - 4\beta^2 R^2 - 24R^3 + 18\beta R^3 - 8\beta^2 R^3 + 2\beta^3 R^3 - 4R^4 + 7\beta R^4 - 4\beta^2 R^4 + \beta^3 R^4)} \\
&= 0.0593.
\end{aligned}$$

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

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

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

We get:

$$\begin{aligned}
D^{(0,0)}(2) &= 0.0448, \\
D^{(1,0)}(2) &= -0.0602 < 0, \\
D^{(0,1)}(2) &= 0.0451, \\
D^{(1,1)}(2) &= -0.0561 < 0.^{17}
\end{aligned}$$

We reject the assumption that all  $D(2) > 0$ . Therefore, under certain states of the LOB, HFTs would not supply the second share of liquidity. We start from the worst state for liquidity suppliers, (1,0), in which a BAT undercuts HFTs on the same side of the LOB, but no BAT undercuts HFTs on the other side of LOB.<sup>18</sup> Therefore,  $D^{(1,0)}(2) = 0$  and all other  $D^{(i,j)}(2) > 0$ .

Thus we solve:

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

---

<sup>17</sup> For brevity, the closed-form solution is not presented, but it is available upon request.

<sup>18</sup> In this state, an HFT liquidity supplier on the ask side cannot trade with the next non-HFT buyer, because a BAT buyer chooses to supply liquidity and changes the state to (1,1), and a non-algo buyer chooses to take the BAT seller's liquidity and changes the state to (0,0).

We obtain:

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

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

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

However,  $D^{(1,1)}(2)$  is still smaller than 0. We further assume that  $D^{(1,1)}(2)$  is also 0, i.e.,

HFTs cancel the second order when BATs submit limit orders on both sides. Therefore,

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

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

We obtain:

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

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

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

#### Proof of Proposition 4

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

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

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

$$LP_{\frac{5d}{6}}(1) = \frac{(1-\beta)\lambda_l}{\lambda_l + 2\lambda_j} \frac{5d}{6} - \frac{\lambda_j}{\lambda_l + 2\lambda_j} \frac{dN-1}{6N}$$

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

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

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

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

1. If there is no limit sell order on  $\frac{d}{6}, \frac{d}{2}$ , and  $\frac{5d}{6}$ , submit a limit sell order at  $\frac{5d}{6}$ .
2. Else, if there is no limit buy order on  $-\frac{d}{6}$ , submit a limit sell order at  $-\frac{d}{6}$ .
3. Else, there is a limit buy order on  $-\frac{d}{6}$  (this is an off-equilibrium path, there are two possible responses, same intuition as the proof of proposition 2)
  - i> If  $R > 2$ , submit a limit sell order at  $\frac{d}{6}$ , costs  $C(1,1)$ .
  - ii> Else, submit a limit sell order at  $-\frac{d}{6}$ , costs  $\frac{d}{6}$ .

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

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

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

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

Since in flash crash equilibrium  $R(1 - \beta) < \frac{1}{5}$ , the BAT's cost is at least  $\frac{d}{2} \cdot \frac{4/5}{6/5} = \frac{d}{3} > \frac{d}{6} =$

*Cost of flash order*. Thus, it is never optimal to submit a limit order at  $\frac{d}{2}$ .

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

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

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

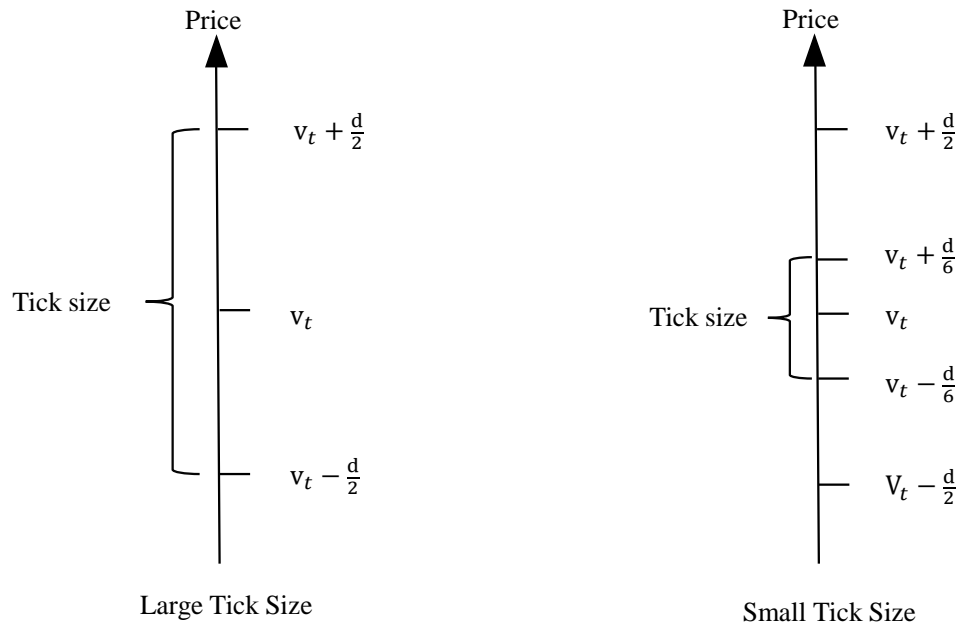
Finally, the BAT seller will post a sell limit order at  $\frac{5d}{6}$ . Her cost is:

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

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

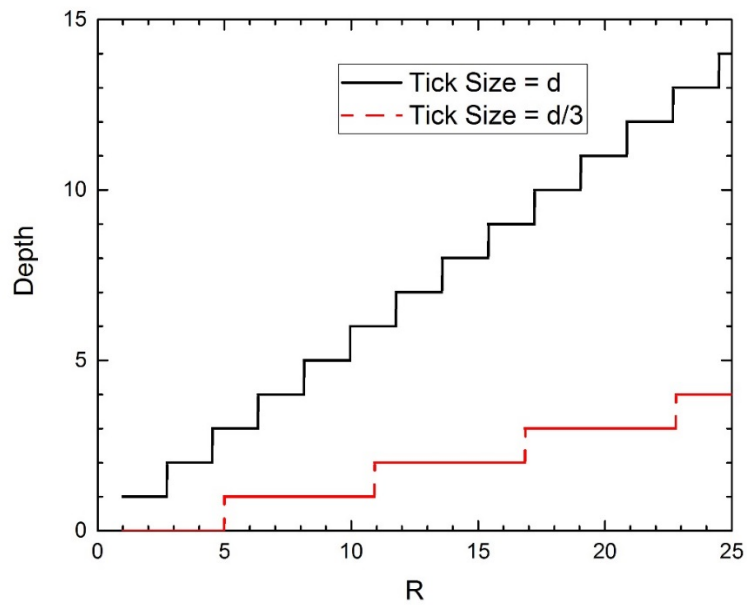
### Figure 1: Pricing Grid under Large vs. Small Tick Sizes

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



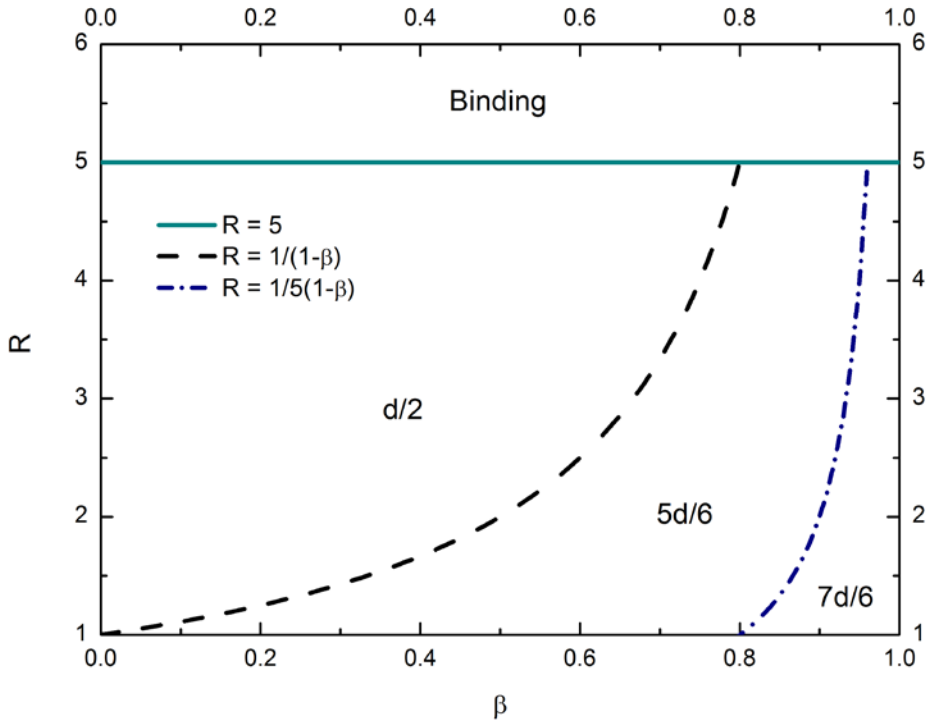
### Figure 2: Depth and the Adverse Selection Risk under a Binding Tick Size

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



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

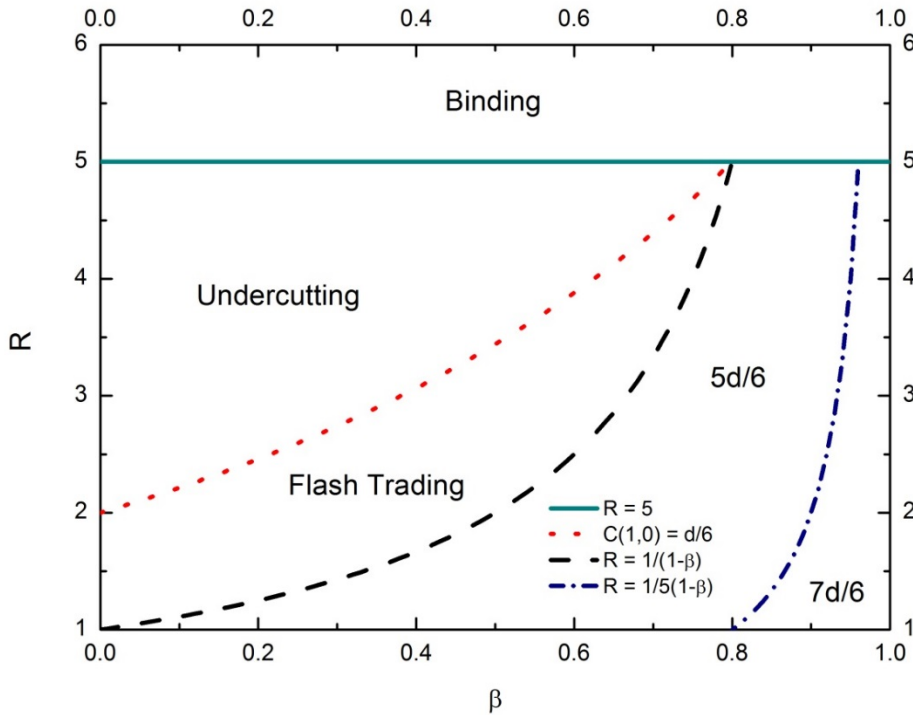
This figure demonstrates the half bid-ask spread quoted by HFTs as a function of  $\beta$  (the fraction of BATs) and  $R \equiv \frac{\lambda_I}{\lambda_J}$  (the arrival intensity of non-HFTs relative to the value jump, a measure of adverse selection risk). When  $R \geq 5$ , adverse selection risk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





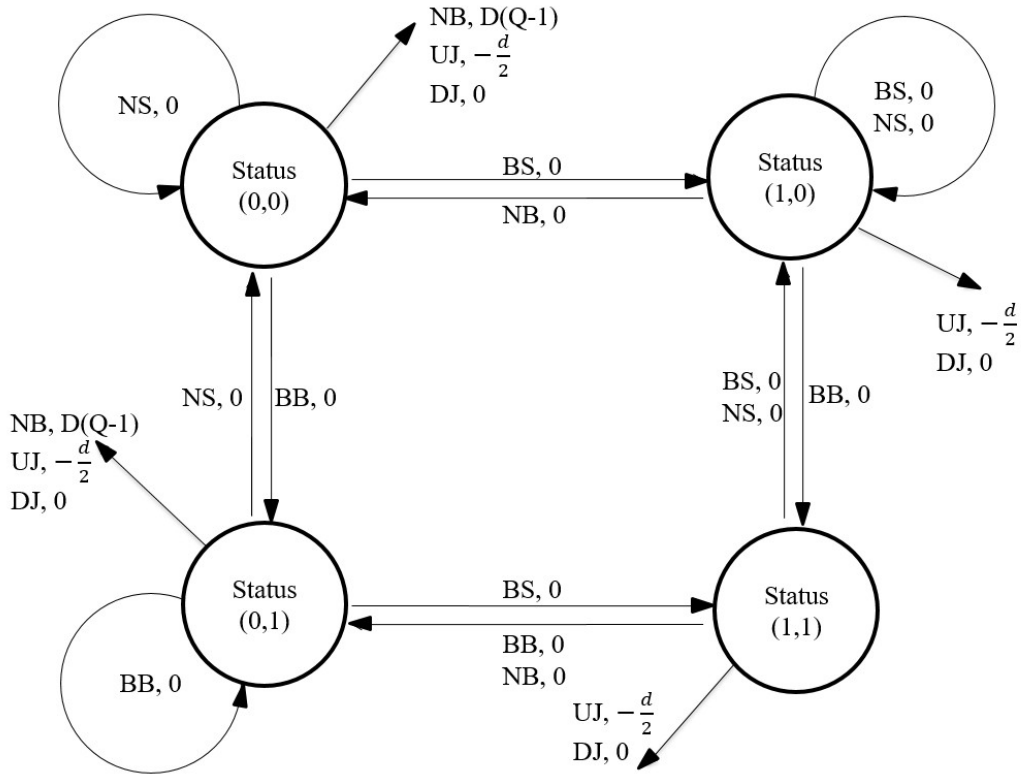
### Figure 4: The Undercutting and the Flash Trading Equilibrium

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



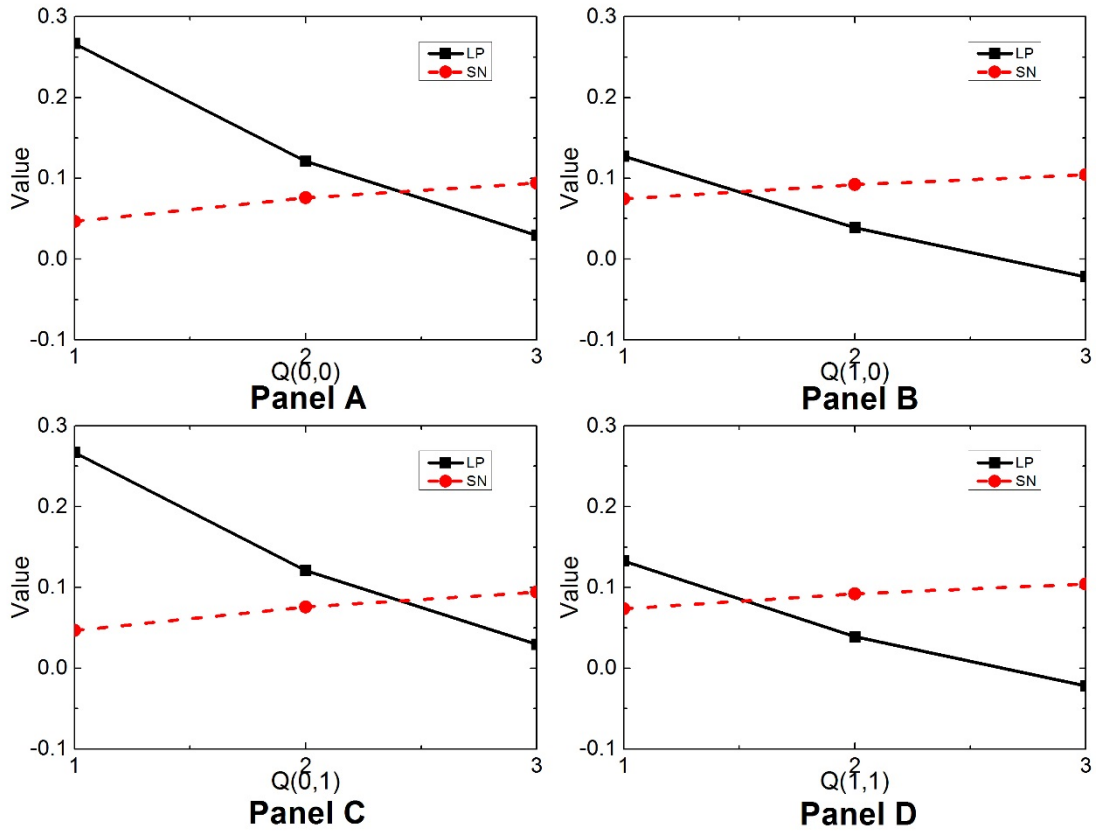
**Figure 5: States and Profits for HFT Liquidity Suppliers with the  $Q^{th}$  Position on the Ask Side**

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



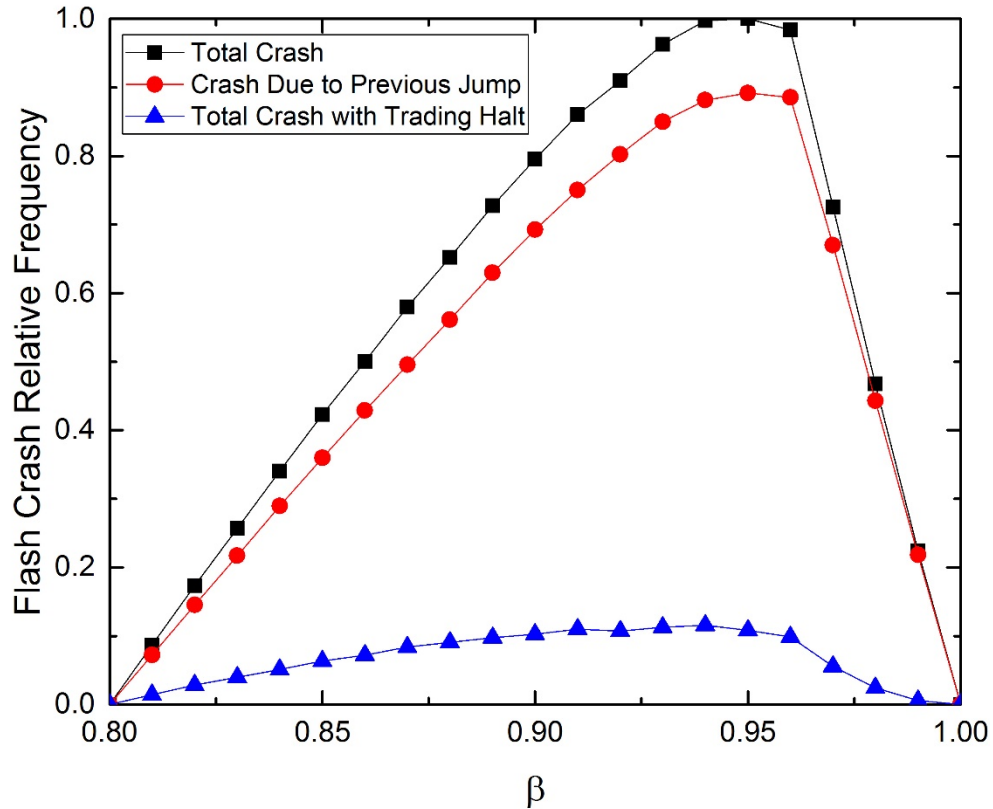
**Figure 6: Value of Liquidity Supply and Stale-Queue Sniping and Queue Length**

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



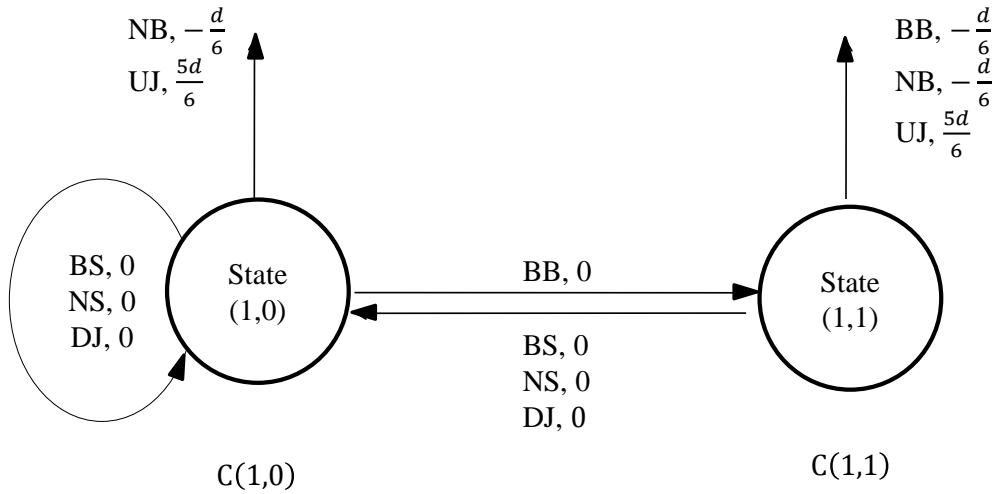
**Figure 7. Flash Crash Intensity**

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



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

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



# Designing Clearinghouse Default Funds\*

Agostino Capponi  
Columbia University

Jessie Jiayu Wang  
Arizona State University

Hongzhong Zhang  
Columbia University

February 21, 2018

## Abstract

A prominent post financial crisis reform to reduce counterparty risk in over-the-counter markets is the adoption of clearinghouses. Current standards require clearing members to contribute to a loss-mutualizing default fund so as to cover the liquidation costs imposed by the default of two members, the “Cover II” rule. We show that such an arrangement is intrinsically vulnerable: although the default funds allow members to share risk ex-post, an inherent externality induces members to take excessive risk ex-ante. We design a default fund level that trades off ex-post risk-sharing with ex-ante risk-shifting, thus providing regulators an optimal cover rule for default fund collection.

**Keywords:** Central counterparties (CCPs), default funds, loss mutualization, externality, risk-taking

**JEL:** G20, G23, G28, D61, D82

---

\*Contact information, Agostino: [ac3827@columbia.edu](mailto:ac3827@columbia.edu); Wang: [Jessie.Jiayu.Wang@asu.edu](mailto:Jessie.Jiayu.Wang@asu.edu); Zhang: [h2244@columbia.edu](mailto:h2244@columbia.edu)

# 1 Introduction

Heightened counterparty risk during the great recession has led policymakers to mandate central counterparties (CCPs) in the over-the-counter (OTC) derivatives markets. Through a process called novation, after a trade is established between two parties, the contractual obligations are replaced by equivalent positions between the two original parties and the clearinghouse. In the event of one party’s default, the original counterparty is insulated from losses as his contractual position is now with the clearinghouse. In this respect, the clearinghouse is referred to as a central counterparty (CCP); see also [Pirrong \(2011\)](#) for a comprehensive review of clearinghouse functioning mechanisms. Both the Dodd-Frank Wall Street Reform Act in the United States and the European Market Infrastructure Regulation in Europe mandate the central clearing of all standardized OTC derivatives contracts. The class of products being centrally cleared is rising steadily—around 80% of all interest rate and credit derivatives in the United States are now centrally cleared ([Financial Stability Board Report, 2015](#)).

There is, however, still considerable debate over the optimal design of clearinghouse arrangements (see, e.g., [Dudley, 2014](#), and [Economist, 2014](#)). One aspect of the arrangements is that members are required to contribute to a loss-mutualizing default fund. Currently, the required contribution by each member is such that the default fund is sufficient to cover the liquidation costs of two defaulting members, the “Cover II” rule.<sup>1</sup> The default loss of an institution that exceeds its initial margins and its default fund contribution are absorbed by the CCP equity capital and the default fund contributions of the surviving members. Typically, the CCP’s equity capital is the first to absorb losses. Residual losses are then allocated to surviving members on a pro-rata basis.<sup>2</sup>

---

<sup>1</sup>CPSS-IOSCO regulatory guidelines require CCPs to maintain a default fund sufficient to cover the liquidation costs caused by the default of two members, in extreme yet plausible market scenarios. The recent European Market Infrastructure Regulation requires each CCP to cover the default of the clearing member to which it has the largest exposure, or of the second and third largest clearing members if the sum of their exposures is larger.

<sup>2</sup>There is not yet a universally agreed upon loss allocation rule. Ice Clear Credit, the leading US CDS clearinghouse, distributes losses to non-defaulting clearing members on a pro-rata basis for cleared futures and options contracts (see [ICE, 2016](#)). Similarly, for cleared credit default swaps, it allocates losses among members on a pro-rata basis corresponding to the uncollateralized stress losses of each individual member.

In this paper, we study the optimal design of the default fund contributions. First, we show that the loss-mutualization arrangement by the CCPs is intrinsically vulnerable: While the default funds allow members to share risk ex post, an inherent externality induces members to take excessive risk ex-ante. Second, we show that the excessive risk-taking behavior can be mitigated by regulating the amount in the default fund. Thirdly, we design a default fund level that trades off ex post risk-sharing with ex-ante risk-shifting, thus providing regulators an optimal cover rule for default fund collection. In particular, we show that as the number of clearing members grows large, the optimal default fund should be designed to cover the default costs of a fixed fraction of the members rather than a fixed number of clearing members, as is done, for instance, in the currently implemented “Cover II” rule.

The economic forces under the loss-mutualization arrangement work as follows. Under the pro-rata rule, CCPs redistribute counterparty risk through mutualization among all members. This achieves *ex post* risk-sharing. However, this risk-sharing benefit comes with a flip side. Sharing the common pool of default funds creates a dependency among members. When members can choose to take excessive risk, a typical negative externality arises. Notably, the size of the default fund contribution directly determines the extent of the externality: While a lower default fund reduces the opportunity costs of the members, it leads to larger negative externalities among members, reducing total welfare through excessive risk-taking. As such, a regulator faces a tradeoff in collecting default fund contributions between (1) reducing the counterparty risk of clearing members and (2) generating excessive risk taking. We study this tradeoff and identify the right balance in designing the clearinghouse default funds. To the best of our knowledge, our study is the first to address this mechanism.

We develop a game-theoretical model to study the optimal design of default funds. In the model, a regulator decides what the size of the default fund contribution should be in order to maximize social welfare, defined as the aggregate value of clearing members and CCP. Given the required default fund amount, members of the clearinghouse decide on the riskiness of the undertaken projects. More specifically, each clearing member maximizes its expected total utility, taking into account the costs the member incurs to absorb losses generated by other



members' defaults. Thus the strategic interaction between the regulator and its members is modeled through a Stackelberg game with a coordination problem, in which the regulator is the leader and the members of the CCP are the followers.

Using the model, we study the positive and normative implications of default fund requirements. Our positive analysis focuses on the social welfare implications of the “Cover II” rule.<sup>3</sup> We show that a collective default fund under the “Cover II” rule might lead to moral hazard problems by reducing the incentives of a clearing member to avoid default. A member may decide to engage in excessively risky activities because the cost of its own default would be jointly borne by the other participants through their default fund contributions. This generates an inefficiency, especially in the setting of central clearing where mitigating systemic risk is critical. From a normative perspective, our results shed light on the optimal cover number for the default fund rule—i.e., the one that is most socially desirable. We illustrate that, under some circumstances, the inefficiency could be mitigated if the regulator were allowed to decide the size of the contribution by the clearing members to the default fund. By regulating the design of default funds, the regulator can give members an incentive to take less risk, hence improving social welfare.

A major novelty of our study is to analytically derive an optimal default fund level. Our model predicts that, as the number of clearing members increases, the optimal default fund level converges to a fixed fraction of the default cost, provided that the opportunity cost of default fund is not too high. In this case, it is socially optimal to mandate a default fund sufficiently high to cover a proportion of the members in the network. The Pareto dominating equilibrium associated with this default fund level prescribes that all members behave safely; i.e., they run away from the externality imposed by the risk-taking activities of the others. To the extent that the opportunity cost of default fund is associated with the cost of funding the collateral position and thus with the prevailing interest rates, our model predicts that default fund levels should be higher than those prescribed by the Cover II rule in the current

---

<sup>3</sup>The CPSS-IOSCO's Principles for Financial Market Infrastructures (PFMI) published in 2012 state that CCPs should maintain financial resources to cover the default of two participants that would potentially cause the largest aggregate credit exposure for the CCP, in extreme but plausible market conditions.

low interest rate environment. Our results are in line with [ISDA \(2013\)](#), which shows that default fund levels are quite conservative for their sample, and that on average less than 20% of the default funds are used to cover defaults that occur under stressed scenarios.

Our results support the Cover II rule when the costs of funding collateral are high. In this case, the socially optimal choice is to require members to contribute a lower amount to the default fund. Such an action induces members to engage in risky activities, resulting in a higher expected number of defaults than what could be covered by the members' default fund levels. Hence, our analysis suggests that the Cover II rule should be viewed in a different perspective: it is optimal when funding illiquidity (associated with marginal opportunity costs of default fund) is so high that it becomes socially preferable to induce a higher number of defaults in the system, than to subject members to the very high costs of raising collateral.

The paper proceeds as follows. [Section 2](#) reviews the literature. [Section 3](#) introduces the baseline model with binary risk and demonstrates members' incentives for excess risk-taking. [Section 4](#) analyzes the game between the regulator and the clearing members; we demonstrate that although members' risk-taking is unobservable, it can be supervised when the regulator strategically chooses the default fund contribution. [Section 5](#) generalizes the environment to the case of continuous risk choice and compares the social benefit and cost of increasing the size of the default fund. [Section 6](#) concludes. Proofs of technical results are in the Appendix.

## 2 Literature Review

Our main contribution to the literature on clearinghouses is to develop a tractable model that delivers explicit “Cover type” rules, accounting for the main economic forces at play—i.e., default costs, opportunity costs of posting collateral, and the risk-return trade-offs of the investments. These rules can be readily employed by clearinghouse supervising authorities to perform macro-stress testing, and compared with the currently employed Cover II rule. To the best of our knowledge, our study is the first to theoretically investigate the design of collateral resources further down the waterfall, namely the default levels.

Following the Dodd-Frank mandatory clearing for standardized OTC derivatives, several studies have analyzed the extent to which central clearing reduces counterparty credit risk. The seminal paper by [Duffie and Zhu \(2011\)](#) shows that netting benefits exist only if a clearinghouse nets across different asset classes, while counterparty credit risk may arise if the clearing process is fragmented across multiple clearinghouses. These predictions are empirically confirmed by [Duffie, Scheicher, and Vuillemeys \(2015\)](#) using bilateral exposures data from the credit default swap market.<sup>4</sup> [Biais, Heider, and Hoerova \(2016\)](#) study how hedging with derivatives can introduce moral hazard originating from the fact that the trading counterparties neglect risk management. They show that margin calls can be optimally designed to mitigate the moral hazard problem thereby enhancing risk-sharing. While these studies consider initial margins—i.e. collateral resources designed to absorb the losses of an individual member—our paper focuses on the optimal degree of risk-sharing for the determination of default fund requirements.

Other studies have analyzed the risk-management implications of transparency introduced by a central clearinghouse. [Acharya and Bisin \(2014\)](#) illustrate one type of counterparty externality arising from the lack of portfolio transparency in OTC markets and show that it can be corrected if trades are centrally cleared. [Zawadowski \(2013\)](#) shows that the establishment of a clearinghouse can improve efficiency as it effectively forces banks to contribute ex-ante to bail-out defaulting counterparties, thus reducing the hedging losses of a bank. In his model, defaults are caused by informational effects that induce runs on banks if they have experienced hedging losses. [Antinolfi, Carapella, and Carli \(2016\)](#) show that central clearing can be socially inefficient because loss-mutualization may weaken the incentives to acquire and reveal information about counterparty risk, whereas bilateral trading typically encourages assessment of counterparty risk. Different from these works, we consider a symmetric information model and highlight a different form of inefficiency due to loss mutualization. Like [Zawadowski \(2013\)](#), we show that default fund requirement is a tool that can be used to correct for inefficiency; unlike him, however, we find that the domino effects of defaults

---

<sup>4</sup>[Loon and Zhong \(2014\)](#) and [Bernstein, Hughson, and Weidenmier \(2014\)](#) both find evidence consistent with central clearing reducing counterparty risk, using CDS spreads data and historical data, respectively.

plays the prominent role, and the inefficiency is corrected by balancing the benefits of ex-post risk-sharing with the costs of ex-ante risk-shifting.

Our paper contributes to a nascent literature focusing on default fund requirements. [Menkveld \(2017\)](#) analyzes systemic liquidation within a crowded trades setting and sets the default fund as the minimum level of funds needed to cover default losses in extreme yet plausible conditions. [Ghamami and Glasserman \(2017\)](#) provide a calibration framework and show that lower default fund requirements reduce the cost of clearing but make CCPs less resilient. While these studies follow risk-measure based rules to determine default funds, the latter are endogenously determined in our model. Other studies focus on the default fund design explicitly. [Amini, Filipović, and Minca \(2015\)](#) analyze systemic risk under central clearing in an Eisenberg-Noe type clearing network, and propose an alternative structure for a default fund to reduce liquidation costs. In contrast, we solve for the optimal default fund in a symmetric equilibrium setting, taking members' incentives into account. [Capponi, Cheng, and Sethuraman \(2017\)](#) analyze the incentives behind the determination of default fund and clearinghouse equity in the default waterfall structure. In their model, the mass of safe and risky members is exogenously specified, and their objective is to study the optimal balance of equity and default fund requirements from the clearinghouse point of view. In our model, members instead endogenously choose the risk profile, and we focus on the socially optimal cover rule.

In emphasizing an intrinsic vulnerability from loss mutualization, our paper joins the literature that highlights various aspects of inefficiency and “unintended consequences” associated with central clearing. The model proposed by [Koepl and Monnet \(2010\)](#) predicts that, though CCP clearing can induce traders to take the socially optimal level of counterparty risk, it affects liquidity across the OTC markets in a way that not all traders universally benefit. Building on this model, [Koepl \(2012\)](#) shows that higher collateral requirements lower default risk, but also reduce market liquidity, which in turn amplifies collateral costs. [Pirrong \(2014\)](#) argues that central clearing reforms may redistribute risk rather than reduce it, and that expanding clearing makes the financial system more connected and transforms credit

risk into liquidity risk. [Arnold \(2017\)](#) shows that, under current market regulations, central clearing may have unintended consequences such as a higher number of issued loans, but a lower credit quality of these loans. A closely related paper by [Biais, Heider, and Hoerova \(2012\)](#) shows that the main advantage of centralized clearing is loss mutualization, which fully insures members against idiosyncratic risk, but not against aggregate risk. They argue that CPP should be designed to incentivize members to search for solid counterparties under aggregate risk. In contrast, we show that the inefficiency arises even when the source of risk is idiosyncratic, and we demonstrate how this inefficiency can be mitigated by the optimal design of default fund contributions.

### 3 Baseline Model

In this section we introduce our baseline model in which clearing members have a binary choice of risk. By comparing the (first-best) risk level that maximizes social welfare with the one that maximizes individual members' profit, we demonstrate that members have an incentive for excessive risk-taking due to an inherent externality associated with loss mutualization.

#### 3.1 The Environment

We consider a two-period model,  $t = 0, 1$ . The economy consists of  $N$  homogeneous clearing members and the regulator. The clearing members are risk-neutral. At  $t = 0$ , the regulator chooses the required level of the default fund and the members make investment decisions. At  $t = 1$ , payoffs are realized.

*Clearing Members.* Clearing members may differ in their choice of investments such as whether to invest in a high-risk or low-risk project. Project types can model engagement in risky investments, choice of weak trading counterparties before novation, lower effort in risk management, or reduced hedging of counterparty exposure. Such choices are unobservable and capture the risk-return tradeoff faced by member institutions.

The payoffs from investing in a high-risk project and a low-risk project are denoted, respectively, by  $R_h$  and  $R_l$ . The expected payoff from the investment depends on the risk level. Let  $\mu = \mathbb{E}[R]$  and assume  $\mu_h > \mu_l$ . We denote by  $q_h$  and  $q_l$  the default probability of a member who chooses, respectively, the high-risk and low-risk project. Defaults are costly, and we use  $c$  to denote the constant default cost. We analyze the economic role of default funds, and do not model initial margins which usually serve as the first line of defense against default losses. Hence, we can view the default costs as describing the losses that exceed initial margin requirements. The following assumption formalizes the risk-return tradeoff.

**Assumption 1** *A member who chooses the high-risk project has a higher expected return but a higher default probability than a member who chooses the low-risk project – i.e.*

$$0 < \mu_l < \mu_h, \quad 0 < q_l < q_h. \quad (1)$$

**Default Fund.** Clearing members derive risk-sharing benefits from entering into a *loss mutualization* arrangement with other member’s through the CCP. A *positive default fund*,  $F > 0$ , sets an upper bound on the loss given the default of a member. We consider a default waterfall structure in which the clearinghouse capital has seniority over the default funds of surviving members in absorbing losses.<sup>c0</sup> After the funds of a defaulted clearing member are exhausted, the loss mutualization mechanism ensures that the remaining losses are allocated proportionally to the available default funds of the surviving members. This is the risk-sharing mechanism.<sup>c0</sup> On the other hand, the default fund increases a surviving members chances of incurring a loss when others default, creating externality among members.

Specifically, the CCP charges a default fund amount  $F$  to each member upfront. Because the default fund is segregated, each member will incur an opportunity cost of  $r_f F$ , where

---

<sup>c0</sup>A well known example is the Korean CCP KRX. The default of a clearing member in December 2013 generated losses that exceeded the defaulter’s collateral. According to the KRX’s rules, the remaining losses were allocated first to the default fund contributions of surviving participants.

<sup>c0</sup>In the contractual agreement between the clearinghouse and the clearing members, the initial margins of a defaulted member are first used to cover losses arising at liquidation. In our analysis, we do not model the initial margins in the clearinghouse default waterfall to focus entirely on the risk-shifting incentives triggered by the default fund resources. This comes without loss of generality, however, as we can view the default costs as the losses that exceed initial margin requirements.

$r_f \in (0, 1)$  denotes the risk-free rate.

A member is willing to participate in such a loss mutualization fund only if the sum of his default fund contribution and the resulting opportunity cost does not exceed his default cost. To summarize, it is rational for individual members to participate in loss mutualization only if

$$0 < F + r_f F < c. \quad (2)$$

**Investment Choice of a Clearing Member.** For a given default fund requirement  $F$ , and for given investment choice  $a^{-i} = (a_1, a_2, \dots, a_{i-1}, a_{i+1}, \dots, a_N)$  made by all clearing members except  $i$ , member  $i$  decides on the riskiness of his investment to maximize the following objective function:

$$V_i(a^{-i}) = \max_{a_i \in \{h, l\}} \mathbb{E} \left[ R_{a_i} - \mathbf{1}_{i \text{ defaults}} F - \mathbf{1}_{i \text{ survives}} \min \left( F, \frac{\sum_{j \neq i} \mathbf{1}_{j \text{ defaults}}}{1 + \sum_{j \neq i} \mathbf{1}_{j \text{ survives}}} (c - F) \right) - r_f F \right]. \quad (3)$$

The objective function (3) can be understood by analyzing two possible scenarios:

1. If  $i$  defaults,  $i$  does not contribute to the loss mutualization, but  $i$ 's default fund  $F$  will be used to cover its loss.  $i$ 's total cash flow will thus include the value of its investment minus the default fund and the associated opportunity cost:

$$R_{a_i} - F - r_f F \quad (4)$$

2. If  $i$  survives, the default fund of  $i$  will be used to absorb the losses generated by defaulted members, if any; hence, the cash flow is given by

$$R_{a_i} - F \min \left( 1, \frac{\sum_{j \neq i} \mathbf{1}_{j \text{ defaults}} (c - F)}{F + \sum_{j \neq i} \mathbf{1}_{j \text{ survives}} F} \right) - r_f F. \quad (5)$$

The cost to a surviving member is the highest if all other clearing members default because that member will have to bear all costs totaling  $\min[F, \sum_{j \neq i} (c - F)]$ . If no clearing member defaults, the cost incurred by  $i$  will be zero. More generally, if  $k$  members default,  $k < N$ , all  $N$  members will collectively contribute an amount equal to  $\min(NF, kc)$  to absorb the default

losses; if  $k \geq \frac{F}{c}N$ , then the default funds of all members, totaling  $NF$ , will be used to cover the default costs.

**Assumption 2** *The aggregate default fund contribution satisfies a Cover II constraint:*

$$2c \leq NF. \quad (6)$$

*The Cover II rule provides that the default fund contribution collected from all members,  $NF$ , is always sufficient to cover the default costs of two members.*

Assumption 2 and the participation constraint in Eq. (2) together set an upper and lower bound for  $F$ :  $\frac{2c}{N} < F < \frac{c}{1+r_f}$ . The *Cover II* constraint is the currently imposed requirement that CCPs should maintain financial resources sufficient to cover a wide range of potential stress scenarios, including the default of two members.

If more than two members default, the total default fund contribution,  $NF$ , may not be enough to cover the default costs. In this case, the rest of the default costs are covered by CCP's capital.

**Assumption 3** *The CCP contributes with his own capital and incurs an equity loss of  $(\mathcal{N}_d \cdot c - N \cdot F)^+$ , where  $\mathcal{N}_d$  is the number of defaults.*

Assumption 3 guarantees that there are always enough resources in the system to pay for the default costs of members in every possible state.

### 3.2 First-best Benchmark: optimal investment

For a given default fund contribution  $F$ , the choice of risky investments  $\{a_i^*\}_{i=1,\dots,N}$  that maximizes the aggregate values of all agents in the economy (clearing members and the CCP) is given by

$$\{a^*(F)\} = \arg \max \mathbb{E} \left[ \sum_i (R_{a_i} - c \mathbf{1}_{i \text{ defaults}} - r_f F) \right]. \quad (7)$$



To solve for the socially optimal risk profile, it is sufficient to consider the risk and return tradeoff of a representative clearing member by comparing the return differential  $\mu_h - \mu_l$  with the expected default cost differential  $c(q_h - q_l)$ .

**Proposition 1** *The socially optimal risk profile is given by*

$$a_i^*(F) = \begin{cases} l, & \frac{\mu_h - \mu_l}{q_h - q_l} \leq c \\ h, & \frac{\mu_h - \mu_l}{q_h - q_l} > c \end{cases} \quad (8)$$

The focus of this paper is on the members' *ex ante* incentive to take excessive risk. Therefore, for the rest of Section 3, we consider the case that the socially optimal risk profile is the low-risk project and impose the following:

**Assumption 4** *The parameters  $\{\mu_h, \mu_l, q_h, q_l, c\}$  satisfy the condition*

$$\frac{\mu_h - \mu_l}{q_h - q_l} \leq c. \quad (9)$$

### 3.3 Equilibrium Investment

In this section, we solve for the best response of members to a given choice of default fund  $F$ . Our objective is to compare the investment choice of the profit-maximizing members in equilibrium with the socially optimal choice. We call *risk-shifting* the situation in which the first-best outcome prescribes that all members choose the low-risk project but the equilibrium response of members is to instead undertake the high-risk project. This is socially inefficient in the sense that by taking excessive risk members obtain a lower total value.

**Definition 1** *Under Assumption 2, and if  $F$  satisfies (2), the risk profile  $(a_1, a_2, \dots, a_N) \in \{h, l\}^N$  is a Nash equilibrium if for all  $i$ , it holds that*

$$V_i(a^{-i}) = \mathbb{E} \left[ R_i^{a_i} - \mathbf{1}_{i \text{ defaults}} F - \mathbf{1}_{i \text{ survives}} \min \left( F, \frac{\sum_{j \neq i} \mathbf{1}_{j \text{ defaults}}}{1 + \sum_{j \neq i} \mathbf{1}_{j \text{ survives}}} (c - F) \right) - r_f F \right].$$

*A Nash equilibrium  $(a_1, a_2, \dots, a_N)$  is Pareto dominating, if, for any other Nash equilib-*

rium  $(b_1, b_2, \dots, b_N)$ , it holds that

$$V_i(a^{-i}) \geq V_i(b^{-i}), \quad i = 1, 2, \dots, N,$$

and the above inequality holds strictly for at least one  $i$ .

Suppose member  $i$  survives, and  $g$  of the remaining  $N - 1$  members choose the low-risk project. Then, for any given admissible choice of default fund,  $F \in [\frac{2c}{N}, \frac{c}{1+r_f}]$ , the expected contribution of member  $i$  to other members' defaults is given by

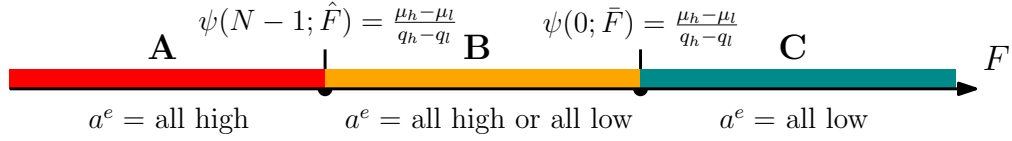
$$\mathbb{E} \left[ \min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right) \middle| \text{member } i \text{ survives} \right] := F - \psi(g; F), \quad (10)$$

where  $\mathcal{N}_s$  is the number of surviving members. The left-hand side of (10) is the expected contribution of member  $i$  to other members' defaults, given that  $i$  survives: If  $N - \mathcal{N}_s$  is the number of defaulted members, the total cost of their defaults is  $c(N - \mathcal{N}_s)$ . Each defaulted member will first absorb the losses using his own default fund, and the remaining cost will be shared equally by the surviving members. Hence, each member  $i$  will be charged, on average, a cost equal to  $(c - F) \frac{N - \mathcal{N}_s}{\mathcal{N}_s}$  capped at maximum amount  $F$  that  $i$  can contribute. The quantity  $F - \psi(g; F)$  on the right-hand side of Eq. (10) is the cost due to loss mutualization imposed by the members of the network on  $i$ . A lower value of  $\psi(g; F)$  means that the negative externalities on  $i$  caused by the risky project choices of the other members are high. The next proposition summarizes our main results on the investment choice of clearing members.

**Proposition 2** *For a given default fund requirement  $F$  satisfying Assumption 2 and condition (2), if Assumption 4 holds, then all possible equilibrium risk profiles are given by*

$$a^e = \begin{cases} h, \forall i & F < \hat{F} \\ h, \forall i, \text{ or } l, \forall i & \hat{F} \leq F \leq \bar{F} \\ l, \forall i & \bar{F} < F \end{cases} \quad (11)$$

Moreover,  $\psi(\cdot; \cdot)$  is a strictly increasing function in both the first and second argument. The



**Figure 1. Equilibrium Investment Choices.** This figure illustrates how the equilibrium investment choices change as we vary the size  $F$  of the default fund.  $a^e$  is the equilibrium risk profile from individual members’ strategic behavior. In region A, the unique equilibrium differs from the first-best benchmark and features risk-shifting; in region B, we have multiple equilibria; in region C, the unique equilibrium coincides with the first-best.

parameters  $\hat{F}$  and  $\bar{F}$  are implicitly defined as

$$\psi(N-1; \hat{F}) = \frac{\mu_h - \mu_l}{q_h - q_l}, \quad \psi(0; \bar{F}) = \frac{\mu_h - \mu_l}{q_h - q_l}. \quad (12)$$

If  $\hat{F} \leq F \leq \bar{F}$ , the “all low”-risk equilibrium Pareto dominates the “all high”-risk equilibrium.

The result in Proposition 2 states that the members’ incentives in shifting risk decrease as the required default level of  $F$  rises. If the profitability of risk-shifting measured by excess return per unit of default probability,  $\frac{\mu_h - \mu_l}{q_h - q_l}$ , is exceeded by the negative externalities generated from the risk-shifting behavior,  $\psi(0, F)$ , or equivalently  $F > \bar{F}$ , then each member decides to run away from the externalities and chooses the low-risk project. Figure 1 illustrates the equilibrium risk profile as a function of  $F$ .

### 3.4 Inefficiency in Investment Choice: risk-shifting

In this section, we show that there exist cases in which the loss mutualization mechanism induces members to take excessive risk *ex ante* due to an inherent externality among members.

**Corollary 2** *If  $F < \hat{F}$ , the first-best benchmark is not an equilibrium, and in fact, “all high”-risk is the unique inefficient equilibrium. More generally, if  $F \leq \bar{F}$ , “all high”-risk is an inefficient equilibrium.*

Under  $F < \hat{F}$ , if all other members choose low-risk, member  $i$  strategically deviates to choose high-risk. Under loss mutualization, the expected liquidation cost  $\psi(N-1; F)$  is lower

than  $c$ . This wedge shifts a member’s incentive from choosing low-risk to high-risk, thereby creating a *risk-shifting* problem.

In the presence of risk-shifting incentives, it is still possible to achieve the socially optimal risk-taking in equilibrium. This happens when the default fund contribution  $F$  is close enough to the default cost such that the externality is restrained.

**Corollary 3** *If  $\bar{F} < F$ , the “all low”-risk equilibrium is the unique equilibrium and is first-best. Moreover, the “all low”-risk equilibrium is the unique Pareto dominant equilibrium if  $\hat{F} < F$ .*

## 4 Equilibrium between Clearing Members and the Regulator

We analyze how to design a default fund that mitigates the inefficiency arising from excessive risk-taking. Corollary 3 indicates that a default fund can potentially correct members’ risk-shifting incentives. In this section, we develop a game theoretic analysis to show that the regulator can correct the inherent externality by optimally choosing a default fund, balancing the *ex post* risk-sharing benefit and *ex ante* risk-shifting cost.

**Definition 4** *A Nash equilibrium between clearing members and the regulator is a set of members’ risk profiles  $a^e := (a_1^e, \dots, a_N^e)$  and a default fund contribution  $F^e$  set by the regulator, such that:*

1. *Taking the default fund  $F^e$  and other members’ risk profile  $a_{-i}^e$  as given,  $a_i^e$  solves the optimization problem of clearing member  $i$  given in (3).*
2. *Taking as given the risk profile of clearing members  $a^e$ , the regulator chooses a feasible default fund level  $F^e$ , satisfying assumptions (2)–(3) and condition (2) to maximize the aggregate value of the members and the equity of CCP:*

$$F^e = \arg \max_F \sum_i V_i - \mathbb{E} [(\mathcal{N}_d \cdot c - N \cdot F)^+]. \quad (13)$$

In the above equation,  $V_i$  is given by Eq. (3), and  $\mathcal{N}_d = \sum_i \mathbf{1}_i$  defaults is the number of defaulted members.

#### 4.1 Default Fund: a tool to mitigate risk-shifting

We solve for the optimal default fund level that satisfies the criterion in Equation (13), taking the functional dependence of  $a^e$  on  $F$  into account. Given the assumption of independent defaults, the objective function of the regulator in (13) may be rewritten as follows:

$$W(F) = \mathbb{E} \left[ \sum_i V_i - (\mathcal{N}_d \cdot c - N \cdot F)^+ \right] = \mathbb{E} \left[ \sum_i R_{a_i} - \mathcal{N}_d \cdot c - N r_f F \right]. \quad (14)$$

Thanks to Proposition 2, it suffices to consider either “all low” or “all high”-risk profiles. We base our analysis on the equilibrium refinement concept of Pareto dominance: From Proposition 2 and Corollary 3, members all choose low-risk when  $\hat{F} \leq F$  because it is either the unique equilibrium (region C in Figure 1) or the Pareto dominating equilibrium (in region B). Hence the equilibrium risk profile *switches* from high to low at the boundary between Regions A and B. To obtain the threshold value of  $\hat{F}$  at this boundary, define the linear function of  $F$ :  $I(F) = \psi(N - 1; F) - \frac{\mu_h - \mu_l}{q_h - q_l}$ . By Lemma 8 in Appendix A,  $I(F)$  is a strictly increasing function. Moreover,  $\psi(N - 1; c) = c$ . Thus if the following condition holds,

$$\psi \left( N - 1; \frac{2c}{N} \right) < \frac{\mu_h - \mu_l}{q_h - q_l} c^0, \quad (15)$$

there is a unique threshold value  $\hat{F}$  such that  $I(\hat{F}) = 0$ , i.e.  $\psi(N - 1; \hat{F}) = \frac{\mu_h - \mu_l}{q_h - q_l}$ .

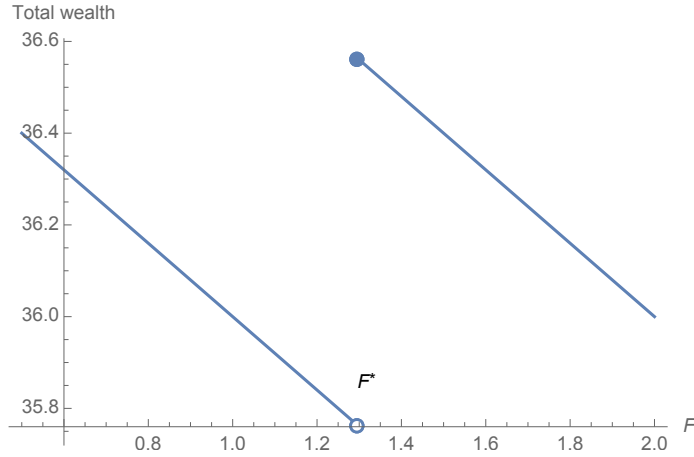
The function  $W(F)$  may be defined piecewisely, with just one discontinuity at  $\hat{F}$ , as follows:

$$W(F) = N \left( W^l(F) \mathbf{1}_{F \geq \hat{F}} + W^h(F) \mathbf{1}_{F < \hat{F}} \right), \quad W^a(F) = \mu_a - c q_a - r_f F, \quad a = l, h. \quad (16)$$

Because the ‘low-risk’ equilibrium is socially optimal,  $W(F)$  exhibits a positive jump as we increase  $F$ : the equilibrium switches from “all high” to “all low”-risk at  $\hat{F}$ . Denote the size

---

<sup>c0</sup>When the inequality fails to hold, the aforementioned switch from “all low”-risk equilibrium to “all high”-risk equilibrium does not occur. The reason is that the lower bound set by the Cover II rule may already be large enough that choosing high-risk is not attractive for an individual member. In the sequel, we only focus on the (interesting) case that the inequality (15) holds.



**Figure 2. Regulator’s Objective Function  $W(F)$ .** This figure plots the total wealth  $W(F)$  as a function of  $F$ , where  $F$  ranges from the lower bound  $\frac{2c}{N}$  that satisfies the Cover II requirement to the upper bound that equals the default cost  $c$ . The graph shows that a default fund given by the Cover II requirement may not be socially optimal. Rather, a higher value of  $F$  that corrects the risk-shifting yields the highest total value. Model parameters:  $\mu_h = 5, \mu_l = 4.8, N = 8, c = 2, q_h = 0.2, q_l = 0.05$ , and  $r_f = 0.1$ . The optimal default fund  $F^e = \hat{F} = 1.30$ , and  $W(F^e) - W(\frac{2}{N}c) = 0.16$ .

of the upward jump at  $\hat{F}$  by  $\Delta$ . Then

$$\Delta \equiv W^l(\hat{F}) - W^h(\hat{F}) = \mu_l - \mu_h - c(q_l - q_h) > 0. \quad (17)$$

Moreover,  $W^h(F)$  and  $W^l(F)$  are both strictly decreasing in  $F$ . Thus the maximum of  $W(F)$  over the set of feasible values for  $F$  is attained either at the lower bound  $\frac{2}{N}c$ , or at the switch point  $\hat{F}$ . We summarize this result in the following proposition (see also Figure 2 for an illustration).

**Proposition 3** *Suppose the inequality (15) holds. Among all feasible values of  $F$  – i.e., those satisfying Assumption 2 and condition (2) – the default fund level that maximizes the regulator’s objective function (13) is given by*

$$F^e = \begin{cases} \hat{F}, & \text{if } \Delta > r_f \left( \hat{F} - \frac{2}{N}c \right), \\ \frac{2}{N}c, & \text{else.} \end{cases} \quad (18)$$

In conclusion, the objective of the regulator is to maximize the total value of the agents

in the system, taking members' incentives for risk-shifting into consideration. The choice of default fund yields the following tradeoff. On the one hand, the benefit of having a higher value of  $F$  is to correct for risk-shifting: as  $F$  increases to  $\hat{F}$ , we move from region A to region B in Figure 1, where members' risk choices switch from high to low. Hence, by mandating a high enough default fund  $F$ , the regulator can give the member an incentive to choose risk with a total social value equal to  $N\Delta$ . On the other hand, increasing  $F$  raises the opportunity cost of each member from  $r_f \frac{2}{N}c$  to  $r_f \hat{F}$ . If the benefit exceeds the cost, it is optimal to set the default fund at  $\hat{F}$ , a higher value than required by the Cover II rule; otherwise, the regulator will find it optimal to use the Cover II rule.

## 4.2 Cover X: the optimal covering number

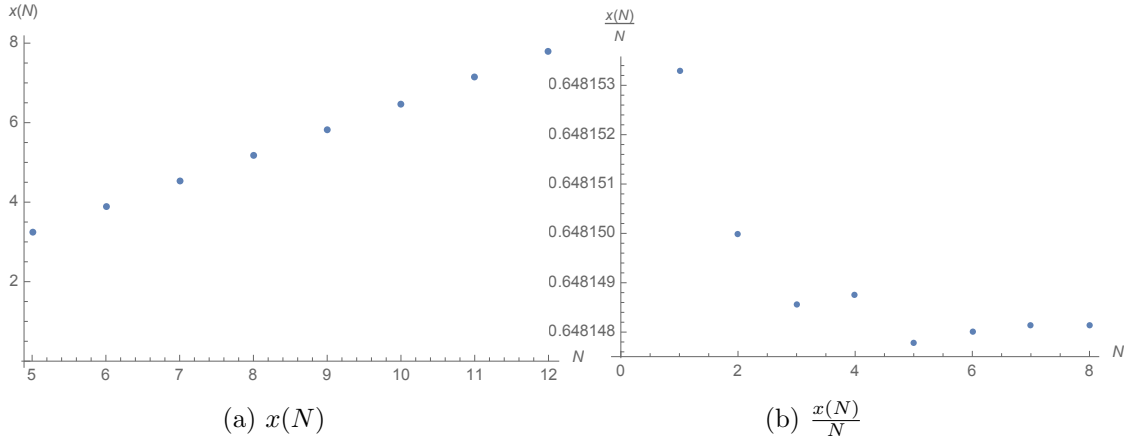
The previous sections have shown that the Cover II rule is not necessarily socially optimal, when we consider the role of mitigating risk-shifting. As the number of clearing members increases, how many members' defaults should the default funds be ready to cover? We design a Cover X rule so that the implied default fund achieves the efficiency of  $\hat{F}$  in Equation (18).

The *generalized Cover X rule* for any given number  $N$  of participating clearing members is

$$x(N) := \frac{N \cdot F^e(N)}{c}, \quad (19)$$

where  $F^e(N)$  is the default fund level that maximizes the regulator's objective given in Proposition 3, when  $N$  is the number of members in the CCP. When  $x(N) > 2$ , our model provides a rationale to charge a default fund more than the current regulatory requirement prescribed by the Cover II rule; see also Figure 3 for an illustration.

Interestingly, the exact Cover X rule depends on the number of members. If  $N = 3$ , then  $x(N)$  is approximately equal to two, so Cover II is able to prevent risk-shifting. However, if, for example,  $N = 8$ , the optimal requirement becomes Cover 5.18. Under the Cover II requirement, clearing members would engage in excessive risk-taking and deviate from the social optimum, resulting in a higher total cost of default. In contrast, the proposed Cover 5.18 requirement induces members to choose low-risk projects, thereby effectively mitigating



**Figure 3. Optimal Covering Number.** This figure shows the optimal covering number  $x(N)$  given in equation (19) as a function of the number of clearing members  $N$ . We use the same parameter settings for project return and default probability as in Figure 2. Left panel: the optimal covering numbers  $x(N)$ ; Right panel: the ratio  $\frac{x(N)}{N}$ .

risk-shifting.

While the optimal covering number clearly depends on the number of members  $N$ , the ratio  $x(N)/N$  shows little variation with respect to  $N$ , if  $N$  is sufficiently large. The next proposition characterizes the asymptotic behavior of the optimal coverage ratio, as the number of clearing members grows large. Specifically, it shows that the ratio between the optimal default fund level  $F^e(N)$  and the default cost  $c$ , or equivalently, the optimal proportion of covered members,  $x(N)/N$ , converges to a constant as the size of the CCP network tends to infinity. If the marginal opportunity cost  $r_f$  is sufficiently low, then this limit is a positive number in  $(0, 1)$ , meaning that the optimal covering number should be proportional to the size of network  $N$  (at least for large  $N$ ); otherwise, this limit is 0, implying that it is optimal to cover only a small portion of the CCP network.

Hence, as the number of clearing members grows large, the cover rule that the regulator should adopt is simple—rather than covering a fixed number of clearing members as prescribed, for instance, by the Cover II rule, the regulator should cover a fixed fraction of the members. Considering that the major U.S. derivative clearinghouses consist of more than 30 members (this is the case, for instance, for the major CDS derivative clearinghouse ICE Clear Credit, and interest rates swaps clearinghouse LCH), our result implies that the default fund



rule is robust with respect to entry and exit of the members in the clearing business.<sup>c0</sup>

**Proposition 4** *In the large CCP network limit – i.e., as  $N \rightarrow \infty$  – we have*

$$\frac{F^e(N)}{c} = \frac{x(N)}{N} \rightarrow \begin{cases} q_l + (1 - q_l) \frac{1}{c} \frac{\mu_h - \mu_l}{q_h - q_l}, & \text{if } \mu_l - \mu_h - c(q_l - q_h) > r_f(q_l c + (1 - q_l) \frac{\mu_h - \mu_l}{q_h - q_l}), \\ 0, & \text{else.} \end{cases}$$

It is immediate from Proposition 4 that if the differential ratio  $\frac{\mu_h - \mu_l}{q_h - q_l}$  is much smaller than the default cost  $c$  (i.e., the risky project is not sufficiently profitable to compensate for the costs incurred at default), then the members are more inclined to switch from risky to safe investments as the default fund level increases. Thus the regulator can use a lower default fund level to prevent the risk-shifting.

## 5 Continuous Choice of Risk-Taking

Having illustrated the simple logic of how a default fund can alleviate risk-taking for two possible choices of risk, in the rest of the paper we extend the analysis to the more general case of a continuous choice of taken risks. Such a setup allows us to track the marginal impact of setting a higher default fund contribution on the risk-taking behavior of a member.

### 5.1 The Environment

Member  $i$  has a continuous choice of risk-levels. The member invests a fraction  $a_i \in [0, 1]$  of its resources in the risky project and allocates the remaining fraction  $1 - a_i$  to a risk-free project. The risk-free project has a guaranteed payoff of  $(1 + r_f)$  times the invested amount and can be thought as an investment in risk-free assets (e.g., U.S. Treasury bonds). The risky project could include a portfolio of loans to firms in the corporate sector or a portfolio of mortgages, exposing the member to volatile returns. The payoff, denoted by  $\tilde{R}_i$ , is realized

---

<sup>c0</sup>For instance, in May 2014, the Royal Bank of Scotland announced the wind down of its clearing business due to increasing operational costs. This was followed by State Street, BNY Mellon, and more recently Nomura, each of whom shut down part or all of their clearing business.

at  $t = 1$  and assumed to be a random variable:

$$\tilde{R}_i = \begin{cases} R, & \text{with probability } 1 - a_i \\ r, & \text{with probability } a_i, \end{cases} \quad (20)$$

$R$  can be viewed as the notional value of the loan/mortgage. Let  $R > 1 + r_f > r$ : in the good state, the realized payoff is higher than that of the risk-free project, whereas in the bad state the payoff is lower than the return from the same investment in the risk-free project. Similar to the setup in [Holmstrom and Tirole \(2001\)](#) and [Acharya, Shin, and Yorulmazer \(2010\)](#), the risky technology has diminishing returns to scale with risk-taking; i.e., the probability of a good state decreases with  $a_i$ .<sup>c0</sup> In particular, the probability of observing the good state is  $1 - a_i$ .

Member  $i$  defaults if the value of the realized risky project is  $r$ :

$$\mathbf{1}_{i \text{ defaults}} \Leftrightarrow a_i \tilde{R}_i + (1 - a_i)(1 + r_f) < 1 + r_f \Leftrightarrow \tilde{R}_i = r. \quad (21)$$

The default probability of member  $i$  is equal to  $a_i$ : the fraction invested by  $i$  in the risky project. Default can always be avoided if member  $i$  invests entirely in the risk-free project. Defaults are costly, and we use  $c > 0$  to denote the cost of a default. Let  $R > 1 + r_f + c$ , indicating that members prefer to invest a non-zero fraction in the risky project. Notice that at  $a = 0$ , the marginal profit of risk-taking is  $R$ , whereas the marginal cost is  $1 + r_f$  (forgone return from the risk-free project) plus  $c$  (marginal cost of default). We assume that the realizations  $\tilde{R}_i$ ,  $i = 1, \dots, N$ , are independent across members, which implies that defaults are independent.

***Strategic Investment Choice of a Member.*** Consider a CCP, and assume a default fund level  $F$  satisfying the Cover II requirement: given investment strategies  $a_{-i}$  chosen by

---

<sup>c0</sup>The assumption of diminishing returns to scale is not essential for our results, but helps to obtain a neat expression for the default probability.

other members except  $i$ , the expected payoff of member  $i$  is given by

$$V_i(a_{-i}) = \sup_{a_i \in [0,1]} E \left[ a_i \tilde{R}_i + (1-a_i) - \mathbf{1}_{i \text{ defaults}} F - \mathbf{1}_{i \text{ survives}} F - r_f F \right]. \quad (22)$$

Each clearing member chooses his own investment strategically to maximize his expected payoff. If all other members choose the same investment strategy  $a_{-i} \in [0, 1]$ , we have

$$V_i(a_{-i}) = \sup_{a \in [0,1]} [-(R-r)a^2 + (R-1)a + 1 - F + (1-a)\phi(a_{-i}; F) - r_f F], \quad (23)$$

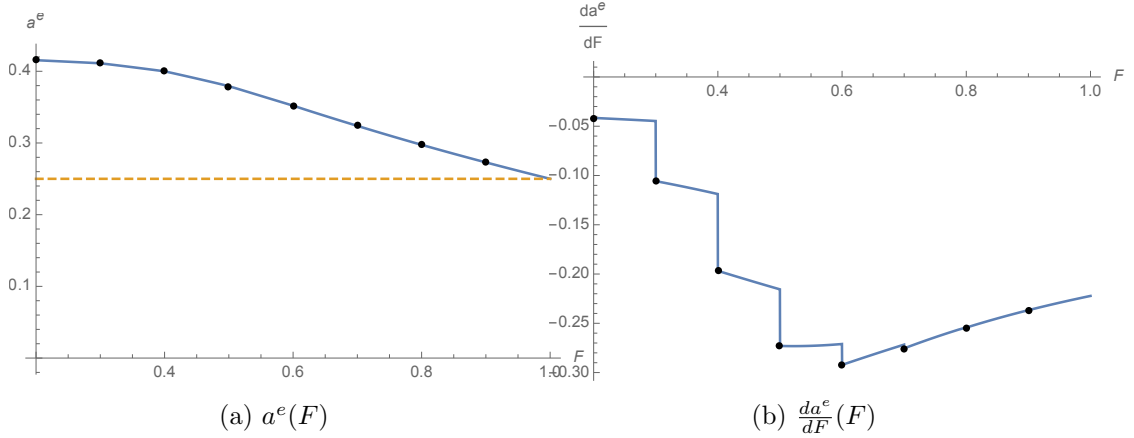
where  $\phi(a_{-i}; F) \equiv \psi(0; F)$ , and  $\psi(0; F)$  is obtained by evaluating the function  $\psi(g; F)$  defined in Eq. (A1) choosing  $q_h \equiv a_{-i}$  therein. Because of our symmetric configuration of clearing members, we restrict our attention to symmetric equilibria. A *symmetric equilibrium* among members is  $a^e \in [0, 1]$  such that, member  $i$ 's best response when all other members choose  $a^e$  is also  $a^e$ ; i.e., there is no unilateral deviation. Taking the first-order condition of the objective function in (23), and then setting  $a_{-i}$  with  $a_i$ , we obtain the following result.

**Proposition 5** *For a large enough differential return  $R - r > 0$ , given a default fund level  $F \in [\frac{2c}{N}, c)$ , there exists a unique symmetric response of members; i.e.,  $a_i = a^e \forall i$ , which satisfies*

$$\frac{R - 1 - \phi(a^e, F)}{2(R - r)} - a^e = 0. \quad (24)$$

Moreover,  $a^e$  is a strictly decreasing function of  $F$ ; for  $F \neq \frac{lc}{N}$ ,  $l = 2, 3, \dots, N - 1$ ,  $a^e$  is an infinitely differentiable function of  $F$  and  $\frac{da^e}{dF} < 0$ ; for  $F = \frac{lc}{N}$ ,  $l = 3, 4, \dots, N - 1$ ,  $\frac{da^e}{dF}(F+) - \frac{da^e}{dF}(F-) < 0$ , where  $\frac{da^e}{dF}(F+)$  and  $\frac{da^e}{dF}(F-)$  are respectively the right and left derivatives of  $a^e(F)$  at  $F$ .

**Risk-shifting.** We demonstrate that the risk-shifting pattern shown in the binary case also manifests in the continuous risk-taking setup. As a first-best benchmark, we solve for the socially optimal risky asset investment  $a^*$  that maximizes the aggregate value of all members



**Figure 4. Members' Strategic Response to the Default Fund Level.** Panel 4a plots the symmetric strategic response  $a^e$  as a function of  $F$  (blue solid line).  $a^e$  is strictly decreasing in  $F$ . The socially optimal invested fraction in the risky project is shown in the amber dashed line and is equal to  $a^* = 0.25$ . Panel 4b plots the derivative of  $a^e$  with respect to  $F$ . There are downward jumps at the kinks  $\frac{l}{N}$  for  $l = 3, 4, \dots, N - 1$ . Kinks are indicated by black dots. The parameters used are:  $R = 3.5, r = 0.5, N = 10$ , and  $c = 1$ .

net of the CCP's equity loss. Equivalently,  $a^*$  maximizes the expected payoff of a representative member:

$$a^* = \arg \max_{a_i} \mathbb{E}[a_i \tilde{R}_i + (1 - a_i) - c \mathbf{1}_{i \text{ defaults}} - r_f F]. \quad (25)$$

While  $a^*$  balances the socially desirable tradeoff between risk and return, members have incentives for risk-shifting under the loss mutualization scheme, as shown by the following proposition:

**Proposition 6** *Assume  $R > 1 + c$ . Then the socially optimal investment  $a^*$  in the risky project satisfies*

$$0 < a^* = \frac{R - 1 - c}{2(R - r)} < \frac{1}{2}. \quad (26)$$

*The privately optimal investment choice  $a^e$  of the clearing members is given by the solution to Eq. (24) and satisfies*

$$a^* < a^e. \quad (27)$$

A direct comparison of the privately optimal investment choice  $a^e$  (solution to Eq. (24)) and the socially optimal outcome  $a^*$  (solution to Eq. (26)) immediately reveals the economic

mechanism. In the risk-return tradeoff faced by the social planner, a member faces marginal cost-due-to-default  $c$ ; in the decentralized risk-return tradeoff under loss mutualization, the marginal cost-due-to-default becomes  $\phi(a; F)$ . As  $\phi(a; F) < c$  (which directly results from risk-sharing via the CCP), members strategically chooses higher risk than what is socially optimal. Figure 4 plots  $a^e(F)$ ,  $a^*$ , and  $\frac{da^e}{dF}$ : the decreasing pattern of  $a^e(F)$  is clearly seen from the figure.

## 5.2 Equilibrium between Clearing Members and the Regulator

In this section, we show that a default fund higher than the Cover II requirement can be used to regulate members' risk-taking. As in Section 4, the regulator selects the default fund level  $F$ , which maximizes the social value of the system (including all members and the CCP), anticipating the risk-taking activities chosen by the members in response to his choice. The members' response has been characterized in Proposition 5. Next, we describe the game theoretical setting:

**Definition 5** *For a given number  $N$  of clearing members, a symmetric Nash equilibrium between all clearing members and the CCP is the set of members' risk profiles  $\{a_i^e\}_{i=1}^n$ , and the default fund contribution  $F^e$  set by the CCP such that,*

1. *Taking the default fund  $F^e$  and other members' risk profile  $a_{-i}^e$  as given,  $a_i^e$  solves the optimization problem of clearing member  $i$  given in (22).*
2. *Taking as given the risk profile of clearing members  $a^e$ , the regulator chooses a feasible default fund level  $F^{e c_0}$  to maximize the total value of the system  $W(F)$ :*

$$W(F) = \sum_i V_i - \mathbb{E}[(\mathcal{N}_d \cdot c - N \cdot F)^+] = N \cdot (B(F) - r_f F), \quad (28)$$

$$F^e = \arg \max_F W(F) = \arg \max_F (B(F) - r_f F), \quad (29)$$

---

<sup>c0</sup>We recall that feasible means that the assumptions (2)-(3) and condition (2) are satisfied.

where  $V_i$  is the payoff of member  $i$  given by (22),  $N_d$  is the number of defaults, and  $B(F) = -(R - r)(a^e(F))^2 + (R - 1 - c)a^e(F) + 1$  is the part of the representative member's value which does not account for the opportunity cost of default fund  $F$ .

To solve for the tradeoff in choosing the default fund, we analyze the differential properties of the various components of the objective function. First, recall from Proposition 5 that  $a^e(F)$  is continuously differentiable in  $F$  except over the set of kinks  $\{\frac{lc}{N}; l = 2, \dots, N - 1\}$ . Thus the same property holds for  $B(F)$ . To see how  $B(F)$  changes with  $F$ , consider the quadratic function  $U(a) = -(R - r)a^2 + (R - 1 - c)a + 1$ , which achieves its maximum at  $a = a^*$ . An application of the chain rule shows that the marginal benefit of increasing  $F$  is given by

$$B'(F) = \frac{\partial U}{\partial a} \frac{\partial a^e}{\partial F} = -2(R - r)(a^e(F) - a^*) \frac{da^e}{dF}, \quad \forall F \neq \frac{lc}{N}, l = 2, \dots, N - 1. \quad (30)$$

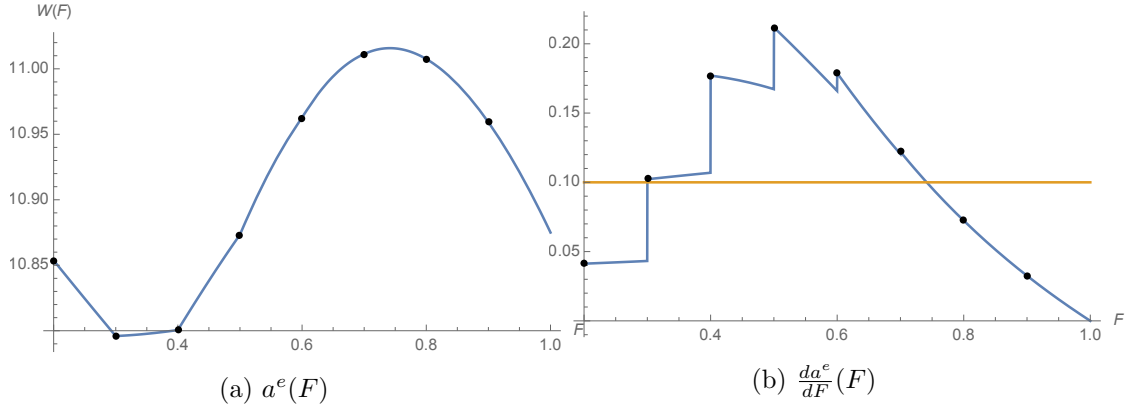
Recall from propositions 5 and 6 that members exhibit risk-shifting ( $a^e(F) > a^*$ ), and that a higher default fund can mitigate risk-shifting ( $\frac{da^e}{dF} < 0, \forall F \neq \frac{lc}{N}, l = 2, \dots, N - 1$ ). Therefore,  $B(F)$  increases with  $F$ ; that is, a higher default fund increases value  $B(F)$  by steering members' risk-taking closer to the socially optimal level.<sup>c0</sup>

**Proposition 7** *The equilibrium default fund  $F^e$  set by the regulator is either the Cover II level  $\frac{2c}{N}$  or a solution to the first-order condition of Equation (29). Formally,  $F^e \in \{\bar{F}^e, \frac{2c}{N}\}$  where  $\bar{F}^e$  satisfies*

$$B'(\bar{F}^e) = r_f, \quad B''(\bar{F}^e) \leq 0, \quad \bar{F}^e \neq \frac{lc}{N}, l = 2, \dots, N. \quad (31)$$

Proposition 7 offers a simple algorithm for the regulator to pin down the equilibrium (constrained optimal) default fund. To maximize the social objective value function (29), the regulator balances the marginal benefit  $B'(\bar{F}^e)$  and marginal cost of capital  $r_f$ . While a higher default fund increases the social value by mitigating risk-shifting, it is also costly because the fund needs to be segregated. If the marginal benefit and marginal cost cross paths

<sup>c0</sup>Formally, Equation (30) implies that the marginal benefit of increasing  $F$  is positive; i.e.,  $B'(F) > 0, \forall F \in [\frac{2c}{N}, c) \setminus \{\frac{lc}{N}; l = 2, \dots, N - 1\}$ . Moreover, as  $F$  crosses a kink  $\frac{lc}{N}$  from below, Proposition 6 indicates that  $B'(\frac{lc}{N}+) - B'(\frac{lc}{N}-) = -2(R - r)(a^e(\frac{lc}{N}) - a^*)[\frac{da^e}{dF}(\frac{lc}{N}+) - \frac{da^e}{dF}(\frac{lc}{N}-)] > 0$ . Hence, at each kink value of  $F$ , the marginal benefit of increasing  $F$  increases with a positive jump. Taken together,  $B(F)$  increases with  $F$ .



**Figure 5. The Social Value Function and the Marginal Value.** Panel (a) plots the total wealth  $W(F)$ . Panel (b) plots the marginal values: the marginal benefit  $B'(F)$  (in blue), and the marginal cost  $r_f$  (in amber). The parameters used are:  $R = 3.5, r = 0.5, N = 10, c = 1$  and  $r_f = 0.1$ . The optimal default fund is  $F^e \approx 0.74$ .

at a fund value higher than the Cover II lower bound, and the condition  $B''(\bar{F}^e) \leq 0$  holds, then  $\bar{F}^e$  is the constrained optimal level. In other words, if the marginal value of increasing  $F$  is higher than the opportunity cost at the lower bound  $\frac{2c}{N}$  ( $B'(\frac{2c}{N}) > 0$ ), then  $\frac{2c}{N} < F^e < c$ . That is, the optimal default fund level exceeds the critical level for the Cover II requirement. Figure 5 considers a scenario in which the Cover II rule is not socially optimal. As the default fund level increases, the social value function increases and peaks at  $F^e = 0.74$ . This value corresponds to a Cover 7.4 rule instead for a 10-member clearing arrangement.

**An Example with Three Members** We consider a simplified setting consisting of three clearing members, in which closed-form expressions for  $a^e(F)$  and  $F^e$  can be obtained. For a given choice of  $F$  that satisfies the Cover II requirement, the equilibrium investment in the risky asset is given by  $a^e(F) = \frac{R-r}{c-F} - \frac{1}{2} - \sqrt{\left(\frac{R-r}{c-F} - \frac{1}{2}\right)^2 - \frac{R-1-F}{c-F}}$ , and obtained by solving Equation (24). It is immediate to see that  $a^e(F)$  is strictly decreasing in  $F$ . We can explicitly compute the unique equilibrium default fund  $F^e$ . In particular, let  $h(a) = a - a^* - r_f \frac{1+a^*+2a^*a-a^2}{(1+a+a^2)^2}$ , which is a strictly increasing function over  $[0, 1]$ , and satisfies  $h(a^*) < 0$ . Then

$$F^e = \begin{cases} \frac{2c}{3}, & \text{if } h(a^e(\frac{2c}{3})) \leq 0, \\ c - 2(R-r) \frac{a_0 - a^*}{1 + a_0 + a_0^2}, & \text{if } h(a^e(\frac{2c}{3})) > 0. \end{cases} \quad (32)$$

where  $a_0$  is the unique root to  $h(a_0) = 0$  over the interval  $(a^*, a^e(\frac{2c}{3}))$ . This example serves to illustrate that even in a simple central clearing setup consisting of three members, neither Cover II nor Cover III may be the optimal default fund allocation. In fact,  $2c/3 < F^e < c$ , and thus an intermediary level for the default fund level might be optimal.

Analogous to the binary case, we can show that the ratio between the optimal default fund level  $F^e(N)$  and the default cost  $c$ , or equivalently, the optimal proportion of covered members,  $x(N)/N$ , also converges to a constant as the number of members  $N$  grows large. This limit is a positive number in  $(0, 1)$  when the marginal opportunity cost  $r_f$  is sufficiently low. In particular, paralleling the result in Proposition 4, we have the following asymptotic result.

**Proposition 8** *In the large CCP network limit – i.e., as  $N \rightarrow \infty$  – the unique symmetric response of members,  $a_i = a^e$  is given by*

$$a^e(F) \rightarrow \begin{cases} \frac{R-1}{2(R-r)}, & \text{if } \frac{F}{c} \leq \frac{R-1}{2(R-r)}, \\ \frac{(1+a^*) - \sqrt{(1+a^*)^2 - 2\frac{R-1-F}{R-r}}}{2}, & \text{if } \frac{F}{c} > \frac{R-1}{2(R-r)}. \end{cases} \quad (33)$$

Letting  $F(a) = R - 1 + \frac{R-r}{2}[(1+a^* - 2a)^2 - (1+a^*)^2]$  be the inverse of line two in (33),  $a_\infty := \frac{(1+r_f)a^* + r_f}{1+2r_f}$ , and  $\hat{F}_\infty := F(a_\infty)$ . Then we have

$$\frac{F^e(N)}{c} = \frac{x(N)}{N} \rightarrow \begin{cases} \frac{\hat{F}_\infty}{c}, & \text{if } B(\hat{F}_\infty) - r_f \hat{F}_\infty > B(0) \text{ and } a_\infty < \frac{R-1}{2(R-r)}, \\ 0, & \text{else,} \end{cases} \quad (34)$$

where  $B(\cdot)$  is defined in Definition 5.

From Proposition 8, we see that for a CCP consisting of many members, the highest risk choice  $a^e$  converges  $\frac{R-1}{2(R-r)}$ . When members take this level of risk, then the optimal response of the regulator is to set the default fund level to zero because the members' risk choice is independent of  $F$ . If the marginal opportunity cost of default fund is low, the regulator may attain a higher per member social welfare by charging a larger default fund requirement  $\hat{F}_\infty$  to members, that in turn incentivize them to take a lower level risk level  $a_\infty$ .



## 6 Conclusion and Policy Implications

The problem of the optimal determination of default fund levels contributed by members of a clearinghouse has been the subject of extensive regulatory debate. Current regulatory requirements prescribe that default fund contributions should guarantee that the clearinghouse is able to continue its services in the event that its two largest clearing members default. There is, however, no economic analysis of the conditions under which this rule is socially optimal, or of alternative designs that are welfare improving. Our paper fills this important gap and introduces a parsimonious model to study the main economic incentives behind the determination of the default fund requirements. While default funds allow members to effectively share counterparty risk *ex post*, we highlight a novel mechanism related to loss mutualization that induces members to take excessive risk *ex ante* owing to an inherent externality among them. Our analysis shows that the CCPs can mitigate the inefficiency generated by members' excessive risk-taking activities through an optimal choice of the default fund level. Such a choice balances the *ex post* risk-sharing and the *ex ante* risk-taking of members.

Our analysis shows that if the clearinghouse consists of a sufficiently high number of clearing members, then the optimal cover rule should be to cover the default costs of a constant fraction of members. This finding contrasts with the currently imposed Cover II requirements, whose optimality is supported by our analysis only if the marginal opportunity costs of collateral posting are high.

Our results have important policy implications. They point to the need to design a simple rule that guarantees the coverage of the costs generated by the default of a proportion of clearing members. This simple covering requirement is robust to the size of the participating member base. The optimal proportion depends, in an explicit way, on the relation between the premium earned by the member who undertakes high-risk projects and the costs incurred at default. These parameters can be accessed by clearinghouses and their supervisory authorities who typically have detailed information on the risk profile of their members. Owing to its simplicity, the proposed rule can also serve as a benchmark against more complex rules based

on simulated scenario stress testing.

## References

- Acharya, V., and A. Bisin. 2014. Counterparty risk externality: Centralized versus over-the-counter markets. *Journal of Economic Theory* 153–82.
- Acharya, V. V., H. S. Shin, and T. Yorulmazer. 2010. Crisis resolution and bank liquidity. *Review of Financial Studies* 24:2166–205.
- Amini, H., D. Filipović, and A. Minca. 2015. Systemic risk and central clearing counterparty design. Working paper.
- Antinolfi, G., F. Carapella, and F. Carli. 2016. Transparency and collateral: central versus bilateral clearing. Working paper.
- Arnold, M. 2017. The impact of central clearing on banks lending discipline. *Journal of Financial Markets* 36.
- Bernstein, A., E. Hughson, and M. D. Weidenmier. 2014. Counterparty risk and the establishment of the new york stock exchange clearinghouse. NBER working paper.
- Biais, B., F. Heider, and M. Hoerova. 2012. Clearing, counterparty risk and aggregate risk. *IMF Economic Review* 60:193–222.
- . 2016. Risk-sharing or risk-taking? counterparty risk, incentives and margins. *Journal of Finance* 71:1669–98.
- Capponi, A., W. A. Cheng, and J. Sethuraman. 2017. Clearinghouse default waterfalls: Risk-sharing, incentives, and systemic risk. Working paper.
- Duffie, D., M. Scheicher, and G. Vuillemeys. 2015. Central clearing and collateral demand. *Journal of Financial Economics* 116:237–56.
- Duffie, D., and H. Zhu. 2011. Does a central clearing counterparty reduce counterparty risk? *Review of Asset Pricing Studies* 1:74–95.
- Ghamami, S., and P. Glasserman. 2017. Does otc derivatives reform incentivize central clearing? *Journal of Financial Intermediation* 32:76–87.
- Holmstrom, B., and J. Tirole. 2001. Lapm: A liquidity-based asset pricing model. *The Journal of Finance* 56:1837–67. ISSN 1540-6261.
- ICE. 2016. Risk management. Available at <https://www.theice.com/clear-europe/risk-management>.
- ISDA. 2013. Risk sensitive capital treatment for clearing member exposure to central counterparty default funds. *International Swaps and Derivatives Association Working paper*.
- Koepl, T., and C. Monnet. 2010. The emergence and future of central counterparties. federal reserve bank of philadelphia. Working paper.
- Koepl, T. V. 2012. Central counterparty clearing: incentives, market discipline and the cost of collateral. Queens university working paper.
- Loon, Y. C., and Z. K. Zhong. 2014. The impact of central clearing on counterparty risk, liquidity, and trading: Evidence from the credit default swap market. *Journal of Financial Economics* 112:91 – 115.

- Menkveld, A. J. 2017. Crowded positions: An overlooked systemic risk for central clearing parties. *The Review of Asset Pricing Studies* 7:209-242.
- Pirrong, C. 2011. *The economics of central clearing: theory and practice*. International Swaps and Derivatives Association.
- . 2014. A bill of goods: Central counterparties and systemic risk. *The Journal of Financial Market Infrastructures* 2:55–85.
- Zawadowski, A. 2013. Entangled financial systems. *Review of Financial Studies* 1291–323.

## A Proof of Proposition 2

In this Appendix, we prove Proposition 2. We first present some technical lemmas to fix preliminary results and notations.

**Lemma 6** *Suppose member  $i$  is alive, and  $g$  of the remaining  $N - 1$  members choose the low risk project. Then, for any given  $F \in [\frac{2c}{N}, \frac{c}{1+r_f}]$ , we have that*

$$\psi(g; F) := \sum_{k=N-1-\lfloor \frac{NF}{c} \rfloor}^{N-1} f_g(k) \left( c - \frac{N(c-F)}{k+1} \right). \quad (\text{A1})$$

Here  $\lfloor \cdot \rfloor$  denotes the floor function (giving the greatest integer less than or equal to the argument), and

$$f_g(k) := \sum_{m=0}^k \binom{g}{m} (1-q_l)^m q_l^{g-m} \times \binom{N-1-g}{k-m} (1-q_h)^{k-m} q_h^{N-1-g-(k-m)}$$

are positive constant.

**Proof of Lemma 6.** Suppose that the default fund  $F$  is such that

$$\frac{lc}{N} \leq F < \frac{(l+1)c}{N} \text{ or equivalently } 1 - \frac{1+l}{N} < 1 - \frac{F}{c} \leq 1 - \frac{l}{N}, \quad (\text{A2})$$

for some integer  $l = 2, 3, \dots, N - 1$ . Then member  $i$ 's contribution to other members' default when himself does not default is given by

$$\begin{aligned} \min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right) &= F \min \left( 1, \left( \frac{N}{\mathcal{N}_s} - 1 \right) \left( \frac{c}{F} - 1 \right) \right) \\ &= (c - F) \left( \frac{N}{\mathcal{N}_s} - 1 \right) 1_{\mathcal{N}_s \geq (1 - \frac{F}{c})N} + F 1_{\mathcal{N}_s < (1 - \frac{F}{c})N} \end{aligned}$$

For  $F$  in the range (A2), we have

$$N - (l+1) < \left( 1 - \frac{F}{c} \right) N \leq N - l$$

Thus  $\mathcal{N}_s \geq (1 - \frac{F}{c})N$  if and only if  $\mathcal{N}_s \geq n - l$ . In other words, among the remaining  $N - 1$  members, if there are less than or equal to  $l$  defaults, member  $i$  will pay less than  $F$ . But if there are  $l$  or more defaults and  $F = \frac{lc}{N}$  then member  $i$ 's default fund will be exhausted completely.

Suppose all members except member  $i$  choose the low risk project, then if member  $i$

survives, his expected contribution is

$$\begin{aligned}
& (c - F) \mathbb{E} \left[ \left( \frac{N}{\mathcal{N}_s} - 1 \right) 1_{\mathcal{N}_s \geq N-l} \middle| \text{member } i \text{ survives} \right] + F \cdot \Pr(\mathcal{N}_s < N - l | \text{member } i \text{ survives}) \\
&= (c - F) \sum_{k=N-(l+1)}^{N-1} \binom{N-1}{k} \frac{N-1-k}{k+1} (1-q_l)^k q_l^{N-1-k} + F \sum_{k=0}^{N-(l+2)} \binom{N-1}{k} (1-q_l)^k q_l^{N-1-k} \\
&= (c - F) \sum_{k=N-(l+1)}^{N-2} \binom{N-1}{k+1} (1-q_l)^k q_l^{N-1-k} + F \sum_{k=0}^{N-(l+2)} \binom{N-1}{k} (1-q_l)^k q_l^{N-1-k}. \quad (\text{A3})
\end{aligned}$$

Likewise, if all members except member  $i$  choose the high risk project, then if member  $i$  survives, his expected contribution is

$$\begin{aligned}
& (c - F) \mathbb{E} \left[ \left( \frac{N}{\mathcal{N}_s} - 1 \right) 1_{\mathcal{N}_s \geq N-l} \middle| \text{member } i \text{ survives} \right] + F \cdot \Pr(\mathcal{N}_s < N - l | \text{member } i \text{ survives}) \\
&= (c - F) \sum_{k=N-(l+1)}^{N-2} \binom{N-1}{k+1} (1-q_h)^k q_h^{N-1-k} + F \sum_{k=0}^{N-(l+2)} \binom{N-1}{k} (1-q_h)^k q_h^{N-1-k}. \quad (\text{A4})
\end{aligned}$$

In general, if there are  $g$  members among the remaining  $N - 1$  choosing the low risk project, for  $g = 0, 1, \dots, N - 1$ , then the number of surviving ones among these  $N - 1$  members,  $\mathcal{N}_s - 1$ , is the sum of the number of the survived ones choosing the low risk project and that of the survived choosing the high risk project. Specifically, the probability that there are  $k$  survived ones is given by

$$f_g(k) := \sum_{m=0}^k \binom{g}{m} (1-q_l)^m q_l^{g-m} \times \binom{N-1-g}{k-m} (1-q_h)^{k-m} q_h^{N-1-g-(k-m)}.$$

It follows that, if member  $i$  survives, his expected contribution is

$$\begin{aligned}
& (c - F) \mathbb{E} \left[ \left( \frac{N}{\mathcal{N}_s} - 1 \right) 1_{\mathcal{N}_s \geq N-l} \middle| \text{member } i \text{ survives} \right] + F \cdot \Pr(\mathcal{N}_s < N - l | \text{member } i \text{ survives}) \\
&= (c - F) \sum_{k=N-(l+1)}^{N-1} f_g(k) \frac{N-1-k}{1+k} + F \sum_{k=0}^{N-(l+2)} f_g(k) \\
&= \sum_{k=N-(l+1)}^{N-1} f_g(k) \left( \frac{N(c-F)}{k+1} - c \right) + F, \quad (\text{A5})
\end{aligned}$$

where the last line comes from the total probability  $\sum_{k=0}^{N-1} f_g(k) = 1$ . ■

**Lemma 7** For any given  $F \in [\frac{2c}{N}, c)$ , the function  $\psi(g; F)$  is strictly decreasing in  $g$ , i.e.

$$0 < \psi(0; F) < \psi(1; F) < \dots < \psi(N-1; F) < F < c.$$

**Proof of Lemma 7.** Suppose member  $i$  survives, and  $\mathcal{N}_s - 1$  is the number of survivals

except member  $i$ . Given how we define a default event, we have

$$\mathcal{N}_s - 1 = \sum_{j \neq i} \mathbf{1}_{\text{member } j \text{ defaults}} \quad (\text{A6})$$

By Lemma 6, we only need to show that

$$g \mapsto \mathbb{E} \left[ \min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right) \right] \equiv F - \psi(g; F)$$

is strictly decreasing in  $g$  for  $g = 0, 1, 2, \dots, n - 1$ , where  $g$  is the number of members other than member  $i$ , who chooses the low risk project. Because the expression  $\psi(g; F)$  only depends on the default probabilities  $p_h, p_l$ , not how defaults occur, we can choose a probability model for the defaults that is convenient to our analysis. More precisely, suppose for each member  $i$ , there is an independent random variable  $\epsilon_i$ , with a uniform distribution on  $(0, 1)$ , such that, if this member has chosen the low risk project, he will default at time 1 if and only if  $\epsilon_i < p_l$ ; on the other hand, if this member has chosen the high risk project, then he will default at time 1 if and only if  $\epsilon_i < p_h$ . Because the event  $\{\epsilon_i < p_l\}$  implies  $\{\epsilon_i < p_h\}$ , we see that, in this probability model of defaults, increasing  $g$ , the number of remaining members (other than member  $i$ ) choosing the low risk project, always makes  $\mathcal{N}_s$  non-increasing, and can make  $\mathcal{N}_s$  strictly increasing with a positive probability (equal to  $p_h - p_l$  in this particular specification of default).

Similarly, there is a positive chance that  $\mathcal{N}_s$  may decrease from 1 to  $N$  as  $g$  increases from 0 to  $N - 1$ . As  $\mathcal{N}_s$  varies between 1 and  $N$ ,  $\frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F)$  varies between  $(N - 1)(c - F)$  and 0, hence the random variable  $\min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right)$  is non-decreasing with  $g$ , and there is a positive chance that it is strictly decreasing with  $g$ . As a result, we know that the mapping

$$g \mapsto \mathbb{E} \left[ \min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right) \right], \quad g = 0, 1, \dots, N - 1$$

is strictly decreasing.

Lastly, because the expected contribution  $F - \psi(N - 1; F)$  is positive (member  $i$  has to contribute as long as there is at least one other member default), we have  $\psi(N - 1; F) < F$ . Moreover, the expectation contribution  $F - \psi(0; F)$  is strictly less than  $F$  (member  $i$  contributes less than  $F$  when there is no default). ■

**Lemma 8** *For any fixed  $g = 0, 1, \dots, N - 1$ , the function  $\psi(g; F)$  is piecewise linear and strictly increasing in  $F$ , in the interval  $[\frac{2}{N}c, \frac{c}{1+r_f}]$ . In particular,  $\psi(g; c) = c$ .*

**Proof of Lemma 8.** From (A1),  $\psi(g; F)$  is linear and strictly increasing for  $F \in (\frac{l}{N}, \frac{l+1}{N})$  with  $l = 2, 3, \dots, N - 1$ . Moreover, the nonnegative random variable  $\min(F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F))$  is almost surely continuous in  $F$ , and is bounded by  $c$ . By the dominated convergence theorem, we know that  $\psi(g; F)$  is continuous. Therefore, the function  $\psi(g; F)$  is strictly increasing for all  $F \in [\frac{2}{N}c, \frac{c}{1+r_f}]$ . The value of  $\psi(g; c)$  follows directly from (10). ■

To prove Proposition 2, we next analyze different scenarios for a member's investment choice, taken as given the default fund and other members' investment choices. In particular,

we characterize conditions such that member  $i$  chooses high risk given all possible combinations of risk profile of other members. When we check those conditions for all members, we will see that “all high risk” and “all low risk” strategy are the only possible equilibria.

Suppose that the other members do not choose the same risk level: There are  $g$  members who choose low risk projects and the rest  $N - 1 - g$  choose high risk projects, where  $g \in \{0, \dots, N - 1\}$ .

If member  $i$  chooses low risk, his expected utility is (by Lemma 6)

$$\begin{aligned} \mathbb{E} \left[ R_i^l - \mathbf{1}_{i \text{ defaults}} F - \mathbf{1}_{i \text{ survives}} \min \left( F, \frac{\sum_{j \neq i} \mathbf{1}_{j \text{ defaults}} (c - F)}{1 + \sum_{j \neq i} \mathbf{1}_{j \text{ survives}}} \right) - r_f F \right] \\ = \mu_l - q_l \psi(g; F) - F + \psi(g; F) - r_f F. \end{aligned}$$

If member  $i$  chooses instead high risk, his expected utility is

$$\begin{aligned} \mathbb{E} \left[ R_i^h - \mathbf{1}_{i \text{ defaults}} F - \mathbf{1}_{i \text{ survives}} \min \left( F, \frac{\sum_{j \neq i} \mathbf{1}_{j \text{ defaults}} (c - F)}{1 + \sum_{j \neq i} \mathbf{1}_{j \text{ survives}}} \right) - r_f F \right] \\ = \mu_h - q_h \psi(0; F) - F + \psi(0; F) - r_f F. \end{aligned}$$

Hence, member  $i$  chooses high (low, resp.) risky project when  $g$  members choose low and  $N - 1 - g$  choose high if and only if

$$\frac{\mu_h - \mu_l}{q_h - q_l} > (<, \text{resp.}) \psi(g; F). \quad (\text{A7})$$

When (A7) takes an equality, member  $i$  is indifferent to choosing high or low risky project. As a consequence,

1. If  $\frac{\mu_h - \mu_l}{q_h - q_l} > \psi(N - 1; F)$ , every member will choose the high risky project, regardless of other members' choice. Hence, the “all high risk” strategy is the unique equilibrium among members.
2. If  $\frac{\mu_h - \mu_l}{q_h - q_l} < \psi(0; F)$ , every member will choose the low risky project, regardless of other members' choice. Hence, the “all low risk” strategy is the unique equilibrium among members.
3. If  $\psi(0; F) \leq \frac{\mu_h - \mu_l}{q_h - q_l} \leq \psi(N - 1; F)$ , it is straightforward to verify that both the “all high risk” strategy and the “all low risk” strategy are equilibriums among members. To prove there cannot be any other forms of equilibrium, suppose there is an equilibrium which is consisted of  $g$  low and  $(N - g)$  high, for some  $g = 1, 2, \dots, N - 1$ . Then for any member choosing high, he faces  $g$  choosing low and  $(N - g - 1)$  choosing high, so in order for him to stay high as well, it must hold that

$$\frac{\mu_h - \mu_l}{q_h - q_l} \geq \psi(g; F). \quad (\text{A8})$$

Yet, for any member choosing low, he faces  $g - 1$  low and  $N - g$  high, so for this member to stay low, it must holds that

$$\frac{\mu_h - \mu_l}{q_h - q_l} \leq \psi(g - 1; F). \quad (\text{A9})$$



However, (A8) and (A9) cannot hold simultaneously because  $\psi(g-1; F) < \psi(g; F)$  (see Lemma 7).

Finally, we prove that the all low risk profile is Pareto dominating the all high risk profile when  $\psi(0; F) \leq \frac{\mu_h - \mu_l}{q_h - q_l} \leq \psi(N-1; F)$  holds. Recall that member  $i$ 's expected utility under the all low risk profile is

$$\mu_l - q_l \psi(N-1; F) - F + \psi(N-1) - r_f F.$$

Likewise, his expected utility under the all high risk profile is

$$\mu_h - q_h \psi(0; F) - F + \psi(0; F) - r_f F.$$

The difference is then given by

$$\begin{aligned} & [\mu_l - q_l \psi(N-1; F) - F + \psi(N-1) - r_f F] - [\mu_h - q_h \psi(0; F) - F + \psi(0; F) - r_f F] \\ &= \mu_l - \mu_h + (1 - q_l) \psi(N-1; F) - (1 - q_h) \psi(0; F). \end{aligned}$$

When  $\mu_h - \mu_l \leq (q_h - q_l) \psi(N-1; F)$ , the above expression is bounded from below by  $(1 - q_h) (\psi(N-1; F) - \psi(0; F)) > 0$ , due to Lemma 7. Therefore, among the two possible equilibriums, the all low risk profile is Pareto dominating.

## B Proof of Proposition 5

Recall that a symmetric equilibrium under cover-II requirement is given by  $a_i^e = a^e$  for all  $i = 1, 2, \dots, N$ , with  $a^e$  being the root to the follow equation:

$$f(a^e; F) = 0, \text{ where } f(a; F) = \frac{R-1-\phi(a; F)}{2(R-r)} - a.$$

We first show the existence of such a root. To that end, recall that (10) implies

$$\mathbb{E} \left[ \min \left( F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F) \right) \middle| \text{member } i \text{ survives} \right] = F - \phi(a; F),$$

where  $\mathcal{N}_s$  is the number of surviving members, each of who has a default probability of  $a$ . It follows that when  $a = 0$ , we have  $\mathcal{N}_s = N$  almost surely, so  $0 = F - \phi(0; F)$ , which implies that

$$\phi(0; F) = F.$$

Therefore we have

$$f(0; F) = \frac{R-1-F}{2(R-r)} > \frac{R-1-c}{2(R-r)} = a^* > 0. \quad (\text{A10})$$

On the other hand, because  $\min(F, \frac{N - \mathcal{N}_s}{\mathcal{N}_s} (c - F)) \leq F$  holds almost surely, we know that  $\phi(a; F) \geq 0$  for all  $a \in [0, 1]$ . Therefore, we have

$$f(1; F) = \frac{R-1-\phi(a; F)}{2(R-r)} - 1 \leq \frac{R-1}{2(R-r)} - 1 < \frac{1}{2} - 1 = -\frac{1}{2} < 0. \quad (\text{A11})$$

Hence, from (A10) and (A11) we can conclude that, for any fixed  $F \in [\frac{2c}{N}, \frac{c}{1+r_f})$ , there exists at least one  $a^e$  such that  $f(a^e; F) = 0$ .

To prove the uniqueness of  $a^e$ , we demonstrate that, when  $R - r > 0$  is sufficiently large,  $f(a; F)$  is strictly decreasing in  $q$  over  $(0, 1)$  for every  $F$ . Indeed, Lemma 7 indicates that  $\phi(q_h; F) = \psi(0; F) < \psi(N - 1; F) = \phi(q_l; F)$  for  $0 < q_l < q_h < 1$ , so  $\phi(a; F)$  is strictly decreasing in  $a$ . Therefore, the monotonicity of  $f(a; F)$  in  $a$  is completely determined by the strictly increasing function  $-\frac{1}{2(R-r)}\phi(a; F)$  and the strictly decreasing function  $-a$ . Intuitively, as  $(R - r)$  becomes large, we have  $-\frac{1}{2(R-r)}\frac{\partial}{\partial a}\phi(a; F) - 1 < 0$ , meaning  $f(a; F)$  is strictly decreasing in  $a$ . In Lemma 9 below we formally bound  $|\frac{\partial}{\partial a}\phi(a; F)|$  from above, which effectively gives a sufficient lower bound for  $(R - r)$  such that  $f(a; F)$  is strictly decreasing, which implies the uniqueness of  $a^e$ .

To prove the monotonicity of  $a^e(F)$  in  $F$ , we use lemma 8 to know that, for each fixed  $a \in (0, 1)$ ,  $\phi(a; F)$  is strictly increasing for all  $F \in [\frac{2c}{N}, \frac{c}{1+r_f})$ , so  $f(a; F)$  is strictly decreasing in  $F$  over the same domain. Let  $\frac{2c}{N} \leq F_1 < F_2 < c$ , we have

$$0 = f(a^e(F_2); F_2) = f(a^e(F_1); F_1) > f(a^e(F_1); F_2).$$

Because  $f(q; F_2)$  is strictly decreasing in  $q$  when  $R - r > 0$  is sufficiently large, we must have  $a^e(F_2) < a^e(F_1)$ . This proves the monotonicity of  $a^e(F)$  on  $F$ .

The infinite differentiability of  $a^e(F)$  in  $F$  follows from implicit differentiation and Lemma 8, from which we know that  $f(a; F)$  is infinitely differentiable in  $F$  if  $F \neq \frac{lc}{N}$  for  $l = 2, 3, \dots, N - 1$ . This, in conjunction with the monotonicity of  $a^e(F)$  in  $F$ , implies that  $\frac{da^e}{dF}(F) < 0$  for any  $F \in (\frac{2c}{N}, c)$  but  $F \neq \frac{lc}{N}$ ,  $l = 3, 4, \dots, N - 1$ .

Lastly, to prove that  $a^e(F)$  is not differentiable at  $F = \frac{lc}{N}$ , and with a greater left derivative, we use implicit differentiation to obtain that, for any  $F \in (\frac{lc}{N}, \frac{(l+1)c}{N})$  with  $l = 3, \dots, N - 2$ ,

$$\frac{da^e}{dF}(F) = -\left. \frac{\frac{\partial f}{\partial F}}{\frac{\partial f}{\partial a}} \right|_{(a,F)=(a^e(F),F)} = -\frac{u_l^{(N)}(a^e(F))}{2(R-r) + c \cdot v_l^{(N)'}(a^e(F)) - (c-F)u_l^{(N)'}(a^e(F))},$$

where  $v_l^{(N)}$  and  $u_l^{(N)}$  are functions defined in the proof of Lemma 9 below. Thus, the right derivative at  $F = \frac{lc}{N}$  is given by

$$\frac{da^e}{dF}\left(\frac{lc}{N}+\right) = -\frac{u_l^{(N)}(a^e(\frac{lc}{N}))}{2(R-r) + c \cdot v_l^{(N)'}(a^e(\frac{lc}{N})) - (c - \frac{lc}{N})u_l^{(N)'}(a^e(\frac{lc}{N}))}. \quad (\text{A12})$$

Similarly, for the left derivative at  $F = \frac{lc}{N}$  we obtain that

$$\frac{da^e}{dF}\left(\frac{lc}{N}-\right) = -\frac{u_{l-1}^{(N)}(a^e(\frac{lc}{N}))}{2(R-r) + c \cdot v_{l-1}^{(N)'}(a^e(\frac{lc}{N})) - (c - \frac{lc}{N})u_{l-1}^{(N)'}(a^e(\frac{lc}{N}))}. \quad (\text{A13})$$

To compare (A12) with (A13), we first show that the denominators for  $\frac{da^e}{dF}(\frac{lc}{N}+)$  and  $\frac{da^e}{dF}(\frac{lc}{N}-)$

are the same. To that end, we have

$$\begin{aligned}
& \left( c \cdot v_l^{(N)'}(a^e(\frac{lc}{N})) - c \frac{N-l}{N} u_l^{(N)'}(a^e(\frac{lc}{N})) \right) - \left( c \cdot v_{l-1}^{(N)'}(a^e(\frac{lc}{N})) - c \frac{N-l}{N} u_{l-1}^{(N)'}(a^e(\frac{lc}{N})) \right) \\
&= c \left( \binom{N-1}{N-1-l} [(1-a)^{N-1-l} a^l]' - \frac{N-l}{N} \binom{N}{N-l} [(1-a)^{N-l-1} a^l]' \right) \Big|_{a=a^e(\frac{lc}{N})} \\
&= c [(1-a)^{N-1-l} a^l]' \Big|_{a=a^e(F)} \cdot \left( \binom{N-1}{l} - \frac{N-l}{N} \binom{N}{l} \right) = 0.
\end{aligned}$$

On the other hand, recall that when  $R-r$  is large enough, both  $\frac{da^e}{dF}(\frac{lc}{N}+)$  and  $\frac{da^e}{dF}(\frac{lc}{N}-)$  are negative. Given that  $u_{l-1}^{(N)}(a^e(\frac{lc}{N})), u_l^{(N)}(a^e(\frac{lc}{N})) > 0$ , we know from (A12) and (A13) that the common denominator for  $\frac{da^e}{dF}(\frac{lc}{N}+)$  and  $\frac{da^e}{dF}(\frac{lc}{N}-)$  must be positive. Thus,

$$\begin{aligned}
\frac{da^e}{dF}(\frac{lc}{N}+) - \frac{da^e}{dF}(\frac{lc}{N}-) &= \frac{-[u_l^{(N)}(a^e(\frac{lc}{N})) - u_{l-1}^{(N)}(a^e(\frac{lc}{N}))]}{2(R-r) + c \cdot v_l^{(N)'}(a^e(\frac{lc}{N})) - c \frac{N-1}{N} u_l^{(N)'}(a^e(\frac{lc}{N}))} \\
&= \frac{-(\binom{N}{N-1})(1-a)^{N-l-1} a^l \Big|_{a=a^e(\frac{lc}{N})}}{2(R-r) + c \cdot v_l^{(N)'}(a^e(\frac{lc}{N})) - c \frac{N-1}{N} u_l^{(N)'}(a^e(\frac{lc}{N}))} < 0.
\end{aligned}$$

**Lemma 9** For any  $F \in [\frac{2c}{N}, \frac{c}{1+r_f})$ , we have that

$$\sup_{a \in [0,1]} \left| \frac{\partial}{\partial a} \phi(a; F) \right| < \infty.$$

**Proof.**

Without loss of generality, let us suppose  $F \in [\frac{lc}{N}, \frac{(l+1)c}{N})$  for some  $l = 2, 3, \dots, N-1$ , so that  $\lfloor \frac{NF}{c} \rfloor = l$ . Using (A1), we have

$$\begin{aligned}
\phi(a; F) &= \sum_{k=N-1-l}^{N-1} \binom{N-1}{k} (1-a)^k a^{N-1-k} \left( c - \frac{N(c-F)}{k+1} \right) \\
&= c \cdot v_l^{(N)}(a) - (c-F) u_l^{(N)}(a),
\end{aligned} \tag{A14}$$

where

$$v_l^{(N)}(a) = \sum_{k=N-1-l}^{N-1} \binom{N-1}{k} (1-a)^k a^{N-1-k}, \quad u_l^{(N)}(a) = \sum_{k=N-l}^N \binom{N}{k} (1-a)^{k-1} a^{N-k}.$$

Functions  $v_l^{(N)}(a)$  and  $u_l^{(N)}(a)$  do not depend on  $F$  and  $c$ , and apparently both of them have continuous first order derivative in  $a$  over  $[0, 1]$ . This completes the proof. ■

## C Proof of Proposition 6

We first study the first best. To that end, recall from (25) that

$$a^* = \arg \max[-(R-r)a^2 + (R-1-c)a + 1 - r_f F].$$

Using the first order condition, we obtain that  $a^* = \frac{R-1-c}{2(R-r)}$ .

To prove (27), the upper bound follows immediately from (24) and the fact that  $\phi(a; F) \geq 0$ :

$$a^e = \frac{R-1-\phi(a^e; F)}{2(R-r)} \leq \frac{R-1}{2(R-r)}.$$

To prove the lower bound, let us suppose for the moment that  $F$  satisfies a cover- $(N-1)$  rule, i.e.  $\frac{(N-1)c}{N} \leq F < c$ . In this case, we have  $\phi(a; F) = F - (c-F)\frac{a(1-a^{N-1})}{1-a} < c$  for all  $0 < a < 1$  and  $\frac{(N-1)c}{N} \leq F < c$ . On the other hand, using Lemma 8 we know that, for each fixed  $a \in (0, 1)$ ,  $\phi(a; F)$  is strictly increasing for all  $F \in [\frac{2c}{N}, \frac{c}{1+r_f})$ , so that  $\phi(a; F) < c$  for all  $\frac{2c}{N} \leq F < c$ . Comparing  $a^*$  and  $a^e$  in expressions (26) and (24), we have that

$$a^e = \frac{R-1-\phi(a^e; F)}{2(R-r)} > \frac{R-1-c}{2(R-r)} = a^*.$$

This completes the proof.

## D Proof of Proposition 7

The objective function  $B(F) - r_f F$  is differentiable in  $F$  off the set of kinks  $\{\frac{lc}{N}; l = 2, 3, \dots, N-1\}$ . Thus, if the equilibrium  $F^e$  is not at one of the kinks, it must solve the regulator's tradeoff between the marginal benefit and marginal cost, i.e. the marginal benefit of increase  $F$  in mitigating risk-shifting should equal to the marginal cost of opportunity of default fund segregation. However, as  $F$  increases over the kink  $\frac{lc}{N}$  for some  $l = 3, 4, \dots, N-1$ , the marginal cost  $r_f$  increases continuously, but the marginal benefit  $B'(F)$  increases abruptly. Thus, it is never optimal to choose a default fund level  $F$  at one of such kinks.

Next we show that if the marginal value of increasing  $F$  is higher than the opportunity cost at the lower bound  $\frac{2c}{N}$  ( $B'(\frac{2c}{N}) > r_f$ ), then  $\frac{2c}{N} < F^e < c$ . We claim that

$$B'(\frac{2c}{N}) > r_f \Leftrightarrow \frac{(a^e(\frac{2c}{N}) - a^*) \cdot u_2^{(N)}(a^e(\frac{2c}{N}))}{1 + \frac{c}{2(R-r)}[v_2^{(N),'}(a^e(\frac{2c}{N})) - \frac{N-2}{N}u_2^{(N),'}(a^e(\frac{2c}{N}))]} > r_f. \quad (\text{A15})$$

where  $v_2^{(N)}(a) = \sum_{k=N-3}^{N-1} \binom{N-1}{k} (1-a)^k a^{N-1-k}$ ,  $u_2^{(N)}(a) = \sum_{k=N-2}^N \binom{N}{k} (1-a)^{k-1} a^{N-k}$ . To see (A15), notice that for any  $F \neq \frac{lc}{N}$ ,

$$B'(F) = [-2(R-r)a^e(F) + R-1-c] \frac{da^e}{dF}(F) = -2(R-r)(a^e(F) - a^*) \frac{da^e}{dF}(F). \quad (\text{A16})$$

Moreover, using implicit differentiation we have

$$-\frac{da^e}{dF}(F) = \frac{\frac{\partial \phi}{\partial F}}{2(R-r) + \frac{\partial \phi}{\partial a}} = \frac{u_2^{(N)}(a^e(F))}{2(R-r) + c[v_2^{(N)}(a^e(F)) - \frac{N-2}{N}u_2^{(N)}(a^e(F))]}.$$
 (A17)

From (A16) and (A17) we have

$$B'(\frac{2c}{N}+) = \frac{u_2^{(N)}(a^e(\frac{2c}{N}))}{1 + \frac{c}{2(R-r)}[v_2^{(N)}(a^e(\frac{2c}{N})) - \frac{N-2}{N}u_2^{(N)}(a^e(\frac{2c}{N}))]}.$$
 (A18)

Therefore, the condition in (A15) is equivalent to  $B'(\frac{2c}{N}+) - r_f > 0$ , so the objective function  $B(F) - r_f F$  is locally increasing in a small right neighborhood of  $\frac{2c}{N}$ . Similarly, because  $a^e(c-) = a^*$  (see Proposition 6), we know that  $B'(c-) - r_f = -r_f < 0$ . That is, the objective function  $B(F) - r_f F$  is locally decreasing in a small left neighborhood of  $c$ . It follows that the maximum  $F^e < c$ .

## E Proof of Equation (32)

When  $N = 3$ , the first order equation is a quadratic equation

$$f(a; F) = \frac{R - 1 - F + (c - F)(a + a^2)}{2(R - r)} - a = 0.$$
 (A19)

Its solution is given by

$$a_{\pm} = \frac{R - r}{c - F} - \frac{1}{2} \pm \sqrt{\left(\frac{R - r}{c - F} - \frac{1}{2}\right)^2 - \frac{R - 1 - F}{c - F}}.$$

Since  $R - r > \frac{c}{2}$ , we know that  $a_- > 0$  and  $a_+ < 0$ . Moreover, one can easily verify that  $a_- < 1$  by plugging  $a = 1$  into (A19) and using the condition that  $2c - 3F \leq 0$  (the cover-II requirement). Hence,  $a_-$  gives the formula for  $a^e(F)$ .

To demonstrate the monotonicity of  $a^e(F)$ , we notice that, for any  $\frac{2c}{3} \leq F < c$  and  $0 \leq a \leq 1$ ,

$$\begin{aligned} \frac{\partial f}{\partial a} &= \frac{(c - F)(1 + 2a)}{2(R - r)} - 1 \leq \frac{\frac{c}{3}(1 + 2)}{2(R - r)} - 1 = \frac{c}{2(R - r)} - 1 < 0, \\ \frac{\partial f}{\partial F} &= -\frac{1}{2(R - r)}(1 + a + a^2) < 0. \end{aligned}$$

By the implicit differentiation theorem, we know that  $a^e(F)$  is strictly decreasing and differentiable in  $F$ .

Now that we have established that the mapping  $F \mapsto a^e(F)$  is one-to-one and decreasing, to obtain the equilibrium  $F^e$ , we consider a change of variable:  $a^e \mapsto F(a^e)$ , namely, the inverse of  $a^e(F)$ :

$$F(a) = c - 2(R - r) \frac{a - a^*}{1 + a + a^2}.$$

Then we know that  $F^e = F(\hat{a})$ , where

$$\begin{aligned}\hat{a} &= \arg \max_{a \in [a^*, a^e(\frac{2c}{3})]} [-(R-r)a^2 + (R-1-c)a + 1 - r_f F(a)] \\ &= \arg \max_{a \in [a^*, a^e(\frac{2c}{3})]} [-(R-r)a^2 + (R-1-c)a + 1 - r_f F(a)].\end{aligned}\quad (\text{A20})$$

To fix  $\hat{a}$ , we calculate the derivative of the objective function of (A20) with respect to  $a$ :

$$-2(R-r)a + R-1-c - 2r_f(R-r) \frac{1-a^2 + a^* + 2a^*a}{(1+a+a^2)^2} = -2(R-r)h(a). \quad (\text{A21})$$

Moreover, notice that

$$h'(a) = 1 + r_f \frac{1-a^3 + 3a^*a^2 + 2(1+a^*)a}{(1+a+a^2)^3} > 1, \forall a \in [0, 1]. \quad (\text{A22})$$

Therefore, the first order condition equation  $h(a) = 0$  can have at most one root. On the other hand, one clearly has  $h(a^*) < 0$ , so if  $h(a^e(\frac{2c}{3})) \leq 0$ , then the objective function of (A20) is strictly increasing in  $a$  over  $[a^*, a^e(\frac{2c}{3})]$ . It follows that  $\hat{a} = a^e(\frac{2c}{3})$  and  $F^e = F(a^e(\frac{2c}{3})) = \frac{2c}{3}$ . On the other hand, if  $h(a^e(\frac{2c}{3})) > 0$ , then there is a unique maximizer  $\hat{a} = a_0 \in (a^*, a^e(\frac{2c}{3}))$ , which solves the first order condition equation  $h(a_0) = 0$ . Hence, the equilibrium default fund  $F^e = F(a_0)$ .

## F The large CCP network limit: proofs of Proposition 4 and Proposition 8

In this section, we derive the limit of  $\psi(N-1; F)$  and  $\phi(a; F)$  as  $N \rightarrow \infty$ , which will imply the limit of  $\hat{F}$  and  $a^e(F)$  as the number of members in the CCP network grows without bound. To that end, we recall (A14) that,

$$\phi(a; F) = c v_l^{(N)}(a) - (c - F) u_l^{(N)}(a),$$

where  $l = \lfloor \frac{NF}{c} \rfloor$ , and functions  $v_l^{(N)}$  and  $u_l^{(N)}$  have the following representation: suppose  $X$  follows a Binomial distribution with parameter  $(N-1, a)$  and  $Y$  follows a Binomial distribution with parameter  $(N, a)$ , then

$$v_l^{(N)}(a) = \mathbb{P}(X \leq l) = \mathbb{P}(\sqrt{N-1}(\frac{X}{N-1} - a) \leq \sqrt{N-1}(\frac{l}{N-1} - a)), \quad (\text{A23})$$

$$u_l^{(N)}(a) = \frac{\mathbb{P}(Y \leq l)}{1-a} = \mathbb{P}(\sqrt{N}(\frac{Y}{N} - a) \leq \sqrt{N}(\frac{l}{N} - a)). \quad (\text{A24})$$

By the central limit theorem, both  $\sqrt{N-1}(\frac{X}{N-1} - a)$  and  $\sqrt{N}(\frac{Y}{N} - a)$  converges in distribution to a normal distribution with mean 0 and variance  $a(1-a)$ . On the other hand, from

$\frac{NF-c}{c} < \lfloor \frac{NF}{c} \rfloor \leq \frac{NF}{c}$  we know that,

$$\lim_{N \rightarrow \infty} \sqrt{N-1} \left( \frac{l}{N-1} - a \right) = \lim_{N \rightarrow \infty} \sqrt{N} \left( \frac{l}{N} - a \right) = \begin{cases} \infty, & \text{if } \frac{F}{c} > a, \\ 0, & \text{if } \frac{F}{c} = a, \\ -\infty, & \text{if } \frac{F}{c} < a \end{cases}$$

It follows that

$$\lim_{N \rightarrow \infty} v_l^{(N)}(a) = 1_{\{\frac{F}{c} \geq a\}} + \frac{1}{2} 1_{\{\frac{F}{c} = a\}}, \quad \lim_{N \rightarrow \infty} u_l^{(N)}(a) = \frac{1_{\{\frac{F}{c} \geq a\}} + \frac{1}{2} 1_{\{\frac{F}{c} = a\}}}{1-a},$$

and

$$\lim_{N \rightarrow \infty} \phi(a; F) = c \cdot 1_{\{\frac{F}{c} > a\}} \left( 1 - \frac{1 - \frac{F}{c}}{1-a} \right).$$

Returning to the definition of  $\phi(a; F)$ , we know that

$$\lim_{N \rightarrow \infty} \psi(N-1; F) = c \cdot 1_{\{\frac{F}{c} > q_l\}} \left( 1 - \frac{1 - \frac{F}{c}}{1-q_l} \right).$$

Therefore, in the binary case, the switching point  $\hat{F}$ , which is defined as the unique root to  $\psi(N-1; \hat{F}) = \frac{\mu_h - \mu_l}{q_h - q_l}$ , converges to the solution to the equation:

$$c \cdot 1_{\{\frac{F}{c} > q_l\}} \left( 1 - \frac{1 - \frac{F}{c}}{1-q_l} \right) = \frac{\mu_h - \mu_l}{q_h - q_l}.$$

We notice that, as the number of members  $N \rightarrow \infty$ , the cover-II requirement for  $F$ , now becomes  $F > 0$ . Hence, if  $\frac{\mu_h - \mu_l}{q_h - q_l} \leq c$ , then we obtain that

$$\hat{F} = q_l c + (1 - q_l) \frac{\mu_h - \mu_l}{q_h - q_l}.$$

Hence, by (18) we know that, as  $N \rightarrow \infty$ ,

$$\frac{x^e(N)}{N} = \frac{F^e(N)}{c} \rightarrow \begin{cases} \frac{\hat{F}}{c}, & \text{if } \mu_l - \mu_h - c(q_l - q_h) > r_f(q_l c + (1 - q_l) \frac{\mu_h - \mu_l}{q_h - q_l}), \\ 0, & \text{else.} \end{cases} \quad (\text{A25})$$

This proves Proposition 4.

In the continuous case, recall that the optimal risk preference  $a^e(F)$  and  $F$  satisfies the first order condition equation:

$$a = \frac{R - 1 - \phi(a; F)}{2(R - r)}.$$

As  $N \rightarrow \infty$ , we notice that the above equation converges to

$$a = \begin{cases} \frac{R-1}{2(R-r)}, & \text{if } \frac{F}{c} \leq a, \\ a^* + \frac{c-F}{2(R-r)(1-a)}, & \text{if } \frac{F}{c} > a. \end{cases}$$

In other words,

$$a^e(F) = \begin{cases} \frac{R-1}{2(R-r)}, & \text{if } \frac{F}{c} \leq \frac{R-1}{2(R-r)}, \\ \frac{(1+a^*) - \sqrt{(1+a^*)^2 - 2\frac{R-1-F}{R-r}}}{2}, & \text{if } \frac{F}{c} > \frac{R-1}{2(R-r)}, \end{cases} \quad (\text{A26})$$

which is a continuous function that is strictly decreasing over  $(\frac{R-1}{2(R-r)}c, c)$ , with limits  $a^e(c-) = a^*$ . From (29) we know that, as  $N \rightarrow \infty$ , we have

$$\frac{x^e(N)}{N} = \frac{F^e(N)}{c} \rightarrow \frac{1}{c} \arg \max_F \left( -(R-r)(a^e(F))^2 + (R-1-c)a^e(F) + 1 - r_f F \right). \quad (\text{A27})$$

To find the maximizer for the right hand side of (A27), we first notice that

$$\arg \max_{0 < F \leq \frac{R-1}{2(R-r)}c} \left( -(R-r)(a^e(F))^2 + (R-1-c)a^e(F) + 1 - r_f F \right) = 0,$$

with maximum

$$-(R-r) \left( \frac{R-1}{2(R-r)}c \right)^2 + (R-1-c) \frac{R-1}{2(R-r)}c + 1 = -\frac{(R-1)^2 c^2}{4(R-r)} + a^*(R-1)c + 1. \quad (\text{A28})$$

On the other hand, the range of the objective function in (A27) for  $F \in (\frac{R-1}{2(R-r)}c, c)$  is the same as that of

$$G(a) := -(R-r)a^2 + (R-1-c)a + 1 - r_f F(a), \quad a \in (a^*, \frac{R-1}{2(R-r)}),$$

where  $F(a)$  is the inverse of  $a^e(F)$  in (A26) for  $F \in (\frac{R-1}{2(R-r)}c, c)$ , given by

$$F(a) = R-1 + \frac{[(1+a^*-2a)^2 - (1+a^*)^2](R-r)}{2}. \quad (\text{A29})$$

From

$$G'(a) = -2(R-r)[(1+2r_f)a - ((1+r_f)a^* + r_f)], \quad (\text{A30})$$



we know that  $G'(a^*) > 0$ . So

$$\sup_{a \in (a^*, \frac{R-1}{2(R-r)})} G(a) = \begin{cases} G\left(\frac{(1+r_f)a^* + r_f}{1+2r_f}\right), & \text{if } \frac{(1+r_f)a^* + r_f}{1+2r_f} < \frac{R-1}{2(R-r)}, \\ G\left(\frac{R-1}{2(R-r)}\right), & \text{else.} \end{cases}$$

Thus, we know that

$$\begin{aligned} & \arg \max_F (-(R-r)(a^e(F))^2 + (R-1-c)a^e(F) + 1 - r_f F) \\ &= \begin{cases} F\left(\frac{(1+r_f)a^* + r_f}{1+2r_f}\right), & \text{if } G\left(\frac{(1+r_f)a^* + r_f}{1+2r_f}\right) > -\frac{(R-1)^2 c^2}{4(R-r)} + a^*(R-1)c + 1 \text{ and} \\ & \frac{(1+r_f)a^* + r_f}{1+2r_f} < \frac{R-1}{2(R-r)}, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{A31}$$

This proves Proposition 8. Finally, since  $a^* < \frac{R-1}{2(R-r)}$  and  $a^*$  maximizes  $G(a)$  when  $r_f = 0$ , we know that we are in the first case of (A31) if  $r_f > 0$  is sufficiently small.

# Trading in Crowded Markets

STEPAN GORBAN, ANNA A. OBIZHAEVA, AND YAJUN WANG\*

First Draft: November, 2017

This Draft: March 2, 2018

*We study crowded markets using a symmetric continuous-time model with strategic informed traders. We model crowdedness by assuming that traders may have incorrect beliefs about the number of smart traders in the market and the correlation among private signals, which distort their inference, trading strategies, and market prices. If traders underestimate the crowdedness, then markets are more liquid, both permanent and temporary market depths tend to be higher, traders take larger positions and trade more on short-run profit opportunities. In contrast, if traders overestimate the crowdedness, then traders believe markets to be less liquid, they are more cautious in both trading on their information and supplying liquidity to others; fears of crowded markets may also lead to “illusion of liquidity” so that the actual endogenous market depth is even lower than what traders believe it to be. Crowdedness makes markets fragile, because flash crashes, triggered whenever some traders liquidate large positions at fire-sale rates, tend to be more pronounced.*

*JEL: B41, D8, G02, G12, G14*

*Keywords: Asset Pricing, Market Liquidity, Market Microstructure, Crowding, Price Impact, Strategic Trading, Transaction Costs*

\* Gorban: New Economic School, 100A Novaya Street, Skolkovo, Moscow, 143026, Russia,

With a dramatic growth in the asset management industry, financial markets have become platforms where the sophisticated institutional players trade intensively with each other, while retail investors are of much less importance. Even though traders usually seek for overlooked opportunities and try to add diversity into their portfolios, many trading strategies often become crowded. Traders are concerned about crowded markets, because these markets tend to be more fragile and prone to crashes.

Traders come up with investment ideas for generating alphas, evaluate transaction costs of implementing these strategies in real markets, and try to assess, often informally, to what extent strategies might be crowded, i.e., how many other traders might be simultaneously entering the same strategy space and to what extent their private signals might be correlated. While there has been lots of academic research in finance on anomalies in asset returns and liquidity, until recently the question about crowding has received little formal attention. In the 2009-presidential address, Jeremy Stein emphasizes this point and notes that “for a broad class of quantitative trading strategies, an important consideration for each individual arbitrageur is that he cannot know in real time exactly how many others are using the same model and taking the same position as him.” Recognizing the importance of this issue, some firms started to provide tools for identifying and measuring crowdedness of trades and strategies, for example, such as the “crowding scorecard” offered by the MSCI. In this paper, we fill the gap and study theoretically the crowded-market problem, analyzing how thinking about crowdedness interacts with other aspects of trading, such as private information and liquidity.

We consider a stationary continuous-time model of trading among oligopolistic traders. Traders observe flows of private information about asset’s fundamental value and trade on their disagreement about the precision of private information. Traders are of two types. “Smart” traders observe private information with high precision and other traders observe private information with low precision; yet, each trader believes that he observes private information with high precision. All traders trade strategically. They take into account how their trades affect prices and smooth out the execution of their bets over time. This modelling structure is borrowed from the smooth trading model of Kyle, Obizhaeva and Wang (2017) due to its convenience and tractability.

We model crowding by assuming that traders make informed guesses about how many of

their peers might be investing in the same trading strategies, how correlated their private signals might be, and how many of them are smart traders. The perceived subjective characteristics about the number of traders, the correlation among private signals, and the number of smart traders can differ from true characteristics defining the market. Actual characteristics are hard to observe, and trader may either underestimate or overestimate these parameters. Our approach differs from the approach in Callahan (2004) and Stein (2009), who propose to model crowding as the uncertainty about the number of traders, but assume that market participants have unbiased estimates about model parameters.

Each trader trades toward a target inventory, which is proportional to the difference between his own valuation and the average valuation of other market participants, inferred from prices and dividends. The price-based mechanism works properly in our model, as traders do learn from history of prices and condition their strategies on their estimates of fundamental values. Trading strategies are not required to be “unanchored,” this is in sharp contrast with Stein (2009). In the equilibrium, since traders optimally choose their consumption path together with trading strategies using their subjective beliefs, strategies depend only on traders’ subjective parameters, not the actual model parameters. Yet, the equilibrium price also reflects the true number of traders, since it is obtained through the actual market-clearing mechanism, which aggregates demand functions of all traders.

Can traders learn about their mistakes by observing price dynamics? For the case when traders might mis-estimate the total number of traders, traders can learn the average of other traders’ signals from prices, but it is impossible for them to figure out the average of exactly how many signals get into the pricing formula. We show that under the *consistency condition* that imposes a restriction on the relationship between traders’ beliefs about the number of peers and the correlation among private signals, traders cannot learn about their possibly wrong beliefs from observable prices and price volatility. The consistency condition requires that traders either simultaneously underestimate or overestimate both the number of traders and correlation among private signals, though the adjustment in correlation estimates satisfying the consistency condition tends to be very small. The main impact on market liquidity and trading strategies is coming from mis-estimation of the number of traders. Thus, we view this consistency condition as a reasonable one for real-world markets. The intuition is simple. For example, if traders simultaneously overestimate the number of peers and correlation among private signals, i.e., they overestimate the crowdedness of the market, then traders would expect a relatively lower volatility due to a larger number of

peers and a relatively higher volatility due to a higher correlation among private signals. When both effects perfectly balance each other, traders can not learn from price dynamics about their mistakes. Similar arguments apply for the case when traders underestimate the crowdedness.

In fact, traders can learn about their mistakes only by experimenting and deviating from equilibrium strategies or from one-time off-equilibrium events, which may allow traders to learn about actual slope of residual demand function. In practice, this type of experiment can be expensive to implement. Even if traders could learn about the actual total number of traders by obtaining some data on residual demand schedules, they still cannot know in real time exactly how many smart traders are trading in the same direction. We study market properties in this situation as well.

In our model, there is a temporary market depth and a permanent market depth that depend on the execution speed and the size of executed orders, respectively. Since traders build their calculations based on *subjective* market-clearing condition, the perceived market depth may differ from actual market depth in the market. Perceived market depth differs from actual market depth by a factor approximately equal to the ratio of the perceived number of traders to the actual number of traders.

Fear of a crowded market may lead to *illusion of liquidity*. We show that when traders overestimate how crowded the market is, they overestimate both temporary and permanent market depth in comparison with actual market depth. However, fear of crowded markets tends to decrease both perceived and actual market depth. Traders trade less intensively, take smaller positions, and are less willing to supply liquidity to other traders. In contrast, when traders underestimate how crowded the market is, they trade more aggressively, take larger positions, and readily supply liquidity to others.

Crowded markets dominated by institutional investors are often blamed for increased fragility and instability of financial markets, see for example Basak and Pavlova (2013). Market crashes often occur when some market participants are liquidating substantial positions at a fast pace ( e.g., Kyle and Obizhaeva (2016)). We model one-time off-equilibrium execution of large orders and study how the market reaction changes depending on traders' beliefs about market crowdedness. The more traders overestimate the number of their peers, the less they are willing to provide liquidity to others, and the more pronounced are flash-crash patterns.

The crowded-trade hypothesis is often mentioned in discussions about some important

finance episodes. During the market-neutral “quant meltdown” in August of 2007, some of the most successful hedge funds suddenly experienced massive losses, even though the overall market itself did not move much. Khandani and Lo (2010) and Pedersen (2009) discuss a popular hypothesis that attributes this event to unprecedentedly large number of hedge funds investing in similar quantitative strategies. The anecdotal evidence shows that crowding in strategies may play roles during the unwinding of carry trades as well as during the momentum crashes. Stein (2009) also illustrates the effect of crowding using a case study about announced changes in the construction of MSCI indices in 2001-2002 that created a profit opportunity for arbitrageurs. In anticipation of trading by index fund managers in response to changes in index weights, arbitrageurs could in theory buy stocks whose weights were known to increase and sell stocks whose weights were known to decrease. This strategy though did not result in predicted profits in practice, perhaps because too many arbitrageurs rushed into this opportunity at the beginning and this led to price overshooting followed by correction.

Our paper contributes to the existing literature on crowded markets. Stein (2009) proposes a one-period model, in which some traders underreact to their private signals, and uncertain number of arbitrageurs chase to profit on this opportunity. To keep things simple, he makes a number of simplifying assumptions by hard-wiring existence of anomalies, restricting strategies, and considering limiting cases. Arbitrageurs do not condition their strategies on their own estimates of fundamental values and their demand functions may be a non-decreasing functions of asset prices. In contrast, in our model, except for overconfidence, traders apply Bayes Law consistently, optimize correctly, and dynamically update their estimates of both alphas and target inventories. Stein (2009) suggests that the effect of crowding among arbitrageurs on market efficiency is likely to exhibit complicated patterns. When there is uncertainty about the degree of crowding, in some cases prices might be pushed further away from fundamentals.

Another related paper is Callahan (2004), who analyze the model of Kyle (1985) with added uncertainty about the number of informed traders. Under specific assumptions about signals of informed traders, he obtains a solution for the case when the total number of informed traders is some unknown number less or equal to two. In contrast, we model crowded markets in oligopolistic setting and the number of strategic informed traders can be any number greater than two. Kondor and Zawadowski (2016) study another issue related to crowding. They analyze how learning induced by competition affects capital allocation

and welfare. They find that additional potential entrants do not improve efficiency of capital allocation and decrease social welfare.

Thinking about crowded markets has recently become important in public policy discussions. Regulators are increasingly concerned about whether some strategies and market segments become crowded and whether any of them are currently at risk of unwinding. For example, crowded trades and concentration on a small set of risk factors may create a systemic risk for a central clearing party and financial system, when some traders are forced to liquidate their positions, as discussed in Menkveld (2017). Our model suggests that market turns to be more vulnerable of crashes when traders overestimate the fraction of traders who are trading in the same direction.

It is difficult to identify and track crowded trades. A number of studies propose and test some measures of crowdedness. Pojarliev and Levich (2011) measure the style crowdedness in currency trades as the percentage of funds with significant positive exposure to a given style less the percentage of funds with significant negative exposure to the same style. Polk and Lou (2013) gauge the level of arbitrageurs crowdedness in momentum strategies from high-frequency (daily or weekly) abnormal returns correlations among stocks in the winner and/or loser portfolios. Sokolovski (2016) applies both measures to analyze dynamics in returns of carry trades. Hong et al. (2013) suggest using days-to-cover metrics, defined as the ratio of a stock's short interest to trading volume, which is expected to be a proxy for the cost of exiting crowded trade. Yan (2013) measures the crowdedness by combining the short interest ratio and the exit rate of institutional investors, defined as the number of shares liquidated; he shows that momentum losses can often be avoided by shorting only non-crowded losers. Usually researchers find empirically that these measures provide useful information about following up performance of strategies. Strategies may work well as long as they are not crowded, and they tend to crash or revert when crowdedness increases.

This paper is structured as follows. Section 1 describes a continuous-time model of crowded markets. Section 2 presents some comparative statics and studies the implications of crowding. Section 3 examines how crowdedness may affect the magnitude of flash crashes and implementation shortfalls. Section 4 concludes. All proofs are in the Appendix.

## 1. A Model of Crowded Market

We consider a dynamic model of trading among  $N$  oligopolistic traders. There is a risky security with zero net supply, which pays out dividends at continuous rate  $D(t)$ . The

dividend  $D(t)$  is publicly observable and follows a stochastic process with mean-reverting stochastic growth rate  $G^*(t)$ . The dividend has a constant instantaneous volatility  $\sigma_D > 0$  and constant rate of mean reversion  $\alpha_D > 0$ ,

$$(1) \quad dD(t) := -\alpha_D D(t) dt + G^*(t) dt + \sigma_D dB_D(t),$$

where  $G^*(t)$  is unobservable growth rate. The growth rate  $G^*(t)$  follows an AR-1 process with mean reversion  $\alpha_G > 0$  and volatility  $\sigma_G > 0$ ,

$$(2) \quad dG^*(t) := -\alpha_G G^*(t) dt + \sigma_G dB_G(t).$$

Each trader  $n$  observes a continuous stream of private information  $I_n(t)$  defined by

$$(3) \quad dI_n(t) := \tau_n^{1/2} \frac{G^*(t)}{\sigma_G \Omega^{1/2}} dt + \rho^{1/2} dZ(t) + (1 - \rho)^{1/2} dB_n(t).$$

Since its drift is proportional to  $G^*(t)$ , each increment  $dI_n(t)$  in the process  $I_n(t)$  is a noisy observation of  $G^*(t)$ . The denominator  $\sigma_G \Omega^{1/2}$  scales  $G^*(t)$  so that its conditional variance is one. The parameter  $\Omega$  measures the steady-state error variance of the trader's estimate of  $G^*(t)$  in units of time; it is defined algebraically below (see equation (8)). The precision parameter  $\tau_n$  measures the informativeness of the signal  $dI_n(t)$  as a signal-to-noise ratio describing how fast new information flows into the market. The error terms are correlated, and  $Cov(dI_n, dI_m) = \rho dt$  for  $m \neq n$ , where  $\rho < 1$ .

The stream of dividends contains some information about the growth rate as well. Define  $dI_0(t) := [\alpha_D D(t) dt + dD(t)] / \sigma_D$  and  $dB_0 := dB_D$ . Then,  $dI_0(t)$  can be written

$$(4) \quad dI_0(t) := \tau_0^{1/2} \frac{G^*(t)}{\sigma_G \Omega^{1/2}} dt + dB_0(t), \quad \text{where} \quad \tau_0 := \frac{\Omega \sigma_G^2}{\sigma_D^2},$$

so that public information  $dI_0(t)$  in the divided stream  $D(t)$  has a form similar to the notation for private information. The process  $I_0(t)$  is informationally equivalent to the dividend process  $D(t)$ . The quantity  $\tau_0$  measures the precision of the dividend process. The Brownian motions  $dB_0(t), dZ(t), dB_1(t), \dots, dB_N(t)$  are independently distributed.

To model trading, we assume that all traders agree about the precision of the public signal  $\tau_0$ , but agree to disagree about the precisions of private signals  $\tau_n$ . Each trader  $n$  is certain that his own private information has high precision  $\tau_n = \tau_H$  and  $N - 1$  other



traders can be of two types:  $N_I - 1$  traders have private information with high precision  $\tau_H$  and the other  $N_U := N - N_I$  traders have private information with low precision  $\tau_L$ , where  $\tau_H > \tau_L \geq 0$ .

Denote the fraction of other traders (except trader  $n$  himself) with high precision in the market as

$$(5) \quad \theta := \frac{N_I - 1}{N_U + N_I - 1}.$$

This implies that  $1 - \theta$  fraction of other traders' private information has low precision. Traders do not know each others' type.

To model crowded markets, we make the following two assumptions that capture two different aspects of these markets. First, traders might make incorrect estimates about the total number of traders; we assume that all traders symmetrically think that there are  $N_s := N_{I_s} + N_{U_s}$  participants. Second, traders might have incorrect beliefs about correlations in private signals (3); we assume that traders symmetrically believe that  $Cov(dI_n, dI_m) = \rho_s dt$  for  $m \neq n$ . Trader may also have subjective beliefs  $\theta_s$  about the fraction of informed traders. We use subscripts  $s$  to denote subjective beliefs to differentiate them from the actually correct parameters  $N$ ,  $\rho$ , and  $\theta$ . We assume that traders' beliefs about the number of traders and the correlation of signals are some known constants. There is no uncertainty about the number of traders and correlation. We study how mistakes in traders' views about these parameters affect trading, prices, and liquidity.

We refer to the model with crowding as  $(N_{I_s}, N_{U_s}, \rho_s; N_I, N_U, \rho)$ -model, where  $N_I$ ,  $N_U$ , and  $\rho$  are objective parameters describing the environment, and  $N_{I_s}$ ,  $N_{U_s}$ , and  $\rho_s$  are subjective parameters describing traders' beliefs. We refer to the model without crowding as  $(N_I, N_U, \rho; N_I, N_U, \rho)$ -model, where traders have correct beliefs about the correlation in private signals of market participants and the number of traders. For any  $(N_{I_s}, N_{U_s}, \rho_s; N_I, N_U, \rho)$ -model, equilibrium strategies depend only on the parameters  $N_{I_s}$ ,  $N_{U_s}$ , and  $\rho_s$ , since traders make their decisions based only on subjective beliefs, not the actual parameters. The equilibrium price though is a result of the correct market clearing based on the actual total number of traders in the market  $N = N_I + N_U$ . In spite of the fact that the equilibrium strategies depend only on the parameters  $N_{I_s}$ ,  $N_{U_s}$ , and  $\rho_s$ .

Let  $S_n(t)$  denote the inventory of trader  $n$  at time  $t$ . Each trader  $n$  chooses a consumption intensity  $c_n(t)$  and trading intensity  $x_n(t)$  to maximize an expected constant-absolute-risk-aversion (CARA) utility function  $U(c_n(s)) := -e^{-A c_n(s)}$  with risk aversion parameter

A. Letting  $\beta > 0$  denote a time preference parameter, trader  $n$  solves the maximization problem

$$(6) \quad \max_{\{c_n(t), x_n(t)\}} E_t^n \left\{ \int_{s=t}^{\infty} e^{-\beta(s-t)} U(c_n(s)) ds \right\},$$

where trader  $n$ 's inventories follow the process  $dS_n(t) = x_n(t) dt$  and his money holdings  $M_n(t)$  follow the stochastic process

$$(7) \quad dM_n(t) = (r M_n(t) + S_n(t) D(t) - c_n(t) - P(t) x_n(t)) dt.$$

Each trader trades ‘‘smoothly’’ in the sense that  $S_n(t)$  is a differentiable function of time with trading intensity  $x_n(t) = dS_n(t)/dt$ . Each trader explicitly takes into account how both the level of his inventory  $S_n(t)$  and the derivative of his inventory  $x_n(t)$  affect the price of a risky asset  $P(t)$ .

Each trader dynamically adjusts his estimates and their error variance. We use  $E_t^n\{\dots\}$  to denote the expectation of trader  $n$  calculated with respect to his information at time  $t$ . The superscript  $n$  indicates that the expectation is taken with respect to the beliefs of trader  $n$ . The subscript  $t$  indicates that the expectation is taken with respect to trader  $n$ 's information set at time  $t$ , which consists of both private information as well as public information extracted from the history of dividends and prices.

Let  $G_n(t) := E_t^n\{G^*(t)\}$  denote trader  $n$ 's estimate of the growth rate. Let  $\Omega$  denote the steady state error variance of the estimate of  $G^*(t)$ , scaled in units of the standard deviation of its innovation  $\sigma_G$ . Stratonovich-Kalman-Bucy filtering implies that, for the beliefs of any trader  $n$ , the total precision  $\tau$  and scaled error variance  $\Omega$  are constants that do not vary over time and given by

$$(8) \quad \Omega := Var \left\{ \frac{G^*(t) - G_n(t)}{\sigma_G} \right\} = (2 \alpha_G + \tau)^{-1},$$

$$(9) \quad \tau = \tau_0 + \tau_H + (N_s - 1) \frac{\left( (\theta_s - \rho_s) \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2} \right)^2}{(1 - \rho_s)(1 + (N_s - 1) \rho_s)}.$$

Define signals of trader  $n$  and the average signal of other traders as

$$(10) \quad H_n(t) := \int_{u=-\infty}^t e^{-(\alpha_G + \tau)(t-u)} dI_n(u), \quad n = 0, 1, \dots, N_s,$$

and

$$(11) \quad H_{-n}(t) := \frac{1}{N_s - 1} \sum_{\substack{m=1 \\ m \neq n}}^{N_s} H_m(t).$$

The importance of each bit of information  $dI_n$  about the growth rate decays exponentially at a rate  $\alpha_G + \tau$ , i.e., the sum of the decay rate  $\alpha_G$  of fundamentals and the speed  $\tau$  of learning about fundamentals.

Trader  $n$ 's estimate  $G_n(t)$  can be conveniently written as the weighted sum of three sufficient statistics  $H_0(t)$ ,  $H_n(t)$ , and  $H_{-n}(t)$ , which summarize the information content of dividends, his private information, and other traders' private information, respectively. The filtering formulas imply that trader  $n$ 's expected growth rate  $G_n(t)$  is a linear combination given by

$$(12) \quad G_n(t) := \sigma_G \Omega^{1/2} \left( \tau_0^{1/2} H_0(t) + (1 - \theta_s) \left( \tau_H^{1/2} - \tau_L^{1/2} \right) / (1 - \rho_s) H_n(t) \right. \\ \left. + \frac{(\theta_s - \rho_s) \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} (H_n(t) + (N_s - 1)H_{-n}(t)) \right).$$

This equation has a simple intuition. Each trader places the same weight  $\tau_0^{1/2}$  on the dividend-information signal  $H_0(t)$ , assigns a larger weight to his own signal  $H_n(t)$  and a lower weight to signals of presumably  $N_s - 1$  other traders, aggregated in variable  $H_{-n}(t)$ .

We focus on a symmetric linear equilibrium. To reduce the number of state variables, it is convenient to replace the three state variables  $H_0(t)$ ,  $H_n(t)$ ,  $H_{-n}(t)$  with two composite state variables  $\hat{H}_n(t)$  and  $\hat{H}_{-n}(t)$  defined using a constant  $\hat{a}$  by

$$(13) \quad \hat{H}_n(t) := H_n(t) + \hat{a} H_0(t), \quad \hat{H}_{-n}(t) := H_{-n}(t) + \hat{a} H_0(t),$$

$$(14) \quad \hat{a} := \frac{(1 + (N_s - 1)\rho_s)\tau_0^{1/2}}{(1 + (N_s - 1)\theta_s)\tau_H^{1/2} + (N_s - 1)(1 - \theta_s)\tau_L^{1/2}}.$$

The trader  $n$  conjectures that the symmetric linear demand schedules for other traders  $m$ ,  $m \neq n, m = 1, \dots, N_s$  is given by

$$(15) \quad x_m(t) = \frac{dS_m(t)}{dt} = \gamma_D D(t) + \gamma_H \hat{H}_m(t) - \gamma_S S_m(t) - \gamma_P P(t).$$

Each trader thinks that his flow-demand  $x_n(t) = dS_n(t)/dt$  must satisfy the following market clearing

$$(16) \quad x_n(t) + \sum_{\substack{m=1 \\ m \neq n}}^{N_s} \left( \gamma_D D(t) + \gamma_H \hat{H}_m(t) - \gamma_S S_m(t) - \gamma_P P(t) \right) = 0,$$

which depends on his estimate  $N_s$  about the number of traders in the market. Using zero net supply restriction  $\sum_{m=1}^{N_s} S_m(t) = 0$ , he solves this equation for  $P(t)$  as a function of his own trading speed  $x_n(t)$  to obtain his estimate about the residual supply function,

$$(17) \quad P(x_n(t)) = \frac{\gamma_D}{\gamma_P} D(t) + \frac{\gamma_H}{\gamma_P} \hat{H}_{-n}(t) + \frac{\gamma_S}{(N_s - 1)\gamma_P} S_n(t) + \frac{1}{(N_s - 1)\gamma_P} x_n(t).$$

Then, each trader  $n$  exercises monopoly power in choosing how fast to demand liquidity from other traders to profit from private information. He also exercises monopoly power in choosing how fast to provide liquidity to the other  $N_s - 1$  traders. Trader  $n$  solves for his optimal consumption and trading strategy by plugging the price impact function (17) into his dynamic optimization problem (6). Although strategies are defined in terms of the average of other traders' signals  $H_{-n}(t)$ , each trader believes that equilibrium prices reveal the average private signal, which enables him to implement his equilibrium strategy by conditioning his trading speed on market prices.

### 1.1. Prices and Consistency Condition

The equilibrium price is determined based on the *actual* market clearing condition which sums up demands of the actual number of traders  $N$  in the market,

$$(18) \quad \sum_{m=1}^N x_m(t) = 0, \quad \text{and} \quad \sum_{m=1}^N S_m(t) = 0.$$

Using equations (15) and (17), we obtain the actual equilibrium price

$$(19) \quad P(t) = \frac{\gamma_D(N_s, \theta_s, \rho_s)}{\gamma_P(N_s, \theta_s, \rho_s)} D(t) + \frac{\gamma_H(N_s, \theta_s, \rho_s)}{\gamma_P(N_s, \theta_s, \rho_s)} \hat{a}(N_s, \theta_s, \rho_s) H_0(t) + \frac{\gamma_H(N_s, \theta_s, \rho_s)}{N \gamma_P(N_s, \theta_s, \rho_s)} \sum_{m=1}^N H_m(t).$$

By contrast, each trader uses in his calculations the *subjective* market clearing condition by summing up demands of the perceived number of traders  $N_s$ ,

$$(20) \quad \sum_{m=1}^{N_s} x_m(t) = 0, \quad \text{and} \quad \sum_{m=1}^{N_s} S_m(t) = 0.$$

Each trader believes that the equilibrium price is determined by

$$(21) \quad P_s(t) = \frac{\gamma_D(N_s, \theta_s, \rho_s)}{\gamma_P(N_s, \theta_s, \rho_s)} D(t) + \frac{\gamma_H(N_s, \theta_s, \rho_s)}{\gamma_P(N_s, \theta_s, \rho_s)} \hat{a}(N_s, \theta_s, \rho_s) H_0(t) + \frac{\gamma_H(N_s, \theta_s, \rho_s)}{N_s \gamma_P(N_s, \theta_s, \rho_s)} \sum_{m=1}^{N_s} H_m(t).$$

The only difference between the two pricing equations (19) and (21) are indices  $N$  and  $N_s$  over which the summation of private signals is done. In the equilibrium, all calculations are done from the perspective of traders, so only their subjective parameters enter equilibrium demands and prices. This is, for example, why parameters  $\rho$  and  $\theta$  are not in the formulas. Among all objective parameters, only the objective number of traders  $N$  sneaks into the pricing formula through the actual market clearing mechanism (18).

Each trader observes the market price  $P(t)$  but thinks that it is his conjectured price  $P_s(t)$ . He infers the average signal of all traders in the model, and (potentially incorrectly) interprets it as the average of  $N_s$  signals, rather than  $N$  signals.

In continuous time, it is not difficult to estimate accurately the diffusion variance of the process  $dP(t)$  by looking at its quadratic variation. If traders simply misinterpret information about the averages in the price, then they would be able to learn about their mistakes from the price dynamics. For example, if traders underestimate the total number of participants in the market ( $N_s < N$ ), then traders would expect to observe a relatively high price volatility comparing to what they see in the market, because errors in private signals would not average out.

Since traders cannot know in real time exactly how many other traders are investing in the same strategies, we make sure that incorrect estimates about the number of traders cannot be easily falsified by observing the price dynamics. This requires that the quadratic

variation of actual price dynamics  $dP(t)$  must coincide with the quadratic variation of perceived price dynamics  $dP_s(t)$ . Using equations (19) and (21), we obtain the consistency condition ensuring that the quadratic variation of  $\rho^{1/2}dZ(t) + (1 - \rho)^{1/2}\frac{1}{N}\sum_{m=1}^N dB_m(t)$  must coincide with the quadratic variation of  $\rho_s^{1/2}dZ(t) + (1 - \rho_s)^{1/2}\frac{1}{N_s}\sum_{m=1}^{N_s} dB_m(t)$ .

COROLLARY 1: *Under the consistency condition*

$$(22) \quad \frac{1 + (N - 1)\rho}{N} = \frac{1 + (N_s - 1)\rho_s}{N_s}$$

such that  $Var_n(dP(t)) = Var_n(dP_s(t))$ , we have

$$(23) \quad Cov(dI_n(t), dP(t)) = Cov(dI_n(t), dP_s(t)).$$

The corollary means that, if the consistency condition (22) is satisfied, then for each trader, the correlation coefficient between his private signal and the actual price change is consistent with the subjective correlation between his private signal and price change. This condition ensures that traders can not learn about their mistakes from price dynamics.

The consistency condition imposes the restriction on  $N, N_s, \rho$ , and  $\rho_s$ . If  $N_s < N$ , then the condition implies that  $\rho_s < \rho$ , and vice versa. If traders underestimate the total number of participants in the market ( $N_s < N$ ), they should simultaneously underestimate the correlation among their private signals ( $\rho_s < \rho$ ) in order to bring downward the overestimated volatility of dollar price changes due to the underestimated number of traders.

## 1.2. Liquidity

Equation (17) defines the subjective permanent market depth  $1/\lambda_s$  and temporary market depth  $1/\kappa_s$ , as inverse slopes of residual demand functions with respect to number of shares traded and the rate of trading,

$$(24) \quad 1/\lambda_s := \frac{(N_s - 1)\gamma_p}{\gamma_S}, \quad 1/\kappa_s := (N_s - 1)\gamma_p,$$

where  $\lambda_s$  is the permanent price impact coefficient and  $\kappa_s$  is the temporary price impact coefficient according to traders' views. Traders believe that markets are deeper when the number of traders is higher ( $N_s$  is high) and they tend to be more willing to provide liquidity to others ( $\gamma_p$  is high).

The subjective estimates of market liquidity may differ from the actual permanent market depth  $1/\lambda$  and temporary market depth  $1/\kappa$ , because in reality the price is determined by the actual market clearing condition (16), but with  $N_s$  replaced by  $N$ . Using the market-clearing condition (16) and equation (17), subjective permanent and temporary market depth  $1/\lambda_s$  and  $1/\kappa_s$  are related to the actual ones as

$$(25) \quad 1/\lambda_s = \frac{N_s - 1}{N - 1} 1/\lambda, \quad 1/\kappa_s = \frac{N_s - 1}{N - 1} 1/\kappa.$$

The subjective market depth is  $\frac{N_s-1}{N-1}$  times of the objective one. If traders overestimate the number of traders in the market ( $N_s > N$ ), they also overestimate both permanent and temporary market depth. We refer to this case as “illusion of liquidity.” If traders underestimate the number of traders, they underestimate market depth and we refer to this case as “illusion of illiquidity.” The subjective and objective market depth differ approximately by a factor of  $N_s/N$ . For example, when traders overestimate the number of total traders by 50 percent, the subjective market depth is larger than actual market depth also by about 50 percent, and vice versa.

Traders do not observe actual residual demand schedules in the equilibrium. They might be able to learn about the actual residual demand schedule’s slopes by implementing a series of experiments and analyzing price responses to executions at some off-equilibrium trading rates. In practice, this type of experiments however are either infeasible or very costly to implement. Even if we assume that traders could learn about  $N$  by obtaining some data on residual demand schedules, they still can not learn about the fraction of informed traders  $\theta$ .

### 1.3. Solution

The following theorem characterizes the equilibrium trading strategies and price. Traders calculate target inventories, defined as inventory levels such that trader  $n$  does not trade ( $x_n(t) = 0$ ). Traders update their targets dynamically and trade toward them smoothly, thus optimizing the market impact of trading.

**THEOREM 1:** *There exists a steady-state equilibrium with symmetric linear flow-strategies and positive trading volume if and only if the six polynomial equations (C-38)–(C-43) have a solution satisfying the second-order condition  $\gamma_P > 0$  and the stationarity condition  $\gamma_S > 0$ . Such an equilibrium has the following properties:*

- 1) There is an endogenously determined constant  $C_L > 0$ , defined in equation (C-32), such that trader  $n$ 's optimal flow-strategy  $x_n(t)$  is given by

$$(26) \quad x_n(t) = \frac{dS_n(t)}{dt} = \gamma_S (S_n^{TI}(t) - S_n(t)),$$

where  $S_n^{TI}(t)$  is trader  $n$ 's "target inventory" defined as

$$(27) \quad S_n^{TI}(t) = C_L \left( \hat{H}_n(t) - \hat{H}_{-n}(t) \right).$$

- 2) There is an endogenously determined constant  $C_G > 0$ , defined in equation (C-32), such that the equilibrium price is

$$(28) \quad P(t) = \frac{D(t)}{r + \alpha_D} + C_G \frac{\bar{G}(t)}{(r + \alpha_D)(r + \alpha_G)},$$

where  $\bar{G}(t) := \frac{1}{N} \sum_{n=1}^N G_n(t)$  denotes the average expected growth rate.

Trader  $n$  targets a long position if his own signal  $\hat{H}_n(t)$  is greater than the average signal of other traders  $\hat{H}_{-n}(t)$  and a short position vice versa. The proportionality constant  $C_L$  in equation (27) measures the sensitivity of target inventories to the difference. The parameter  $\gamma_S$  in equation (26) measures the speed of trading as the rate at which inventories adjust toward their target levels. The price in equation (28) immediately reveals the average of all signals. If  $C_G$  were equal to one, the price in equation (28) would equal the average of traders' risk-neutral buy-and-hold valuations, consistent with the Gordon's growth formula. Aggregation of heterogeneous beliefs in a dynamic model, which we refer to as the Keynesian beauty contest effect, makes the multiplier  $C_G$  less than one.

Obtaining an analytical solution for the equilibrium in Theorem 1 requires solving the six polynomial equations (C-38)–(C-43). While these equations have no obvious analytical solution, they can be solved numerically. Extensive numerical calculations lead us to conjecture that the existence condition for the continuous-time model is exactly the same as the existence condition for the similar one-period model presented in Appendix A:

CONJECTURE 1: **Existence Condition.** *A steady-state equilibrium with symmetric, linear flow-strategies exists if and only if*

$$(29) \quad \theta_s < 1 - \frac{N_s(1 - \rho_s)\tau_H^{1/2}}{(N_s - 1)(2 + (N_s - 2)\rho_s)(\tau_H^{1/2} - \tau_L^{1/2})} < 1.$$



Equation (29) implies that, for an equilibrium with positive trading volume to exist, the fraction of other traders whose private information has high precision  $\theta_s$  cannot be too high. The existence condition is reduced to  $\tau_H^{1/2}/\tau_L^{1/2} > 2 + \frac{N_s}{N_s-2}$  if  $\rho_s = 0$  and  $\theta_s = 0$ , as in the setting of Kyle, Obizhaeva and Wang (2017). This condition requires  $N_s \geq 3$  and  $\tau_H^{1/2}$  to be sufficiently more than twice as large as  $\tau_L^{1/2}$ .

## 2. Properties of Crowded Markets

In this section, we study how changes in correlation of private signals and the number of traders whose private information has high precision affect market liquidity and traders' trading strategies.

### 2.1. Effects of Changes in Correlations of Private Signals

To develop intuition, we next consider the  $(N_I, N_U, \rho; N_I, N_U, \rho)$ -model with traders making no mistakes about the level of crowdedness, but the correlation  $\rho$  among private information potentially may take different values. We study how changes in  $\rho$  affect the market.

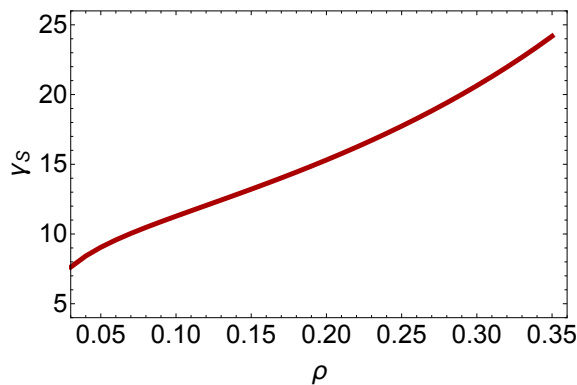


FIGURE 1.  $\gamma_S$  AGAINST  $\rho$ .

Figure 1 illustrates that the speed of trade  $\gamma_S$  increases with the correlation coefficient  $\rho$ .<sup>1</sup> When traders observe private signals with highly correlated errors, they engage in a rat race with each other, as in Foster and Vishwanathan (1996), and trade more aggressively at a higher speed  $\gamma_S$  toward their target inventory levels. Figure 2 shows that as  $\rho$

<sup>1</sup>In Figures 1, 2, 3, 4, and 6 parameter values are  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\theta = 0.1$ ,  $\tau_H = 1$ ,  $\tau_L = 0.2$ .

increases, the total precision of information  $\tau$  decreases, the error variance of the growth rate estimates increases, and the coefficient  $\gamma_P$  increases, i.e., each trader is more willing to provide liquidity to others.<sup>2</sup>

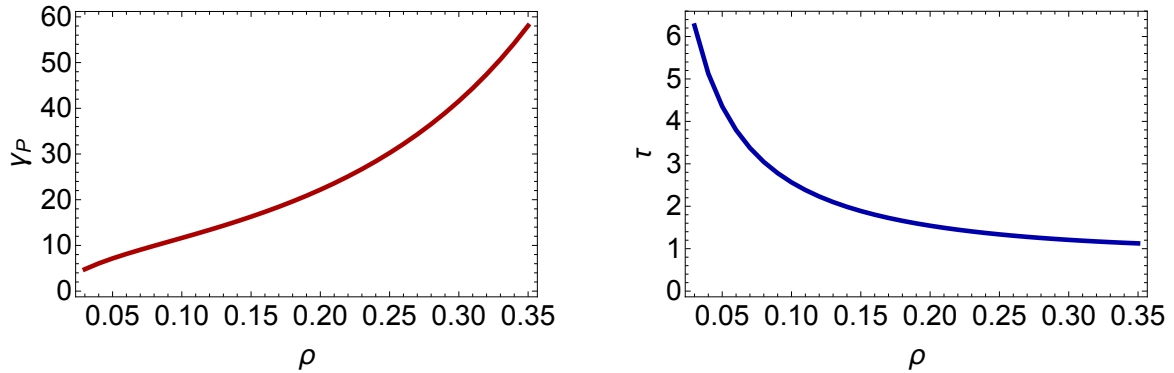


FIGURE 2. VALUES OF  $\gamma_P$  AND  $\tau$  AGAINST  $\rho$ .

Figure 3 shows that both permanent market depth  $1/\lambda$  and temporary market depth  $1/\kappa$  increase, as  $\rho$  increases. The market becomes deeper. The perceived depth coincides with the actual market depth, because traders do not misestimate the number of their peers in the market.

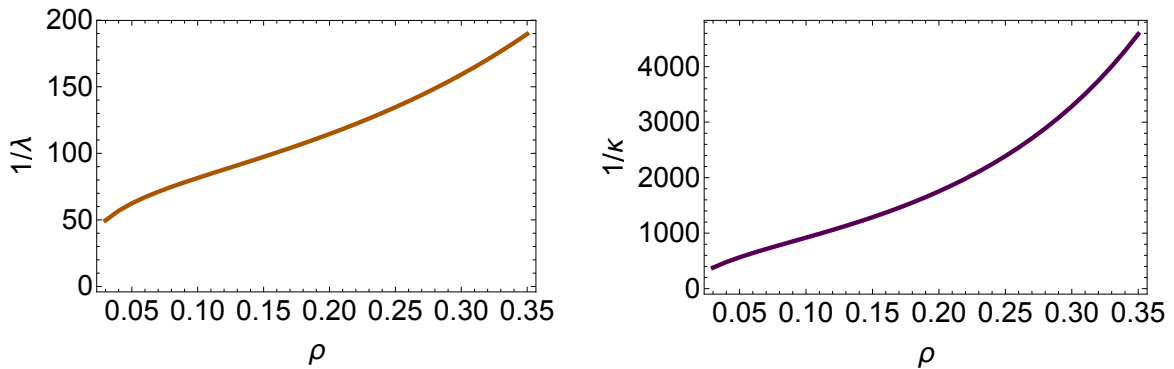


FIGURE 3. VALUES OF  $1/\lambda$  AND  $1/\kappa$  AGAINST  $\rho$ .

Traders believe that trading becomes more valuable and the value of trading on innovations to future information (built into the constant term  $-\psi_0$  defined in trader's value function (C-20)) increases in correlation  $\rho$ , as shown in Figure 4.

<sup>2</sup>Total precision  $\tau$  decreases with  $\rho$  as long as  $\rho$  is not very close to 1.

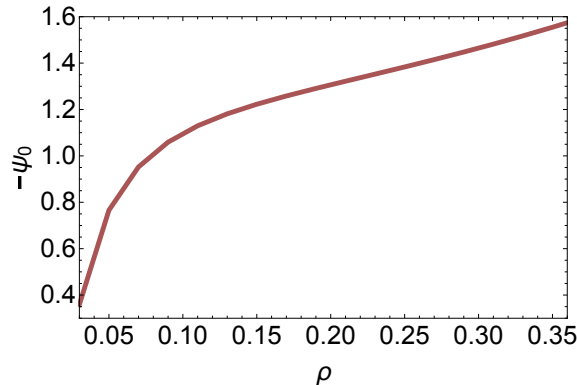


FIGURE 4. VALUE OF TRADING ON INNOVATIONS TO FUTURE INFORMATION  $-\psi_0$  AGAINST  $\rho$ .

Figure 5 presents two simulated paths for target inventories (dashed lines) and actual inventories (solid lines).<sup>3</sup> In panel (a) where correlation  $\rho$  is small, the market is less liquid, traders adjust their inventories at a lower rate to reduce transaction costs, and actual inventories may deviate significantly from target inventories. In panel (b) where correlation  $\rho$  is larger, the market is more liquid, traders adjust their inventories at a faster rate, and actual inventories closely track target inventories.

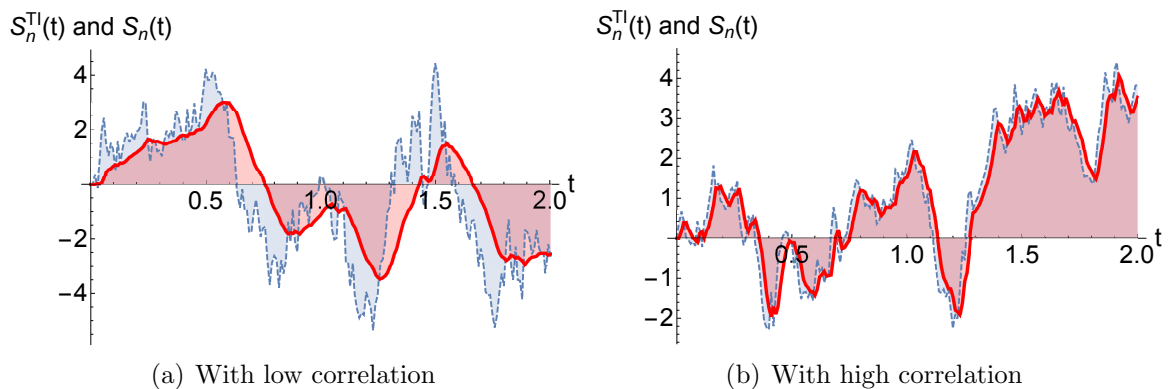
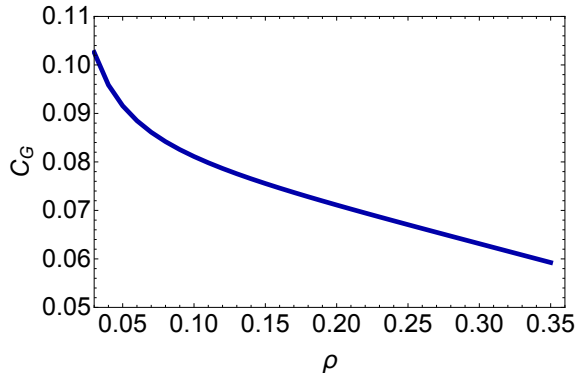


FIGURE 5. SIMULATED PATHS OF  $S_n^{TI}(t)$  (DASHED) AND  $S_n(t)$  (SOLID).

Figure 6 shows that the coefficient  $C_G$  in the equilibrium pricing rule decreases, when the correlation coefficient  $\rho$  increases. Higher correlation among private signals leads to

<sup>3</sup>The paths are generated using equations (27), (C-18), and (C-19), which describe the dynamics of  $\hat{H}_n(t)$ ,  $\hat{H}_{-n}(t)$ , and  $S_n^{TI}(t)$ . Numerical calculations in Figure 5 are based on the exogenous parameter values  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\theta = 0.1$ ,  $\tau_H = 1$ ,  $\tau_L = 0.2$  in both (a) and (b);  $\rho = 0.05$  in (a);  $\rho = 0.5$  in (b).

FIGURE 6.  $C_G$  AGAINST  $\rho$ .

more pronounced price dampening effect ( $C_G < 1$ ). Indeed, in the model each trader believes that other traders make mistakes and they will revise their forecasts in the future. Due to highly correlated signals and a lot of liquidity, traders have greater incentives to engage in short-term speculative trading and take advantage of this predictability in short-term trading patterns of other traders, which could be quite different from expected price dynamics in the long run, because each trader believes that at some point in the future the price will converge to his own estimates of fundamentals.

## 2.2. Effects of Changes in Crowdedness

We next study properties of the  $(N_{I_s}, N_{U_s}, \rho_s; N_I, N_U, \rho)$ -markets where parameters  $N_I$ ,  $N_U$ , and  $\rho$  describe the trading environment, and parameters  $N_{I_s}, N_{U_s}, \rho_s$  describe traders' subjective beliefs. Traders believe that there are  $N_{I_s}$  and  $N_{U_s}$  traders whose private information has high precision (e.g., “smart traders”) and low precision, respectively, and that the correlation among innovations in private signals is equal to  $\rho_s$ . We study how beliefs of traders about the crowdedness of smart traders affect the market and its properties.

We consider two cases. In both cases, we fix the trading environment  $N_I, N_U$ , and  $\rho$ . In the first case, we vary beliefs of traders about the number of smart traders  $N_{I_s}$  in the same market, but fix the number of traders whose private information has low precision  $N_{U_s} = N_U$ . In the second case, we vary  $N_{I_s}$  but fix the total number of traders  $N_s = N_{I_s} + N_{U_s} = N$ .<sup>4</sup>

<sup>4</sup>As we discussed, traders might implement a series of experiments and analyze price response to executions at some off-equilibrium trading rates to estimate the total number of traders in the market  $N$ . However, traders cannot know in real time exactly how many of these traders are smart, i.e., they don't know  $N_{I_s}$ .

Since traders' estimate of the total number of traders  $N_s$  may differ from actual parameter  $N$  in the first case, we consider two subcases. In the base case, we change both  $N_{I_s}$  and  $\rho_s$  in lockstep to satisfy the consistency condition (22) so that traders can not learn about their mistakes by observing price dynamics. The subjective correlation is calculated as  $\rho_s = \frac{1}{N_s-1} \left( \frac{N_s}{N} (1 + (N-1)\rho) - 1 \right)$ . In another subcase, we change only  $N_{I_s}$ , but keep  $\rho_s = \rho$  fixed. These subcases allow us to disentangle the effects of changes in the perceived correlation  $\rho_s$  and the estimate of the number of traders  $N_{I_s}$ . The first subcase is presented by solid lines, and the second subcase is presented in dashed lines in Figures 8, 9, 10, 13, 14, 15 and 16 below.

Figure 7 shows how  $\rho_s$  must change with changes in  $N_{I_s}$  in order to satisfy the consistency condition. When  $N_{I_s}$  is the same as the actual number of traders  $N_I$  ( $N_I = N_{I_s} = 30$ ), the subjective correlation  $\rho_s$  converges to the actual correlation  $\rho$  ( $\rho = \rho_s = 0.20$ ). If  $N_{I_s}$  drops from 30 to 10, the subjective correlation  $\rho_s$  changes only slightly from 0.20 to about 0.195. If  $N_{I_s}$  raises from 30 to 50, the subjective correlation  $\rho_s$  changes from 0.20 to about 0.202.<sup>5</sup>

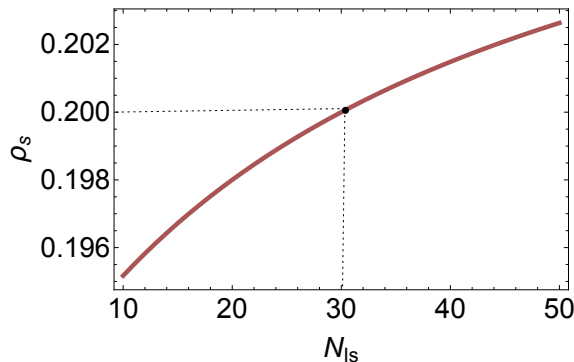


FIGURE 7.  $\rho_s$  AGAINST  $N_{I_s}$ .

It means that the consistency condition requires only small changes in subjective correlations in response to large changes in subjective estimates of the number of traders. Since it is difficult to estimate the correlation among private signals in practice, this consistency condition is practically realistic, because potentially incorrect beliefs of traders cannot be easily falsified by observing the price dynamics. Also, in both subcases, all variables exhibit very similar patterns, so we do not discuss these cases separately.

<sup>5</sup>Parameter values are  $r = 0.01$ ,  $\beta = 0.05$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $N_U = 40$ ,  $N_I = 30$ ,  $\rho = 0.2$ ,  $\tau_H = 1$ ,  $\tau_L = 0$ .

Figure 8 plots the speed of trading  $\gamma_S$  against  $N_{I_s}$  for fixed number of traders with low precision  $N_{U_s} = N_U$  (left panel)<sup>6</sup> and fixed total number of traders  $N_s = N$  (right panel).<sup>7</sup> When traders overestimate the number of smart traders in the market, they tend to trade less aggressively. If  $N_s = N$  is fixed in panel (b), then traders also underestimate the number of traders with low-precision signals, which makes them to trade less aggressively as well. If  $N_{U_s} = N_U$  is fixed, then there are two effects. More smart traders imply that competition among traders becomes more fierce and information decays at a faster rate which also increases traders' trading speed. However, more traders with high precision also imply that adverse price impact increases, this tends to slow down traders' trading speed. These two opposite effects explain why  $\gamma_S$  may first decrease and then increase slightly when  $N_{I_s}$  is getting larger with fixed  $N_U$ , since first the adverse price impact effect dominates and then the competition effect dominates. In panel (a) of Figure 8 the difference between dashed and solid lines is hardly noticeable, this suggests that the decrease in trading speed mainly comes from overestimating the number of smart traders, not from misestimating the correlation.

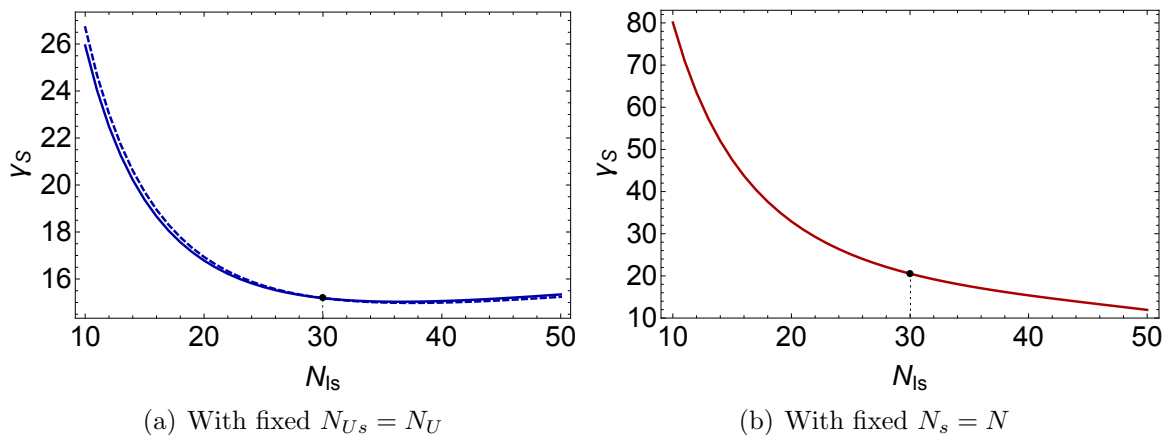


FIGURE 8. VALUES OF  $\gamma_S$  AGAINST  $N_{I_s}$ .

Figure 9 shows that what traders think about crowdedness also affects how large positions they are willing to take. Traders target smaller positions when they overestimate the crowdedness of the smart traders, since the profit opportunities get smaller. The effect is

<sup>6</sup>In Figure 10 and in the left panel of Figures 8, 9, 13, 14, 15 and 16, parameter values are  $r = 0.01$ ,  $\beta = 0.05$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $N_U = 40$ ,  $N_I = 30$ ,  $\rho = 0.2$ ,  $\tau_H = 1$ ,  $\tau_L = 0$ .

<sup>7</sup>In Figure 11 and in the right panel of Figures 8, 9, 13, 14, 15, and 16, parameter values are  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $N = N_s = 70$ ,  $\tau_H = 1$ ,  $\tau_L = 0$ .

slightly more pronounced when the total number of traders is fixed, because the adverse price impact is more significant with an increased fraction of smart traders while fixing the total number of traders in the market.

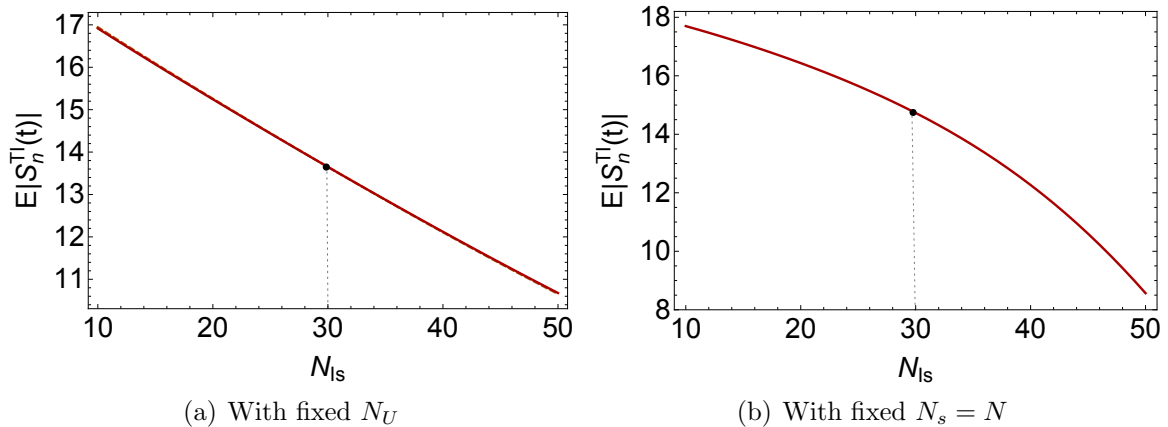


FIGURE 9. VALUES OF  $E|S_n^{TI}(t)|$  AGAINST  $N_{I_s}$ .

Figure 10 plots permanent market depth  $1/\lambda$  and temporary market depth  $1/\kappa$  against  $N_{I_s}$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid curve) satisfying the consistency condition (22). It also plots subjective estimates of market depths  $1/\lambda_s$  and  $1/\kappa_s$ . As before, the figure suggests that the change in market depth comes mainly from misestimation of the number of traders, not correlation among private signals.

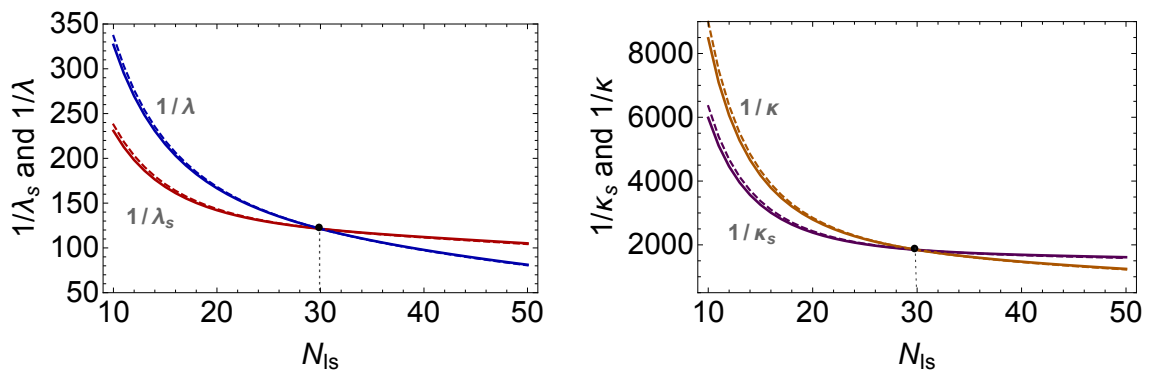


FIGURE 10. VALUES OF  $1/\lambda$ ,  $1/\kappa$ ,  $1/\lambda_s$ , AND  $1/\kappa_s$  AGAINST  $N_{I_s}$  FOR THE CASE WITH FIXED  $N_U$ .

Fear of crowding of smart traders reduces market liquidity. Indeed, when traders overestimate the crowdedness of smart traders ( $N_{I_s} > N_I$ ), they also expect that the market depth is somewhat low, because everybody is less willing to provide liquidity to each other.

In reality, the actual market depth is even lower than what traders think,  $1/\lambda < 1/\lambda_s$  and  $1/\kappa < 1/\kappa_s$ . In contrast, when traders underestimate the number of smart traders ( $N_{I_s} < N_I$  and  $\rho_s < \rho$ ), all types of market depth increase, because traders are more aggressive in trading on private information and providing liquidity to others. The actual market depth is even higher than the perceived one ( $1/\lambda > 1/\lambda_s$  and  $1/\kappa > 1/\kappa_s$ ).

Figure 11 plots permanent market depth  $1/\lambda$  and temporary market depth  $1/\kappa$  against  $N_{I_s}$  with fixed  $N$ . For this case, the perceived market depth is the same as the actual market depth since traders correctly estimate the total number of market participants. This figure shows that underestimating the number of smart traders  $N_{I_s}$  with fixed  $N$  tends to increase market liquidity by a larger magnitude comparing to the case with fixed  $N_U$ .

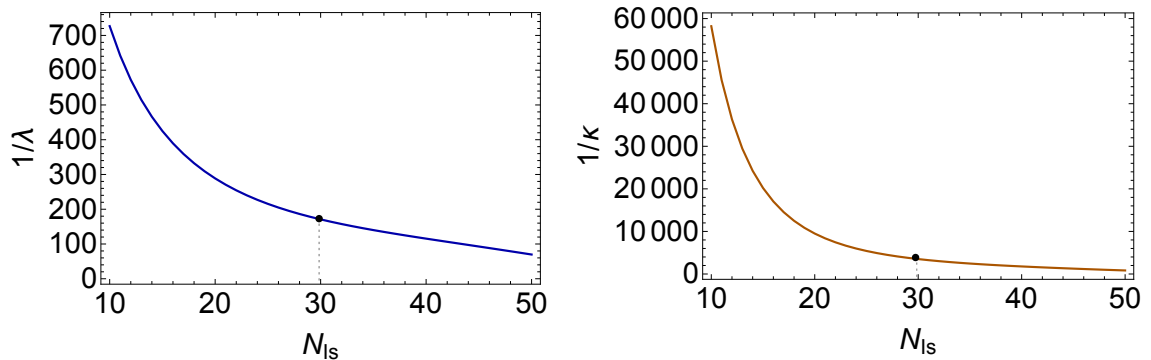


FIGURE 11. VALUES OF  $1/\lambda$ ,  $1/\kappa$ ,  $1/\lambda_s$ , AND  $1/\kappa_s$  AGAINST  $N_{I_s}$  FOR THE CASE WITH FIXED  $N_s = N$ .

Figure 12 presents two simulated paths for target inventories (dashed curve) and actual inventories (solid curve).<sup>8</sup> When traders underestimate the number of smart traders—and the market is more liquid—actual inventories deviate less significantly from target inventories since traders trade at a higher rate, as in panel (a). When traders overestimate the number of smart traders—and the market is less liquid—actual inventories deviate more significantly from target inventories, as in panel (b).

The left panel of Figure 13 plots  $\gamma_P$  against  $N_{I_s}$  for fixed  $N_U$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid curve) satisfying the consistency condition (22). The right panel of Figure 13 plots  $\gamma_P$  against  $N_{I_s}$  for fixed  $N$ . As we can see from Figure 13,  $\gamma_P$  is lower (higher) when

<sup>8</sup>The paths are generated using equations (27), (C-18), and (C-19), which describe the dynamics of  $\hat{H}_n(t)$ ,  $\hat{H}_{-n}(t)$ , and  $S_n^{TI}(t)$ . Numerical calculations in Figure 12 are based on the exogenous parameter values  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\tau_H = 1$ ,  $\tau_L = 0$ ,  $N_I = 30$ ,  $N_U = 40$ ,  $\rho_s = \rho = 0.2$  in both (a) and (b);  $N_{I_s} = 20$  and  $N_{U_s} = 50$  in (a);  $N_{I_s} = 40$  and  $N_{U_s} = 30$  in (b).



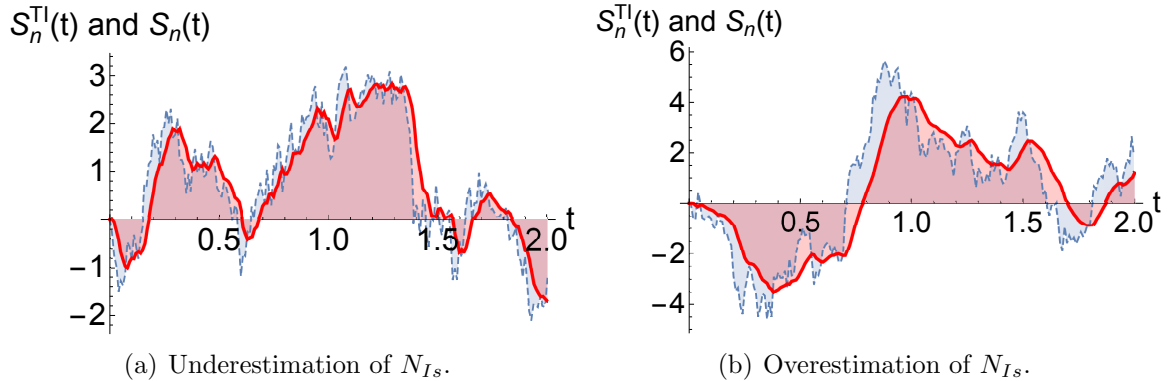


FIGURE 12. SIMULATED PATHS OF  $S_n^{TI}(t)$  (DASHED) AND  $S_n(t)$  (SOLID).

traders overestimate (underestimate) the number of traders whose private information has low precision.

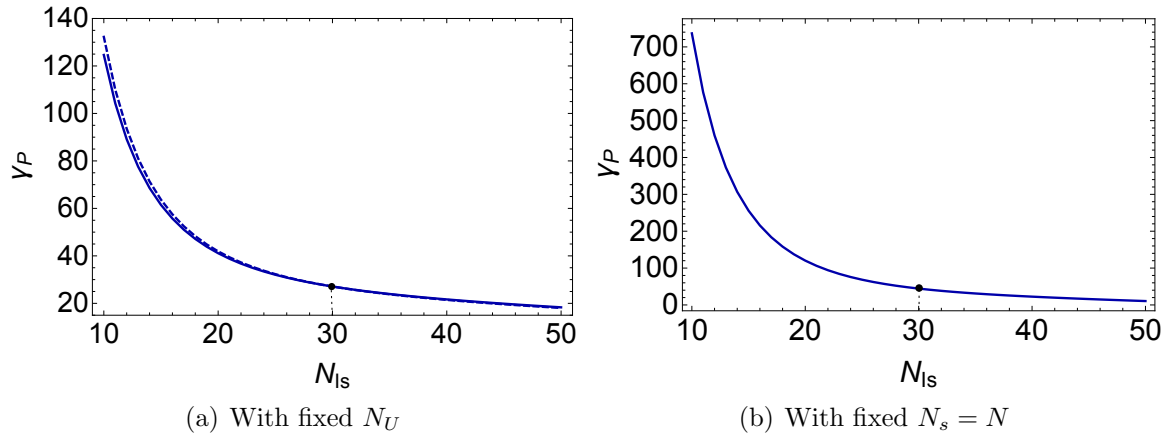
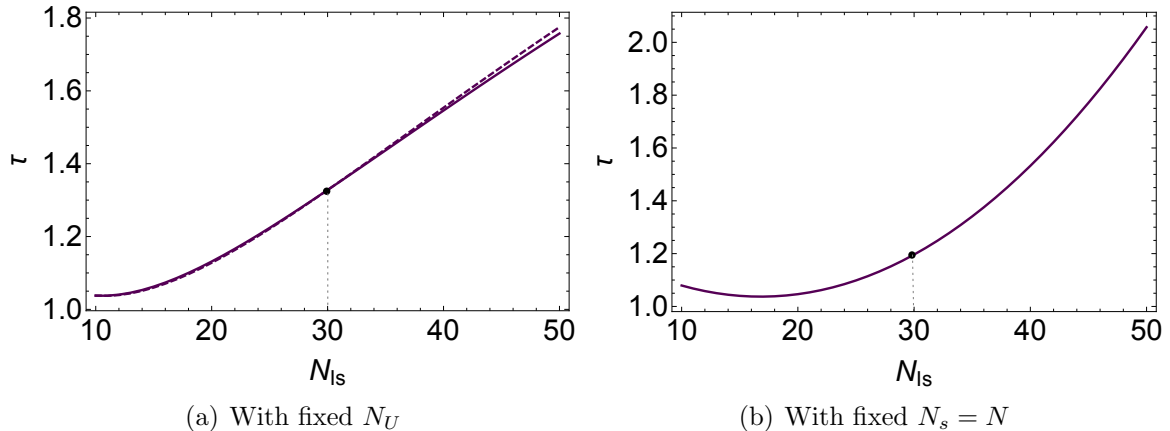
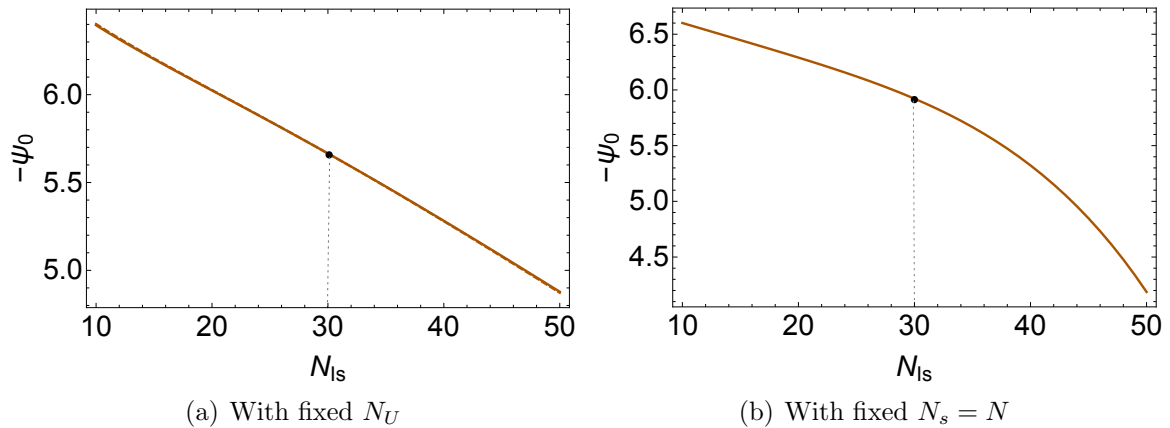


FIGURE 13. VALUES OF  $\gamma_P$  AGAINST  $N_{Is}$ .

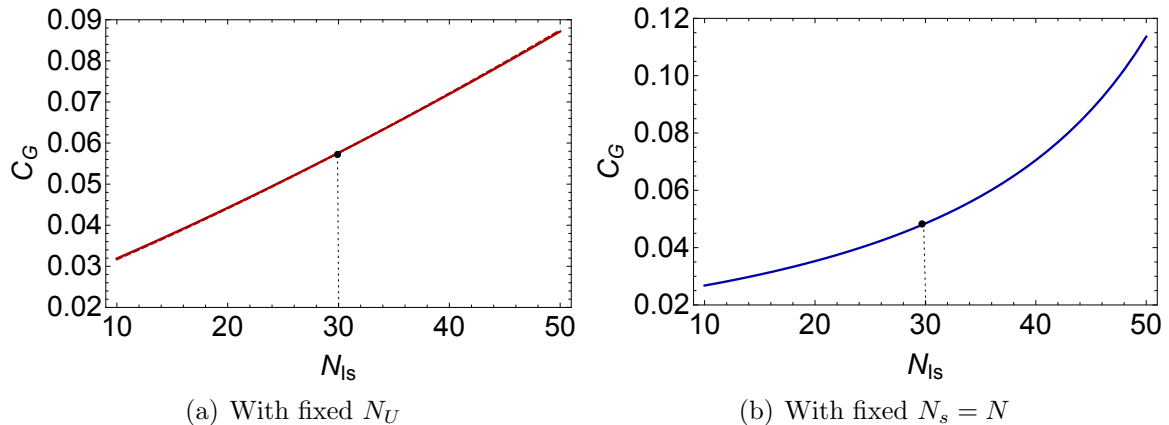
Figure 14 illustrates changes in the total precision  $\tau$ . A higher  $\rho$  tends to decrease the total precision. Overestimating the number of traders tends to increase total precision, as illustrated by the dashed curve in Figure 14. The net effect of overestimating the number of traders and correlation increases the total precision (as shown by the solid curve in Figure 14) and decreases the error variance of the estimate of the growth rate. This makes trading due to agreeing to disagreement less valuable, as shown in Figure 15, the value of trading on innovations to future information ( $-\psi_0$ ) decreases.

Figure 16 presents how  $C_G$  changes with  $N_{Is}$  with fixed  $N_U$  and fixed  $N$ . Figure 16

FIGURE 14. VALUES OF  $\tau$  AGAINST  $N_{I_s}$ .FIGURE 15. VALUES OF  $-\psi_0$  AGAINST  $N_{I_s}$ .

illustrates that  $C_G$  is higher when traders overestimate the number of smart traders. Overestimating the number of smart traders results in less pronounced price dampening (a larger  $C_G$ ), as traders are less willing to engage in short-term speculation due to greater adverse price impacts.

To summarize, when traders overestimate the crowdedness of the smart traders, they tend to have smaller target inventories, trade less aggressively toward target levels, trade less on short-run opportunities, expect less liquidity, and believe that trading is less valuable. When traders underestimate how many other smart traders who are trading in the same direction as them, they tend to have larger target inventories, adjust actual inventories faster toward target levels, trade more on short-run profit opportunities, expect higher liquidity, and are more willing to provide liquidity to others.

FIGURE 16. VALUES OF  $C_G$  AGAINST  $N_{Is}$ .

### 3. Crowding and Fire Sales

It is believed that crowding may make markets more fragile. As we discussed in the previous section, when traders underestimate the crowdedness of smart traders, they tend to take larger positions, trade more on short-run profit opportunities, and are more willing to provide liquidity to others. When traders are concerned that they might have underestimated the crowdedness of the traders who are trading in the same direction. They would liquidate some of their inventories and market becomes less liquid at the same time. This tends to make market more fragile.

Our model allows us to study what would happen in the crowded market if some traders suddenly have to liquidate large positions in a “fire sale” mode. For example, this analysis will help us examine theoretically how the market is expected to respond to events similar to quant meltdown in August 2007. We show that these unexpected off-equilibrium fire sales would create flash crashes. When traders are concerned about the crowding in their trading strategies, they trade less aggressively toward their targets and provide less liquidity to others. We show that this makes flash crashes more substantial.

We present a numerical example of how the market would respond to an off-equilibrium fire sale of a trader. For simplicity, suppose at time 0, a trader observes a private signal  $H_n(0)$  and holds some positive inventory, which is consistent with his target inventory. We also assume that he thinks signals of other traders are at their long-term mean  $H_{-n}(0) = 0$  and dividends  $D(0) = 0$  (with  $H_0(0) = 0$ ). It follows that his inventory at time 0 is

$$(30) \quad S_n(0) = S_n^{TI}(0) = C_L(N_{Is}, \rho_s) H_n(0) > 0.$$

We explicitly state the argument  $(N_{Is}, \rho_s)$  on which the coefficient  $C_L$  depends to emphasize that this coefficient—as well as some other coefficients—depends on subjective beliefs of a trader about the number of traders and correlations  $\rho_s$  of signals. From equations (17) and (24), we get

$$(31) \quad P(0) = \frac{\gamma_S(N_{Is}, \rho_s)}{(N-1)\gamma_P(N_{Is}, \rho_s)} S_n(0) > 0.$$

A trader is not able to learn from the price about mistakes. Equations (19) and (21) therefore imply that the perceived average of private signals must coincide with the actual average.<sup>9</sup> This is a starting consistent-with-equilibrium point of our example.

Next, assume that at time  $t = 0^+$ , all traders receive new private information, so that trader  $n$ 's signal  $H_n(0)$  and other traders' signal  $H_{-n}(0^+)$  suddenly drop to zero, reducing his target inventory from  $S_n^{TI}(0)$  to  $S_n^{TI}(0^+) = 0$ . Since  $H_n(0^+) = H_{-n}(0^+) = 0$ , the new equilibrium price is  $E_0^n[P(t)] = 0$ . Suppose also that a trader has to trade toward his target inventory at a fire-sale speed  $\bar{\gamma}_S$ , which is much faster than the equilibrium rate  $\gamma_S$ ,

$$(32) \quad \bar{x}_n(t) = \bar{\gamma}_S (S_n^{TI}(t) - \bar{S}_n(t))$$

at each point  $t > 0$ . Since  $\bar{\gamma}_S > \gamma_S$ , the trader moves to his target inventory  $S_n^{TI}(t)$  more aggressively. This captures the idea of a sudden rushed sale in the market.

After date  $t = 0$ , off-equilibrium inventory  $\bar{S}_n(t)$  is expected to evolve according to

$$(33) \quad \bar{S}_n(t) = e^{-\bar{\gamma}_S t} \left( S_n(0) + \int_{u=0}^t e^{\bar{\gamma}_S u} \bar{\gamma}_S C_L (H_n(u) - H_{-n}(u)) du \right).$$

A rushed sale leads to execution at a heavy discount. Trader  $n$  can calculate the impulse-response functions of how market prices  $E_0^n[\bar{P}(t)]$  are expected to change in response to his sales, described by  $E_0^n[\bar{S}_n(t)]$ ,

$$(34) \quad E_0^n[\bar{S}_n(t)] = e^{-\bar{\gamma}_S t} S_n(0),$$

<sup>9</sup>For the case with fixed number of “noise traders”  $N_U$ , traders' estimate about the total number of traders is  $N_s$  while the actual number of traders is  $N$ . The actual average of all private signals is  $1/N (H_n(0) + (N-1)\check{H}_{-n}(0))$ , whereas the trader believes that the perceived average of all private signals is equal to  $1/N_s H_n(0)$ , since there are  $N_s$  traders and signals of other traders are zero. Matching these two averages, we get the average of other traders' signals  $\check{H}_{-n}(0)$  such that a trader does not learn from the price about his misestimation of the total number of traders in the market. For the case with fixed  $N$ , traders correctly estimate the total number of traders, then  $\check{H}_{-n}(0) = H_{-n}$ .

$$(35) \quad E_0^n[\bar{P}(t)] = -\frac{\bar{\gamma}_S - \gamma_S(N_{I_s}, \rho_s)}{(N_s - 1)\gamma_P(N_{I_s}, \rho_s)} e^{-\bar{\gamma}_S t} S_n(0).$$

Figure 17 shows expected paths of future prices based on equation (35) for the case without crowding ( $N_{I_s} = N_I$ ) and with crowding ( $N_{I_s} > N_I$ ). Figure 18 shows paths of trader  $n$ 's future inventories based on equation (34) for the case without crowding ( $N_{I_s} = N_I$ ) and with crowding ( $N_{I_s} > N_I$ ).<sup>10</sup> There are two cases in each figure. The first baseline case is shown by the solid red lines: If trader  $n$  liquidates his inventory at an equilibrium rate ( $\bar{\gamma}_S = \gamma_S$ ), then the price immediately drops to the equilibrium level of zero, but the trader continues to trade out of his inventories smoothly over time.

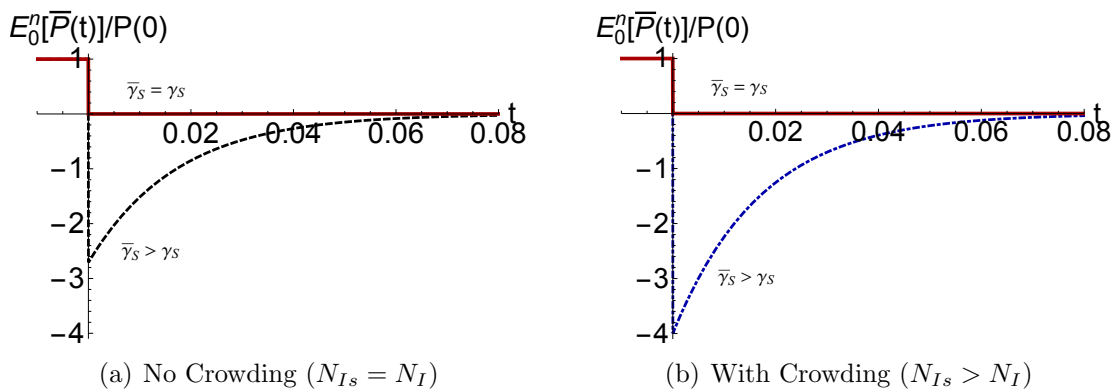


FIGURE 17. THE DYNAMICS OF EXPECTED PRICES WITH AND WITHOUT CROWDING.

The other case show what happens when trader  $n$  liquidates his position at an off-equilibrium fire-sale rate, which is five times faster than normal rate ( $\bar{\gamma}_S = 5 \gamma_S$ ). In panel (a) of Figure 17, black dashed line corresponds to price dynamics for the case with no crowding ( $N_I = N_{I_s}$ ), and blue dashed line in panel (b) of Figure 17 corresponds to price dynamics for the case when traders are concerned about crowding ( $N_I < N_{I_s}$ ).

In panel (a) of Figure 18, black dashed line corresponds to inventory dynamics for the case with no crowding ( $N_I = N_{I_s}$ ), and blue dashed line in panel (b) corresponds to inventory dynamics for the case when traders are concerned about crowding ( $N_I < N_{I_s}$ ).

In both cases with and without crowding, price paths exhibit distinct V-shaped patterns, i.e., after a sharp initial drop the price changes its direction and converges to the new equilibrium level. As explained in Kyle, Obizhaeva and Wang (2017), faster-than-equilibrium trading generates “flash crashes” by increasing temporary price impact.

<sup>10</sup>Parameter values are  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $N = N_s = 70$ ,  $N_I = 30$ ,  $N_{I_s} = 40$ ,  $\rho = 0.2$ ,  $\tau_H = 1$ ,  $\tau_L = 0$ , and  $D(0^+) = 0$ ,  $H_0(0^+) = 0$ . The endogenous parameter

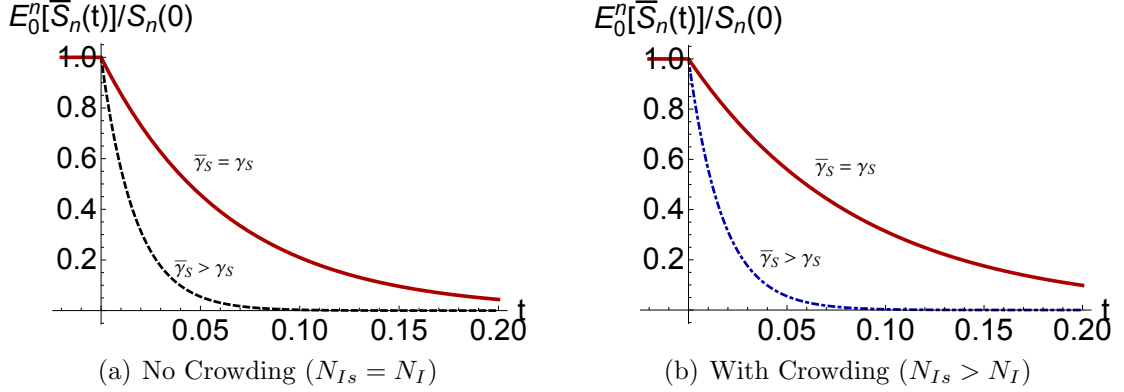


FIGURE 18. THE DYNAMICS OF EXPECTED INVENTORIES WITH AND WITHOUT CROWDING.

When traders are concerned about crowding in their trading strategies, traders are more cautious and slower in trading on their information and providing liquidity to others, therefore flash crashes may be more likely to occur and their price patterns may be more pronounced, as indeed confirmed by more significant price changes in panel (b) of Figure 17 when  $N_{Is} > N_I$ .

When traders overestimate the number of smart traders in the market, both temporary and permanent market depth are smaller, and thus transaction costs are larger. We next present two simple examples to illustrate how overestimating the fraction of smart traders affects execution costs.

Suppose a “new” trader  $n = N + 1$  silently enters the market and liquidates inventories  $\bar{S}_{N+1}(t)$  at a rate  $\bar{x}_{N+1}(t)$ , unbeknownst to the other  $N$  traders. We can explicitly calculate the effect on prices if a trader deviates from his optimal inventory policy  $S_n^*(t)$  and instead holds inventories denoted  $S_n(t)$ . As a result of the deviation, the old equilibrium price path  $P^*(t)$  will be changed to a different price path, denoted  $P(t)$ , given by

$$(36) \quad P(t) = P^*(t) + \lambda (S_n(t) - S_n^*(t)) + \kappa (x_n(t) - x_n^*(t)).$$

Since the new trader does not trade in actual equilibrium, we assume  $S_{N+1}^*(t) = x_{N+1}^*(t) = 0$

We measure his execution costs  $C$  using the concept of implementation shortfall, as described by Perold (1988). The expected price impact costs are given by

$$(37) \quad E\{C\} = E \left\{ \int_{u=t}^{\infty} (P(u) - P^*(u)) \bar{x}(u) du \right\}.$$

values are  $\gamma_S(N_I) = 15.635$ , for  $N_I = 30$  and  $\gamma_S(N_{Is}) = 11.6015$  for  $N_{Is} = 40$ ,  $\bar{\gamma}_S = 5$   $\gamma_S(N_{Is}) = 58$ .

The expected implementation shortfall depends on how the new trader trades. Here are two simple examples.

*Example 1:* Suppose the new trader  $N + 1$  enters the market at date  $t = 0$  and liquidates a random block of shares  $B$ , uncorrelated with signals  $H_n(t)$ ,  $n = 1, \dots, N$ , by trading at the constant rate  $\bar{x}(t) = B/T$  over some interval  $[0, T]$ . Then his expected implementation shortfall is given by

$$(38) \quad E\{C_1\} = \left(\lambda + \frac{\kappa}{T/2}\right) \frac{B^2}{2}.$$

*Example 2:* Suppose instead that the new trader enters the market at date  $t = 0$  and liquidates the random inventory  $B$  by trading at rate  $x_{N+1}(t) = \gamma_S(B - \bar{S}_{N+1}(t))$ . Then his inventory evolves as  $\bar{S}(t) = B(1 - e^{-\gamma_S t})$ , with  $\bar{S}(t) \rightarrow B$  as  $t \rightarrow \infty$ , and the implementation shortfall is given by

$$(39) \quad E\{C_2\} = \left(\lambda + \kappa \gamma_S\right) \frac{B^2}{2}.$$

When traders are concerned about crowdedness of their trading strategies, market becomes less liquid and the implementation shortfall increases for a trader who enters the market and acquires certain shares of the stock. Since faster execution leads to larger temporary price impact, overestimating the fraction of smart traders tends to have bigger impact on the implementation shortfall when a trader needs to acquire or liquidate certain inventory level faster.

## 4. Conclusion

After the Quant Meltdown of August 2007, institutional traders are increasingly concerned about crowded markets, because this factor may impede their efforts to deliver good performance and make them vulnerable to externalities imposed by other market participants.

In this paper, we develop a continuous-time model with strategic informed traders to study the phenomenon of crowded markets. Traders may have incorrect views about the correlation among traders' private signals and the number of traders chasing similar investment strategies.

Even though equilibrium trading strategies depend only on traders' subjective beliefs, the equilibrium prices are determined by the actual market clearing condition, and thus the perceived market depth may differ from the actual market depth available in the market.

Underestimation of the crowdedness of smart traders in the market increases both the perceived and actual market depth. Traders trade more intensively, take larger positions, and are more willing to supply liquidity to other traders. Overestimation of the crowdedness of the market tends to increase both temporary and permanent price impact and thus increase traders' implementation shortfall. Traders trade less aggressively, take smaller positions, and are less willing to supply liquidity to others.

When some traders are forced to liquidate large positions at a suboptimal fire-sale pace, then flash crashes happen. Our paper suggests that flash-crash price patterns may be more pronounced when traders become more concerned about the crowdedness of the market. To reduce the risks, it is important to understand the mechanisms that drive these patterns in crowded markets. Our analysis also implies that it is important that regulators carefully monitor the crowding risk of many investment strategies.



## REFERENCES

- Basak, Suleyman, and Anna Pavlova.** 2013. “Asset Prices and Institutional Investors.” *American Economic Review*, 103(5): 1728–1758.
- Callahan, Tyrone.** 2004. “Speculative Markets With an Unknown Number of Insiders.” *5th Annual Texas Finance Festival*.
- Foster, Douglas F., and S. Vishwanathan.** 1996. “Strategic Trading When Agents Forecast the Forecasts of Others.” *Journal of Finance*, 51(4): 1437–1478.
- Hong, Harrison G., Frank Weikai Li, Sophie X. Ni, Jose A. Scheinkman, and Philip Yan.** 2013. “Days to Cover and Stock Returns.” Working Paper.
- Khandani, Amir, and Andrew Lo.** 2010. “What Happened to the Quants in August 2007? .”
- Kondor, Peter, and Adam Zawadowski.** 2016. “Learning in Crowded Markets .” *2016 Meeting Papers, Society for Economic Dynamics*, 338.
- Kyle, Albert S.** 1985. “Continuous Auctions and Insider Trading.” *Econometrica*, 53(6): 1315–1335.
- Kyle, Albert S., and Anna A. Obizhaeva.** 2016. “Large Bets and Stock Market Crashes.”
- Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang.** 2017. “Smooth Trading with Overconfidence and Market Power.” *Review of Economic Studies*, Posted March 8: <http://www.restud.com/paper/smooth-trading-with-overconfidence-and-market-power/>.
- Menkveld, Albert.** 2017. “Crowded Trades: An Overlooked Systemic Risk for Central Clearing Parties .” *Review of Asset Pricing Studies*, 7: 209–242.
- Pedersen, Lasse H.** 2009. “When Everyone Runs for the Exit.” *The International Journal of Central Banking*, 5: 177–199.
- Perold, André.** 1988. “The Implementation Shortfall: Paper vs. Reality.” *Journal of Portfolio Management*, 14(3): 4–9.
- Pojarliev, Momtchil, and Richard M. Levich.** 2011. “Detecting Crowded Trades in Currency Funds.” *Financial Analysts Journal*, 67(1): 26–39.

- Polk, Christopher, and Dong Lou.** 2013. “Comomentum: Inferring Arbitrage Activity from Return Correlations.” Working Paper.
- Sokolovski, Valeri.** 2016. “Crowds, Crashes, and the Carry Trade.” Working Paper.
- Stein, Jeremy.** 2009. “Presidential Address: Sophisticated Investors and Market Efficiency.” *Journal of Finance*, VOL.LXIV: 1517–1548.
- Yan, Phillip.** 2013. “Crowded Trades, Short Covering, and Momentum Crashes.” Working Paper.

## A. One-period Model

There are two assets. A risk free asset and a risky asset that has random liquidation value  $v \sim N(0, 1/\tau_v)$ . Both assets are in zero net supply. Trader  $n$  is endowed with inventory  $S_n$  with  $\sum_{n=1}^N S_n = 0$ . Traders observe signals about the normalized liquidation value  $\tau_v^{1/2} v \sim N(0, 1)$ . All traders observe a public signal  $i_0 := \tau_0^{1/2} (\tau_v^{1/2} v) + e_0$  with  $e_0 \sim N(0, 1)$ . Each trader  $n$  observes a private signal  $i_n := \tau_n^{1/2} (\tau_v^{1/2} v) + \rho^{1/2} z + (1 - \rho)^{1/2} e_n$  with  $e_n \sim N(0, 1)$ , where  $v, z, e_0, e_1, \dots, e_N$  are independently distributed.

Traders agree about the precision of the public signal  $\tau_0$  and agree to disagree about the precisions of private signals  $\tau_n$ . Each trader is certain that his own private information has a high precision  $\tau_n = \tau_H$  and  $N - 1$  other traders can be of two types:  $N_I - 1$  traders' private information has high precision  $\tau_H$  and the other  $N_U := N - N_I$  traders' private information has low precision  $\tau_L$ , with  $\tau_H > \tau_L \geq 0$ .

Denote the fraction of other traders (except trader  $n$  himself) with high precision in the market as

$$(A-1) \quad \theta := \frac{N_I - 1}{N_U + N_I - 1}.$$

Each trader submits a demand schedule  $X_n(p) := X_n(i_0, i_n, S_n, p)$  to a single-price auction. An auctioneer clears the market at price  $p := p[X_1, \dots, X_N]$ . Trader  $n$ 's terminal wealth is

$$(A-2) \quad W_n := v (S_n + X_n(p)) - p X_n(p).$$

Each trader  $n$  maximizes the same expected exponential utility function of wealth  $E^n[-e^{-A W_n}]$  using his own beliefs to calculate the expectation.

An *equilibrium* is a set of trading strategies  $X_1, \dots, X_N$  such that each trader's strategy maximizes his expected utility, taking as given the trading strategies of other traders. Let  $i_{-n} := \frac{1}{N-1} \sum_{m=1, m \neq n}^N i_m$  denote the average of other traders' signals. When trader  $n$  conjectures that other traders submit symmetric linear demand schedules

$$(A-3) \quad X_m(i_0, i_m, S_m, p) = \alpha i_0 + \beta i_m - \gamma p - \delta S_m, \quad m = 1, \dots, N, \quad m \neq n,$$

he infers from the market-clearing condition

$$(A-4) \quad x_n + \sum_{\substack{m=1 \\ m \neq n}}^N (\alpha i_0 + \beta i_m - \gamma p - \delta S_m) = 0$$

that his residual supply schedule  $P(x_n)$  is a function of his quantity  $x_n$  given by

$$(A-5) \quad P(x_n) = \frac{\alpha}{\gamma} i_0 + \frac{\beta}{\gamma} i_{-n} + \frac{\delta}{(N-1)\gamma} S_n + \frac{1}{(N-1)\gamma} x_n.$$

Let  $E^n[\dots]$  and  $\text{Var}^n[\dots]$  denote trader  $n$ 's expectation and variance operators conditional on all signals  $i_0, i_1, \dots, i_N$ . Define "total precision"  $\tau$  by

$$(A-6) \quad \tau := (\text{Var}^n[v])^{-1} = \tau_v \left( 1 + \tau_0 + \tau_H + (N-1) \frac{\left( (\theta - \rho)\tau_H^{1/2} + (1 - \theta)\tau_L^{1/2} \right)^2}{(1 - \rho)(1 + (N-1)\rho)} \right).$$

The projection theorem for jointly normally distributed random variables implies

$$(A-7) \quad E^n[v] = \frac{\tau_v^{1/2}}{\tau} \left( \tau_0^{1/2} i_0 + \frac{1 - \theta}{1 - \rho} \left( \tau_H^{1/2} - \tau_L^{1/2} \right) i_n + \frac{(\theta - \rho)\tau_H^{1/2} + (1 - \theta)\tau_L^{1/2}}{(1 - \rho)(1 + (N-1)\rho)} (i_n + (N-1)i_{-n}) \right).$$

Conditional on all information, trader  $n$ 's terminal wealth  $W_n$  is a normally distributed random variable with mean and variance given by

$$(A-8) \quad E^n[W_n] = E^n[v] (S_n + x_n) - P(x_n) x_n, \quad \text{Var}^n[W_n] = (S_n + x_n)^2 \text{Var}^n[v].$$

Maximizing this function is equivalent to maximizing  $E^n[W_n] - \frac{1}{2} A \text{Var}^n[W_n]$ . Oligopolistic trader  $n$  exercises market power by taking into account how his quantity  $x_n$  affects the price  $P(x_n)$  on his residual supply schedule (A-5). The following Theorem characterizes the equilibrium in this one-period model.

**THEOREM 2:** *There exists a unique symmetric equilibrium with linear trading strategies and nonzero trade if and only if the second-order condition*

$$(A-9) \quad \theta < 1 - \frac{N(1 - \rho)\tau_H^{1/2}}{(N-1)(2 + (N-2)\rho)(\tau_H^{1/2} - \tau_L^{1/2})}$$

*holds. The equilibrium satisfies the following:*

1. Trader  $n$  trades the quantity  $x_n^*$  given by

$$(A-10) \quad x_n^* = \delta \left( \frac{1}{A} \frac{1-\theta}{1-\rho} \left( 1 - \frac{1}{N} \right) \tau_v^{1/2} (\tau_H^{1/2} - \tau_L^{1/2}) (i_n - i_{-n}) - S_n \right),$$

where the inventory adjustment factor  $\delta$  is

$$(A-11) \quad 0 < \delta = \frac{2 + (N-2)\rho}{1 + (N-1)\rho} - \frac{N(1-\rho)\tau_H^{1/2}}{(N-1)(1-\theta)(1+(N-1)\rho)(\tau_H^{1/2} - \tau_L^{1/2})} < 1.$$

2. The price  $p^*$  is the average of traders' valuations:

$$(A-12) \quad p^* = \frac{1}{N} \sum_{n=1}^N \mathbb{E}^n[v] = \frac{\tau_v^{1/2}}{\tau} \left( \tau_0^{1/2} i_0 + \frac{(1 + (N-1)\theta)\tau_H^{1/2} + (N-1)(1-\theta)\tau_L^{1/2}}{N(1 + (N-1)\rho)} \sum_{n=1}^N i_n \right).$$

3. The parameters  $\alpha > 0$ ,  $\beta > 0$ , and  $\gamma > 0$ , defining the linear trading strategies in equation (A-3), have unique closed-form solutions defined in (C-2).

For an equilibrium with positive trading volume to exist, the fraction of traders whose private information has high precision must satisfy condition (A-9). Each trader trades in the direction of his private signal  $i_n$ , trades against the average of other traders' signals  $i_{-n}$ , and hedges a fraction  $\delta$  of his initial inventory. Equation (A-12) implies that the equilibrium price is a weighted average of traders' valuations with weights summing to one.

Define a trader's "target inventory"  $S_n^{TI}$  as the inventory such that he would not want to trade ( $x_n^* = 0$ ). From equation (A-10), it is equal to

$$(A-13) \quad S_n^{TI} = \frac{1}{A} \frac{1-\theta}{1-\rho} \left( 1 - \frac{1}{N} \right) \tau_v^{1/2} (\tau_H^{1/2} - \tau_L^{1/2}) (i_n - i_{-n}).$$

Then trader  $n$ 's optimal quantity traded can be written

$$(A-14) \quad x_n^* = \delta (S_n^{TI} - S_n).$$

The parameter  $\delta$ , defined in equation (A-11), is the fraction by which traders adjust positions toward target levels. It can be proved analytically that  $\delta$  increases in correlation  $\rho$  and decreases in  $\theta$  while fixing everything else.

From the perspective of trader  $n$ , equation (A-5) implies that price impact can be written as a function of both  $x_n$  and  $S_n$ ,

$$(A-15) \quad P(x_n, S_n) := p_{0,n} + \lambda S_n + \kappa x_n,$$

where  $p_{0,n}$  is a linear combination of random variables  $i_0$  and  $i_{-n}$ , and constants  $\lambda$  and  $\kappa$  are given by

$$(A-16) \quad \lambda := \frac{\delta}{(N-1)\gamma} = \frac{A(1-\rho) \left( (1+(N-1)\theta)\tau_H^{1/2} + (N-1)(1-\theta)\tau_L^{1/2} \right)}{\tau(N-1)(1+(N-1)\rho)(1-\theta)(\tau_H^{1/2} - \tau_L^{1/2})},$$

$$(A-17) \quad \kappa := \frac{\lambda}{\delta} = \frac{1}{(N-1)\gamma}.$$

It can be proved analytically that both  $\lambda$  and  $\kappa$  both decrease in correlation  $\rho$  and increase in the fraction of traders whose private information has high precision while fixing everything the same. Market becomes more liquid if traders' private information are highly correlated and less liquid if the fraction of traders whose private information has high precision increases.

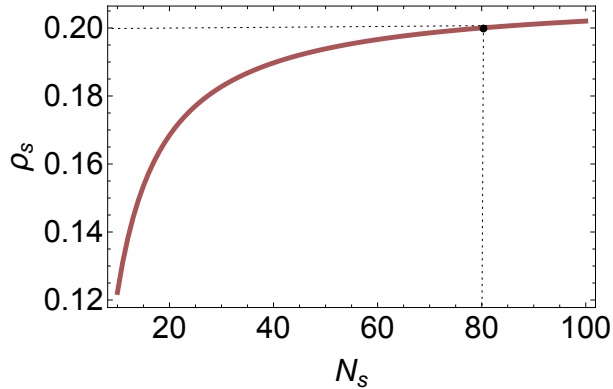
## B. Effects of Changes in Crowding of the Total Market

In this section we focus on the case when traders might misestimate the total number of traders while correctly estimate the fraction of smart traders in the market. We consider two situations: (1) the base case when both  $N_s$  and  $\rho_s$  are changing in lockstep satisfying the consistency condition so that traders can not learn about their mistakes by observing price dynamics, and (2) another case when only  $N_s$  is changing, but  $\rho_s$  remains fixed. The first case is presented by solid lines and the second case is presented in dashes lines below.

Figure B.1 shows that how  $\rho_s$  changes with changes in  $N_s$  in order to satisfy the consistency condition. When  $N_s$  is the same as the actual number of traders  $N$  ( $N = N_s = 80$ ), the subjective correlation  $\rho_s$  converges to the actual correlation  $\rho$  ( $\rho = \rho_s = 0.20$ ). If  $N_s$  drops by a half to 40, the subjective correlation  $\rho_s$  changes only slightly to about 0.19.<sup>11</sup>

Figure B.1 shows that satisfying the consistency condition only requires small changes in subjective correlations in response to large changes in subjective estimates of the number of traders. Since it is difficult to estimate the correlation among private signals in actual financial markets, this consistency condition is a very reasonable one to ensure that traders' potentially incorrect beliefs cannot be easily falsified by observing the price dynamics.

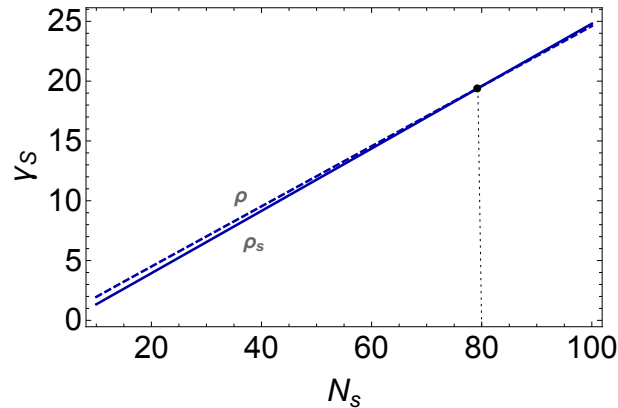
<sup>11</sup>In Figures B.1, B.2, B.3 and B.7, parameter values are  $r = 0.01$ ,  $\beta = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\theta = 0.1$ ,  $N = 80$ ,  $\rho = 0.2$ ,  $\tau_H = 1$ ,  $\tau_L = 0.1$ .

FIGURE B.1.  $\rho_s$  AGAINST  $N_s$ .

We now study how crowding affects market liquidity and traders' trading strategies. Figure B.2 plots the speed of trading  $\gamma_S$  against traders' subjective belief about the number of total market participants  $N_s$ . To separate the impact of the number of traders from the impact of correlation among private signals on the trading speed, we first plot  $\gamma_S$  against changing  $N_s$  while keeping  $\rho_s = \rho$  fixed. The dashed curve illustrates that traders trade less aggressively toward target inventory when there are fewer traders in the market (fixing  $\rho$ ); if traders believe that there are fewer of them in the market and the competition is less intensive, then traders trade less aggressively.

When correlation  $\rho_s$  is adjusted to satisfy the consistency condition, traders trade toward target inventories even slower, as depicted by the solid line in Figure B.2. The subjective correlation is calculated as  $\rho_s = \frac{1}{N_s - 1} \left( \frac{N_s}{N} (1 + (N - 1)\rho) - 1 \right)$  to satisfy the consistency condition (22), it decreases with lower  $N_s$ , and a lower subjective correlation among private signals leads to a slower trading rate  $\gamma_S$ , as shown previously in Figure 1. Figure B.2 also suggests that the decrease in trading speed comes mainly from underestimating the number of traders, not from underestimating of the correlation among private signals. Indeed, the difference between dashed line and solid lines are hardly noticeable. This difference is small for most of other variables, so we will next discuss only our base case when both  $N_s$  and  $\rho_s$  are changing to satisfy the consistency condition.

The left panel of Figure B.3 plots  $\gamma_P$  against  $N_s$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid curve) satisfying the consistency condition (22).  $\gamma_P$  is lower (higher) when traders underestimate (overestimate) the number of market participants. The right panel of figure B.3 illustrates changes in the total precision  $\tau$ . A lower  $\rho$  tends to increase the total precision. Underestimating the number of traders tends to decrease total precision, as

FIGURE B.2. VALUES OF  $\gamma_S$  AGAINST  $N_s$ .

illustrated by the dashed curve in Figure B.3. The net effect of underestimating the number of traders and correlation tends to increase the total precision (as shown by the solid curve in Figure B.3) and decrease the error variance of the estimate of the growth rate. This makes trading due to agreeing to disagreement less valuable, as shown in figure B.4, the value of trading on innovations to future information ( $-\psi_0$ ) decreases.

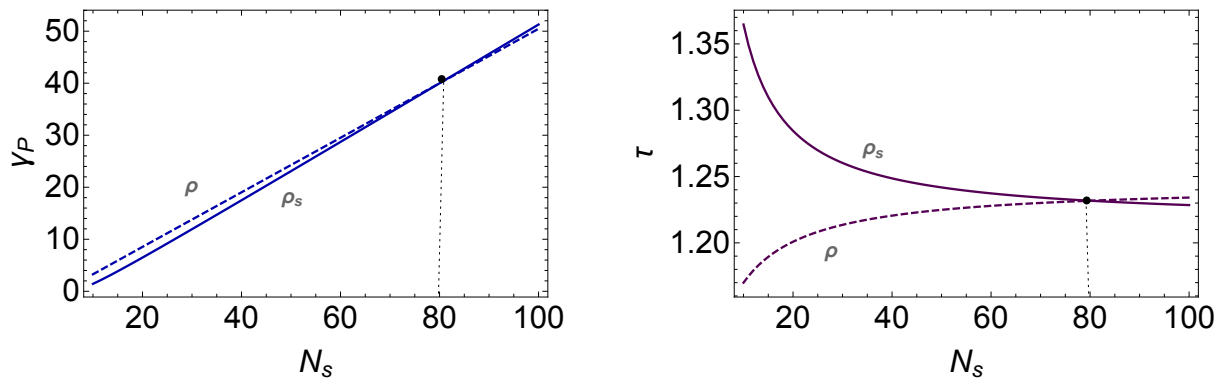
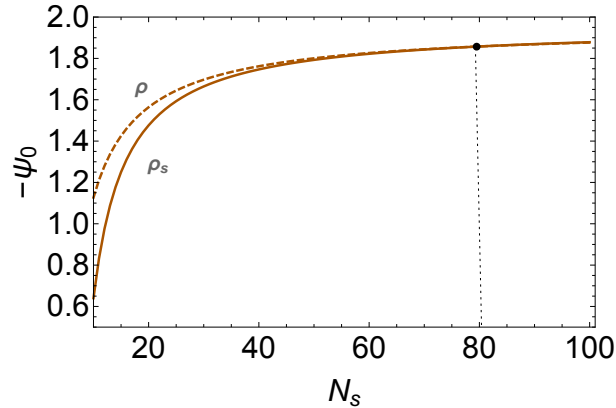
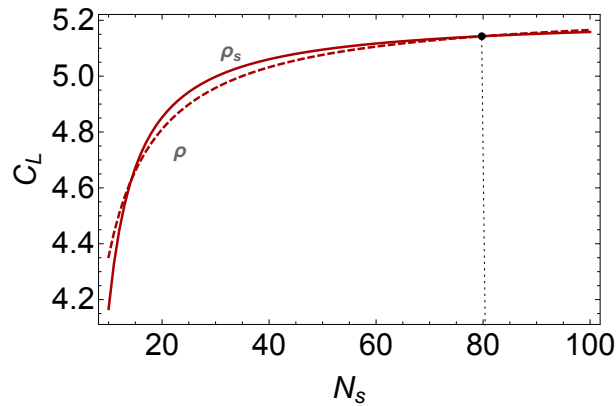
FIGURE B.3. VALUES OF  $\gamma_P$  AND  $\tau$  AGAINST  $N_s$ .

Figure B.5 presents how  $C_L$  changes with  $N_s$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid curve) satisfying the consistency condition (22). It shows that  $C_L$  is lower when traders underestimate the total number of participants and correlation among private signals since traders trade less aggressively with fewer number of traders. This implies traders tend to hold smaller positions when they underestimate the number of traders.

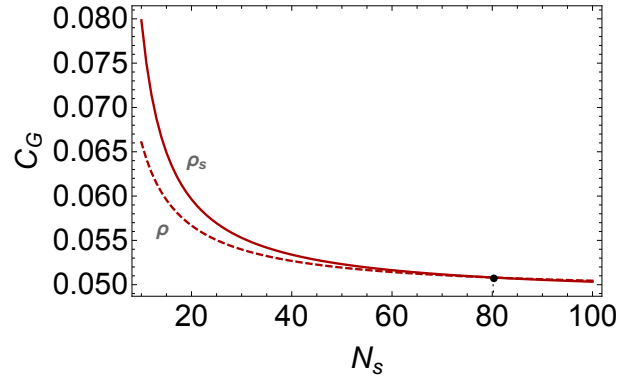
Figure B.6 presents how  $C_G$  changes with  $N_s$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid



FIGURE B.4. VALUES OF  $-\psi_0$  AGAINST  $N_s$ .FIGURE B.5. VALUES OF  $C_L$  AGAINST  $N_s$ .

curve) satisfying the consistency condition (22). Figure B.6 illustrates that  $C_G$  is higher when traders underestimate the total number of participants and correlation among private signals. A lower  $\rho$  leads to a higher  $C_G$  since traders trade less aggressively while fewer number of traders also result in a higher  $C_G$  due to less intensive competition. Therefore, underestimating the number of traders and correlation results in less pronounced price dampening (a larger  $C_G$ ) and traders are less willing to engage in short-term speculation.

Figure B.7 plots permanent market depth  $1/\lambda$  and temporary market depth  $1/\kappa$  against  $N_S$  using  $\rho_s = \rho$  (dashed curve) and  $\rho_s$  (solid curve) satisfying the consistency condition (22). It also plots subjective estimates of market depths  $1/\lambda_s$  and  $1/\kappa_s$ . As before, the figure suggests that the change in market depth comes mainly from misestimation of the number of traders, not correlation among private signals.

FIGURE B.6. VALUES OF  $C_G$  AGAINST  $N_s$ .

When traders overestimate the crowdedness ( $N_s > N$  and  $\rho_s > \rho$ ), traders trade more intensively. Fear of crowding leads to illusion of liquidity in the market, and indeed market depth increases. However, the actual market depth is much lower than the perceived one ( $1/\lambda < 1/\lambda_s$  and  $1/\kappa < 1/\kappa_s$ ). The actual permanent market depth  $1/\lambda$  does not change much comparing to the case without crowding. When traders underestimate the crowdedness ( $N_s < N$  and  $\rho_s < \rho$ ), all types of market depth decrease, because traders trade less aggressively on their signals and are less willing to provide liquidity. Underestimating the crowdedness tends to reduce market liquidity, but the actual market depth is higher than the perceived one ( $1/\lambda > 1/\lambda_s$  and  $1/\kappa > 1/\kappa_s$ ). In this example, the drop in the actual permanent market depth is not as substantial as the drop in the actual temporary market depth. When traders underestimate the crowdedness of the market by a half (e.g.,  $N_s = 40$  and  $N = 80$ ), then  $1/\lambda$  changes only slightly from about 150 to 140, whereas  $1/\kappa$  drops by about a half from approximately 3000 to 1500.

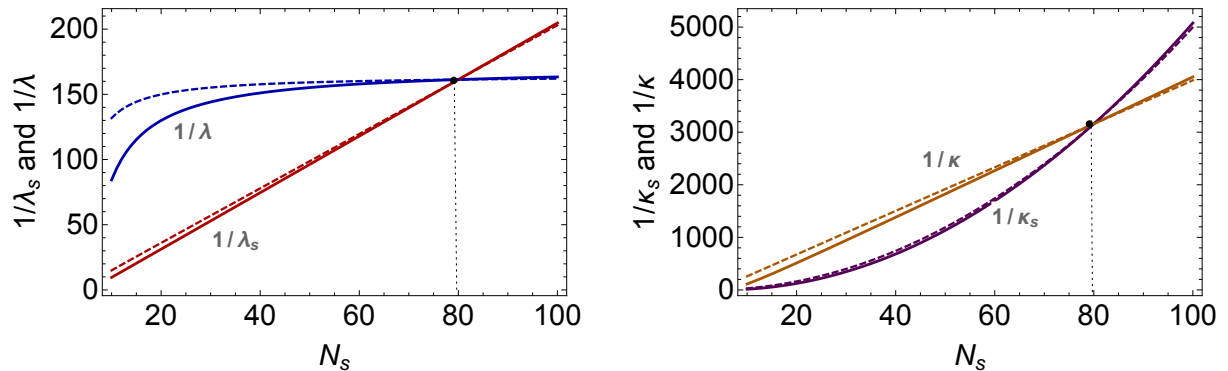
FIGURE B.7. VALUES OF  $1/\lambda$ ,  $1/\kappa$ ,  $1/\lambda_s$ , AND  $1/\kappa_s$  AGAINST  $N_s$ .

Figure B.8 presents two simulated paths for target inventories (dashed curve) and actual inventories (solid curve).<sup>12</sup> When traders underestimate the number of traders and the correlation among private signals—and the market is less liquid—actual inventories deviate more significantly from target inventories since traders trade at a lower rate, as in panel (a). When traders correctly estimate the number of traders and correlation among private signals—and the market is more liquid—actual inventories deviate less significantly from target inventories, as in panel (b).

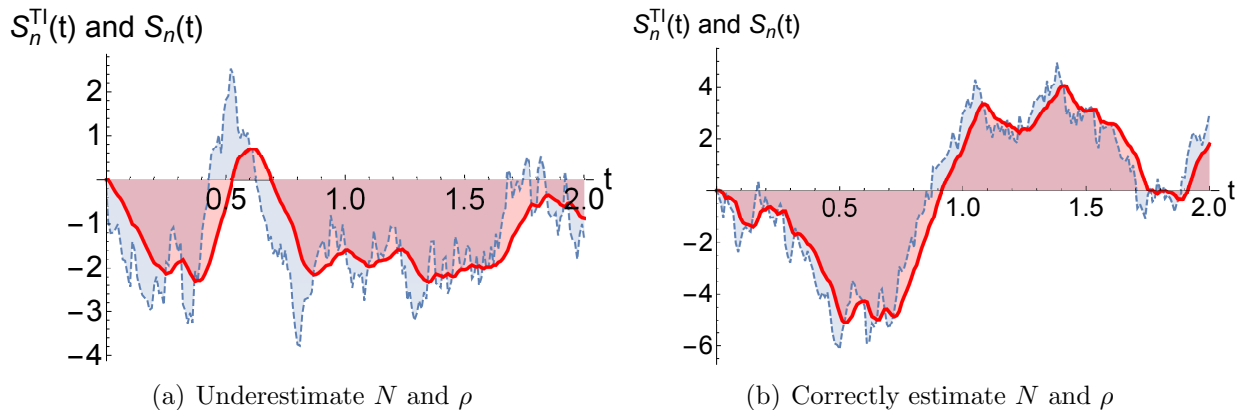


FIGURE B.8. SIMULATED PATHS OF  $S_n^{TI}(t)$  (DASHED) AND  $S_n(t)$  (SOLID).

To summarize, when traders overestimate crowding in the market, they tend to have larger target inventories, trade more aggressively toward target levels, trade more on short-run opportunities, expect more liquidity, and believe that trading is more valuable. When traders underestimate how crowded strategies are, they tend to have smaller target inventories, adjust actual inventories more slowly toward target levels, trade less on short-run profit opportunities, expect less liquidity, and are less willing to provide liquidity to others.

Figure B.9<sup>13</sup> suggests that, in crowded markets, flash crashes may be more likely to occur and their price patterns may be more pronounced, as indeed confirmed by more significant price changes in Figure B.9 when  $N_s < N$ .

<sup>12</sup>The paths are generated using equations (27), (C-18), and (C-19), which describe the dynamics of  $\hat{H}_n(t)$ ,  $\hat{H}_{-n}(t)$ , and  $S_n^{TI}(t)$ . Numerical calculations in Figure B.8 are based on the exogenous parameter values  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\theta = 0.1$ ,  $\tau_H = 1$ ,  $\tau_L = 0.2$  in both (a) and (b);  $N_s = 40 < N = 80$  and  $\rho_s = 0.19 < \rho = 0.2$  in (a);  $N_s = N = 80$  and  $\rho_s = \rho = 0.2$  in (b).

<sup>13</sup>Parameter values are  $r = 0.01$ ,  $A = 1$ ,  $\alpha_D = 0.1$ ,  $\alpha_G = 0.02$ ,  $\sigma_D = 0.5$ ,  $\sigma_G = 0.1$ ,  $\theta = \theta_s = 0$ ,  $\tau_H = 1$ ,  $\tau_L = 0.1$ , and  $D(0^+) = 0$ ,  $H_0(0^+) = 0$ . The endogenous parameter values are  $\gamma_S(N, \rho) = 24.04$ , for  $N = 80$  and  $\rho = 0.2$ ; and  $\gamma_S(N_s, \rho_s) = 11.23$  for  $N_s = 40$  and  $\rho_s = 0.19$ ,  $\tilde{\gamma}_S = 5$   $\gamma_S(N_s, \rho_s) = 56.14$ .

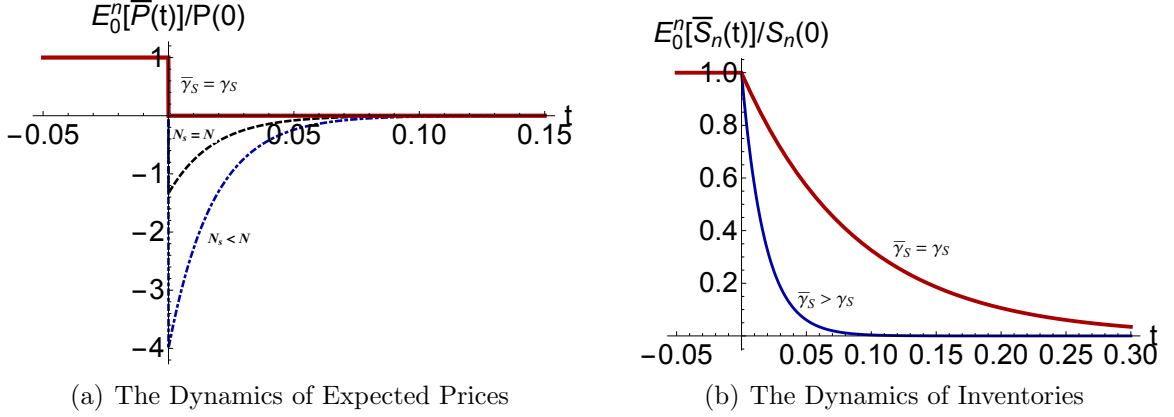


FIGURE B.9. UNDERESTIMATING THE CROWDEDNESS OF THE TOTAL MARKET LEADS TO MORE PRO-  
NOUNCED FLASH-CRASH PRICE PATTERNS.

## C. Proofs

### C.1. Proof of Theorem 2

The first-order condition yields the optimal demand:

$$(C-1) \quad x_n = \frac{E^n[v] - \left(\frac{\alpha}{\gamma} i_0 + \frac{\beta}{\gamma} i_{-n}\right) - \left(\frac{\delta}{(N-1)\gamma} + \frac{A}{\tau}\right) S_n}{\frac{2}{(N-1)\gamma} + \frac{A}{\tau}}.$$

Solving for  $i_{-n}$  instead of  $p$  in the market-clearing condition (A-4), substituting this solution into equation (C-1) above, and then solving for  $x_n$ , yields a demand schedule  $X_n(i_0, i_n, S_n, p)$  for trader  $n$  as a function of price  $p$ . In a symmetric linear equilibrium, the strategy chosen by trader  $n$  must be the same as the linear strategy (A-3) conjectured for the other traders. Equating the corresponding coefficients of the variables  $i_0$ ,  $i_n$ ,  $p$ , and  $S_n$  yields a system of four equations in terms of the four unknowns  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ . The unique solution is

$$(C-2) \quad \alpha = \frac{\tau_0^{1/2} \tau_v^{1/2}}{\tau} \gamma, \quad \beta = \frac{(1-\theta)(\tau_H^{1/2} - \tau_L^{1/2})}{A(1-\rho)} \tau_v^{1/2} \delta,$$

$$(C-3) \quad \gamma = \frac{\tau(1+(N-1)\rho)}{(1+(N-1)\theta)\tau_H^{1/2} + (N-1)(1-\theta)\tau_L^{1/2}} \frac{\beta}{\tau_v^{1/2}},$$

$$(C-4) \quad \delta = \frac{2+(N-2)\rho}{1+(N-1)\rho} - \frac{N(1-\rho)\tau_H^{1/2}}{(N-1)(1-\theta)(1+(N-1)\rho)(\tau_H^{1/2} - \tau_L^{1/2})}.$$

Substituting (C-2) into (C-1) yields trader  $n$ 's optimal demand (A-10). Substituting (A-10) into (A-5) yields the equilibrium price (A-12).

The second-order condition has the correct sign if and only if  $\frac{2}{(N-1)\gamma} + \frac{A}{\tau} > 0$ . This is equivalent to

$$(C-5) \quad \theta < 1 - \frac{N(1-\rho)\tau_H^{1/2}}{(N-1)(2+(N-2)\rho)(\tau_H^{1/2}-\tau_L^{1/2})}.$$

## C.2. Proof of Theorem 1

We assume that all traders believe that there are  $N_s$  traders in the market, and that their private signals are pairwise positively correlated with correlation coefficient of  $\rho_s$ .

Apply the Stratonovich–Kalman–Bucy filter to the filtering problem. This yields trader  $n$ 's filtering estimate of the growth rate  $G_n(t)$  defined by the Itô differential equation

$$(C-6) \quad \begin{aligned} dG_n(t) = & -\alpha_G G_n(t) dt + \tau_0^{1/2} \sigma_G \Omega^{1/2} \left( dI_0(t) - \frac{\tau_0^{1/2} dt}{\sigma_G \Omega^{1/2}} G_n(t) \right) \\ & + \frac{\left( (1+(N_s-2)\rho_s)\tau_H^{1/2} - (N_s-1)\rho_s\tau_L^{1/2} \right) \sigma_G \Omega^{1/2}}{(1-\rho_s)(1+(N_s-1)\rho_s)} \left( dI_n(t) - \frac{\tau_H^{1/2} dt}{\sigma_G \Omega^{1/2}} G_n(t) \right) \\ & + \frac{(\tau_L^{1/2} - \rho_s\tau_H^{1/2})\sigma_G \Omega^{1/2}}{(1-\rho_s)(1+(N_s-1)\rho_s)} \left( \sum_{\substack{m=1 \\ m \neq n}}^{N_s} dI_m(t) - \frac{(N_s-1)\tau_L^{1/2} dt}{\sigma_G \Omega^{1/2}} G_n(t) \right). \end{aligned}$$

Rearranging terms yields

$$(C-7) \quad \begin{aligned} dG_n(t) = & -(\alpha_G + \tau) G_n(t) dt + \tau_0^{1/2} \sigma_G \Omega^{1/2} dI_0(t) + \frac{(\tau_L^{1/2} - \rho_s\tau_H^{1/2})\sigma_G \Omega^{1/2}}{(1-\rho_s)(1+(N_s-1)\rho_s)} \sum_{\substack{m=1 \\ m \neq n}}^{N_s} dI_m(t) \\ & + \frac{\left( (1+(N_s-2)\rho_s)\tau_H^{1/2} - (N_s-1)\rho_s\tau_L^{1/2} \right) \sigma_G \Omega^{1/2}}{(1-\rho_s)(1+(N_s-1)\rho_s)} dI_n(t). \end{aligned}$$

The mean-square filtering error of the estimate  $G(t)$ , denoted  $\sigma_G^2 \Omega(t)$ , is defined by the

Riccati differential equation

(C-8)

$$\sigma_G^2 \frac{d\Omega(t)}{dt} = -2\alpha_G \sigma_G^2 \Omega(t) + \sigma_G^2 - \sigma_G^2 \Omega(t) \left( \tau_0 + \tau_H + (N_s - 1) \frac{\left( (\theta_s - \rho_s) \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2} \right)^2}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} \right).$$

Let  $\Omega$  denote the steady state of the function of time  $\Omega(t)$ . Using the steady-state assumption  $d\Omega(t)/dt = 0$ , solve the second equation for the steady state value  $\Omega = \Omega(t)$  to obtain equation (8). The error variance  $\Omega$  corresponds to a steady state that balances an increase in error variance due to innovations  $dB_G(t)$  in the true growth rate with a reduction in error variance due to (1) mean reversion of the true growth rate at rate  $\alpha_G$  and (2) arrival of new information with total precision  $\tau$ .

Note that  $\Omega$  is not a free parameter but is instead determined as an endogenous function of the other parameters. Equation (8) implies that  $\Omega$  turns out to be the solution to the quadratic equation  $\Omega^{-1} = 2\alpha_G + \tau$ . In equations (3) and (4), we scaled the units with which precision is measured by the endogenous parameter  $\Omega$ . This leads to simpler filtering expressions. The estimate  $G_n(t)$  can be conveniently written as the weighted sum of  $N_s + 1$  sufficient statistics  $H_n(t)$  corresponding to  $N_s + 1$  information flows  $dI_n$ . The sufficient statistics  $H_n(t)$  is defined by equation (10).  $G_n(t)$  becomes a linear combination of sufficient statistics  $H_n(t)$  as given by equation (12). Using the two composite signals,  $\hat{H}_n(t)$  and  $\hat{H}_{-n}(t)$ , defined in equation (13), trader  $n$ 's estimate of the dividend growth rate can be expressed as a function of the two composite signals  $\hat{H}_n(t)$  and  $\hat{H}_{-n}(t)$  as

$$(C-9) \quad G_n(t) := \sigma_G \Omega^{1/2} \left( \left( \frac{(1 - \theta_s)(\tau_H^{1/2} - \tau_L^{1/2})}{1 - \rho_s} + \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} \right) \hat{H}_n(t) \right. \\ \left. + \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} (N_s - 1) \hat{H}_{-n}(t) \right).$$

Define the processes  $dB_0^n$ ,  $dB_n^n$ , and  $dB_m^n$ ,  $m = 1, \dots, N_s$ ,  $m \neq n$ , by

$$(C-10) \quad dB_0^n(t) = \tau_0^{1/2} \frac{G^*(t) - G_n(t)}{\sigma_G \Omega^{1/2}} dt + dB_0(t),$$

$$(C-11) \quad dB_n^n(t) = \tau_H^{1/2} \frac{G^*(t) - G_n(t)}{\sigma_G \Omega^{1/2}} dt + \rho_s^{1/2} dZ(t) + (1 - \rho_s)^{1/2} dB_n(t),$$

and

$$(C-12) \quad dB_m^n(t) = (\theta_s \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2}) \frac{G^*(t) - G_n(t)}{\sigma_G \Omega^{1/2}} dt + \rho_s^{1/2} dZ(t) + (1 - \rho_s)^{1/2} dB_m(t).$$

The superscript  $n$  indicates conditioning on beliefs of trader  $n$ . These  $N_s + 1$  processes are correlated distributed Brownian motions from the perspective of trader  $n$ . Trader  $n$  believes that signals change as follows:

$$(C-13) \quad dH_0(t) = -(\alpha_G + \tau) H_0(t) dt + \tau_0^{1/2} \frac{G_n(t)}{\sigma_G \Omega^{1/2}} dt + dB_0^n(t),$$

$$(C-14) \quad dH_n(t) = -(\alpha_G + \tau) H_n(t) dt + \tau_H^{1/2} \frac{G_n(t)}{\sigma_G \Omega^{1/2}} dt + dB_n^n(t),$$

$$(C-15) \quad dH_{-n}(t) = -(\alpha_G + \tau) H_{-n}(t) dt + (\theta_s \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2}) \frac{G_n(t)}{\sigma_G \Omega^{1/2}} dt + \frac{1}{N_s - 1} \sum_{\substack{m=1 \\ m \neq n}}^{N_s} dB_m^n(t).$$

Note that each signal drifts toward zero at rate  $\alpha_G + \tau$  and drifts toward the optimal forecast  $G_n(t)$  at a rate proportional to the square root of the signal's precision  $\tau_0^{1/2}$ ,  $\tau_H^{1/2}$ , or  $\theta_s \tau_H^{1/2} + (1 - \theta_s) \tau_L^{1/2}$ , respectively.

In terms of the composite variables  $\hat{H}_n$  and  $\hat{H}_{-n}$ , we conjecture (and verify below) a steady-state value function of the form  $V(M_n, S_n, D, \hat{H}_n, \hat{H}_{-n})$ . Letting  $(c_n(t), x_n(t))$  denote consumption and investment choices, write

$$(C-16) \quad V(M_n, S_n, D, \hat{H}_n, \hat{H}_{-n}) := \max_{[c_n(t), x_n(t)]} \mathbf{E}_t^n \left[ \int_{s=t}^{\infty} -e^{-\beta(s-t) - A c_n(s)} ds \right],$$

where  $P(x_n(t))$  is given by equation (17), dividends follow equation (1), inventories follow  $dS_n(t) = x_n(t) dt$ , the change in cash holdings  $dM_n(t)$  follows

$$(C-17) \quad dM_n(t) = (r M_n(t) + S_n(t) D(t) - c_n(t) - P(x_n(t)) x_n(t)) dt,$$

and signals  $\hat{H}_n$  and  $\hat{H}_{-n}$  are given by

$$(C-18) \quad d\hat{H}_n(t) = -(\alpha_G + \tau) \hat{H}_n(t) dt + \frac{\tau_H^{1/2} + \hat{a} \tau_0^{1/2}}{\sigma_G \Omega^{1/2}} G_n(t) dt + \hat{a} dB_0^n(t) + dB_n^n(t),$$

(C-19)

$$d\hat{H}_{-n}(t) = -(\alpha_G + \tau)\hat{H}_{-n}(t)dt + \frac{\theta_s\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2}}{\sigma_G\Omega^{1/2}}G_n(t)dt + \hat{a}dB_0^n(t) + \frac{1}{N_s - 1} \sum_{\substack{m=1 \\ m \neq n}}^{N_s} dB_m^n(t).$$

The dynamics of  $\hat{H}_n$  and  $\hat{H}_{-n}$  in equations (C-18) and (C-19) can be derived from equations (C-13), (C-14), and (C-15).

Note that the coefficient  $\tau_H^{1/2} + \hat{a}\tau_0^{1/2}$  in the second line of equation (C-18) is different from the coefficient  $\theta_s\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2}$  in the second line of equation (C-19). This difference captures the fact that—in addition to disagreeing about the value of the asset in the present—traders also disagree about the dynamics of their future valuations.

We conjecture and verify that the value function  $V(M_n, S_n, D, \hat{H}_n, \hat{H}_{-n})$  has the specific quadratic exponential form

(C-20)

$$V\left(M_n, S_n, D, \hat{H}_n, \hat{H}_{-n}\right) = -\exp\left(\psi_0 + \psi_M M_n + \frac{1}{2}\psi_{SS}S_n^2 + \psi_{SD}S_n D + \psi_{S_n} S_n \hat{H}_n + \psi_{S_x} S_n \hat{H}_{-n} + \frac{1}{2}\psi_{nn} \hat{H}_n^2 + \frac{1}{2}\psi_{xx} \hat{H}_{-n}^2 + \psi_{nx} \hat{H}_n \hat{H}_{-n}\right).$$

The nine constants  $\psi_0, \psi_M, \psi_{SS}, \psi_{SD}, \psi_{S_n}, \psi_{S_x}, \psi_{nn}, \psi_{xx}$ , and  $\psi_{nx}$  have values consistent with a steady-state equilibrium. The Hamilton–Jacobi–Bellman (HJB) equation corresponding to the conjectured value function  $V(M_n, S_n, D, \hat{H}_n, \hat{H}_{-n})$  in equation (C-16) is

(C-21)

$$\begin{aligned} 0 = \max_{c_n, x_n} & \left[ U(c_n) - \beta V + \frac{\partial V}{\partial M_n} (rM_n + S_n D - c_n - P(x_n) x_n) + \frac{\partial V}{\partial S_n} x_n \right] \\ & + \frac{\partial V}{\partial D} (-\alpha_D D + G_n(t)) + \frac{\partial V}{\partial \hat{H}_n} \left( -(\alpha_G + \tau)\hat{H}_n(t) + \frac{\tau_H^{1/2} + \hat{a}\tau_0^{1/2}}{\sigma_G\Omega^{1/2}}G_n(t) \right) + \frac{1}{2}\frac{\partial^2 V}{\partial D^2}\sigma_D^2 \\ & + \frac{\partial V}{\partial \hat{H}_{-n}} \left( -(\alpha_G + \tau)\hat{H}_{-n}(t) + \frac{\theta_s\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2}}{\sigma_G\Omega^{1/2}}G_n(t) \right) + \frac{1}{2}\frac{\partial^2 V}{\partial \hat{H}_n^2} (1 + \hat{a}^2) \\ & + \frac{1}{2}\frac{\partial^2 V}{\partial \hat{H}_{-n}^2} \left( \rho_s + \frac{1 - \rho_s}{N_s - 1} + \hat{a}^2 \right) + \left( \frac{\partial^2 V}{\partial D \partial \hat{H}_n} + \frac{\partial^2 V}{\partial D \partial \hat{H}_{-n}} \right) \hat{a}\sigma_D + \frac{\partial^2 V}{\partial \hat{H}_n \partial \hat{H}_{-n}} (\rho_s + \hat{a}^2). \end{aligned}$$

For the specific quadratic specification of the value function in equation (C-20), the HJB



equation becomes

$$\begin{aligned}
\text{(C-22)} \quad 0 = \min_{c_n, x_n} & \left[ -\frac{e^{-Ac_n}}{V} - \beta + \psi_M(rM_n + S_n D - c_n - P(x_n) x_n) \right. \\
& \left. + (\psi_{SS}S_n + \psi_{SD}D + \psi_{S_n}\hat{H}_n + \psi_{S_x}\hat{H}_{-n})x_n \right] + \psi_{SD}S_n(-\alpha_D D + G_n(t)) \\
& + \left( \psi_{S_n}S_n + \psi_{nn}\hat{H}_n + \psi_{nx}\hat{H}_{-n} \right) \left( -(\alpha_G + \tau)\hat{H}_n(t) + \frac{\tau_H^{1/2} + \hat{a}\tau_0^{1/2}}{\sigma_G \Omega^{1/2}} G_n(t) \right) \\
& + \left( \psi_{S_x}S_n + \psi_{xx}\hat{H}_{-n} + \psi_{nx}\hat{H}_n \right) \left( -(\alpha_G + \tau)\hat{H}_{-n}(t) + \frac{\theta_s \tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2}}{\sigma_G \Omega^{1/2}} G_n(t) \right) \\
& + \frac{1}{2}\psi_{SD}^2 S_n^2 \sigma_D^2 + \frac{1}{2} \left( (\psi_{S_n}S_n + \psi_{nn}\hat{H}_n + \psi_{nx}\hat{H}_{-n})^2 + \psi_{nn} \right) (1 + \hat{a}^2) \\
& + \frac{1}{2} \left( (\psi_{S_x}S_n + \psi_{xx}\hat{H}_{-n} + \psi_{nx}\hat{H}_n)^2 + \psi_{xx} \right) \left( \rho_s + \frac{1 - \rho_s}{N_s - 1} + \hat{a}^2 \right) \\
& + \left( (\psi_{S_n} + \psi_{S_x})S_n + (\psi_{nn} + \psi_{nx})\hat{H}_n + (\psi_{xx} + \psi_{nx})\hat{H}_{-n} \right) \psi_{SD}S_n \hat{a} \sigma_D \\
& + \left( (\psi_{S_n}S_n + \psi_{nn}\hat{H}_n + \psi_{nx}\hat{H}_{-n}) (\psi_{S_x}S_n + \psi_{xx}\hat{H}_{-n} + \psi_{nx}\hat{H}_n) + \psi_{nx} \right) (\rho_s + \hat{a}^2).
\end{aligned}$$

The solution for optimal consumption is

$$\text{(C-23)} \quad c_n(t) = -\frac{1}{A} \log \left( \frac{\psi_M V(t)}{A} \right).$$

The optimal trading strategy is a linear function of the state variables given by

$$\begin{aligned}
\text{(C-24)} \quad x_n(t) = & \frac{(N_s - 1)\gamma_P}{2\psi_M} \left( \left( \psi_{SD} - \frac{\psi_M \gamma_D}{\gamma_P} \right) D(t) + \left( \psi_{SS} - \frac{\psi_M \gamma_S}{(N_s - 1)\gamma_P} \right) S_n(t) \right. \\
& \left. + \psi_{S_n} \hat{H}_n(t) + \left( \psi_{S_x} - \frac{\psi_M \gamma_H}{\gamma_P} \right) \hat{H}_{-n}(t) \right).
\end{aligned}$$

Trader  $n$  can infer from the market-clearing condition (16) that  $\hat{H}_{-n}$  is given by

$$\text{(C-25)} \quad \hat{H}_{-n}(t) = \frac{\gamma_P}{\gamma_H} \left( P(t) - D(t) \frac{\gamma_D}{\gamma_P} \right) - \frac{1}{(N_s - 1)\gamma_H} x_n(t) - \frac{\gamma_S}{(N - 1)\gamma_H} S_n(t).$$

Plugging equation (C-25) into equation (C-24) yields  $x_n(t)$  as a linear demand schedule

given by

$$(C-26) \quad x_n(t) = \frac{(N_s - 1)\gamma_P}{\psi_M} \left(1 + \frac{\psi_{Sx} \gamma_P}{\psi_M \gamma_H}\right)^{-1} \cdot \left( \left(\psi_{SD} - \psi_{Sx} \frac{\gamma_D}{\gamma_H}\right) D(t) + \left(\psi_{SS} - \psi_{Sx} \frac{\gamma_S}{(N_s - 1)\gamma_H}\right) S_n(t) + \psi_{Sn} \hat{H}_n(t) + \left(\psi_{Sx} \frac{\gamma_P}{\gamma_H} - \psi_M\right) P(t) \right).$$

Equating the coefficients of  $D(t)$ ,  $\hat{H}_n(t)$ ,  $S_n(t)$ , and  $P(t)$  in equation (C-26) to the conjectured coefficients  $\gamma_D$ ,  $\gamma_H$ ,  $-\gamma_S$ , and  $-\gamma_P$  results in the following four equations:

$$(C-27) \quad \frac{(N_s - 1)\gamma_P}{\psi_M} \left(1 + \frac{\psi_{Sx} \gamma_P}{\psi_M \gamma_H}\right)^{-1} \left(\psi_{SD} - \psi_{Sx} \frac{\gamma_D}{\gamma_H}\right) = \gamma_D,$$

$$(C-28) \quad \frac{(N_s - 1)\gamma_P}{\psi_M} \left(1 + \frac{\psi_{Sx} \gamma_P}{\psi_M \gamma_H}\right)^{-1} \psi_{Sn} = \gamma_H,$$

$$(C-29) \quad \frac{(N_s - 1)\gamma_P}{\psi_M} \left(1 + \frac{\psi_{Sx} \gamma_P}{\psi_M \gamma_H}\right)^{-1} \left(\psi_{SS} - \psi_{Sx} \frac{\gamma_S}{(N_s - 1)\gamma_H}\right) = -\gamma_S,$$

$$(C-30) \quad \frac{(N_s - 1)\gamma_P}{\psi_M} \left(1 + \frac{\psi_{Sx} \gamma_P}{\psi_M \gamma_H}\right)^{-1} \left(\psi_{Sx} \frac{\gamma_P}{\gamma_H} - \psi_M\right) = -\gamma_P.$$

We obtain

$$(C-31) \quad \psi_{Sx} = \frac{N_s - 2}{2} \psi_{Sn}, \quad \gamma_H = \frac{N_s \gamma_P}{2\psi_M} \psi_{Sn}, \quad \gamma_S = -\frac{(N_s - 1)\gamma_P}{\psi_M} \psi_{SS}, \quad \gamma_D = \frac{\gamma_P}{\psi_M} \psi_{SD}.$$

Define the constants  $C_L$  and  $C_G$  by

$$(C-32) \quad C_L := -\frac{\psi_{Sn}}{2\psi_{SS}}, \quad C_G := \frac{\psi_{Sn}}{2\psi_M} \frac{N_s(r + \alpha_D)(r + \alpha_G) \left(1 + (N_s - 1)\rho_s\right)}{\sigma_G \Omega^{1/2} \left( (1 + (N_s - 1)\theta_s) \tau_H^{1/2} + (N_s - 1)(1 - \theta_s) \tau_L^{1/2} \right)}.$$

Substituting equation (C-31) into equation (C-24) yields the solution for optimal strategy.

$$(C-33) \quad x_n^*(t) = \gamma_S \left( C_L (H_n(t) - H_{-n}(t)) - S_n(t) \right).$$

Define the average of traders' expected growth rates  $\bar{G}(t)$  by

$$(C-34) \quad \bar{G}(t) := \frac{1}{N_s} \sum_{n=1}^{N_s} G_n(t),$$

Then, the equilibrium price is

$$(C-35) \quad P^*(t) = \frac{D(t)}{r + \alpha_D} + \frac{C_G \bar{G}(t)}{(r + \alpha_D)(r + \alpha_G)}.$$

Plugging (C-23) and (C-24) back into the Bellman equation and setting the constant term and the coefficients of  $M_n$ ,  $S_n D$ ,  $S_n^2$ ,  $S_n \hat{H}_n$ ,  $S_n \hat{H}_{-n}$ ,  $\hat{H}_n^2$ ,  $\hat{H}_{-n}^2$ , and  $\hat{H}_n \hat{H}_{-n}$  to be zero, we obtain nine equations. Using the first equation (C-31) to substitute  $\psi_{S_n}$  for  $\psi_{S_x}$ , there are in total nine equations in nine unknowns  $\gamma_P$ ,  $\psi_0$ ,  $\psi_M$ ,  $\psi_{SD}$ ,  $\psi_{SS}$ ,  $\psi_{S_n}$ ,  $\psi_{nn}$ ,  $\psi_{xx}$ , and  $\psi_{nx}$ .

By setting the constant term, coefficient of  $M$ , and coefficient of  $S_n D$  to be zero, we obtain

$$(C-36) \quad \psi_M = -rA, \quad \psi_{SD} = -\frac{rA}{r + \alpha_D},$$

$$(C-37) \quad \psi_0 = 1 - \ln r + \frac{1}{r} \left( -\beta + \frac{1}{2}(1 + \hat{a}^2)\psi_{nn} + \frac{1}{2} \left( \hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1} \right) \psi_{xx} + (\hat{a}^2 + \rho_s)\psi_{nx} \right).$$

In addition, by setting the coefficients of  $S_n^2$ ,  $S_n \hat{H}_n$ ,  $S_n \hat{H}_{-n}$ ,  $\hat{H}_n^2$ ,  $\hat{H}_{-n}^2$  and  $\hat{H}_n \hat{H}_{-n}$  to be zero, we obtain six polynomial equations in the six unknowns  $\gamma_P$ ,  $\psi_{SS}$ ,  $\psi_{S_n}$ ,  $\psi_{nn}$ ,  $\psi_{xx}$ , and  $\psi_{nx}$ . Defining the constants  $a_1$ ,  $a_2$ ,  $a_3$ , and  $a_4$  by

$$\begin{aligned} a_1 &:= -(\alpha_G + \tau) + (\tau_H^{1/2} + \hat{a}\tau_0^{1/2}) \left( \frac{(1 - \theta_s)(\tau_H^{1/2} - \tau_L^{1/2})}{1 - \rho_s} + \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} \right), \\ a_2 &:= -(\alpha_G + \tau) + (N_s - 1) \left( \theta_s \tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2} \right) \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)}, \\ a_3 &:= (\tau_H^{1/2} + \hat{a}\tau_0^{1/2})(N_s - 1) \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)}, \\ a_4 &:= \left( \theta_s \tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2} + \hat{a}\tau_0^{1/2} \right) \left( \frac{(1 - \theta_s)(\tau_H^{1/2} - \tau_L^{1/2})}{1 - \rho_s} + \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} \right), \end{aligned}$$

these six equations in six unknowns can be written

(C-38)

$$0 = -\frac{1}{2}r\psi_{SS} - \frac{\gamma_P(N_s - 1)}{rA}\psi_{SS}^2 + \frac{r^2A^2\sigma_D^2}{2(r + \alpha_D)^2} + \frac{1}{2}(1 + \hat{a}^2)\psi_{sn}^2 + \frac{1}{2}\left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\frac{(N_s - 2)^2}{4}\psi_{Sn}^2 - \frac{rA}{r + \alpha_D}\hat{a}\sigma_D\frac{N_s}{2}\psi_{Sn} + \frac{N_s - 2}{2}\psi_{Sn}^2(\hat{a}^2 + \rho_s)$$

(C-39)

$$0 = -r\psi_{Sn} - \frac{rA}{r + \alpha_D}\sigma_G\Omega^{1/2}\left(\frac{(1 - \theta_s)(\tau_H^{1/2} - \tau_L^{1/2})}{1 - \rho_s} + \frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)}\right) + a_1\psi_{Sn} - \frac{\gamma_P(N_s - 1)}{rA}\psi_{SS}\psi_{Sn} + \frac{N_s - 2}{2}a_4\psi_{Sn} + (1 + \hat{a}^2)\psi_{nn}\psi_{Sn} + \frac{N_s - 2}{2}\left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\psi_{nx}\psi_{Sn} - \frac{rA}{r + \alpha_D}\hat{a}\sigma_D(\psi_{nn} + \psi_{nx}) + (\hat{a}^2 + \rho_s)\left(\psi_{nx}\psi_{Sn} + \frac{N_s - 2}{2}\psi_{nn}\psi_{Sn}\right),$$

(C-40)

$$0 = -r\frac{N_s - 2}{2}\psi_{Sn} + \frac{\gamma_P(N_s - 1)}{rA}\psi_{SS}\psi_{Sn} - \frac{rA}{r + \alpha_D}\sigma_G\Omega^{1/2}(N_s - 1)\frac{(\theta_s - \rho_s)\tau_H^{1/2} + (1 - \theta_s)\tau_L^{1/2}}{(1 - \rho_s)(1 + (N_s - 1)\rho_s)} + (a_3 + \frac{N_s - 2}{2}a_2)\psi_{Sn} + (1 + \hat{a}^2)\psi_{Sn}\psi_{nx} + \frac{N_s - 2}{2}\left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\psi_{xx}\psi_{Sn} - \frac{rA}{r + \alpha_D}\hat{a}\sigma_D(\psi_{xx} + \psi_{nx}) + (\hat{a}^2 + \rho_s)\left(\psi_{xx}\psi_{Sn} + \frac{N_s - 2}{2}\psi_{nx}\psi_{Sn}\right),$$

(C-41)

$$0 = -\frac{r}{2}\psi_{nn} - \frac{\gamma_P(N_s - 1)}{4rA}\psi_{Sn}^2 + a_1\psi_{nn} + a_4\psi_{nx} + \frac{1}{2}(1 + \hat{a}^2)\psi_{nn}^2 + \frac{1}{2}\left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\psi_{nx}^2 + (\hat{a}^2 + \rho_s)\psi_{nn}\psi_{nx},$$

(C-42)

$$0 = -\frac{r}{2}\psi_{xx} - \frac{\gamma_P(N_s - 1)}{4rA}\psi_{Sn}^2 + a_2\psi_{xx} + a_3\psi_{nx} + \frac{1 + \hat{a}^2}{2}\psi_{nx}^2 + \frac{1}{2}\left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\psi_{xx}^2 + (\hat{a}^2 + \rho_s)\psi_{xx}\psi_{nx},$$

$$\begin{aligned}
(C-43) \quad 0 = & -r\psi_{nx} + \frac{\gamma_P(N_s - 1)}{2rA}\psi_{S_n}^2 + a_3\psi_{nn} + a_4\psi_{xx} + (a_1 + a_2)\psi_{nx} \\
& + (1 + \hat{a}^2)\psi_{nn}\psi_{nx} + \left(\hat{a}^2 + \frac{1 + (N_s - 2)\rho_s}{N_s - 1}\right)\psi_{xx}\psi_{nx} + (\hat{a}^2 + \rho_s)(\psi_{nn}\psi_{xx} + \psi_{nx}^2).
\end{aligned}$$

We solve equations (C-38)–(C-43) numerically. For a solution to the six polynomial equations to define a stationary equilibrium, a second-order condition implying  $\gamma_P > 0$ , a stationarity condition implying  $\gamma_S > 0$ , and a transversality condition requiring  $r > 0$ .

The transversality condition for the value function  $V(\dots)$  is

$$(C-44) \quad \lim_{T \rightarrow +\infty} \mathbf{E}_t^n \left[ e^{-\rho_s(T-t)} V \left( M_n(T), S_n(T), D(T), \hat{H}_n(T), \hat{H}_{-n}(T) \right) \right] = 0.$$

The transversality condition (C-44) is satisfied if  $r > 0$ . Under the assumptions  $\gamma_P > 0$  and  $\gamma_S > 0$ , analytical results imply  $\gamma_D > 0$ ,  $\psi_M < 0$ ,  $\psi_{SD} < 0$ , and  $\psi_{SS} > 0$ . The numerical results indicate that  $\gamma_H > 0$ ,  $\psi_{S_n} < 0$ ,  $\psi_{S_x} < 0$ ,  $\psi_{nn} < 0$ ,  $\psi_{xx} < 0$  and the sign of  $\psi_{nx}$  is intuitively and numerically ambiguous.

### C.3. Proof of Corollary 1

The consistency condition in equation (22) implies that

$$(C-45) \quad \rho_s - \rho = \frac{(N_s - N)(1 - \rho)}{N(N_s - 1)}.$$

Equation (C-45) implies that, if  $N_s < N$ , then  $\rho_s < \rho$ .

(C-46)

$$\begin{aligned}
Cov(dI_n(t), dP(t)) &= Cov(\rho^{1/2}dZ(t) + (1 - \rho)^{1/2}dB_n(t), \rho^{1/2}dZ(t) + (1 - \rho)^{1/2}\frac{1}{N}\sum_{m=1}^N dB_m(t)) \\
&= \rho + \frac{1}{N}(1 - \rho).
\end{aligned}$$

(C-47)

$$\begin{aligned}
Cov(dI_n(t), dP_s(t)) &= Cov(\rho_s^{1/2}dZ(t) + (1 - \rho_s)^{1/2}dB_n(t), \rho_s^{1/2}dZ(t) + (1 - \rho_s)^{1/2}\frac{1}{N_s}\sum_{m=1}^{N_s} dB_m(t)) \\
&= \rho_s + \frac{1}{N_s}(1 - \rho_s).
\end{aligned}$$

Therefore, if the consistency condition (22) is satisfied, then for each trader, the correlation coefficient between his private signal and the actual price is consistent with the correlation between his private signal and his “subjective” price.



BANK OF ENGLAND

# Staff Working Paper No. 711

## Judgement Day: algorithmic trading around the Swiss franc cap removal

Francis Breedon, Louisa Chen, Angelo Ranaldo and  
Nicholas Vause

February 2018

Staff Working Papers describe research in progress by the author(s) and are published to elicit comments and to further debate. Any views expressed are solely those of the author(s) and so cannot be taken to represent those of the Bank of England or to state Bank of England policy. This paper should therefore not be reported as representing the views of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Committee.



BANK OF ENGLAND

# Staff Working Paper No. 711

## Judgement Day: algorithmic trading around the Swiss franc cap removal

Francis Breedon,<sup>(1)</sup> Louisa Chen,<sup>(2)</sup> Angelo Ranaldo<sup>(3)</sup> and Nicholas Vause<sup>(4)</sup>

### Abstract

A key issue raised by the rapid growth of computerised algorithmic trading is how it responds in extreme situations. Using data on foreign exchange orders and transactions that includes identification of algorithmic trading, we find that this type of trading contributed to the deterioration of market quality following the removal of the cap on the Swiss franc on 15 January 2015, which was an event that came as a complete surprise to market participants. In particular, we find that algorithmic traders withdrew liquidity and generated uninformative volatility in Swiss franc currency pairs, while human traders did the opposite. However, we find no evidence that algorithmic trading propagated these adverse effects on market quality to other currency pairs.

**Key words:** Swiss franc, algorithmic trading, liquidity, volatility, price discovery, arbitrage opportunities.

**JEL classification:** G14, G23.

---

(1) School of Economics and Finance, Queen Mary University of London. Email: f.breedon@qmul.ac.uk

(2) School of Business, Management and Economics, University of Sussex. Email: l.x.chen@sussex.ac.uk. Chen was with the Bank of England when the paper was written.

(3) Swiss Institute of Banking and Finance, University of St. Gallen. Email: angelo.ranaldo@unisg.ch

(4) Bank of England. Email: nicholas.vause@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to EBS for providing their anonymised and aggregated market data. EBS has not contributed to any of the analysis in the paper and does not endorse or support any of its conclusions. We are also grateful to Patrick Schaffner for excellent research assistance, and participants at the 2016 conference on Microstructure of Foreign Exchange Markets at Cambridge University, the 2016 conference on Financial Determinants of Foreign Exchange Rates at the Bank of England and a 2017 seminar at the Swiss National Bank for their comments. Our title is inspired by that of a related paper, 'Rise of the Machines: algorithmic trading in the foreign exchange market' (Chaboud *et al* (2014)). 'Judgement Day' followed 'Rise of the Machines' in the Terminator film franchise.

The Bank's working paper series can be found at [www.bankofengland.co.uk/working-paper/Working-papers](http://www.bankofengland.co.uk/working-paper/Working-papers)

Publications and Design Team, Bank of England, Threadneedle Street, London, EC2R 8AH  
Telephone +44 (0)20 7601 4030 email [publications@bankofengland.co.uk](mailto:publications@bankofengland.co.uk)



# Contents

1	Introduction	3
2	Literature review	5
3	Market structure and data	6
	3.1 Market structure	6
	3.2 Data	7
4	Overview of trading patterns around the Swiss franc event	8
	4.1 The Swiss franc event	8
	4.2 Overview of trading patterns	9
5	Liquidity contributions of computer and human trading	13
	5.1 Volumes of liquidity provided and consumed	13
	5.2 Effective spreads	15
6	Impact of computer and human trading on volatility	16
	6.1 Contributions to realised volatility	17
	6.2 Contributions to efficient pricing	18
7	Arbitrage opportunities and market efficiency	21
8	Non-CHF foreign exchange rates	22
9	Conclusion	24
10	References	26

# 1 Introduction

New technologies have dramatically changed financial markets. One of the main innovations of recent years is computerised algorithmic trading (AT), which broadly refers to the direct use of computers to implement trades. AT is now widely used by financial institutions such as banks and hedge funds, and has important effects on the operation of financial markets. It can improve market liquidity by reducing transaction costs (Hendershot *et al.*, 2011) and the reliance on financial intermediaries (Menkveld, 2013). It can also make security prices more efficient, in the sense that they better reflect fundamental values (O’Hara, 2015). It can even reduce risks by lessening the impact of human feelings on overall investor behaviour, such as panic reactions and herding behaviour. On the other hand, AT may have less desirable implications because it can increase market power over slower traders (Hoffmann, 2014), raise adverse selection (Biais *et al.*, 2015), excess volatility or extreme market movements (Foucault *et al.*, 2016) and so potentially harm financial stability. It is this last issue we focus on in this paper.

We analyse the role of AT in foreign exchange (FX) markets in a period containing the 15 January 2015 announcement by the Swiss National Bank that it had discontinued its policy of capping the value of the Swiss franc against the euro. This ‘Swiss franc event’ represents a natural experiment as one of the largest shocks to the FX market in recent years and probably the most significant ‘black swan’ event in the period in which AT has been a prominent force in FX markets.<sup>1</sup> In particular, we study the contribution of AT and human traders to two important dimensions of market quality, namely liquidity and price efficiency. Our analysis is based on a unique dataset with a detailed identification of AT obtained from EBS Market, which is the leading platform for electronic spot FX trading in many of the major currencies.<sup>2</sup>

A detailed understanding of AT in distressed situations is important for at least two reasons. First, a better comprehension of whether AT is beneficial or detrimental for market quality in extreme situations would help inform the ongoing reform of trading venues, as pursued by Regulation NMS in the United States and MiFID I and II in Europe. Second, the resilience of an exchange system depends on the behaviour of different types of market participant and their reciprocal influence on each other. For instance, a tendency of AT to offer liquidity in calm markets and withdraw it in distressed situations could lead less sophisticated agents to become reliant on high levels of market liquidity only to find it in short supply when they most needed it. If these adverse consequences of AT were predominant or not offset by other traders, then AT could represent a systemic threat to the whole trading system. To shed light on this key issue for financial stability, we analyse whether human traders and AT substitute for or complement each other in supplying and consuming liquidity.

---

<sup>1</sup> A ‘black swan’ is a metaphor that describes an event that comes as a complete surprise and has a major effect. The term is based on an ancient saying that presumed black swans did not exist, but the saying was rewritten after black swans were discovered in the wild.

<sup>2</sup> While EBS has supplied the relevant market data, it does not endorse or support any conclusions made in this paper and has not contributed to any of the analysis in it.

We proceed in three steps. First, we describe the EBS Market platform and our sample of data from it. The platform is the central limit order book for spot FX operated by EBS Service Company Limited, which is part of NEX Markets, a business division of the NEX Group plc. To introduce our analysis, we provide an overview of trading patterns conducted by AT and human traders around the Swiss franc event. Second, we perform an in-depth analysis of market liquidity and price movements by decomposing order flow, effective spreads and intraday volatility by type of trader. This enables us to highlight the contribution of AT and human traders to liquidity provision and consumption, transaction costs and realised volatility. Third, we study the contribution to efficient pricing of AT and human traders. To do this, we analyse the formation of efficient prices by performing a vector autoregression (VAR) as in Hasbrouck (1991(a), 1991(b), 2007), but conditioning on traders' types as in Hendershott *et al.* (2011). We substantiate the analysis of price efficiency by looking at arbitrage opportunities (Chaboud *et al.*, 2014) and variance ratios (O'Hara and Ye, 2011).

Our study delivers two important findings. First, in reaction to the Swiss franc event, we find that AT tended to consume liquidity and reinforce the price disruption. Opposite and offsetting patterns apply for human traders, who supported market quality by providing liquidity and aiding price discovery. Second, we find that this market quality degradation coming from AT was concentrated in the shocked FX rate (EUR/CHF) and, to a lesser extent, USD/CHF. Non-CHF currency pairs (USD/JPY, EUR/JPY and EUR/USD in our sample) were essentially unaffected.<sup>3</sup> This suggests that AT models were somewhat compartmentalised, which, along with human trading, helped to sustain market quality beyond the CHF currency pairs.

Of course, like Tolstoy's comment on unhappy families, each period of market distress is distressed in its own way. This limits our ability to draw general conclusions. However, as in previous papers investigating important shocks such as the failure of Lehman Brothers in 2008 (Afonso *et al.*, 2011) or the 'Flash Crash' in 2010 (Kirilenko *et al.*, 2017), our analysis of a single event should provide indicative evidence on these broader issues.

Our paper contributes to the growing literature on algorithmic trading, which we survey in the next section. Except for the thorough analysis by Chaboud *et al.* (2014), prior research on FX algorithmic trading is scant. While that work shows AT improves price efficiency in 'normal' market conditions, we find that AT reactions to an extreme event were detrimental to market liquidity and price resilience, whereas human traders sustained market quality during the event. FX markets have a different structure to other markets where AT is prevalent, with more trading conducted bilaterally and on multilateral platforms, and less on centralised exchanges. Hence, it is important to build the literature on FX markets.

The paper proceeds as follows. After briefly describing related literature (Section 2) and our data (Section 3), we give an overview of trading patterns around the Swiss franc event (Section 4). We then undertake a more formal analysis of liquidity (Section 5), volatility

---

<sup>3</sup> The currency codes are: CHF for Swiss francs, EUR for euros, JPY for Japanese yen and USD for US dollars.

(Section 6) and market efficiency (Section 7) for Swiss franc currency pairs, followed by analysis of non-CHF FX rates (Section 8).

## 2 Literature review

The literature on AT has grown substantially in the last few years.<sup>4</sup> The theoretical literature focuses on how high-frequency trading (HFT), which is the most commonly discussed form of AT, affects liquidity and price efficiency. The HFT community's ability to revise their quotes quickly after news arrives reduces the winner's curse problem but creates disincentives for trading by slower traders (Hoffmann, 2014; and Jovanovich and Menkveld, 2016), including by increasing adverse selection and price impact (Foucault *et al.*, 2016). Aït-Sahalia and Saglam (2014) and Rosu (2016) model HFT that is averse to inventory risk and predict that volatility will lead high-frequency traders to reduce their provision of liquidity. Biais *et al.* (2015) find a role for HFT in fragmented markets that produces adverse selection, negative externalities and over-investment in equilibrium.<sup>5</sup>

The focus of prior empirical research has been the stock market. As described by O'Hara (2015), the consensus is that HFT market making enhances market quality by reducing spreads and raising informational efficiency.<sup>6</sup> However, it is not clear whether HFTs go with or lean against the wind, that is amplify price falls (rises) by actively selling (buying) or dampening them by actively buying (selling) (see Korajczyk and Murphy, 2016; van Kervel and Menkveld, 2016; Breckenfelder, 2013; and Tong, 2015).

By analysing a 'black swan' event, our paper is related to the stock market literature, which provides mixed evidence on the role of HFT in distressed markets. On the one hand, HFTs are found to withdraw from their market-making role during 'flash crashes' (see *e.g.* CFTC-SEC, 2010; Easley *et al.*, 2012; Kirilenko *et al.*, 2017; and Menkveld and Yueshen, 2015) or when market conditions become unfavourable (see *e.g.* Raman *et al.*, 2014; Anand and Venkataraman, 2015; and Korajczyk and Murphy, 2015). On the other hand, HFT provides liquidity and absorbs imbalances created by non-high frequency traders around large price movements (Brogaard *et al.*, 2017) and the short-interval return volatility of most stocks varies inversely with a market-wide measure of correlated HFT strategies (Boehmer, Li, and Saar, 2016).

The only earlier paper that provides an in-depth analysis of AT in the FX market is Chaboud *et al.* (2014). They show that algorithmic trading is associated with a reduction in arbitrage opportunities while AT liquidity provision decreases return autocorrelation.

---

<sup>4</sup> There are many excellent surveys on AT and HFT, including recent papers by Biais and Woolley (2011), Chordia *et al.* (2013), Easley *et al.* (2013), Gomber *et al.* (2011), Goldstein *et al.* (2014), Jones (2013), Kirilenko and Lo (2013), Biais and Foucault (2014), SEC (2010), O'Hara (2015) and Menkveld (2016, 2017).

<sup>5</sup> The role of AT in fragmented markets with multiple exchanges is studied in Pagnotta and Philippon (2015). Other papers analyse the welfare implications from double auctions (*e.g.* Cespa and Vives, 2015; Du and Zhu, 2015) and asynchronous arrivals (*e.g.* Budish *et al.*, 2015; Bongaerts and Van Achter, 2016). Bernales (2014) and Rojcek and Ziegler (2016) use numerical methods for dynamic models encompassing the endogenous role of HFT, general limit order book and latency.

<sup>6</sup> See, for example, Boehmer *et al.* (2015), Brogaard *et al.* (2015), Carrion (2017), Conrad *et al.* (2015), Hasbrouck and Saar (2013), Hendershott *et al.* (2011), Menkveld (2013) and Malinova *et al.* (2013).

## 3 Market structure and data

### 3.1 Market structure

Two significant global electronic spot trading platforms in major currency pairs are EBS and Reuters. USD/CHF and EUR/CHF, which are the focus of this study, trade primarily on EBS (see King, *et al.*, 2011). Prices on EBS also constitute the reference for derivative pricing in these currencies. Moreover, during the period of the Swiss franc event, EBS was the key trading platform for all Swiss franc positions as trading of futures on the Chicago Mercantile Exchange was suspended and over-the-counter trading had largely disappeared (Hagströmer and Menkveld, 2016).

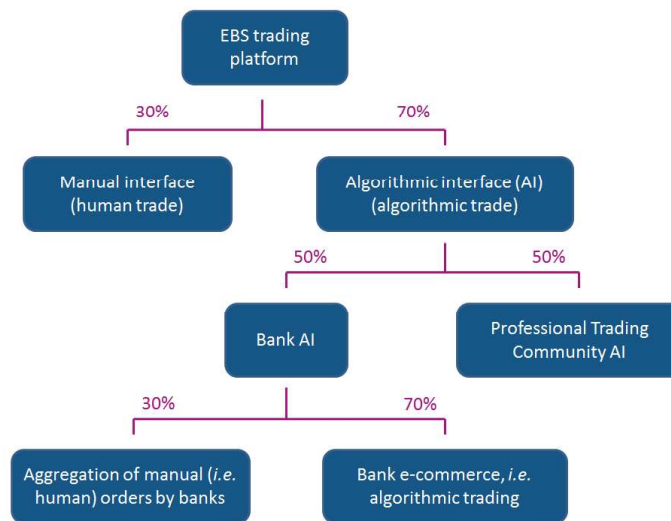
EBS Market is an order-driven electronic trading system, which unites buyers and sellers of spot FX across the globe on a pre-trade anonymous central limit order book. EBS is accessible to foreign exchange dealing banks and, under the auspices of dealing banks (via prime brokerage arrangements), to hedge funds and commodity trading advisors (CTAs). EBS controls the network and each of the terminals on which trading is conducted, and records whether a trade is placed by an ordinary keyboard ('manual' or 'human' trades), or by a direct computer interface ('algorithmic' trades).

As well as this simple distinction by terminal type, EBS requires market participants to identify what types of trading they engage in. This allows EBS to decompose the algorithmic trades into two different categories, referred to as 'bank AI' and 'PTC AI', where PTC stands for Professional Trading Community and AI for algorithmic interface, which is how EBS labels the direct computer interfaces. Market data at this level of granularity is not ordinarily sold or distributed by EBS to third parties. PTC essentially refers to principal trading firms, hedge funds and commodity trading advisors, which can trade directly on EBS under the auspices of dealing banks (via prime brokerage arrangements). Trading through this route is all AT in the sense that trades are initiated by computers. Bank AI is harder to categorise since it includes a significant share (EBS estimates approximately 30%) of aggregators that are computer-based trading systems that simply process orders received from bank customers and then execute them algorithmically. The source of these trades may be described as 'agency algorithms' or 'smart execution algorithms', as discussed by Gomber *et al.* (2011), rather than the proprietary algorithms usually highlighted in the AT literature. Note that the former category also includes 'auto-liquidation' algorithms that respond to margin calls, which have been highlighted as a source of price distortions during crashes (*e.g.* McCann and Yan, 2015). Note that PTC firms or banks may also submit manual orders. The three trade categories and their shares of average transaction volumes in EBS Market are represented in Figure 1.

Computer-based trading classified as PTC or bank AI accounts for approximately 70% of EBS Market transaction volume. However, the system includes features designed to prevent strategies where speed or low latency is the sole contributor to success. First, it imposes a minimum quote life (MQL) for the five core currency pairs on EBS Market, so that once a good-

till-cancelled order is submitted, it cannot be cancelled for 250 milliseconds.<sup>7</sup> Second, and more importantly, EBS Market operates a randomised batching window on all messages that enter its matching engine, referred to as a ‘Latency Floor’. This mechanism generates batching windows of 1-3 milliseconds<sup>8</sup>, in which messages are processed on a randomised basis. As a result, the first message to arrive may not be the first released and sub-millisecond differences in latency become less important for trading on EBS Market. Note that this is not analogous to the frequent batch auction system described by Budish, Crampton and Shim (2015), which eliminates sniping of stale quotes, but is more like the random order delay system of Harris (2012).

**Figure 1: Indicative breakdown of EBS trading volumes**



Source: EBS.

As EBS Market is a ‘wholesale’ trading system, the minimum trade size over our sample period is one million of the base currency<sup>9</sup>, and trade sizes are only allowed in a multiple of millions of the base currency.

### 3.2 Data

Our data consist of both intraday quotes and transactions for EUR/CHF, USD/CHF, EUR/USD, USD/JPY and EUR/JPY during 5-23 January 2015. We specify 15 January as the Swiss franc event day, 5-14 January as the pre-event period, and 16-23 January as the post-event period. Throughout the subsequent analysis, we focus on data between 8.00 GMT and 17.00 GMT,

<sup>7</sup> A MQL of 250 milliseconds was applied in the EUR/USD, USD/JPY, USD/CHF, EUR/CHF and EUR/JPY currency pairs on EBS Market for the dates referenced in this paper.

<sup>8</sup> For the dates referenced in this paper.

<sup>9</sup> The base currency is the first currency displayed in the symbol of the currency pair. For example, the euro is the base currency of EURCHF.

which represent the effective trading hours of the day. We also exclude weekends, when the EBS trading platform is inactive.<sup>10,11</sup>

The transaction data set records the time stamp to the nearest millisecond of each trade that occurred, along with the actual transaction price, the amount transacted, and the direction of the trade. Importantly for our study, the nature of the party providing liquidity (submitting the good-till-cancelled (GTC) order) and consuming liquidity (submitting the immediate-or-cancel (“IOC”) order) is categorised by EBS as human, bank AI or PTC AI.<sup>12</sup> Thus, each trade may be classified according to nine different possible combinations of liquidity provider and consumer. In addition, each trade has an indicator of whether the liquidity provider was the buyer or the seller. We line up these millisecond time-stamped transactions with the quote data, which are available at 100 millisecond intervals, so for a given transaction the top 10 anonymised best bid and ask prices at the nearest previous whole 100 millisecond interval are also available. All quotes are firm and therefore truly represent the market prices at that instant.<sup>13</sup> Unlike in the transactions data, the type of trader that posts each quote is not available to us.

## 4 Overview of trading patterns around the Swiss franc event

### 4.1 The Swiss franc event

The Swiss National Bank (SNB) began intervening in FX markets to cap the value of the Swiss franc against the euro on 6 September 2011. In a press release of the same date, the SNB said “the massive overvaluation of the Swiss franc poses an acute threat to the Swiss economy and carries the risk of a deflationary development”. Therefore, “it will no longer tolerate a EUR/CHF exchange rate below the minimum rate of CHF 1.20 ... and is prepared to buy foreign currency in unlimited quantities” (SNB, 2011).

Following the introduction of this policy, the franc generally traded a little below its cap (Figure 2). It pushed against it for a period in mid-2012, before appreciation pressures again intensified towards the end of 2014 due to weakness in the euro-area economy. In response, the SNB cut its interest rate on sight deposits to *minus* 0.25%. However, commitment to the exchange rate policy appeared to remain firm. On 18 December 2014 the SNB Governor stated that the central bank was “committed to purchasing unlimited quantities of foreign currency to enforce the minimum exchange rate with the utmost determination” (Jordan, 2014). Similarly,

---

<sup>10</sup> See Chaboud *et al.* (2014) for further discussion of trading activity on the EBS system. In addition, EBS indicated to us that their dealing services are ordinarily open for trading 24 hours a day, 7 days a week, with the exception of a maintenance window that ordinarily occurs from 5:50pm New York time on a Friday until Saturday morning. For the purpose of computing market data products, such as ‘highs’ and ‘lows’, EBS regards trades between 5pm Friday New York time and 5am Monday Sydney time as not conducted in normal market conditions or market hours and excludes these trades from the calculations. However, the trades remain in EBS’ data.

<sup>11</sup> In our sample, we drop five transactions and 24 quotes on EUR/CHF that took place between 09:32:29 GMT and 09:32:39 GMT on January 15, where the price was exceptionally low at 0.0015, with volume of one million of the base currency for each transaction. EBS confirmed that those transactions turned out to be errors made by the traders, and the counterparties have settled solutions outside EBS.

<sup>12</sup> GTC and IOC orders respectively align closely with the concepts of limit and market orders in the broader market microstructure literature.

<sup>13</sup> The historical market data provided by EBS is time-sliced at 100 milliseconds and is therefore a snapshot of the activity during previous time-period. Consequently the quote and paid/given trade data provided are a summary and not a full life-cycle of every event.

on 12 January 2015, another member of the SNB Governing Board said that “we are convinced that the minimum exchange rate must remain the cornerstone of our monetary policy” (Reuters, 2015a).

**Figure 2: EUR/CHF exchange rate versus the cap set by the Swiss National Bank**



Source: Bloomberg

Despite that, the policy of capping the value of the franc against the euro was discontinued at 9:30 GMT on 15 January 2015. A press release gave the following explanation: “Recently, divergences between monetary policies of the major currency areas have increased significantly. The euro has depreciated significantly against the US dollar and this, in turn, has caused the Swiss franc to weaken against the US dollar. In these circumstances, the SNB concluded that enforcing and maintaining the minimum exchange rate for the Swiss franc against the euro is no longer justified”. This news came as a complete surprise to market participants, as was reflected both in FX options prices leading up to the announcement (Mirkov *et al.*, 2016) and financial reporting after it (*e.g.* Reuters, 2015b; Bloomberg, 2015).

## 4.2 Overview of trading patterns

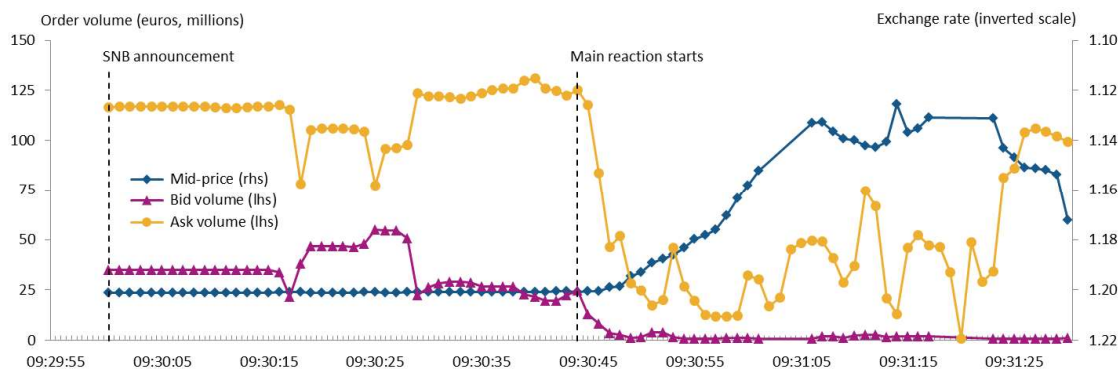
In this section we provide an overview of the market reaction to the SNB announcement on 15 January 2015. In particular, we illustrate graphically some of the key features of human and algorithmic trading around the announcement as a prelude to the more formal analysis in the rest of the paper. We first focus on the seconds around the announcement, then the minutes, and finally the hours.

Figure 3 plots the volume of GTC orders to buy (‘bid’) or sell (‘ask’) euros for Swiss francs as well as the mid-price, which is the mid-point between the best bid and ask prices, during the 90 seconds following the announcement. Not shown, for reasons of sensitivity, is a sharp fall in outstanding orders to sell francs in the seconds approaching 9:30. While our data from EBS does not include the identities of the institutions submitting the orders, we presume this was driven by the SNB in preparation for its 9:30 announcement. The chart does show that it was not until about 44 seconds after the announcement that the market reacted significantly. At this time, the mid-price started to reflect a very rapid appreciation of the franc, as both sides



of the order book collapsed in size.<sup>14</sup> Indeed, for a few seconds during the minute of 9.31, there were no orders to buy euros in exchange for Swiss francs at any price. We take the delayed response to the announcement as further evidence that it was not anticipated.

**Figure 3: EUR/CHF price and orders in the seconds following the SNB announcement**



Sources: EBS and authors' calculations.

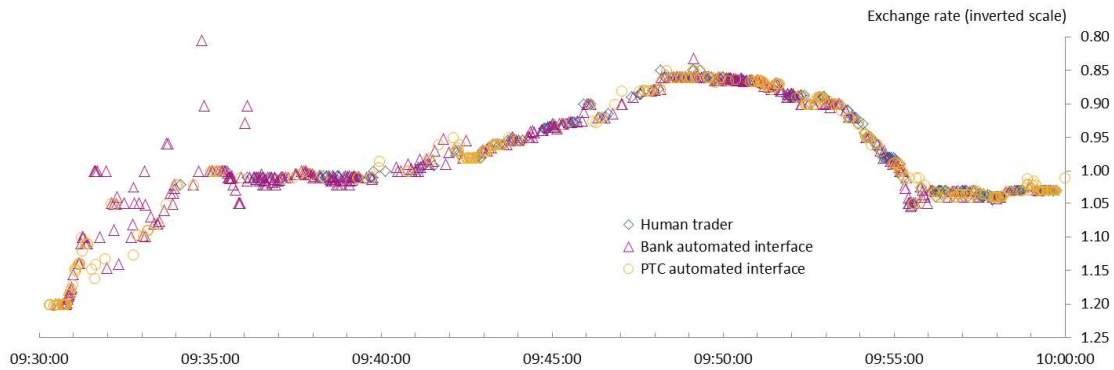
Figure 4 shows the prices at which different types of trader exchanged euros for Swiss francs in the 30 minutes following the SNB announcement depending on whether their trades were consuming liquidity (top panel) or providing it (bottom panel). Trades that consume liquidity result from IOC orders, while those that provide it result from GTC orders. The top panel shows that bank AIs consumed liquidity at extreme prices (prices significantly different to those of immediately preceding trades) on a number of occasions, notably between 9.31 and 9.36. Thus, over 75% of the cumulative appreciation of the franc in the 20 minutes to 9.50 was attributable to bank AIs, which accounted for 61% of the volume of liquidity-consuming trades. Indeed, we show below that bank AIs accounted for an even larger share of the realised variance of the EUR/CHF rate at this time. The lower panel shows that bank AIs also provided liquidity for some of the extreme-price trades. That bank AIs both consumed and provided liquidity at extreme prices may reflect the diverse set of traders from whom these trades may originate. This includes not only the different banks but also their various clients. In addition, a roughly equal number of extreme-price trades were accommodated by human traders. Indeed, human traders accounted for a significantly higher share of liquidity-providing trades (50%) than they did for liquidity-consuming trades (19%) during the 20 minutes to 9.50 when the Swiss franc appreciated sharply.<sup>15</sup>

<sup>14</sup> We thank Alain Chaboud for pointing out to us this ‘Wile E. Coyote’ moment, in which a significant portion of the orders underpinning the value of the euro against the franc were withdrawn but it was not until seconds later that the price started to fall.

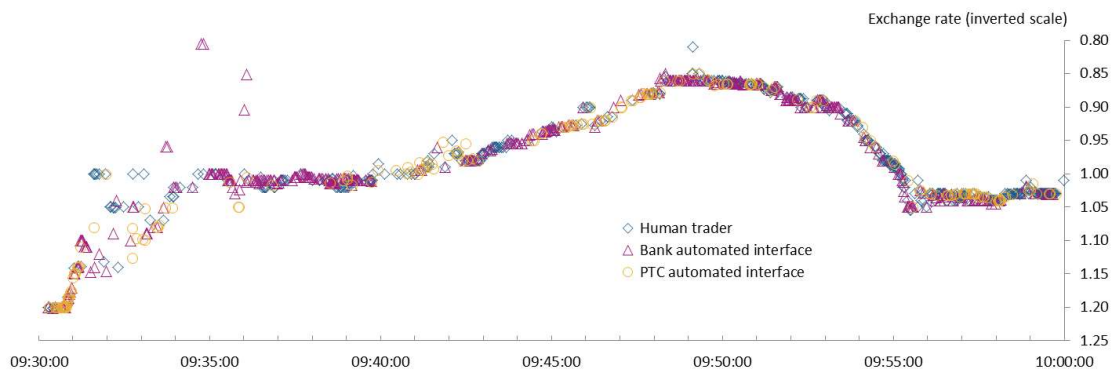
<sup>15</sup> Figure 4 excludes trades that we infer (as discussed later) may have involved the SNB, though we found few such trades in the first twenty minutes after the announcement.

**Figure 4: EUR/CHF trades in the minutes following the SNB announcement<sup>(1)</sup>**

*By type of trader consuming liquidity (i.e. supplied the IOC order)*



*By type of trader providing liquidity (i.e. supplied the GTC order)*

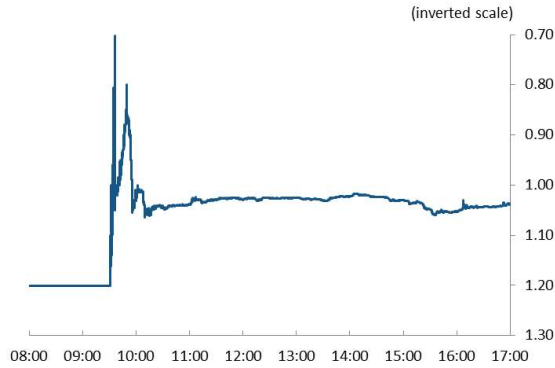


<sup>(1)</sup> Each data point plotted in the charts represents a simple average of trade prices within a given second. Due to averaging, the prices in the top and bottom panels need not be identical.  
Sources: EBS and authors' calculations.

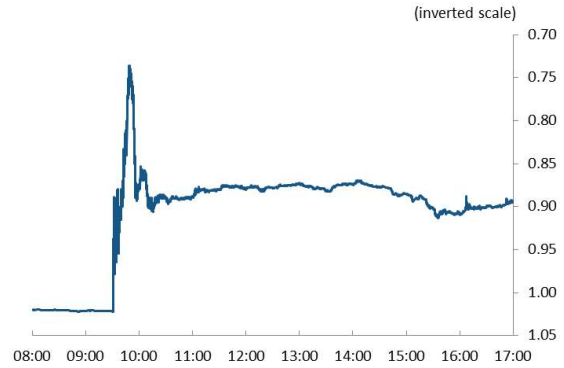
Figure 5 gives an overview of the reaction of both the EUR/CHF and USD/CHF markets to the SNB announcement over the whole trading day of 15 January 2015. The first row shows that the Swiss franc appreciated extremely sharply against both 'base' currencies in the first 20 minutes following the announcement, but that sizeable portions of these gains were reversed in the subsequent hour. After that the two spot rates were much more stable, with the Swiss franc worth about 10% more than at the start of the day. The second row shows that algorithmic traders were net purchasers of Swiss francs over the day, particular bank AIs against the euro and PTC AIs against the US dollar, while human traders were net purchasers of the base currencies. Thus, computers traded 'with the wind', buying the franc as it appreciated, while humans 'leaned against the wind'. Note, however, that human traders did not make net purchases of the base currencies in the key 20-minute period immediately after the announcement. Finally, the third row shows that human traders were consistently net suppliers of liquidity over the day, while PTC AI trades consumed it. However, as we shall see below, net liquidity consumption by PTC AIs is not unusual in these two currency-pairs.

**Figure 5: Market reaction on the day of the SNB announcement**

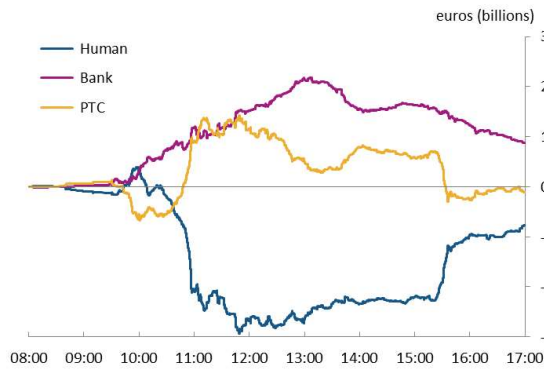
*EUR/CHF: spot exchange rate*



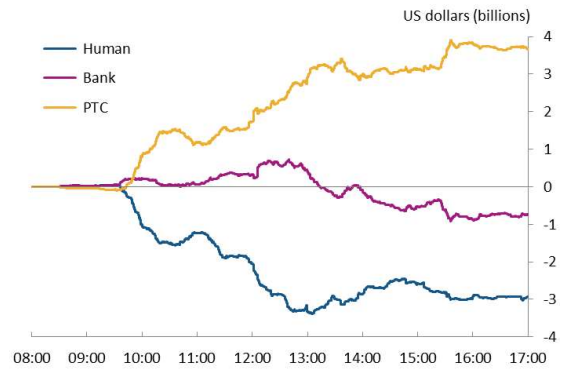
*USD/CHF: spot exchange rate*



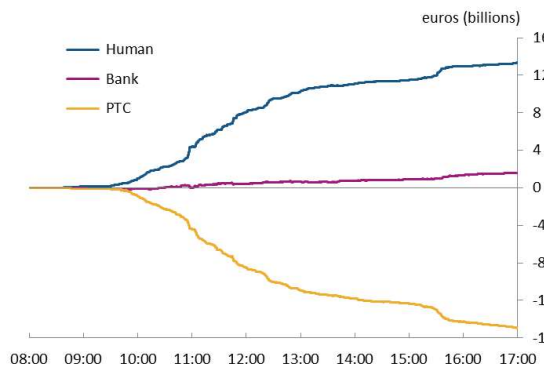
*EUR/CHF: cumulative net purchases of CHF*



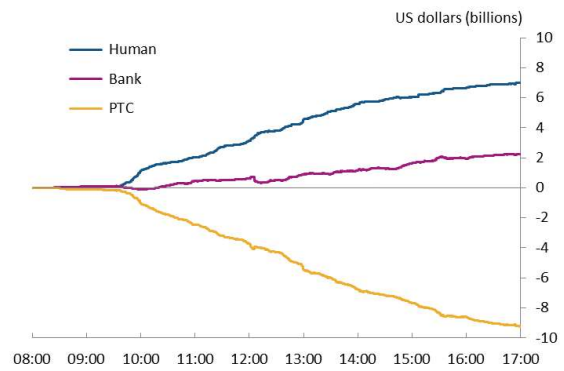
*USD/CHF: cumulative net purchases of CHF*



*EUR/CHF: cumulative net liquidity provision*



*USD/CHF: cumulative net liquidity provision*



Sources: EBS and authors' calculations.

Overall, our graphical overview suggests that algorithmic traders contributed both to the liquidity dry-up (as they were net consumers of liquidity) and the price move (as they were net purchasers of the appreciating currency) after the SNB announcement. In the following sections, we examine these issues in more detail.

## 5 Liquidity contributions of computer and human trading

In this section we investigate the contributions of computer and human traders to the liquidity of EUR/CHF and USD/CHF markets before, during and after the day of the SNB announcement. We study both quantity-based and price-based indicators of liquidity.

### 5.1 Volumes of liquidity provided and consumed

First, we focus on volumes of liquidity. As in Section 4, we identify whether the consumer and provider of liquidity for each trade was a human (H), a bank AI (B) or a PTC AI (P). We then record the shares of total trade volumes in three different periods for which providers and consumers of liquidity were H, B or P. The three periods are a ‘pre-event’ period (5-14 January 2015, excluding weekends), the ‘event day’ (15 January 2015) and a ‘post-event’ period (16-23 January 2015, excluding weekends). Finally, we record ‘net liquidity provision’, which is the difference between the share of trades for which a given type of trader provided liquidity and the share for which it consumed liquidity. The results are reported in Table 1.

**Table 1: Liquidity volumes by trader type**

<i>EUR/CHF</i>									
	Liquidity provider			Liquidity consumer			Net liquidity provision <sup>(1)</sup>		
	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI
<i>Share of trade volume (%)</i>									
Pre-event period	38.8	54.1	7.1	19.4	42.0	38.6	19.4	12.1	-31.5
Event day	68.4	25.1	6.5	34.9	26.7	38.4	33.5	-1.6	-31.9
Post-event period	50.1	35.3	14.5	21.5	23.3	55.1	28.6	12.0	-40.6
<i>Statistical tests (t-statistics)<sup>(2)</sup></i>									
Event day = pre-event?	6.0 <sup>***</sup>	-4.3 <sup>***</sup>	-0.3	11.6 <sup>***</sup>	-5.0 <sup>***</sup>	-0.1	3.1 <sup>***</sup>	-2.2 <sup>**</sup>	-0.1
Post-event = pre-event?	1.9 <sup>**</sup>	-2.5 <sup>***</sup>	3.1 <sup>***</sup>	1.1	-5.3 <sup>***</sup>	4.6 <sup>***</sup>	1.7 <sup>*</sup>	0.0	-2.7 <sup>***</sup>
<i>USD/CHF</i>									
	Liquidity provider			Liquidity consumer			Net liquidity provision <sup>(1)</sup>		
	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI
<i>Share of trade volume (%)</i>									
Pre-event period	3.8	15.4	69.1	16.6	31.5	40.3	-12.8	-16.1	28.9
Event day	8.9	25.6	60.3	34.1	34.6	26.2	-25.1	-9.0	34.1
Post-event period	7.7	24.9	56.3	27.6	39.0	22.4	-19.8	-14.1	34.0
<i>Statistical tests (t-statistics)<sup>(2)</sup></i>									
Event day = pre-event?	17.6 <sup>***</sup>	10.0 <sup>***</sup>	-8.8 <sup>***</sup>	20.3 <sup>***</sup>	2.4 <sup>**</sup>	-11.6 <sup>***</sup>	20.8 <sup>***</sup>	-6.6 <sup>***</sup>	-4.4 <sup>***</sup>
Post-event = pre-event?	8.7 <sup>***</sup>	5.9 <sup>***</sup>	-9.3 <sup>***</sup>	8.1 <sup>***</sup>	3.7 <sup>***</sup>	-9.0 <sup>***</sup>	5.6 <sup>***</sup>	-0.87	-2.9 <sup>***</sup>

<sup>(1)</sup> Share of volume as liquidity provider minus share of volume as liquidity consumer.

<sup>(2)</sup> \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.

Sources: EBS and authors' calculations.

The table shows that net liquidity provision by AIs fell on the event day and in the post-event period compared with the pre-event period, while it increased for humans. The increases in human net liquidity provision were all statistically significant, while the decreases for algorithmic traders were significant for at least one type of AI. In EUR/CHF, net liquidity

provision by bank AIs declined significantly on the event day, while it was that of PTC AIs that declined significantly in the post-event period. In USD/CHF, net liquidity provision by both types of AI declined significantly on the event day and that of PTC AIs declined significantly in the post-event period. Economically, the most significant changes were the decline in the share of bank AI liquidity-providing trades on the event day in EUR/CHF and the offsetting increase in the human share. Thus, although bank AIs also increased their share of liquidity-consuming trades on the event day, and humans reduced their share, this was not enough to change the pattern in net liquidity provision.

One possible explanation for the importance of human traders in supporting liquidity on the event day and in the post-event period is that the SNB was an active trader in this category. If that were the case, it would be hard to draw general conclusions from our analysis about the role of human traders in extreme events. Hence, we have estimated how the SNB may have traded during this period. We need to make estimates, based on assumptions, as our data from EBS is anonymous so we cannot identify from this any trades of the SNB or any other individual institution. We then repeat the analysis underpinning Table 1 excluding these estimated trades to see if the results are materially affected.

Our estimates of SNB trading activity are based on three assumptions. First, that SNB activity over our full sample period cannot have exceeded about 16% of the turnover on EBS. This figure is derived from public data on SNB sight deposits, which is the key liability created by FX interventions. It is an upper bound because other public documents suggest that the SNB favours a diversified approach to FX interventions (Moser, 2016), including substantial use of telephone orders (Fischer, 2004). Hence, only a share of any such activity may take place directly through EBS. The other assumptions are based on Fischer (2004), who notes that SNB interventions cluster around prices ending in 00 or 50, and that they tend to come in bursts of many transactions within short periods. Hence, we assume that bid orders ending in 00 or 50 originated from the SNB if there was a trade at the same price within 100 milliseconds with a human liquidity provider that was buying EUR. Our strategy does not estimate IOC orders placed by the SNB, so we may overestimate net liquidity provision by the SNB.

Table 2 shows adjusted results for liquidity provision and consumption by computers and humans excluding estimated SNB trades. It only shows results for EUR/CHF, as we made relatively more adjustments for this currency pair. These results are very similar to those in Table 1. In particular, we still find that human traders supported liquidity on the event day and in the post-event period, offsetting a reduction in net liquidity provision by computers. Hence, we proceed using the full data set in the rest of the paper.

**Table 2: Adjusted EUR/CHF liquidity volumes by trader type**

	Liquidity provider			Liquidity consumer			Net liquidity provision <sup>(1)</sup>		
	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI
<i>Share of trade volume (%)</i>									
Pre-event period	38.7	54.2	7.1	19.4	41.9	38.6	19.3	12.3	-31.5
Event day	63.9	28.7	7.5	33.5	25.6	40.9	30.4	3.1	-33.4
Post-event period	50.1	35.4	14.5	21.6	23.3	55.2	28.6	12.1	-40.6
<i>Statistical test (t-statistics)<sup>(2)</sup></i>									
Event day = pre-event?	5.1 <sup>***</sup>	-3.8 <sup>***</sup>	0.2	10.6 <sup>***</sup>	-5.4 <sup>***</sup>	0.9	2.4 <sup>**</sup>	-1.5	-0.7
Post-event = pre-event?	1.9 <sup>**</sup>	-2.5 <sup>**</sup>	3.1 <sup>***</sup>	1.1	-5.3 <sup>***</sup>	4.6 <sup>***</sup>	1.7 <sup>*</sup>	-0.0	-2.7 <sup>***</sup>

<sup>(1)</sup> Share of volume as liquidity provider minus share of volume as liquidity consumer.

<sup>(2)</sup> \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.

Sources: EBS and authors' calculations.

## 5.2 Effective spreads

We now turn our focus to a price-based indicator of liquidity. We saw in Section 5.1 that computers reduced their net volume of liquidity provision on the event day and afterwards, but did they widen bid-ask spreads on trades for which they did still provide liquidity? To investigate this question, we calculate a series of effective spreads,  $s$ , for each of our trader types:

$$s_{tk} = \frac{q_{tk}(p_{tk} - m_t)}{m_t}$$

where  $t$  indexes the time of the trade,  $k$  indexes the type of trader providing liquidity, *i.e.* supplying the GTC order (human, bank AI or PTC AI),  $q$  is a binary variable equal to +1 for trades in which the liquidity consumer was buying the base currency and -1 for trades in which it was selling it,  $p$  is the transaction price and  $m$  is the mid-point between the best bid and ask quotes from any type of trader in the 100 millisecond window in which the trade took place.

Table 3 shows median effective spreads for EUR/CHF and USD/CHF. We report median values as our calculated spreads include some extreme observations, particularly on the event day. For the same reason, the tests of equality of the medians reported in the table are non-parametric tests.<sup>16</sup> For EUR/CHF, median effective spreads were very similar across trader types in the pre-event period. They increased very sharply for all trader types on the event day, but significantly more so for computer trades than human trades, and they only contracted a little moving into the post-event period. The results are similar for USD/CHF. In this case, humans offered narrower spreads in the pre-event period, but all spreads again increased very sharply on the event day, particularly those of PTC AIs, and they remained much wider in the post-event period than in the pre-event period.

<sup>16</sup> In particular, we use K-sample tests to investigate equality across periods and Snedecor and Cochran (1989) tests to investigate equality across trader types.

**Table 3: Effective spreads by type of liquidity provider***EUR/CHF*

	Median spread ( <i>basis points</i> )			Statistical tests <sup>(1)</sup> ( <i>p-values</i> )	
	Human	Bank AI	PTC AI	Human = Bank AI?	Human = PTC AI?
Pre-event period	0.21	0.21	0.21	0.45	0.41
Event day	0.92	1.22	1.91	0.03**	0.00***
Post-event period	0.91	0.91	1.70	0.12	0.00***
Statistical tests <sup>(1)</sup> ( $\chi^2$ statistics)					
Event day = pre-event?	303.1***	413.3***	222.9***		
Post-event = pre-event?	830.8***	880.3***	826.6***		

*USD/CHF*

	Median spread ( <i>basis points</i> )			Statistical tests <sup>(1)</sup> ( <i>p-values</i> )	
	Human	Bank AI	PTC AI	Human = Bank AI?	Human = PTC AI?
Pre-event period	0.24	0.30	0.43	0.00***	0.00***
Event day	0.96	0.94	2.00	0.38	0.00***
Post-event period	0.86	1.05	1.80	0.00***	0.00***
Statistical tests <sup>(1)</sup> ( $\chi^2$ statistics)					
Event day = pre-event?	71.7***	73.3***	97.1***		
Post-event = pre-event?	429.6***	666.7***	864.2***		

<sup>(1)</sup> The tests are described in footnote 18. \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.  
Sources: EBS and authors' calculations.

Taking our quantity-based and price-based indicators of market liquidity together, we conclude that those classified as AI traders significantly reduced their net volume of liquidity provision and, where they did still provide liquidity, this was only at much wider spreads. In contrast, human traders significantly increased their net volume of liquidity provision relative to AIs and did so at narrower spreads than algorithmic traders.

## 6 Impact of computer and human trading on volatility

A second important dimension of market quality, alongside liquidity, is pricing efficiency. In theory, in an efficient market any gaps that might arise between the actual price of an asset and its 'efficient price', which reflects its fundamental value, tend to be small and closed quickly by traders drawing on the available information about the fundamental value.<sup>17</sup> As a result there would be little excess volatility in prices, *i.e.* on top of that attributable to changes in the efficient price. In this section, we will examine the contributions of different trader types to efficient pricing of the EUR/CHF and USD/CHF exchange rates around the SNB announcement. However, we begin relatively simply by looking at contributions of different trader types to the realised variance of returns.

<sup>17</sup> As Hendershott and Menkveld (2014) note, the net provision of limit orders need not be a good measure of liquidity supply, since a market order that leans against price pressure (goes against the prevailing market trend) can be thought of as a contribution to liquidity and reducing volatility.

## 6.1 Contributions to realised volatility

Following O’Hara and Ye (2011), we calculate contributions of different trader types to the realised variance of returns in our pre-event, event day and post-event periods as:

$$V_k = \frac{\sum_{n=1}^{N_t} (r_n d_{nk})^2}{\sum_{n=1}^{N_t} r_n^2}$$

where  $k$  indexes our different types of trader (human, bank AI or PTC AI) and  $n$  indexes the return observations, of which there are  $N$  in total;  $r$  denotes the returns, which are logarithmic returns derived from successive transaction prices, and  $d$  is a dummy variable that equals one if the trader initiating the trade (*i.e.* consuming liquidity) is of a particular type and is otherwise zero. So, for example, if prices changed by 2% during a particular period as a result of two trades, one by a computer that moved the price by 1% and another by a human trader that moved the price by a further 1%, then each type of trader would have contributed 50% of the realised variance in that period. Results of this breakdown applied to EUR/CHF and USD/CHF during our three periods are shown in the left-hand panel of Table 4.

**Table 4: Contributions to realised variance**

<i>Share of total variance</i>				<i>Per-trade share of variance</i>			
<i>Per cent</i>	Human	Bank AI	PTC AI	<i>Normalised<sup>(1)</sup></i>	Human	Bank AI	PTC AI
<i>EUR/CHF</i>				<i>EUR/CHF</i>			
Pre-event period	17.4	45.0	37.6	Pre-event period	1.14	1.14	0.83
Event day	1.0	90.7	8.3	Event day	0.05	3.55	0.15
Post-event period	17.7	46.8	35.6	Post-event period	0.94	1.81	0.64
<i>USD/CHF</i>				<i>USD/CHF</i>			
Pre-event period	5.3	17.7	77.1	Pre-event period	0.93	0.95	1.02
Event day	6.8	33.7	59.5	Event day	0.62	1.22	0.97
Post-event period	8.3	26.4	65.3	Post-event period	0.88	0.98	1.03

<sup>(1)</sup> Such that the average variance per trade across all three periods and all trader types is one.  
Sources: EBS and authors’ calculations.

The results for EUR/CHF show a remarkable increase in the variance contribution of bank AIs on the event day. This jumped to over 90% of the total, before the variance shares of all trader types returned close to pre-event levels in the post-event period. The results for USD/CHF show a similar pattern, but the event-day increase in the bank AI share is much smaller and the variance shares of the different trader types do not fully revert to pre-event levels in the post-event period. As bank AIs include both proprietary algorithmic trading by banks (estimated to be about 70% of total bank AI trading) and computer systems that aggregate and transact client orders (estimated to be about 30% of total bank AI trading), either of these could have been responsible for the event-day jumps in variance contributions. To the extent that PTC AIs run similar proprietary trading algorithms to those of banks, the fact we see a different pattern for bank AI and PTC AI suggests that aggregators may have played an important role.

The breakdown in the left-hand panel of Table 4 is effectively a combination of the share of total trading of each type of trader and the per-trade impact on volatility of each type of trade.



Changes in variance contributions could therefore simply reflect changes in trading shares. We therefore calculate a per-trade variance impact coefficient, which simply scales variance contributions by the number of trades undertaken by the different trader types.

The right-hand panel of Table 4 shows the results of this analysis. The remarkably high contribution of bank AIs to EUR/CHF volatility on the event day is still present on a per-trade basis, as is an increase in its contribution to USD/CHF volatility. This panel also shows that the contribution to volatility of human trades declined on a per-trade basis for both EUR/CHF and USD/CHF, particularly on the event day itself. We obtain the same qualitative results if we normalise by trade volume rather than number of trades.

As a complement to our per-trade variance results, Table 5 shows estimated price-impact coefficients (Kyle, 1985). These were derived by regressing five-minute returns on the net order flow for each type of trader during the same five-minute periods. Specifically, returns were computed as logarithmic returns (on the base currency) between successive mid-points of the best bid and ask quotes in the final 100 milliseconds of each period. Net order flow was computed as the difference between liquidity-consuming purchases of the base currency and liquidity-consuming sales of the base currency by each trader type.

**Table 5: Price impact coefficients<sup>(1)</sup>**

	<i>EUR/CHF</i>			<i>USD/CHF</i>		
	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI
Pre-event period	0.001	0.000	0.002 <sup>***</sup>	0.064 <sup>**</sup>	0.073 <sup>***</sup>	0.034 <sup>***</sup>
Event day	-0.102	0.000	0.118	0.716	1.421 <sup>***</sup>	-0.371
Post-event period	0.454 <sup>***</sup>	0.480 <sup>***</sup>	0.151 <sup>***</sup>	0.244 <sup>**</sup>	0.516 <sup>***</sup>	-0.023

<sup>(1)</sup> \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.

Sources: EBS and authors' calculations.

The results share some similarities with the right-hand panel of Table 4. In particular, bank AI order flow had the largest price impacts on both EUR/CHF and USD/CHF in the post-event period, as well as on USD/CHF on the event day itself. However, the price impact of bank AI order flow was effectively zero in EUR/CHF on the event day, while bank AI trades generated the most volatility. This suggests that bank AIs contributed a lot of uninformative ‘noise’ to the EUR/CHF market on the event day. We investigate the information contributions of our three trader types in more detail in the next section.

## 6.2 Contributions to efficient pricing

Since it would be possible to argue that any increase in volatility driven by AI or human trading was valuable if it helped market prices to fully reflect available information, it would be desirable to view the results in Section 6.1 in combination with estimated contributions of different trader types to the formation of efficient prices. We estimate the latter using a variant

of the vector autoregression (VAR) model developed in Hasbrouck (1991(a), 1991(b), 2007) and employed in Hendershott *et al.* (2011).<sup>18</sup> Specifically, we estimate the following model:

$$\begin{aligned}
r_t &= \sum_{k=1}^K \alpha_k r_{t-k} + \sum_{k=0}^K \beta_{k,P} x_{t-k,P} + \sum_{k=0}^K \beta_{k,B} x_{t-k,B} + \sum_{k=0}^K \beta_{k,H} x_{t-k,H} + \varepsilon_{r,t} \\
x_{t,P} &= \sum_{k=1}^K \gamma_k r_{t-k} + \sum_{k=1}^K \delta_{k,P} x_{t-k,P} + \sum_{k=0}^K \delta_{k,B} x_{t-k,B} + \sum_{k=0}^K \delta_{k,H} x_{t-k,H} + \varepsilon_{P,t} \\
x_{t,B} &= \sum_{k=1}^K \zeta_k r_{t-k} + \sum_{k=1}^K \eta_{k,P} x_{t-k,P} + \sum_{k=1}^K \eta_{k,B} x_{t-k,B} + \sum_{k=0}^K \eta_{k,H} x_{t-k,H} + \varepsilon_{B,t} \\
x_{t,H} &= \sum_{k=1}^K \lambda_k r_{t-k} + \sum_{k=1}^K \nu_{k,P} x_{t-k,P} + \sum_{k=1}^K \nu_{k,B} x_{t-k,B} + \sum_{k=1}^K \nu_{k,H} x_{t-k,H} + \varepsilon_{H,t}
\end{aligned}$$

where  $r$  denotes returns on the base currency and  $x$  denotes its order flows, *i.e.* net liquidity-consuming purchases, both are calculated over five-minute periods, indexed by  $t$ . More specifically, returns are logarithmic returns based on the mid-point of the best bid and ask quotes in the final 100 milliseconds of the current and previous periods, and order flows are computed separately for the different types of trader.

The only structural assumptions in this model relate to timings. Thus, we assume that PTC AIs – as the fastest traders in the market – can adjust their net orders to the contemporary order flow of other market participants. Similarly, bank AIs – as the next fastest trader type – can adjust their net orders to the contemporary order flow of human traders, but not to that of PTC AIs. Finally, human traders can only adjust their net orders to the previous order flows of other market participants.<sup>19</sup> These assumptions are in a similar vein to those of Brogaard *et al.* (2014) in their study of HFT and non-HFT trading activity.

We estimate this model, selecting  $K = 5$  as the optimal number of lags, and transform it to a vector moving-average representation by repeatedly substituting for the right-hand side terms.<sup>20</sup> The resulting equation for returns is:

$$r_t = (\varepsilon_{r,t} + \sum_{k=1}^{\infty} a_k \varepsilon_{r,t-k}) + \sum_{k=0}^{\infty} b_{k,P} \varepsilon_{P,t-k} + \sum_{k=0}^{\infty} b_{k,B} \varepsilon_{B,t-k} + \sum_{k=0}^{\infty} b_{k,H} \varepsilon_{H,t-k}$$

As suggested by Hendershott *et al.* (2011), the last three terms may be considered ‘private information’ as they reflect order flows from particular trader types, while the first term may be considered ‘public information’. Thus, we can identify separate contributions to efficient pricing from public information and private information pertaining to each of our three trader types via:<sup>21</sup>

$$\sigma_r^2 = (1 + \sum_{k=1}^{\infty} a_k)^2 \sigma_{\varepsilon,r}^2 + (\sum_{k=0}^{\infty} b_{P,k})^2 \sigma_{\varepsilon,P}^2 + (\sum_{k=0}^{\infty} b_{B,k})^2 \sigma_{\varepsilon,B}^2 + (\sum_{k=0}^{\infty} b_{H,k})^2 \sigma_{\varepsilon,H}^2$$

<sup>18</sup> The model is described in some detail on pages 78-85 of Hasbrouck (2007).

<sup>19</sup> This is the most logical ordering, but even with alternative orderings the pattern of results across periods remains similar.

<sup>20</sup> We did this ten times, by when the marginal effect of each substitution had become small.

<sup>21</sup> To be clear, this follows from the assumptions of the model, and not from information in the market data provided by EBS.

where  $\sigma_r^2$  is the overall variance of returns, and the terms on the right-hand side respectively represent contributions to this from public information and private information pertaining to PTC AI, bank AI and human traders.  $\sigma_{\varepsilon,r}^2$ ,  $\sigma_{\varepsilon,p}^2$ ,  $\sigma_{\varepsilon,B}^2$ ,  $\sigma_{\varepsilon,H}^2$  denote the variances of return shocks and shocks to order flows of each trader type. Table 6 presents the results of this decomposition.

**Table 6: Estimated contributions to variance of efficient returns**

<i>EUR/CHF</i>					<i>USD/CHF</i>				
<i>Per cent</i>	Returns	Order flow			<i>Per cent</i>	Returns	Order flow		
		Human	Bank AI	PTC AI			Human	Bank AI	PTC AI
Pre-event period	63.8	4.4	0.3	31.5	Pre-event period	79.5	4.3	11.0	5.2
Event day <sup>(1)</sup>	11.8	69.2	18.1	0.9	Event day <sup>(1)</sup>	19.2	17.9	26.1	36.9
Post-event period	39.9	27.3	19.4	13.4	Post-event period	85.3	3.1	11.3	0.2

<sup>(1)</sup> A small number of the most extreme returns on the event day were excluded to avoid these driving the results. Sources: EBS and authors' calculations.

The results for EUR/CHF show a striking shift in information contributions across our three periods. In the pre-event period, PTC AIs are estimated to have made by far the largest contribution to the variance of efficient returns of all types of order flows, with bank AIs and human traders contributing very little. On the event day, human trading took over as the most significant contributor, while the influence of PTC AIs all but disappeared. Human trading also maintained a significant contribution in the post-event period. The role of bank AIs on the event day and in the post-event period was similar to that of human traders, though not as dramatic, stepping up from pre-event levels. This was dwarfed by the increased contribution of bank AI to total realised volatility highlighted in Section 6.1.

The pattern for USD/CHF is somewhat less clear, as the informational role of all types of trading is estimated to be relatively small in both the pre-event and post-event periods. When the contribution of public information collapsed on the event day, however, order flows did temporarily become much more informative, particularly those from AI trades.

This section suggests that human traders played an important role in the discovery of efficient prices, notably in the EUR/CHF market on the event day. Here, they substituted for computers, notably PTC AIs, which still contributed less to price discovery than human traders in the post-event period. Combining these results with previous ones, showing that AI traders made large contributions to realised volatility, we conclude that these traders added significant noise to FX rates following the SNB announcement. This may reflect the possibility that many computer trades after the announcement were driven by liquidity needs rather than information. Such 'fire-selling' would be consistent computers being net consumers of liquidity, as we saw in Section 5.

## 7 Arbitrage opportunities and market efficiency

Studies have found that computer trading algorithms sometimes help to iron out market imperfections such as arbitrage opportunities. In this section, we present measures of market efficiency relating to triangular arbitrage to see if there were changes in the efficiency of Swiss franc currency markets as computer traders withdrew liquidity following the SNB announcement on 15 January 2015.

The Swiss franc is one of the few currencies continuously quoted directly against both the euro and the US dollar. Some computer trading in these markets may therefore be engaged in triangular arbitrage. This involves searching for and trading on instances of direct quotes for EUR/CHF that have moved out of line with implied quotes derived from USD/CHF and EUR/USD.

We begin by calculating the frequency and average size of such arbitrage opportunities on the day of the SNB announcement and in the periods preceding and following it. Specifically, we examine the best bid and ask quotes in each 100 millisecond window and record the existence of an arbitrage opportunity if a profit could have been made by buying EUR/CHF directly and selling ‘synthesised’ EUR/CHF via USD/CHF and EUR/USD trades or *vice versa*. The profits have to exceed a *de minimus* one basis point, and we record the average profitability of all arbitrage opportunities meeting this criterion. The results are shown in the left-hand panel of Table 7.

**Table 7: Arbitrage opportunities between EUR/CHF, USD/CHF and EUR/USD**

<i>Size and frequency of opportunities</i>			<i>Trading on arbitrage opportunities<sup>(1)</sup></i>		
	Frequency <sup>(2)</sup> <i>Per cent</i>	Profitability <sup>(3)</sup> <i>Basis points</i>	<i>Coefficients</i>	Pre-event period	Post-event period
Pre-event period	0.004	9.9	Human	-0.0022	0.0010
Event day	0.897	181.2	Bank AI	0.0061***	0.0025
Post-event period	0.046	4.8	PTC AI	0.0089***	-0.0050

(1) \*\*\* / \*\* / \* denotes statistical significance at the 1% / 5% / 10% level.

(2) Percentage of 100 millisecond periods in which the combination of best bid and ask quotes across the three currency-pairs offers a profit in excess of one basis point.

(3) Average profitability of arbitrage opportunities where they exist.

Sources: EBS and authors’ calculations.

Not surprisingly, by far the largest and most frequent arbitrage opportunities occurred during the event day itself. Arbitrage opportunities then remained over ten times more frequent in the post-event period compared with the pre-event period. This suggests that algorithmic trading may have become less active in this latter period, possibly in response to the increased volatility of the two CHF rates in the arbitrage triangle.

To investigate more thoroughly how the role of algorithmic trading in arbitrage in the post-event period compares with that of the pre-event period, we estimate a structural vector autoregression (SVAR) model of the relationship between arbitrage opportunities and the

trading volumes of different types of trader. This analysis closely follows that of Chaboud *et al.* (2014). In particular, we estimate the following model:

$$AY_t = \alpha(L)Y_t + \beta X_t + \delta G_t + \epsilon_t$$

where  $Y$  contains four endogenous variables, the first of which measures the frequency of arbitrage opportunities, while the remaining ones measure the order flow of each trader type relative to total market order flow. These variables are measured over five-minute windows.  $A$  is a  $4 \times 4$  matrix of coefficients governing contemporaneous relationships between the endogenous variables. These were estimated using the approach of Rigobon (1993). Two lags of the endogenous variables are also included in the model, as are six exogenous variables,  $X$ . These are total trade volumes and return volatilities for each of the three currency pairs in the arbitrage triangle, all computed over the preceding ten minutes. Finally, we include nine time dummy variables,  $G$ , one for each hour of the trading day.

The right-hand panel of Table 7 shows the estimated contemporaneous coefficients that reflect how the trading activity of each type of trader responds to arbitrage opportunities. It reveals statistically significant positive responses of human, bank AI and PTC AI trading in the pre-event period, with both bank AI and PTC AI trading responding more than that of human traders. In the post-event period, however, human traders pursued triangular arbitrage opportunities more than bank AIs and PTC AIs, even though the relationship was weaker and no longer statistically significant. Thus, our results suggest a significant reduction in computer resources devoted to triangular arbitrage following the Swiss franc event, with human trading becoming the most important source of arbitrage (though not significantly so). The fact that no type of trading made a significant contribution to arbitrage suggests that most mis-pricings were closed by quote adjustment rather than active trading.

## 8 Non-CHF foreign exchange rates

Although we have some evidence that AI trading adversely affected market liquidity and price formation in exchange rates featuring the Swiss franc after the SNB announcement, an important question for financial stability is whether these effects spread to exchange rates more widely. If so, to what extent did AI trading undermine the quality of these FX markets? In order shed light on this issue, we now focus on three other currency pairs, EUR/USD, USD/JPY and EUR/JPY. We chose these three cross-rates as EUR/USD and USD/JPY are the two most traded currency pairs in all FX markets, and while EUR/USD is associated with EUR/CHF and USD/CHF through triangular arbitrage, USD/JPY is not. We added EUR/JPY to study arbitrage in another triangle of currencies.

First we repeat the analysis of liquidity provision and consumption of Section 5.1 for these three currency pairs. Here, we focus on net liquidity provision, which is shown in Table 8.

**Table 8: Net liquidity provision<sup>(1)</sup>**

	EUR/USD			USD/JPY			EUR/JPY		
	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI	Human	Bank AI	PTC AI
<i>Share of trade volume (%)</i>									
Pre-event period	20.9	5.0	-25.9	20.1	7.7	-27.8	35.3	18.5	-53.8
Event day	17.7	-2.1	-15.6	18.2	-2.4	-15.8	31.9	18.4	-50.4
Post-event period	18.5	4.1	-22.6	19.6	7.7	-27.3	36.6	20.4	-57.0
<i>Statistical tests (t-statistics)<sup>(2)</sup></i>									
Event day = pre-event?	-3.3 <sup>***</sup>	-6.7 <sup>***</sup>	9.2 <sup>***</sup>	-1.6	-8.0 <sup>***</sup>	7.4 <sup>***</sup>	-2.1 <sup>*</sup>	-0.1	1.8
Post-event = pre-event?	-1.3	-0.6	1.4 <sup>*</sup>	-0.3	0.0	0.1	0.3	0.6	-0.8

<sup>(1)</sup> Share of volume as liquidity provider minus share of volume as liquidity consumer.

<sup>(2)</sup> \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.

Sources: EBS and authors' calculations.

**Table 9: Effective spreads by type of liquidity provider***EUR/USD*

	Median spread ( <i>basis points</i> )			Statistical tests <sup>(1)</sup> ( <i>p-values</i> )	
	Human	Bank AI	PTC AI	Human = Bank AI?	Human = PTC AI?
Pre-event period	0.17	0.21	0.34	0.00 <sup>***</sup>	0.00 <sup>***</sup>
Event day	0.22	0.30	0.51	0.00 <sup>***</sup>	0.00 <sup>***</sup>
Post-event period	0.20	0.25	0.42	0.00 <sup>***</sup>	0.00 <sup>***</sup>
<i>Statistical tests<sup>(1)</sup> (<math>\chi^2</math> statistics)</i>					
Event day = pre-event?	42.6 <sup>***</sup>	62.3 <sup>***</sup>	84.4 <sup>***</sup>		
Post-event = pre-event?	51.5 <sup>***</sup>	101.8 <sup>***</sup>	160.7 <sup>***</sup>		

*USD/JPY*

	Median spread ( <i>basis points</i> )			Statistical tests <sup>(1)</sup> ( <i>p-values</i> )	
	Human	Bank AI	PTC AI	Human = Bank AI?	Human = PTC AI?
Pre-event period	0.14	0.24	0.35	0.00 <sup>***</sup>	0.00 <sup>***</sup>
Event day	0.18	0.27	0.47	0.00 <sup>***</sup>	0.00 <sup>***</sup>
Post-event period	0.14	0.26	0.41	0.00 <sup>***</sup>	0.00 <sup>***</sup>
<i>Statistical tests<sup>(1)</sup> (<math>\chi^2</math> statistics)</i>					
Event day = pre-event?	18.7 <sup>***</sup>	11.8 <sup>***</sup>	45.6 <sup>***</sup>		
Post-event = pre-event?	0.6	13.2 <sup>***</sup>	57.5 <sup>***</sup>		

*EUR/JPY*

	Median spread ( <i>basis points</i> )			Statistical tests <sup>(1)</sup> ( <i>p-values</i> )	
	Human	Bank AI	PTC AI	Human = Bank AI?	Human = PTC AI?
Pre-event period	0.22	0.36	0.45	0.00 <sup>***</sup>	0.00 <sup>***</sup>
Event day	0.46	0.56	0.92	0.91	0.00 <sup>***</sup>
Post-event period	0.28	0.48	0.64	0.00 <sup>***</sup>	0.00 <sup>***</sup>
<i>Statistical tests<sup>(1)</sup> (<math>\chi^2</math> statistics)</i>					
Event day = pre-event?	28.3 <sup>***</sup>	19.0 <sup>***</sup>	73.9 <sup>***</sup>		
Post-event = pre-event?	14.8 <sup>***</sup>	20.2 <sup>***</sup>	60.0 <sup>***</sup>		

<sup>(1)</sup> The tests are described in footnote 18. \*\*\* / \*\* / \* denote statistical significance at the 1% / 5% / 10% level.

Sources: EBS and authors' calculations.

The table shows no significant change in the net supply of liquidity in non-CHF currency pairs in the post-event period compared with the pre-event period. On the event day, there was some decline in net liquidity provision in EUR/USD and USD/JPY by bank AIs, but this was compensated for by an increase from PTC AIs.

Second, we look at effective spreads, as in Section 5.2. Table 9 shows how effective spreads varied on and after the day of the SNB announcement by type of liquidity provider for our three non-CHF currency pairs. Across these currency pairs, spreads on human trades were always narrower than spreads on computer trades, but all spreads widened on the event day and reverted towards pre-event values in the post-event period, such that the relative positions of spreads with human and computer traders did not change.

Finally, we repeat our analysis of the frequency and size of arbitrage opportunities in Section 8.1 for the USD/JPY-EUR/JPY-EUR/USD triangle. If the Swiss franc event had a widespread impact on computer trading in all FX cross-rates we might expect more arbitrage opportunities to have appeared in this triangle. However, as Table 10 shows, although there was some increase in arbitrage opportunities on the event day, which was statistically significant, they were almost identical in the pre-event and post-event periods.

**Table 10: Arbitrage opportunities between USD/JPY, EUR/JPY and EUR/USD**

	Frequency <sup>(1)</sup> <i>Per cent</i>	Profitability <sup>(2)</sup> <i>Basis points</i>
Pre-event period	0.012	0.2
Event day	0.494	1.9
Post-event period	0.014	0.2

<sup>(1)</sup> Percentage of 100 millisecond periods in which the combination of best bid and ask quotes across the three currency-pairs offers a profit in excess of one basis point.

<sup>(2)</sup> Average profitability of arbitrage opportunities where they exist.

Sources: EBS and authors' calculations.

## 9 Conclusion

The Swiss franc event is probably the most significant shock to FX markets since computerised algorithmic trading has been prominent. Studying the reaction to this shock, we find that algorithmic trading contributed to the decline of EUR/CHF and USD/CHF market quality on the event day and afterwards as they withdrew liquidity and generated uninformative volatility. Human traders took over as the main contributors to efficient pricing, while algorithms tended to amplify price movements by following trends. Trades by bank algorithms, in particular, contributed substantially to non-informative EUR/CHF volatility. This may indicate that the role of trade aggregators, which has been highlighted in other extreme market events (see for example BIS, 2017), was important. Both PTC and bank algorithms, which had traded on

triangular arbitrage opportunities before the event, ceased doing so afterwards. However, adverse effects on non-CHF currency pairs were limited, suggesting that algorithmic trading did not propagate contagion in the FX market, at least on this occasion.

Of course, it is hard to draw general conclusions from one event, not least because we have only studied *how* algorithmic trading reacted and not *why* it did so. Was it due to more stringent capital or trading requirements applied to algorithmic trading or because of the behaviour of the algorithms themselves? If the latter, were there types of algorithm that reacted more and types that reacted less or even partially offset the behaviour of the others? If that were the case, the mix of algorithms operating at the time of future financial market shocks could affect the scale of any amplification that occurs. Indeed, that mix could be affected by the adaptation of algorithms as they experience periods of market stress like the one studied in this paper. Nevertheless, our results contrast with evidence that algorithmic trading in aggregate improves liquidity and price discovery in normal times. This suggests there is some value in maintaining a diversity of trader types to help keep markets resilient through different trading conditions.



## 10 References

- Afonso, G., Kovner, A. and Schoar, A. (2011), ‘Stressed, not frozen: the federal funds market in the financial crisis’, *Journal of Finance*, vol. 66, pp. 1109–1139.
- Ait-Sahalia, Y. and Saglam, M. (2014), ‘High-frequency traders: taking advantage of speed’, *working paper*.
- Anand, A., and Venkataraman, K. (2014), ‘Should exchanges impose market maker obligations?’, *working paper*.
- Baldauf, M. and Mollner, J. (2015), ‘High-frequency trading and market performance’, *working paper*.
- Bernales, A. (2014), ‘How fast can you trade? High-frequency trading in dynamic limit order markets’, *Banque de France working paper*.
- Biais, B. and Foucault, T. (2014), ‘HFT and market quality. Bankers, markets and investors’, *manuscript*.
- Biais, B., Foucault, T. and Moinas, S. (2015), ‘Equilibrium fast trading’, *Journal of Financial Economics*, vol. 116, pp. 292-313.
- Biais B. and Woolley, p. (2011), ‘High-frequency trading’, *Toulouse University manuscript*.
- BIS (2017), ‘The sterling ‘flash event’ of 7 October 2016’, *Bank for International Settlements Market Committee report*.
- Bloomberg (2015), ‘SNB unexpectedly gives up cap on Franc’, 15 January.
- Boehmer, E., Fong, K. and Wu, J. (2015), ‘International evidence on algorithmic trading’, *working paper*.
- Boehmer, E., Li, D. and Saar, G. (2017), ‘The competitive landscape of high-frequency trading firms’, *working paper*.
- Bongaerts, D. and Van Achter, M. (2016), ‘High-frequency trading and market stability’, *working paper*.
- Breckenfelder, J. (2013), ‘Competition between high-frequency traders and market quality’, *working paper*.
- Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A. and Sokolov, K. (2017), ‘High-frequency trading and extreme price movements’, *working paper*.
- Brogaard, J., Hagstromer, B., Norden, L. and Riordan, R. (2015), ‘Trading fast and slow: co-location and market quality’, *Review of Financial Studies*, vol. 28, pp. 3407–3443.
- Brogaard, J., Hendershott, T. and Riordan, R. (2014), ‘High-frequency trading and price discovery’, *Review of Financial Studies*, vol. 27, pp. 2267–2306.

- Budish, E., Cramton, P. and Shim, J. (2015), 'The high-frequency trading arms race: frequent batch auctions as a market design response', *Quarterly Journal of Economics*, vol. 130, pp. 1547–1621.
- Carrion, A. (2017), 'Very fast money: high-frequency trading on NASDAQ', *Journal of Financial Markets*, forthcoming.
- Cespa, G. and Vives, X. (2015), 'The beauty contest and short-term trading', *Journal of Finance*, vol. 70, pp. 2099-2154.
- Chaboud, A., Chiquoine, B., Hjalmarsson, E. and Vega, C. (2014), 'Rise of the machines: algorithmic trading in the foreign exchange market', *Journal of Finance*, vol. 69, pp. 2045-2084.
- CFTC-SEC Staff Report (2010), 'Findings regarding the market events of May 6, 2010'.
- Chordia T., Goyal, A., Lehmann B. and Saar G. (2013), 'High-frequency trading', *Journal of Financial Markets*, vol. 16, pp. 637-645.
- Conrad, J., Wahal S. and Xiang, J. (2015), 'High-frequency quoting, trading, and the efficiency of prices', *Journal of Financial Economics*, vol. 116, pp. 271-291.
- Du, S. and Zhu, H. (2015), 'Welfare and optimal trading frequency in dynamic double auctions', *National Bureau of Economic Research working paper*, no. 20588.
- Easley, D., Lopez de Prado, M. and O'Hara, M. (2012), 'Flow toxicity and liquidity in a high-frequency world', *Review of Financial Studies*, vol. 25, 1457–1493.
- Easley, D., Lopez de Prado, M. and O'Hara, M. (2013), 'High-frequency trading – new realities for traders, markets and regulators', *Risk Books*, London.
- Fischer A. (2004), 'Price clustering in the FX market: a disaggregated analysis using central bank interventions', *Centre for Economic Policy and Research discussion paper*, no. 4529.
- Foucault, T., Hombert, J. and Rosu, I. (2016), 'News trading and speed', *Journal of Finance*, vol. 71, 335-382.
- Goldstein, M., Kumar, P. and Graves, F. (2014), 'Computerized and high-frequency trading', *The Financial Review*, vol. 49, 177-202.
- Gomber, P., Arndt, B., Lutat, M. and Uhle, T. (2011), 'High-frequency trading', *Goethe University working paper*.
- Hagströmer, B. and Menkveld, A. (2016) 'A network map of information percolation', *working paper*.
- Harris, L. (2012) 'Stop the high-frequency trader arms race', *Financial Times*, December 27.
- Hasbrouck, J. (1991a), 'Measuring the information content of stock trades', *Journal of Finance*, vol. 46, pp. 179–207.

- Hasbrouck, J. (1991b), 'The summary informativeness of stock trades: an econometric analysis', *Review of Financial Studies*, vol. 4, 571–595.
- Hasbrouck, J. (2007), 'Empirical market microstructure', *Oxford University Press*, New York.
- Hasbrouck, J. and Saar, G. (2013), 'Low-latency trading', *Journal of Financial Markets*, vol. 16, pp. 646-679.
- Hendershott, T., Jones, C. and Menkveld, A. (2011), 'Does algorithmic trading improve liquidity?' *Journal of Finance*, vol. 66, 1-33.
- Hendershott, T. and Menkveld, A. (2014), 'Price pressures', *Journal of Financial Economics*, vol. 114, 405-423.
- Hoffmann, P. (2014), 'A dynamic limit order market with fast and slow traders', *Journal of Financial Economics*, vol. 113, pp. 156-169.
- Jones, C. (2013), 'What do we know about high-frequency trading?', *Columbia University manuscript*.
- Jovanovic, B. and Menkveld, A. (2016), 'Middlemen in limit-order markets', *working paper*.
- Kirilenko, A., Kyle, A., Samadi, M. and Tuzun, T. (2017), 'The flash crash: the impact of high-frequency trading on an electronic market', *Journal of Finance*, vol. 72, 967-998.
- Kirilenko, A. and Lo, A. (2013) 'Moore's Law versus Murphy's Law: algorithmic trading and its discontents', *Journal of Economic Perspectives*, vol. 27, pp. 51-72.
- Korajczyk, R. and Murphy D. (2016) 'High-frequency market making to large institutional trades', *working paper*.
- Jordan, T. (2014), 'Introductory remarks by Thomas Jordan', speech available [here](#).
- King, M., Osler, C. and Rime, D. (2011), 'Foreign exchange market structure, players and evolution', *Norges Bank working paper*.
- Kyle, A. (1985), 'Continuous auctions and insider trading', *Econometrica*, vol. 53, pp. 1315–1336.
- Lo, A. and MacKinlay, A. (1988) 'Stock market prices do not follow random walks: evidence from a simple specification test', *Review of Financial Studies*, vol. 1, pp. 41–66.
- Malinova, K., Park, A. and Riordan, R. (2013), 'Do retail traders suffer from high-frequency traders?', *working paper*.
- McCann, C. and Yan, M. (2015), 'The recent market turmoil spells trouble for "auto-liquidators" like interactive brokers', Securities Litigation & Consulting Group technical report, available [here](#).
- Menkveld, A. (2013), 'High-frequency trading and the new market makers', *Journal of Financial Markets*, vol. 16, pp. 712-740.

- Menkveld, A. (2016), ‘The economics of high-frequency trading: taking stock’, *Annual Review of Financial Economics*, vol. 8, pp. 1-24.
- Menkveld, A. (2017), ‘High-frequency trading as viewed through an electronic microscope’, *Financial Analysts Journal*, forthcoming.
- Menkveld, A. and Yueshen, B. (2015), ‘The flash crash: a cautionary tale about highly fragmented markets’, *working paper*.
- Mirkov, K., Pozdeev, I. and Söderlind, P. (2016), ‘Toward removal of the Swiss franc cap: market expectations and verbal interventions’, *Swiss National Bank working paper*, no. 10/16.
- Moser, D. (2016), ‘On the pulse of the financial markets’, speech at Geneva Money Market Event, available [here](#).
- Pagnotta, E. and Philippon, T. (2015), ‘Competing on speed’, *working paper*.
- O’Hara, M. (2015), ‘High-frequency market microstructure’, *Journal of Financial Economics*, vol. 116, pp. 257–270.
- O’Hara, M. and Ye, M. (2011), ‘Is market fragmentation harming market quality?’, *Journal of Financial Economics*, vol. 100, pp. 459–474.
- Raman, V., Robe, M. and Yadav, P. (2014), ‘Electronic market makers, trader anonymity and market fragility’, *Chicago Futures Trading Commission manuscript*.
- Reuters (2015a), ‘SNB’s Danthine says cap on franc remains policy cornerstone’, 12 January.
- Reuters (2015b), ‘Swiss franc shock shuts some FX brokers’, 16 January.
- Rigobon, R. (2003), ‘Identification through heteroskedasticity’, *Review of Economics and Statistics*, vol. 85, pp. 777-792.
- Rojcek, J. and Ziegler, A. (2016), ‘High-frequency trading in limit order markets: equilibrium impact and regulation’, *working paper*.
- Rosu, I. (2016), ‘Fast and slow informed trading’, *HEC (Paris) manuscript*.
- Securities Exchange Commission (2010), ‘Concept release on equity market structure’, release no. 34-61358, file no. S7-02-10.
- Swiss National Bank (2011), ‘Swiss National Bank sets minimum exchange rate at CHF 1.20 per euro’, available [here](#).
- Swiss National Bank (2015), ‘SNB discontinues minimum exchange rate and lowers interest to –0.75%’, available [here](#).
- Tong, L. (2015), ‘A blessing or a curse? The impact of high-frequency trading on institutional investors’, *working paper*.
- van Kervel, V. and Menkveld, A. (2016), ‘High-frequency trading around large institutional orders’, *working paper*.

# The Sound of Many Funds Rebalancing\*

Alex Chinco<sup>†</sup> and Vyacheslav Fos<sup>‡</sup>

February 8, 2018

## Abstract

This paper proposes that complexity generates noise in financial markets. A stock's demand might appear random, not because individual investors are behaving randomly, but because it's too computationally complex to predict how a wide variety of simple, deterministic, trading rules will interact with one another—even if you yourself are fully rational. There are two parts to our analysis. First, we illustrate how complexity can generate noise by modeling a particular kind of trading-rule interaction: index-fund rebalancing cascades. An initial shock to stock  $A$  causes an index fund to buy stock  $A$  and sell stock  $B$ , which then causes a second fund following a different benchmark to sell stock  $B$  and buy stock  $C$ , which then causes a third fund following yet another benchmark to... Although it's easy to predict *if* this index-fund rebalancing cascade will eventually affect the demand for an unrelated stock  $Z$ , predicting *how* stock  $Z$  will be affected (buy? or sell?) is computationally intractable. Second, we give empirical evidence that complexity actually does generate noise in real-world financial markets by analyzing the end-of-day holdings of exchange-traded funds (ETFs). We show that ETF rebalancing cascades transmit economically large demand shocks, which are also statistically unpredictable. And, we document market participants behaving as if these demand shocks are noise.

JEL CLASSIFICATION. G00, G02, G14

KEYWORDS. Noise, Indexing, Thresholds

---

\*We thank Kerry Back, Nick Barberis, Zahi Ben-David, James Choi, Adam Clark-Joseph, Tony Cookson, Xavier Gabaix, Itay Goldstein, Sam Hartzmark, Ralph Koijen, Pete Kyle, Chris Parsons, Jeff Pontiff, and Brian Weller, as well as seminar participants at CalTech, Colorado, Illinois, Maryland, Yale, the Young-Scholars Finance Consortium, the FINRA Market-Structure Conference, the SFS Calvalcade, the Conference on the Econometrics of Financial Markets, the Helsinki Behavioral Finance Conference, and the FRA Conference for extremely helpful comments and suggestions.

Current Version: <http://www.alexchinco.com/sound-of-rebalancing.pdf>

<sup>†</sup>University of Illinois at Urbana-Champaign, College of Business; [alexchinco@gmail.com](mailto:alexchinco@gmail.com).

<sup>‡</sup>Boston College, Carroll School of Management; [fos@bc.edu](mailto:fos@bc.edu).

# 1 Introduction

Imagine you're a trader who's just discovered that stock  $Z$  is under-priced. In a market without noise, there's no way for you to take advantage of this discovery. The moment you try to buy a share, other traders will immediately realize that you must have uncovered some good news. And, you won't find anyone willing to sell you a share at the old price (Aumann, 1976; Milgrom and Stokey, 1982).

Noise pulls the rug out from under this no-trade theorem. In a market with noise, someone may always be trying to buy or sell stock  $Z$  for erratic non-fundamental reasons. So, when you try to buy a share, other traders won't immediately realize that you've uncovered good news since your buy order could just be some more random noise. It's this plausible cover story that allows you to both trade on and profit from your discovery. It's this plausible cover story that Fischer Black was referring to when he wrote that "noise makes financial markets possible".

But, where exactly does this all-important noise come from? Who generates it? And, what are their erratic non-fundamental reasons for trading?

The standard answer to these questions is that i) noise comes from individual investors, and that ii) their demand looks erratic and unrelated to fundamentals because individual investors are just plain bad traders. These are the standard answers for a reason. Not only do individual investors suffer from all sorts of behavioral biases when they trade (Barberis and Thaler, 2003) but they trade far too often (Barber and Odean, 2000). So, it's clear that individual investors can generate noise.

But, are they the only source? It seems unlikely. The importance of individual investors has steadily declined over the past few decades. While individual investors held 47.9% of all U.S. equity in 1980, this percentage was down to only 21.5% by 2007 (French, 2008). And, in June 2017 JP Morgan strategists reported that only around "10% of trading is done by traditional, 'discretionary' traders, as opposed to systematic rules-based ones."<sup>1</sup> However, this steady downward trend in the importance of individual investors has not been accompanied by a drop in trading volume.

Motivated by this explanatory gap, we propose an alternative noise-generating mechanism based on computational complexity. A stock's demand might appear random, not because individual investors are behaving randomly, but because it's too computationally complex to predict how a wide variety of simple, deterministic, trading rules will interact with one another—even if you yourself are fully rational. There

---

<sup>1</sup>Financial Times. 6/14/2017. *Not Your Father's Market: Tech Tantrum Shows How U.S. Equities Trading Has Changed.*

are two parts to our analysis. First, we show theoretically how computational complexity can generate noise by modeling a particular kind of trading-rule interaction: index-fund rebalancing cascades. Then, we give empirical evidence that index-fund rebalancing cascades actually generate noise in real-world financial markets using data on the end-of-day holdings of exchange-traded funds (ETFs).

*Theoretical Model.* As individual investors have shrunk in importance, “passive investing—indexing—has become popular as an alternative to active investment management” and “active managers... have become more index-like in their investing (Stambaugh, 2014).” However, these new ‘index-like’ funds have not been created in Jack Bogle’s image. Many choose their holdings “based on custom criteria” that involve threshold-based rules.<sup>2</sup> For instance, the PowerShares S&P 500 Low-Volatility ETF [SPLV] tracks the lowest-volatility quintile of S&P 500 stocks. This fund uses a threshold-based rule because an arbitrarily small change in a stock’s volatility can move it from 101st to 100th place on the low-volatility leaderboard. When this happens, SPLV has to exit its position in one stock and build a new position in another, affecting each stock in equal-but-opposite ways. The price of the stock being added will rise while the price of the ‘stock formerly known as 100th’ will fall.

We begin our analysis by presenting a model where, because there are so many index funds tracking so many different threshold-based benchmarks, a small change in stock  $A$ ’s price can cause one index fund to buy stock  $A$  and sell stock  $B$ , which can then cause a second index fund following a different threshold-based benchmark to sell stock  $B$  and buy stock  $C$ , which can then cause... Our main theoretical result is that, although it’s possible to determine *if* a stock will be affected by one of these index-fund rebalancing cascades, the problem of determining *how* the stock will be affected (buy? or sell?) is computationally intractable. In fact, it’s NP hard. Thus, index-fund rebalancing cascades can generate seemingly random demand shocks even though each index fund involved in the cascade is following a completely deterministic trading rule. In other words, index-fund rebalancing cascades generate noise in a way that does not require traders to suffer from behavioral biases or make cognitive errors.

*Rebalancing Cascades.* Our theoretical model shows how index-fund rebalancing cascades are able to generate seemingly random demand shocks. But, is there any evidence that they are actually doing this in real-world financial markets? To answer this question, we study end-of-day ETF holdings using data from ETF Global, which covers every trading day from January 2010 to December 2015. We focus our attention

---

<sup>2</sup>Bloomberg. 5/12/2017. *There Are Now More Indexes Than Stocks.*

on ETFs that rebalance daily. So, when you look at our results, you should be thinking about the PowerShares S&P 500 Low-Volatility ETF [SPLV] rather than the SPDR S&P 500 ETF [SPY]. ETFs that rebalance daily are smaller than funds that track broad value-weighted market indexes, such as SPY. But, their rebalancing activity matters because these funds tend to do the bulk of their trading during the final 20-to-30 minutes of the trading day.<sup>3</sup> We also net-out changes in ETF holdings due to creations and redemptions. These trades are executed as in-kind transfers for tax reasons (Madhavan, 2016) and so can't contribute to index-fund rebalancing cascades.

Here's how we structure our tests. Our theoretical model studies index-fund rebalancing cascades that stem from an initial shock. So, we study ETF rebalancing cascades that stem from an initial M&A announcement, referring to the target of an M&A announcement as stock  $A$ . Our data on M&A announcements comes from Thomson Financial. M&A deals are a natural choice for the initial shocks because "a profusion of event studies has demonstrated that mergers seem to create shareholder value, with most of the gains accruing to the target company (Andrade et al., 2001)." While M&A targets are not randomly chosen, the exact date of the announcement (Tuesday? Wednesday? or Thursday?) may as well be.

Following each stock  $A$ 's announcement as an M&A target, we examine the set of unrelated stock  $Z$ s. For stock  $A$  and stock  $Z$  to be unrelated, they have to be twice removed in the network of ETF holdings at the time of the M&A announcement. Stock  $Z$  can't have been recently held by any ETF that also held stock  $A$ . And, if stock  $A$  and stock  $B$  both belong to the same benchmark, then stock  $Z$  can't have been recently held by any ETF that also held stock  $B$ . In other words, the chain has to be  $A \rightarrow B \rightarrow C \rightarrow Z$  or longer. Because there are smart-beta ETFs focusing on things like large-cap, value, and industry-specific benchmarks, this unrelatedness criteria also implies that stock  $Z$  doesn't share any well-known characteristics such as size, book-to-market, or industry with stock  $A$ .

Our theoretical model predicts that, all else equal, a stock  $Z$  that's on the cusp of more ETF rebalancing thresholds is more likely to be hit by an ETF rebalancing cascade. So, we split the set of stock  $Z$ s for each M&A announcement into two subsets: those that are on the cusp of rebalancing for an above-median number of ETFs, and those that aren't. Consistent with our theoretical prediction, we find that ETF rebalancing volume grows by 169% more for the above-median group of stock  $Z$ s than for the below-median group in the 5 days immediately following an M&A

---

<sup>3</sup>Wall Street Journal. 5/27/2015. *Stock-Market Traders Pile In at the Close*.

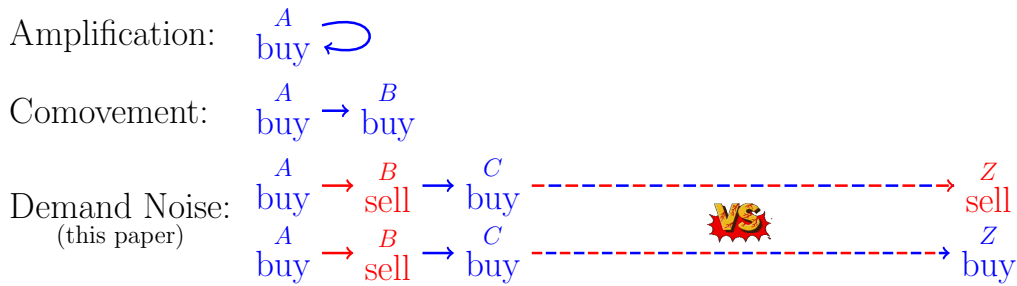


announcement. And, we show that this increase is no more likely to be made up of buy orders than of sell orders. Taken together, these results suggest that it's possible to predict *if* stock  $Z$  will be affected by an ETF rebalancing cascade but not *how* stock  $Z$  will be affected. What's more, because the same stock  $Z$  can be above-median relative to one M&A target but below-median relative to another, our results can't be attributed to unobserved characteristics of stock  $Z$ .

*Market Reaction.* To show that ETF rebalancing cascades are actually generating noise, however, we need to do more than just give statistical evidence that they are unpredictable. To be convincing, we need to show that market participants treat the resulting demand shocks as noise, too. The natural way to investigate whether market participants are treating the demand coming from ETF rebalancing cascades as noise is to look for differences in liquidity. As highlighted in the opening paragraphs, if there is more demand noise, then a randomly selected trade is less likely to be coming from an informed trader. So, we look for otherwise identical stock  $Z$ s and ask: 'Is stock  $Z_1$  more liquid than stock  $Z_2$  at times when stock  $Z_1$  happens to be more susceptible to ETF rebalancing cascades than stock  $Z_2$ ?' The data reveals that the answer is 'Yes'. Above-median stock  $Z$ s are much more liquid than the below-median stock  $Z$ s suggesting that market participants treat the demand shocks coming from ETF rebalancing cascades as noise.

*Broader Implications.* John Maynard Keynes (1921) pointed out that, because a daily national census is logistically impossible, the answer to the question 'Is the population of France an even or an odd number?' is effectively a coin toss. So, economists had intuited a connection between apparent randomness and computational complexity long before ETFs arrived on the scene. But, in the past, this intuitive connection has always been just that—an intuition. The goal of our theoretical model is to make the connection concrete.

By showing precisely why it's computationally intractable to predict ETF rebalancing cascades even if you yourself are fully rational, we make it possible for researchers to identify other situations where the same logic holds. For example, our theoretical model also applies to any other group of funds following a wide variety of threshold-based rebalancing rules. Think about quantitative hedge funds following strategies of the form 'Buy the top 30% and sell the bottom 30% of stocks when sorting on  $X$ ' (Khandani and Lo, 2007). Or, consider pension funds with strict portfolio mandates of the form '15% of our assets will be held in alternative investments' (Pennacchi and Rastad, 2011).



**Figure 1: How This Paper Is Different.** *Papers on index-linked investing fall into two groups. The first studies how trading due to index inclusion can amplify an initial shock to stock A (Row 1). The second studies how stock A’s returns suddenly co-move with stock B’s returns as soon as stock A gets added to stock B’s index (Row 2). By contrast, this paper focuses on the unpredictable consequences of stock A’s index inclusion, not for stock A or for stock B, but for seemingly unrelated stock Zs (Rows 3 and 4).*

## 1.1 Related Literature

This paper connects to three main strands of literature.

*Noise.* The problem we study is motivated by the central role that noise plays in information-based asset pricing models (Grossman and Stiglitz, 1980; Hellwig, 1980; Admati, 1985; Kyle, 1985) and limits-to-arbitrage models (Shleifer and Summers, 1990; Shleifer and Vishny, 1997; Gromb and Vayanos, 2010). We propose an explanation for noise that does not rely on individual investors behaving randomly.

*Indexing.* Our paper also relates to work on index-linked investing (Wurgler, 2010). Some of these papers study how index inclusion can amplify an initial shock to stock A. For instance, Chang et al. (2014) shows how getting added to the Russell 2000 results in further price increases. For examples involving ETFs, see Ben-David et al. (2017), Brown et al. (2016), and Israeli et al. (2017). Other papers study how stock A’s returns suddenly co-move with stock B’s returns as soon as stock A gets added to stock B’s index. For instance, Barberis et al. (2005) shows that a stock’s beta with the S&P 500 jumps sharply after index inclusion. For other examples, see Greenwood and Thesmar (2011), Vayanos and Woolley (2013), and Anton and Polk (2014). By contrast, we focus on the unpredictable consequences of stock A’s index inclusion, not for stocks A or B, but for seemingly unrelated stock Zs.

*Thresholds.* Finally, people use threshold-based rules to make all sorts of decisions (Gabaix, 2014). The literature on heuristic decision making typically measures the cost of using a heuristic rule in *expected*-utility loss (Bernheim and Rangel, 2009). Whereas, we look at how simple decision rules can affect demand *volatility*.

## 2 Theoretical Model

Because there are so many index funds tracking so many different benchmarks, a small change in stock  $A$ 's characteristics can cause one index fund to buy stock  $A$  and sell stock  $B$ , which can then cause a second index fund following a different benchmark to sell stock  $B$  and buy stock  $C$ , which can then cause... This section presents a model showing that, while it's possible to determine *if* a stock will be affected by one of these index-fund rebalancing cascades, predicting *how* the stock will be affected (buy? or sell?) is an NP-hard problem. As a result, demand shocks coming from index-fund rebalancing cascades are effectively random even though each index fund involved in a cascade is following a simple, completely deterministic, rebalancing rule.

### 2.1 Market Structure

Here's how we model index funds transmitting an initial shock from stock  $A$  to stock  $B$ , and then from stock  $B$  to stock  $C$ , and then from stock  $C$  to stock  $D$ , and so on.

*Network.* Imagine a market with a set of stocks  $S = \{1, 2, \dots, S\}$ . Index-fund rebalancing rules define a network over these stocks. There is an edge from stock  $s$  to stock  $s'$ , not if they both currently belong to the same benchmark, but rather if a shock to stock  $s$  would cause an index fund to swap its position in stock  $s$  for a new position in stock  $s'$ . If a positive shock to stock  $s$  would cause some fund to sell stock  $s'$  and buy stock  $s$ , then stock  $s'$  is a negative neighbor to stock  $s$ :

$$\mathbf{N}_s^- = \{s' \in S \mid \text{positive shock to } s \Rightarrow \text{negative shock to } s'\} \quad (1)$$

Whereas, if a negative shock to stock  $s$  would cause some fund to buy stock  $s'$  and sell stock  $s$ , then stock  $s'$  is a positive neighbor of stock  $s$ :

$$\mathbf{N}_s^+ = \{s' \in S \mid \text{negative shock to } s \Rightarrow \text{positive shock to } s'\} \quad (2)$$

The market structure is the set of positive and negative neighbors for each stock,  $\mathbf{M} = \{(\mathbf{N}_1^+ \parallel \mathbf{N}_1^-), (\mathbf{N}_2^+ \parallel \mathbf{N}_2^-), \dots, (\mathbf{N}_S^+ \parallel \mathbf{N}_S^-)\}$ .

*Distortion.* This network of rebalancing rules propagates shocks through the market in discrete rounds, which are indexed by  $t = 0, 1, 2, \dots$ . For this to happen, index-fund rebalancing decisions must have the potential to distort stock characteristics. If one fund decides to sell stock  $s$ , then this decision must have the potential to change stock  $s$  in a way that causes a second fund to rebalance, too. In other words, it's important that demand curves slope down (Shleifer, 1986).

This assumption is consistent both with trader descriptions of index-fund rebalancing and with current academic research (Ben-David et al., 2017). More and more people are talking about how ETF rebalancing “influences trading in individual stocks.”<sup>4</sup> And, there’s a lot of overlap between index-fund portfolios. The same stock is often involved in numerous ETF benchmarks, such as “active beta, momentum, dividend growth, deep value, quality, and total earnings.”<sup>5</sup>

We embed this rebalancing-distortions assumption in our model by using a single variable,  $x_{s,t}$ , to keep track of both index-fund rebalancing decisions and changes in stock characteristics:

$$x_{s,t} \in \{-1, 0, 1\} \quad \Delta x_{s,t} = x_{s,t} - x_{s,t-1} \quad (3)$$

If  $(x_{s,t}, \Delta x_{s,t}) = (1, 1)$ , then stock  $s$  has realized a positive shock because some fund built a new position in stock  $s$ . If  $(x_{s,t}, \Delta x_{s,t}) = (-1, -1)$ , then the opposite outcome has taken place. Stock  $s$  has realized a negative shock because some fund exited an existing position in stock  $s$ . To emphasize that index-fund rebalancing decisions can affect more than just a stock’s price, we refer to changes in stock ‘characteristics’.

*Propagation.* Because we want to illustrate how computational complexity can generate seemingly random demand shocks even in the absence of any random behavior on the part of individual investors, we model how index-fund rebalancing decisions propagate shocks through the market as a mechanical 3-step process. STEP 1 involves identifying the set of stocks that will be affected at time  $(t + 1)$  by index funds’ rebalancing decisions at time  $t$ :

$$\text{Out}_{s,t}^+ = \begin{cases} \{s' \in \mathbf{N}_s^+ \mid s \notin \text{Out}_{s',t-1}^-\} & \text{if } (x_{s,t}, \Delta x_{s,t}) = (-1, -1) \\ \emptyset & \text{otherwise} \end{cases} \quad (4a)$$

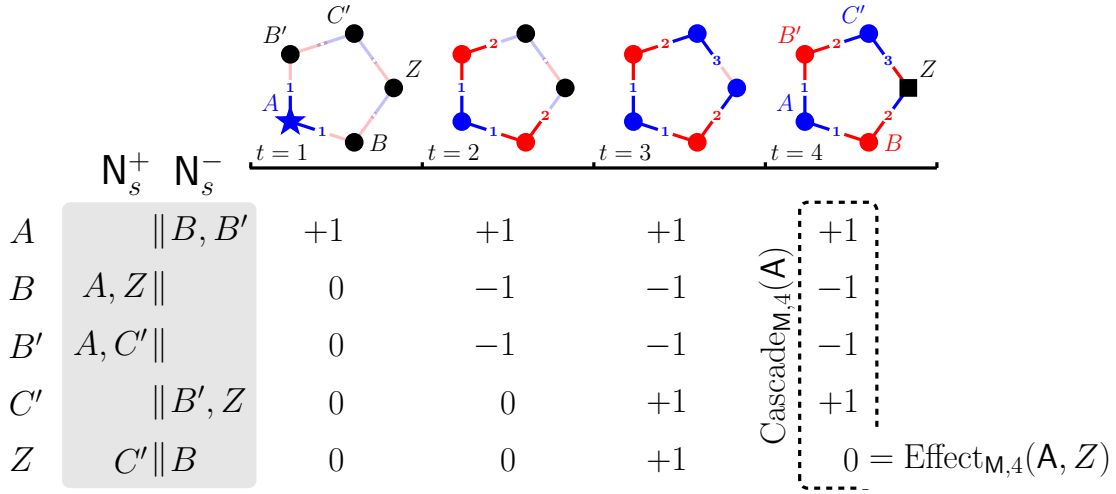
$$\text{Out}_{s,t}^- = \begin{cases} \{s' \in \mathbf{N}_s^- \mid s \notin \text{Out}_{s',t-1}^+\} & \text{if } (x_{s,t}, \Delta x_{s,t}) = (1, 1) \\ \emptyset & \text{otherwise} \end{cases} \quad (4b)$$

$\text{Out}_{s,t}^-$  is the set of stocks that will be negatively affected at time  $(t + 1)$  by some index fund’s decision to buy stock  $s$  at time  $t$ . Likewise,  $\text{Out}_{s,t}^+$  is the set of stocks that will be positively affected at time  $(t + 1)$  by some index fund’s decision to sell stock  $s$  at time  $t$ . The restrictions on  $\text{Out}_{s,t}^+$  and  $\text{Out}_{s,t}^-$  that  $s \notin \text{Out}_{s',t-1}^-$  and  $s \notin \text{Out}_{s',t-1}^+$  respectively ensure that a shock doesn’t just bounce back and forth between stocks  $s$  and  $s'$  over and over again in perpetuity.

---

<sup>4</sup>Bloomberg. 4/10/2015. *Tail Can Wag Dog When ETFs Influence Single Stocks, Goldman Says.*

<sup>5</sup>Financial Times. 10/7/2017. *On The Perverse Economic Effects Created by ETFs.*



**Figure 2: An Example.** An example of an index-fund rebalancing cascade involving 5 stocks that starts with a positive shock to stock A,  $\mathbf{A} = \{A\}$ . Grey box depicts market structure,  $M$ . Columns to the right depict state of each stock,  $x_{s,t}$ , at times  $t = 1, 2, 3, 4$ . Diagram above each column depicts shock propagation through the market. Dots denote stocks. Red dot indicates  $x_{s,t} = +1$ ; blue dot indicates  $x_{s,t} = -1$ ; and, black dot indicates  $x_{s,t} = 0$ . Dashed box reports result of index-fund rebalancing cascade at time  $t = 4$ ,  $\text{Cascade}_{M,4}(A)$ . Notice that cascade has positive effect on stock Z in round  $t = 3$ ,  $\text{Effect}_{M,3}(A, Z) = +1$ . But, in round  $t = 4$ , its net effect on stock Z reverts to  $\text{Effect}_{M,4}(A, Z) = 0$ .

STEP 2 involves identifying all the ways that each stock  $s \in S$  will be affected at time  $(t + 1)$  by this collection of outgoing links at time  $t$ :

$$\text{In}_{s,t+1}^+ = \{s' \in S \mid s \in \text{Out}_{s',t}^+\} \quad (5a)$$

$$\text{In}_{s,t+1}^- = \{s' \in S \mid s \in \text{Out}_{s',t}^-\} \quad (5b)$$

Positive incoming links for stock  $s$  correspond to situations where an index fund sold stock  $s'$  at time  $t$ , and this selling pressure then forced a second index fund following a different benchmark to sell stock  $s'$  and buy stock  $s$  at time  $(t + 1)$ . Negative incoming links for stock  $s$  correspond to the same sequence of events with opposite signs.

Finally, STEP 3 involves calculating how this collection of incoming links will distort the characteristics of each stock at time  $(t + 1)$ :

$$u_{s,t+1} = 1_{\{|\text{In}_{s,t+1}^+| > |\text{In}_{s,t+1}^-|\}} - 1_{\{|\text{In}_{s,t+1}^+| < |\text{In}_{s,t+1}^-|\}} \quad (6a)$$

$$x_{s,t+1} = \text{Sign}[x_{s,t} + u_{s,t+1}] \quad (6b)$$

In the equation above,  $\text{Sign}[y] = y/|y|$ . This updating rule simply says that, if more index funds decided to buy stock  $s$  than sell stock  $s$  at time  $(t + 1)$ , then it will realize a positive shock; whereas, if more index funds decided to sell stock  $s$  than buy stock

$s$ , then it will realize a negative shock.

*Cascades.* An index-fund rebalancing cascade starts in round  $t = 0$  with all stocks at their default levels:

$$(x_{s,0}, \Delta x_{s,0}) = (0, 0) \quad (7)$$

Then, at time  $t = 1$ , nature selects an  $\epsilon$ -small subset of stocks,  $A$ , to receive an initial positive shock:

$$(x_{s,1}, \Delta x_{s,1}) = (1, 1) \quad \text{for each } s \in A \quad (8)$$

We assume that everyone knows the identity of the stocks in  $A$ . We say that  $A$  is  $\epsilon$ -small if there's a positive constant  $\epsilon > 0$  such that  $|A| < \epsilon \cdot S$  as  $S \rightarrow \infty$ . The positive-initial-shock convention is without loss of generality.

Following this initial shock, an index-fund rebalancing cascade is just the iteration of the 3-step updating procedure until a time limit  $T \in \{1, 2, \dots\}$  has been reached:

```

function CascadeM,T(A):
     $t \leftarrow 0$ 
    for all ( $s \in A$ ):
         $(x_s, \Delta x_s) \leftarrow (1, 1)$ 
    while ( $t \leq T$ ):
        for all ( $s \in S$ ):
STEP 1:           $(\text{Out}_s^+, \text{Out}_s^-) \leftarrow \text{Update}[(\text{Out}_s^+, \text{Out}_s^-)|(x_s, \Delta x_s)]$ 
        for all ( $s \in S$ ):
STEP 2:           $(\text{In}_s^+, \text{In}_s^-) \leftarrow \text{Update}[(\text{In}_s^+, \text{In}_s^-)]$ 
STEP 3:           $(x_s, \Delta x_s) \leftarrow \text{Update}[(x_s, \Delta x_s)]$ 
         $t \leftarrow t + 1$ 
    return [ $x_1 \ x_2 \ \dots \ x_S$ ]

```

An index-fund rebalancing cascade's effect on stock  $Z$ ,  $\text{Effect}_{M,T}(A, Z)$ , is the  $Z$ th element of the output from  $\text{Cascade}_{M,T}(A)$ . Notice that how description of an index-fund rebalancing cascade suggests a second interpretation for the symbol  $M$ .  $M$  is not just a description of index-fund rebalancing rules. It's also a description of a machine that computes the effects of index-fund rebalancing cascades.

*An Example.* Figure 2 shows an example of an index-fund rebalancing cascade involving 5 stocks that starts with a positive shock to stock  $A$ . At time  $t = 3$ , the cascade delivers a positive shock to stock  $Z$ ,  $\text{Effect}_{M,3}(\{A\}, Z) = +1$ . But then, at

time  $t = 4$ , a second branch of the cascade hits stock  $Z$ , canceling out the effect of the first shock,  $\text{Effect}_{\mathbf{M},4}(\{A\}, Z) = 0$ . This example highlights the two questions we want to ask about index-fund rebalancing cascades in the following two subsections. First, is there any way for an index-fund rebalancing cascade that starts at stock  $A$  to effect stock  $Z$ ? Second, suppose there is. What will be the net effect of the rebalancing cascade on stock  $Z$ ? In the next two subsections, we're going to investigate the computational complexity of answering each of these questions.

## 2.2 ‘If?’ Problem

How hard is it to figure out whether an index-fund rebalancing cascade triggered by an initial shock to stock  $A$  might eventually affect the demand for stock  $Z$ ?

*Decision Problem.* Solving this decision problem means finding at least one path connecting a particular stock  $A$  to stock  $Z$ . A  $K$ -path connecting stock  $A$  to stock  $Z$  is a sequence of  $K$  stocks  $\{s_1, \dots, s_K\}$  such that the first stock is stock  $A$ , the last stock is stock  $Z$ , and

$$s_k \in \begin{cases} \mathbf{N}_{s_{k-1}}^+ & k \text{ odd} \\ \mathbf{N}_{s_{k-1}}^- & k \text{ even} \end{cases} \quad \text{for all } k \in \{2, \dots, K\} \quad (10)$$

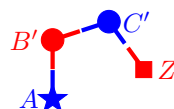
For example, in Figure 2, there are two different paths from stock  $A$  to stock  $Z$ . One travels from stock  $A$  to stock  $B$  to stock  $Z$ :



$$\{(\emptyset \parallel \{B\}), (\{A, Z\} \parallel \emptyset), (\emptyset \parallel \{B\})\} \quad (11)$$

Stock A → Stock B → Stock Z

The other travels from stock  $A$  to stock  $B'$  to stock  $C'$  to stock  $Z$ :



$$\{(\emptyset \parallel \{B'\}), (\{A, C'\} \parallel \emptyset), (\emptyset \parallel \{B', Z\}), (\{C'\} \parallel \emptyset)\} \quad (12)$$

Stock A → Stock B' → Stock C' → Stock Z

If such a path exists, then it's possible that an index-fund rebalancing cascade triggered by an initial shock to stock  $A$  might affect the demand for stock  $Z$ .

Below we give a formal definition of the ‘If?’ problem.

**Problem 2.2a** (If).

- *Instance:* A choice for stock  $Z$ ; a market structure  $\mathbf{M}$ ; a time  $T \geq 1$ ; and, a subset of stocks  $\hat{\mathbf{S}} \subseteq \mathbf{S}$ .
- *Question:* For each stock  $s \in \hat{\mathbf{S}}$ , is there a  $K$ -path connecting stock  $s$  to stock  $Z$  for some  $K \leq T$ ?

If denotes the set of instances where the answer is ‘Yes’. Solving the ‘If?’ problem means deciding whether  $(Z, M, T, \hat{S}) \in \text{If}$ . If  $(Z, M, T, \hat{S}) \in \text{If}$ , then there’s at least one  $K$ -path connecting each stock  $s \in \hat{S}$  to stock  $Z$  in  $K \leq T$  steps.

*If Complexity.* Problems with polynomial-time solutions are considered “tractable problems” that “can be solved in a reasonable amount of time (Moore and Mertens, 2011).” And, the proposition below shows that If can be solved in polynomial time. So, it’s easy to determine which stocks have the potential to trigger an index-fund rebalancing cascade that might affect stock  $Z$ .

**Proposition 2.2a** (If Complexity). *If can be solved in polynomial time.*

We say that  $f(y) = O[g(y)]$  if there exists an  $\alpha > 0$  and a  $y_0 > 0$  such that  $|f(y)| \leq \alpha \cdot |g(y)|$  for all  $y \geq y_0$ . And, we say that  $f(y) = \text{Poly}[y]$  if there exists some  $\beta > 0$  such that  $f(y) = O[y^\beta]$ . The size of an instance of If is governed by the number of stocks in the market,  $S$ . So, a polynomial-time solution for If is an algorithm that decides whether  $(Z, M, T, \hat{S}) \in \text{If}$  in  $\text{Poly}[S]$  steps for every possible choice of  $(Z, M, T, \hat{S})$ —i.e., computational-complexity results typically provide bounds on the time needed to solve worst-case instances.

*Predicting If.* The computational tractability of If also means that you can make useful predictions about the size of  $\hat{S}$  for a given stock  $Z$ . To illustrate, suppose that for any pair of stocks  $(s, s') \in S^2$ , stock  $s'$  is chosen as a positive neighbor to stock  $s$  independently with probability  $\kappa/s$  where  $\kappa > 0$  is some  $O[\log(S)]$  function. Under these assumptions, the number of positive neighbors for each stock,  $N_s^+ = |\mathbf{N}_s^+|$ , obeys a Poisson distribution as  $S \rightarrow \infty$  (Erdos and Rényi, 1960)

$$N_s^+ \sim \text{Poisson}(\kappa, S) \tag{13}$$

which implies that the typical stock has  $E[N_s^+] = \kappa$  positive neighbors. Thus, if  $\kappa \approx 0$ , then the market will be fragmented with most stocks having no neighbors; whereas, if  $\kappa \approx \log(S)$ , then the market will be densely connected with each stock on the cusp of rebalancing for many different funds.

The proposition below shows that it’s easy to predict how many stocks are connected to stock  $Z$  just by counting the number of neighbors for stock  $Z$ .

**Proposition 2.2b** (Predicting If). *If M is a market structure generated using connectivity parameter  $\kappa > 1$  and*

$$\hat{S}_{\max}(Z, M, T) = \max_{\hat{S} \in 2^S} \{ |\hat{S}| \text{ s.t. } (Z, M, T, \hat{S}) \in \text{If} \} \tag{14}$$



denotes the number of stocks with a  $K$ -path to stock  $Z$  for some  $K \leq T$ , then  $E[\hat{S}_{\max}(Z, \mathbf{M}, T)]$  is increasing in the total number of neighbors to stock  $Z$ .

Put differently, stocks with more neighbors are more likely to be affected by index-fund rebalancing cascades. And, you can infer this property about stock  $Z$  without having to trace out each individual path that a rebalancing cascade might take. We will make use of this fact in our empirical analysis below.

## 2.3 ‘How?’ Problem

Although it’s easy to predict *if* a stock is likely to be affected by an index-fund rebalancing cascade, predicting *how* a stock will be affected is computationally intractable.

*Some Intuition.* What does it mean to say that ‘If?’ is an easier question than ‘How?’? To build some intuition, let’s start by looking at Figure 3. Each row depicts a single market with  $S = 25$  stocks and is broken up into 3 panels. Here’s the exercise we have in mind. First, examine the left panel in each row, which depicts the index-fund rebalancing rules that define each market. Then, ask yourself: i) ‘Will stock  $Z$ , which is denoted by the large black square with a question mark in it, be affected by an index-fund rebalancing cascade that starts at stock  $A$ , which is denoted by the large blue star?’ and ii) ‘If so, how exactly will stock  $Z$  be affected (buy vs. sell)?’

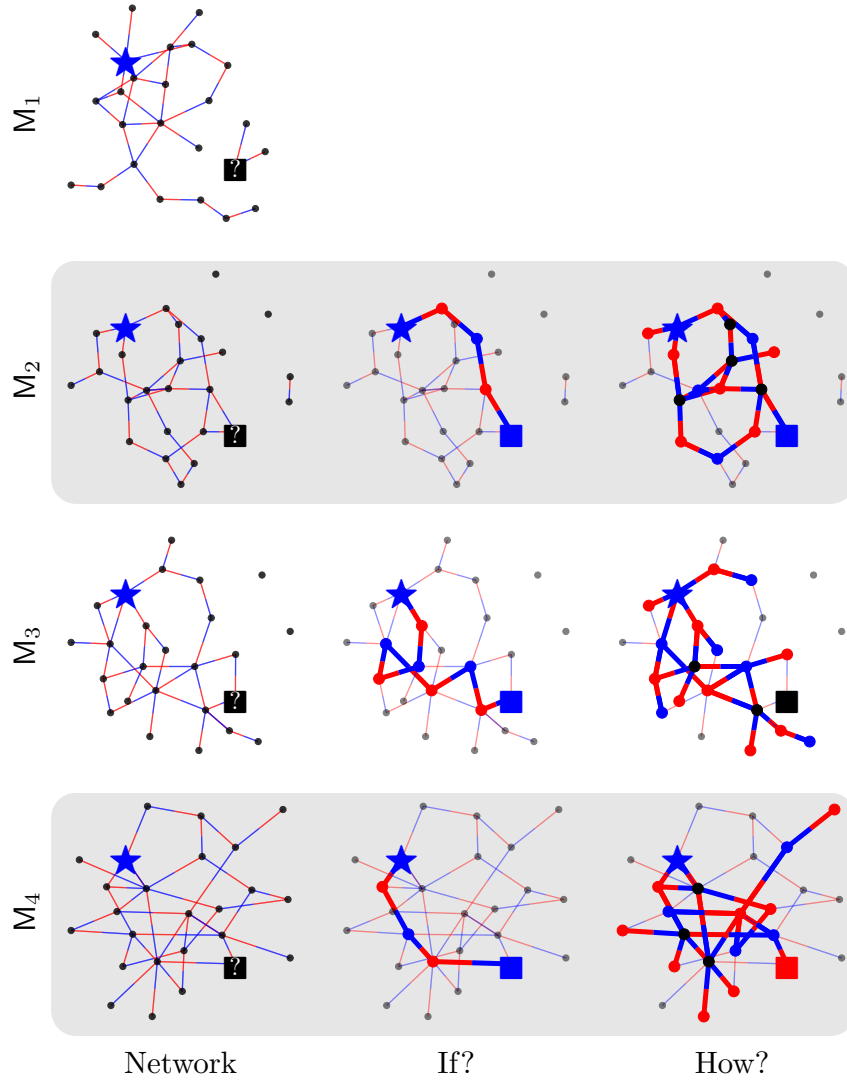
On one hand, you can immediately *see* how easy it is to answer the first question. The middle panels show that there’s a path connecting stock  $A$  to stock  $Z$  in  $\mathbf{M}_2$ ,  $\mathbf{M}_3$ , and  $\mathbf{M}_4$  but not in  $\mathbf{M}_1$ . So, stock  $Z$  might be affected by an index-fund rebalancing cascade starting with stock  $A$  in  $\mathbf{M}_2$ ,  $\mathbf{M}_3$ , and  $\mathbf{M}_4$  but not in  $\mathbf{M}_1$ . Answering this first question gives you a sense of what it means to have a polynomial-time solution.

But, on the other hand, you can also immediately *see* how hard it is to answer the second question. There’s no way to guess how an index-fund rebalancing cascade will affect stock  $Z$  by examining the set of index-fund rebalancing rules involved, even though these rules are completely deterministic.  $\mathbf{M}_2$ ,  $\mathbf{M}_3$ , and  $\mathbf{M}_4$  all have paths connecting stock  $A$  to stock  $Z$  ending positive shocks. But, the effect of the entire index-fund rebalancing cascade only agrees with this naïve prediction in  $\mathbf{M}_2$ .

*Decision Problem.* Below is the formal definition of the ‘How?’ decision problem.

### Problem 2.3a (How).

- *Instance:* A choice for stock  $Z$ ; a market structure  $\mathbf{M}$ ; a time  $T = \text{Poly}[S]$ ; a positive constant  $\epsilon > 0$ ; and, the power set  $\hat{\mathbf{A}} \subseteq 2^S$  of all  $\epsilon$ -small sets  $\mathbf{A} \subseteq S$ .



**Figure 3: Some Intuition.** Each row contains 3 panels and depicts simulated results for a single market with  $S = 25$  stocks—i.e., one market structure per row. Nodes are stocks. Node color denotes effect of index-fund rebalancing cascade: blue=positive, red=negative, black=no effect. Star: stock A. Square: stock Z. Edges denote index-fund rebalancing rules. Blue(s)-to-red(s'): stock  $s'$  is negative neighbor to stock  $s$ . Red(s)-to-blue(s'): stock  $s'$  is positive neighbor to stock  $s$ . Stock A and stock Z are in same position in all panels. Network: Index-fund rebalancing rules. If?: Path connecting stock A to stock Z if one exists. How?: Net effect of index-fund rebalancing cascade if path exists.

- *Question: Is there some  $A \in \hat{A}$  such that  $\text{Effect}_{M,T}(A, Z) \neq +1$ ?*

**How** denotes the set of instances where the answer is ‘Yes’. Here’s what **How** is asking in plain English. Imagine the universe of all index-fund rebalancing cascades that stem from an initial positive shock to an arbitrarily small subset of stocks in the market. Will every single one of these rebalancing cascades have a positive effect on stock  $Z$  after  $T$  rounds of rebalancing?

*How Complexity.* The proposition below gives a mathematical result that mirrors the intuition we built up in Figure 3. Solving **How** is much harder than solving **If**.

**Proposition 2.3a (How Complexity).** *How is an NP-complete problem.*

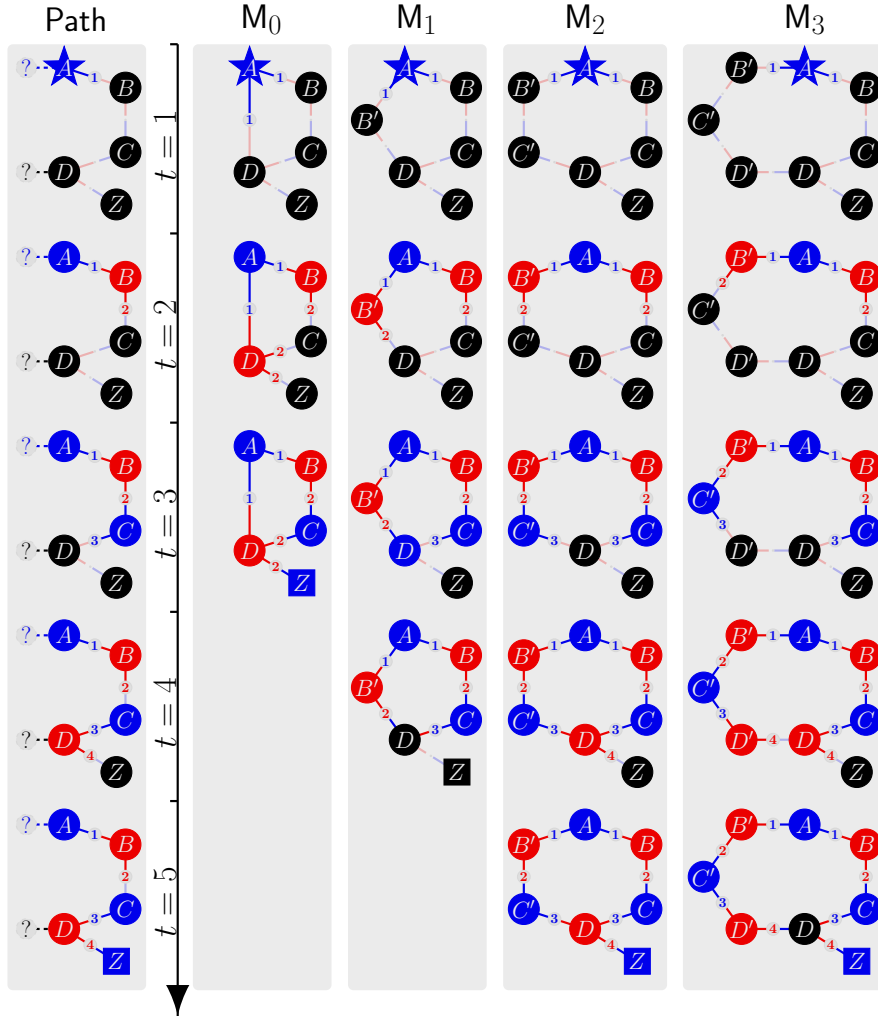
Just like instances of **If**, the size of instances of **How** are governed by the number of stocks in the market,  $S$ . The complexity class **NP** is the set of decision problems with solutions that can be verified in polynomial time. A crossword puzzle is a good example of a problem that’s hard to solve but easy to verify (Garey and Johnson, 2002). Solving this Sunday’s grid might take you an hour, but it will only take a second to verify your guess for 31-down using the answer key.

What does it mean for a decision problem to be **NP** complete? For any pair of decision problems, **Prob<sub>1</sub>** and **Prob<sub>2</sub>**, we say that solving **Prob<sub>2</sub>** can be reduced to solving **Prob<sub>1</sub>** if you can solve **Prob<sub>2</sub>** by just mapping each instance of **Prob<sub>2</sub>** over to a corresponding instance of **Prob<sub>1</sub>** and then simply solving **Prob<sub>1</sub>**. Intuitively, if solving **Prob<sub>2</sub>** can be reduced to solving **Prob<sub>1</sub>**, then solving **Prob<sub>2</sub>** is no harder than solving **Prob<sub>1</sub>**. A decision problem is **NP** complete if every decision problem in **NP** can be reduced to it and it belongs to **NP**.

*Root of the Problem.* Figure 4 illustrates why **How** is so computationally intractable. Each vertical gray region denotes a separate sequence events, starting at the top and ending at the bottom. On the left, there’s a proposed path connecting stock  $A$  to stock  $Z$  that ends with a positive shock to stock  $Z$ :



The trouble is that stocks  $A$  and  $D$  are also connected to other stocks that may not belong to the original path (dotted lines), which means that the market structure could contain a secondary path. The four gray regions to the right show how small changes in the length of this secondary path can change the cascade’s net effect on stock  $Z$ . If stock  $A$  and stock  $D$  are directly connected,  $M_0$ , then the secondary path doesn’t matter. If there is a 1-path connecting stock  $A$  to stock  $D$ ,  $M_1$ , then



**Figure 4: Root of the Problem.** Each vertical gray region denotes a separate sequence events, which starts at the top and ends at the bottom. Each node denotes a stock. Node color denotes effect of cascade: blue=positive, red=negative, black=no effect. Star: initial shock to stock A. Square: final effect for stock Z. Edges denote index-fund rebalancing rules. Blue(s)-to-red(s'): stock  $s'$  is negative neighbor to stock  $s$ . Red(s)-to-blue(s'): stock  $s'$  is positive neighbor to stock  $s$ . Path: path connecting stock A to stock Z. Location of stocks A, B, C, D, and Z remain unchanged in all sequences. Dotted lines: neighbors to stock A and stock Z that could form alternate path.  $M_k$ : market structure that contains alternate path with  $k \in \{0, 1, 2, 3\}$  stocks separating stock A and stock Z.

the secondary path implies that stock  $Z$  will be unaffected by the entire index-fund rebalancing cascade. But, if there’s a 2-path connecting stock  $A$  to stock  $D$ ,  $M_2$ , then the secondary path won’t matter once again. And, if there’s a 3-path connecting stock  $A$  to stock  $D$ ,  $M_3$ , then stock  $Z$  will be positively affected by the index-fund rebalancing cascade even though stock  $D$  will be unaffected. Tiny changes in the structure of a rebalancing cascade can lead to different outcomes for stock  $Z$ .

As a result, determining how a particular index-fund rebalancing cascade will affect stock  $Z$  requires a detailed simulation of how the entire cascade will play out. So, finding an initial shock which results in a negative effect on stock  $Z$  could require checking every possible  $\epsilon$ -small subset. And, the size of this power set scales exponentially with the number of stocks in the market,  $S$ . Suppose you could solve instances of **How** in less than one microsecond when there were only 20 ETFs in the market. Proposition 2.3a implies that this same algorithm would take longer than the current age of the universe to execute in today’s market, which contains roughly 2,000 U.S.-listed ETFs.<sup>6</sup> “A running time that scales exponentially implies a harsh bound on the problems we can ever solve—even if our project deadline is as far away in the future as the Big Bang is in the past (Moore and Mertens, 2011).”

*Predicting How.* Proposition 2.3a says that the problem of figuring out how every single index-fund rebalancing cascade will effect stock  $Z$  is computationally intractable. But, maybe this is an unreasonable goal. What if you only try to figure out how most index-fund rebalancing cascades will affect stock  $Z$ ? We introduce the following decision problem to make this idea precise.

**Problem 2.3b (MajorityHow).**

- *Instance:* A choice for stock  $Z$ ; a market structure  $M$ ; a time  $T = \text{Poly}[S]$ ; a positive constant  $\epsilon > 0$ ; and, the power set  $\hat{A} \subseteq 2^S$  of all  $\epsilon$ -small sets  $A \subseteq S$ .
- *Question:* Is  $\sum_{A \in \hat{A}} 1_{\{\text{Effect}_{M,T}(A, Z) = +1\}} > |\hat{A}|/2$ ?

Compared to **How**, **MajorityHow** seems like a much closer analogue to the problem that real-world traders care about. Traders know which index funds hold each stock. And, they know the rebalancing rules that index funds are following. So, given this information, they would like to determine whether or not some stock  $Z$  will be affected by the majority of index-fund rebalancing cascade that might occur. For a particular market structure, will more than half of all possible index-fund rebalancing cascades

---

<sup>6</sup>Financial Times. 1/14/2017. *ETFs Are Eating The US Stock Market.*

result in buy orders?

At first, **MajorityHow** seems like a much easier problem to solve than **How** because it doesn't involve finding a particular verboten instance. But, this first reaction is wrong. Proposition 2.2b shows that stock  $Z$ s with more neighbors are more likely to be hit by index-fund rebalancing cascades. But, Proposition 2.3b shows that determining whether more than half of all possible index-fund rebalancing cascades will result in buy orders is tantamount to predicting the outcome of a coin flip.

**Proposition 2.3b** (Predicting **How**). *MajorityHow is an NP-hard problem.*

A decision problem is NP hard if every decision problem in NP can be reduced to it but the problem itself might not belong to NP. So, if **MajorityHow** is an NP-hard problem, then it is at least as hard as any decision problem in NP. And, a polynomial-time solution to **MajorityHow** would imply  $P = NP$ .

## 2.4 Key Ingredients

We've just seen that predicting how index-fund rebalancing cascades will affect a stock's demand with accuracy better than a coin flip is an NP-hard problem. As a result, the demand shocks coming from the rebalancing cascades are effectively noise. To make it easier for other researchers to spot other situations where the same mathematical reasoning applies, we now describe three key features of index-fund rebalancing cascades that make them so hard to predict.

*Alternation.* First, index-fund rebalancing cascades are only hard to predict if they involve alternating sequences of buy and sell orders. In a world where a positive shock to stock  $A$  can only ever result in a positive shock to stock  $B$ , predicting how stock  $Z$  will be affected by a rebalancing cascade is easy. In fact, it's equivalent to solving the 'If?' problem.

**Proposition 2.4a** (Necessity of Alternation). *Without alternation, **How** is solvable in polynomial time.*

Index-fund rebalancing cascades necessarily involve an alternating sequence of buy and sell orders. When an index fund rebalances its portfolio, it swaps out an existing position in one stock for a new position in another. But, there are other cascade-like phenomena where this isn't the case. For example, think about bank runs. During a bank run, depositors are choosing whether to withdraw their money—sell only. As a result, equilibrium demand in these models behaves in a predictable way depending

on whether some threshold has been crossed (Diamond and Dybvig, 1983).

*Feedback Loops.* Second, index-fund rebalancing cascades are only hard to predict in a market structure that involves cancellation due to feedback loops. It's important that different cascade paths have the potential to cancel each other out, as shown in Figure 4. To illustrate, think about what would happen if every stock in the market had exactly 2 neighbors. In this setting, if there exists a path connecting stock  $A$  to stock  $Z$ , then you can determine how a rebalancing cascade starting with stock  $A$  will affect stock  $Z$  by counting the number of stocks in the path. If it's an odd number, then stock  $Z$  will realize a positive shock, like in Equation (11). Whereas, if it's an even number, then the shock will be negative, like in Equation (12).

**Proposition 2.4b** (Necessity of Feedback Loops). *Without cancellation due to feedback loops, How is solvable in polynomial time.*

Again, we feel that feedback loops are a natural part of the index-fund universe. There is no central-planning committee that limits the number of indexes that a single stock can belong to. There's nothing stopping 20 different smart-beta ETFs from holding the same stock at the same time.<sup>7</sup> Thus, the associated collections of rebalancing rules will contain market structures with feedback loops. And, it's these loopy instances that make solving **How** computationally intractable.

*Thresholds.* Third, index-fund rebalancing cascades are only hard to predict if their benchmark indexes involve threshold-based rebalancing rules. For example, it's important that the PowerShares S&P 500 Low-Volatility ETF [SPLV] tracks a benchmark consisting of only the 100 lowest-volatility stocks on the S&P 500 and not a benchmark including all S&P 500 stocks with relatively more shares of lower-volatility constituents. In the first case, an arbitrarily small change in a stock's volatility can move it from 101st to 100th place on the low-volatility leaderboard and force SPLV to exit its entire position. In the second case, an arbitrarily small change in a stock's volatility would only lead to an even smaller change in the fund's portfolio position. Without threshold-based rebalancing rules, longer cascade paths would necessarily have smaller effects for the same reason that AR(1) impulse-response functions get weaker at longer horizons. So, you could approximate an index-fund rebalancing cascade's net effect on stock  $Z$  by using the effect of the shortest path to stock  $Z$ .

**Proposition 2.4c** (Necessity of Thresholds). *If index funds don't use threshold-based rebalancing rules, then there's a fully polynomial-time approximation scheme for How.*

---

<sup>7</sup>SeekingAlpha. 6/27/2017. *Smart Beta ETFs Love These Stocks.*

It's a simple fact that index funds use threshold-based rebalancing rules. This is how many index funds operate. But, threshold-based trading rules can be found all over the place in financial markets. A typical stat-arb trading strategy will have the form, 'Buy the top 30% and sell the bottom 30% of stocks when sorting on  $X$ .' where  $X$  is some variable that predicts the cross-section of expected returns. Our goal is not to explain why funds choose to follow these sorts of trading rules; instead, we point out one natural consequence of this choice: noise.

*No-Trade Theorem.* We began this paper with a discussion of Milgrom and Stokey (1982)'s classic no-trade theorem. There's no error in their paper. So, at this point, you might be wondering why doesn't their result apply to the setting we study in our paper. What implicit assumption is being violated?

Milgrom and Stokey (1982) consider a setting where all traders start out with common priors and then one of them gets a private signal. They then prove that, if this lone trader acts on his private signal using a simple deterministic trading rule, then everyone else in the market will be able to figure out what he's learned by studying his trading behavior. We show that this result can break down in modern financial markets because there isn't just one lone trader following a simple deterministic trading rule. There are hordes of them. So, even if each index fund is using a simple deterministic rebalancing rule, the net demand coming from the entire interacting mass of index funds can still appear random.

*Different Application.* Finally, we would like to point out a nice parallel between our main theoretical results and the analysis in Arora et al. (2011). Instead of studying the demand-shock distribution for a single stock, Arora et al. study the loan-quality distribution within a single mortgage-backed security. They too show that the problem of determining whether an asset-backed security contains slightly more bad loans than expected is NP hard. Same mathematical insight. Different financial applications.

### 3 Rebalancing Cascades

We've just seen that index-fund rebalancing cascades can generate seemingly random demand shocks in a theoretical model. We now use data on end-of-day ETF holdings to show that the ETF rebalancing cascades generate unpredictable demand shocks in real-world financial markets.



### 3.1 Index Funds

We study index-fund rebalancing cascades using data on a particular kind of index fund—namely, exchange-traded funds (ETFs). There are three reasons for this choice.

*Reason #1: Diversity.* First, we need a large group of index funds that follows a very heterogeneous collection of benchmark indexes. Prior to January 2008, ETFs all looked like the SPDR S&P 500 ETF [SPY] in that they all tracked some sort of pre-existing market index, like the S&P 500. But, in early 2008, the SEC changed its guidelines so that an ETF could track its own self-defined benchmark. After this change, Invesco PowerShares was free to create an ETF tracking the returns of the quintile of S&P 500 stocks with the lowest historical volatility even though there was no pre-existing low-volatility S&P 500 index. All Invesco had to do was promise to announce the identities and weights involved in the benchmark one day in advance.

Now, there are more ETFs than stocks.<sup>8</sup> “From ProShares we have CLIX (100% long internet retailers and 50% short bricks-and-mortar U.S. retailers) and EMTY (which just bets against bricks-and-mortar retailers)... meanwhile from EventShares, we have policy-factor ETFs... like... GOP (full of oil drillers, gun manufacturers, and so on that would benefit from Republican policies) and DEMS (with companies that should do well under Democrats, such as clean-energy companies). There is also an ETF called TAXR that invests in companies poised to benefit most from a successful attempt to pass a tax reform bill.”<sup>9</sup>

The sheer number and variety of these so-called ‘smart-beta’ ETFs has become something of a hot-button issue of late. To be sure, niche ETFs like DEMS tend to be smaller than broad value-weighted market ETFs, like the SPDR S&P 500 ETF SPY. But, even the rebalancing activity of niche ETFs can affect a stock’s fundamentals because ETFs often execute the bulk of their trades during the final 20 to 30 minutes of the trading day. The numbers are stark: “37% of New York Stock Exchange trading volume now happens in the last 30 minutes of the session, according to JPMorgan. The chief culprit is the swelling exchange-traded fund industry... ETFs are essentially investment algorithms of varying degrees of complexity, and typically automatically rebalance their holdings at the end of the day.”<sup>10</sup>

*Reason #2: Discretion.* Second, ETF managers have less ability to deviate from their stated benchmarks than either mutual- or hedge-fund managers due to the

---

<sup>8</sup>Bloomberg. 5/16/2017. *Mutual Funds Ate the Stock Market. Now ETFs Are Doing It.*

<sup>9</sup>Financial Times. 11/21/2017. *A ROSE by any other ticker symbol...*

<sup>10</sup>Financial Times. 3/17/2017. *Machines and Markets: 5 Areas To Watch.*

underlying structure of the ETF market (Madhavan, 2016; Ben-David et al., 2017). The company running an ETF (its ‘sponsor’) has an obligation to create or redeem shares at the end-of-day market value of its stated benchmark. So, if an ETF’s price is higher than the end-of-day market value of its benchmark, then an arbitrageur can sell shares of the ETF back to the sponsor and use the proceeds to buy shares of the underlying assets in the benchmark index. The reverse logic holds when underpriced.

If arbitrageurs are constantly asking an ETF sponsor to create or redeem lots of shares, then the sponsor must be losing lots of money. So, just like you’d expect, creations and redemptions are only a small fraction of daily trading volume for ETFs, and these trades involve less than 0.5% percent of ETFs’ net assets (Investment Company Institute, 2015). Instead, ETF trading volume primarily comes from managers’ rebalancing activity just prior to market close. This end-of-day trading is how ETF sponsors make sure that there is very little difference between the market value of their end-of-day holdings and the market value of their stated benchmark.

An ETF manager who does the bulk of his rebalancing right at market close will incur higher trading costs. But, the typical investor in a smart-beta ETF is not looking for a cheap way to buy and hold a broad market portfolio. ETF investors traded “\$20 trillion worth of shares last year even though ETFs only have \$2.5 trillion in assets. That’s 800% asset turnover, which is about 3-times more than stocks.”<sup>11</sup> An investor interested in holding a smart-beta ETF is looking for quick access to a very targeted position. He’d rather the ETF manager have slightly higher trading costs and be much more faithful to his stated benchmark. For a niche ETF, the additional trading costs incurred by the end-of-day trading are nothing compared to the costs associated with replicating the entire position from scratch.

*Reason #3: Data.* Third, we can observe end-of-day portfolio positions for ETFs. Specifically, we use data from ETF Global that includes both the assets under management,  $AUM_{f,t}$ , and the relative portfolio weight on each stock,  $\Omega_{f,s,t}$ , for each ETF  $f \in \{1, \dots, F\}$  at the end of each trading day from January 2010 to December 2015. We restrict our sample to include only those ETFs that rebalance their positions daily—think about the PowerShares S&P 500 Low-Volatility ETF [SPLV] rather than the SPDR S&P 500 ETF [SPY]. Thus, if  $P_{s,t}$  is the price of stock  $s$  on day  $t$ , then the actual number of shares of stock  $s$  that the  $f$ th ETF holds on day  $t$ ,  $Q_{f,s,t}$ , is:

$$Q_{f,s,t} = \frac{1}{P_{s,t}} \times \{ \Omega_{f,s,t} \cdot AUM_{f,t} \} \quad (15)$$

---

<sup>11</sup>Bloomberg. 3/3/2017. *5 Ways ‘Passive’ Investing Is Actually Quite Active.*

And, total ETF trading volume for stock  $s$  on day  $t$  is given by  $\sum_{f=1}^F |Q_{f,s,t} - Q_{f,s,t-1}|$ .

This end-of-day data is important. Other papers in the ETF literature, such as Ben-David et al. (2017), impute ETFs' daily portfolio positions from their end-of-quarter financial statements. But, we are interested in how the rebalancing decisions of different ETFs interact with one another over the course of a few days. And, we can't study these interactions if we are forced to impute daily holdings from end-of-quarter data.

*Rebalancing Volume.* We use data on end-of-day ETF holdings to create two main variables of interest. The first is ETF rebalancing volume. This requires a little bit of subtlety because not all ETF trading is due to rebalancing decisions. If money pours into ETF  $f$  on day  $t$ ,  $AUM_{f,t} \gg AUM_{f,t-1}$ , then the fund is going to have to buy a bunch of shares of each stock that it holds. But, this trading volume won't be due to any rebalancing decision. The fund manager is just scaling up his existing holdings. So, to adjust for trading due to inflows and outflows, we first calculate each ETF's predicted holdings on day  $t$  given its portfolio weights on the previous day ( $t-1$ ) and the realized inflows and outflows on day  $t$ :

$$\bar{Q}_{f,s,t} = \frac{1}{P_{s,t}} \times \{ \Omega_{f,s,t-1} \cdot AUM_{f,t} \} \quad (16)$$

Then, for each stock  $s$ , we compute the total difference between every ETF's actual end-of-day holdings and this inflow-adjusted prediction on day  $t$ :

$$\text{etfRebalVlm}_{s,t} = \sum_{f=1}^F |Q_{f,s,t} - \bar{Q}_{f,s,t}| \quad (17)$$

We use this as our daily measure of ETF rebalancing volume for each stock. Note that we will write all regression variables in **teletype font** to distinguish them from estimated parameters. Table 2 shows that the typical stock during our sample period had  $e^{7.62} \approx 2,038$  shares traded each day due to ETF rebalancing decisions.

*Order Imbalance.* Then, to evaluate whether ETFs are trading in different directions, we also compute a corresponding measure of ETF order imbalance:

$$\text{etfOrdImbal}_{s,t} = \sum_{f=1}^F \frac{Q_{f,s,t} - \bar{Q}_{f,s,t}}{\text{etfRebal}_{s,t}} \quad (18)$$

This variable lies on the interval  $[-1, 1]$ . If  $\text{etfOrdImbal}_{s,t} = -1$ , then every share of stock  $s$  traded on day  $t$  was a sell order. Whereas, if  $\text{etfOrdImbal}_{s,t} = 1$ , then every share of stock  $s$  traded on day  $t$  was a buy order. Table 2 contains summary statistics describing the typical ETF order imbalance for each stock during our sample period.

## 3.2 Initial Shocks

We use M&A announcements for our set of initial shocks, referring to the stock that’s the target of the M&A announcement as stock  $A$ .

*M&A Announcements.* Our source for data on M&A deals is Thomson Financial. We use all deals with an announcement date between January 1st, 2010 and December 31st, 2015 where the target is a public company. Table 2 shows that there were 884 such events during our sample period. M&A announcements are a natural choice for our initial shocks because there is solid empirical evidence that the target of an M&A announcement realizes a sharp price increase (Andrade et al., 2001). And, while acquirers do not choose their M&A targets at random, the exact day that a deal is announced can be taken as random. We use  $t_A$  to denote the day of the M&A announcement in which stock  $A$  was the target firm.

*Desired Effect.* Table 3 contains direct evidence that ETFs rebalance in response to these initial shocks. We create a panel dataset containing the ETF rebalancing volume for each M&A target in our sample during the time window  $t \in \{t_A - 20, \dots, t_A + 5\}$ . Then, we regress log ETF rebalancing volume on indicator variables for the date of the M&A announcement:

$$\ln(\text{etfRebal}_{A,t}) = \alpha + \beta \cdot \mathbf{1}_{\{t=t_A-1\}} + \gamma \cdot \mathbf{1}_{\{t=t_A\}} + \delta \cdot \mathbf{1}_{\{t=t_A+1\}} + \dots + \varepsilon_{A,t} \quad (19)$$

In different specifications, the “...” term in the equation above contains year-month fixed effects, stock  $A$  fixed effects, and lagged trading volume. The first column in Table 3 shows that ETF rebalancing volume for stock  $A$  rises by 166% on the day it’s announced as an M&A target. The second column in Table 3 shows that this jump in ETF rebalancing activity is not explained by lagged volume.

*Manager Discretion.* The third column of Table 3 shows the results of the same regression specification but with additional indicator variables for days  $(t_A - 2)$ ,  $(t_A - 3)$ ,  $(t_A - 4)$ , and  $(t_A - 5)$ . This column reveals that there is no pretrend in ETF managers’ reaction to the M&A announcement. ETF rebalancing volume only starts to rise on the day immediately before the announcement, and this 1-day-early effect is due to the way overnight announcements are coded by Thomson Financial. What’s more, after we include lagged volume in the specification, the jump in ETF rebalancing volume is gone the day after. This supports our claim that ETF managers don’t have much discretion when it comes to deviating from their benchmark index overnight. There’s no reason to suspect that ETF managers are slowly rebalancing their position in stock  $Z$  in response to demand shocks coming from ETF rebalancing cascades if

they aren't doing the same thing in response to the initial M&A announcement shock to stock  $A$ .

*Placebo Test.* Finally, the fifth column of Table 3 shows the results of the same regression specification using randomly selected dates for  $t_A$  rather than the actual M&A announcements dates. Just as expected, there is no sharp jump in ETF rebalancing volume on these randomly selected dates. But, more importantly, the coefficients on lagged trading volume are also unchanged. This placebo test suggests that our result isn't being driven by some broader trading-volume anomaly that occurs around the time of each M&A announcement. Things look normal other than the spike in ETF rebalancing volume.

### 3.3 Main Results

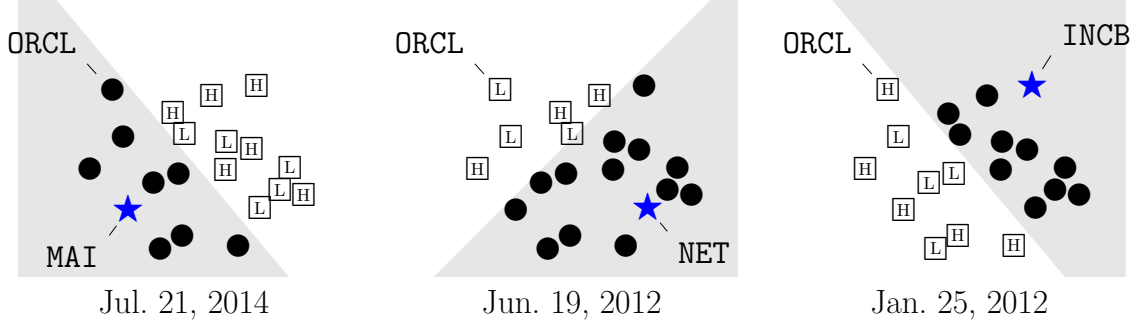
We now give evidence that these M&A announcements lead to ETF rebalancing cascades that result in unpredictable demand shocks for unrelated stock  $Z$ s.

*Diff-in-Diff Approach.* Here's how we set up our tests. First, we create a panel dataset containing the ETF rebalancing volume on days  $t \in \{t_A - 20, \dots, t_A + 5\}$  for each unrelated stock  $Z$ s relative to each M&A target announcement in our sample. We use  $\text{afterAncmt}_{A,t}$  as an indicator variable to flag the 5 days after the M&A announcement for a given stock  $A$ :

$$\text{afterAncmt}_{A,t} = \begin{cases} 1 & \text{if } t \in \{t_A + 1, \dots, t_A + 5\} \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

For stock  $Z$  to be unrelated to stock  $A$ , it has to be twice removed in the network of ETF holdings. It can't have been recently held by any ETF that also recently held stock  $A$ . And, if stock  $B$  and stock  $A$  are both held by the same ETF, then stock  $Z$  can't have been recently held by any ETF that also recently held stock  $B$  either. i.e., the chain has to be  $A \rightarrow B \rightarrow C \rightarrow Z$  or longer. Because there are so many different smart-beta ETFs that are specifically designed to give their investors exposure to things like size and value, this criteria implies that each set of stock  $Z$ s doesn't share well-known characteristics with the associated stock  $A$ .

Proposition 2.2b suggests that, all else equal, stocks on the cusp of more rebalancing thresholds are more likely to be hit by an ETF rebalancing cascade. So, we split the set of stock  $Z$ s for each initial M&A announcement into two subsets: those on the cusp of an above-median number of ETF rebalancing thresholds (i.e., stocks with lots of neighboring stocks in the ETF rebalancing network) and those on the cusp of



**Figure 5: Empirical Design.** Each panel depicts the the same set of stocks during 3 different M&A announcements: Owens & Minor’s purchase of Medical Action Industries [MAI] announced on Jul. 21, 2014; Sonus Networks’ purchase of Network Equip Technologies [NET] announced on Jun. 19, 2012; and, Old National Bancorp’s purchase of Indiana Community Bancorp [INCB] announced on Jan. 25, 2012. The target of each M&A announcement, stock  $A$ , is denoted by a blue star. Each black circle denotes a stock that’s related to stock  $A$  at the time of the announcement. Each white square denotes a stock that’s unrelated to stock  $A$  at the time of the announcement. This is the set of stock  $Z$ s. Unrelated stocks that are neighbors with an above-median number of other stocks are labeled with an “H”; whereas, those that are neighbors with a below-median number are labeled with an “L”. Oracle Corp. is a related stock in the left panel, a below-median stock  $Z$  in the middle panel, and an above-median stock  $Z$  in the right panel.

a below-median number of ETF rebalancing thresholds. We use  $\text{manyNbrs}_{A \rightarrow Z, t}$  as an indicator variable to flag the subset of stock  $Z$ s that are on an above-median number of ETF rebalancing thresholds:

$$\text{manyNbrs}_{A \rightarrow Z, t} = \begin{cases} 1 & \text{if stock } Z \text{ has an above-median number of neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

We say that stock  $s'$  is a neighbor to stock  $s$  if a fund that currently holds stock  $s$  also rebalanced its position in stock  $s'$  at some point during the previous month.

There are two key predictions from Section 2 that we want to test. First, Proposition 2.2b suggests that stock  $Z$ s with more neighbors should be more likely to be hit by ETF rebalancing cascades and so should have proportionally higher ETF rebalancing volume in the days immediately following the initial M&A announcement for stock  $A$ . We test this prediction using a standard diff-in-diff regression:

$$\begin{aligned} \ln(\text{etfRebalVlm}_{Z, t}) = & \alpha + \beta \cdot \text{afterAncmt}_{A, t} \\ & + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z, t} \\ & + \delta \cdot \{\text{afterAncmt}_{A, t} \times \text{manyNbrs}_{A \rightarrow Z, t}\} \\ & + \dots + \varepsilon_{A \rightarrow Z, t} \end{aligned} \quad (22)$$

The null hypothesis is that ETF rebalancing cascades only affect stocks that are closely related to the initial stock  $A$ . If this were the case, then we'd expect to find  $\delta = 0$ . By contrast, if stock  $Z$ s with more neighbors are indeed more likely to be hit by a rebalancing cascade, then we should estimate  $\delta > 0$ .

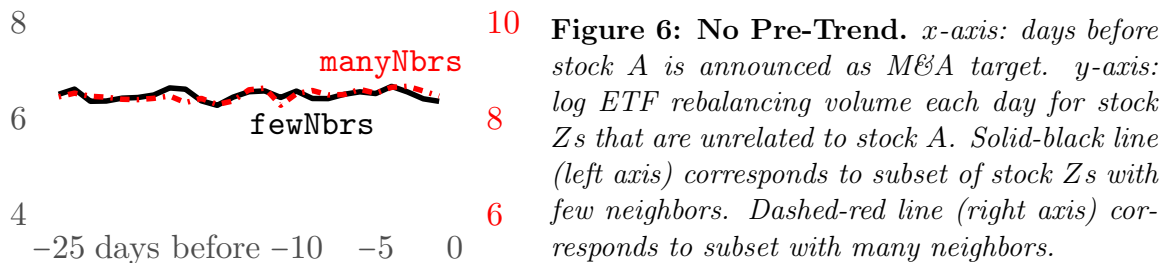
How should we interpret  $\beta$ ? ETF rebalancing cascades have the potential to affect the demand for all stock  $Z$ s; it's just that they're much more likely to affect the demand of stock  $Z$ s with more neighbors. Stocks with only 1 or 2 neighbors can still be included in a rebalancing cascade as shown in Figure 4. So, if ETF rebalancing cascades are taking place, then we should also expect to find  $\beta > 0$ . And, we can use this auxiliary prediction as a way of checking the internal consistency of our results. But, while finding  $\beta > 0$  is consistent with the existence of ETF rebalancing cascades, it's also consistent with general market conditions changing following an M&A announcement. By checking whether  $\delta > 0$ , we can see whether this increase in ETF rebalancing volume for stock  $Z$  is related to the structure of ETF rebalancing rules in a way that's predicted by our theory.

Second, even though stock  $Z$ s with more neighbors are more likely to be hit by ETF rebalancing cascades, Proposition 2.3b suggests that the direction of the resulting demand shock should be a coin flip. We test this prediction using the same diff-in-diff specification as before but with order imbalance as the left-hand-side variable:

$$\begin{aligned} \ln(\text{etfOrdImbal}_{Z,t}) = & \alpha + \beta \cdot \text{afterAncmt}_{A,t} \\ & + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z,t} \\ & + \delta \cdot \{ \text{afterAncmt}_{A,t} \times \text{manyNbrs}_{A \rightarrow Z,t} \} \\ & + \dots + \varepsilon_{A \rightarrow Z,t} \end{aligned} \tag{23}$$

If a stock  $Z$  with many neighbors is no more likely to realize a positive demand shock than a stock  $Z$  with few neighbors, then we should estimate  $\delta = 0$ .

*Empirical Design.* At this point, you might be worried that the unrelated stock  $Z$ s with more neighbors are just different kinds of stocks than the unrelated stock  $Z$ s with few neighbors. And, this is a valid concern. But, there is an important detail about how we set up our diff-in-diff approach that helps us address this concern. Specifically, we define the set of unrelated stock  $Z$ s separately for each initial M&A announcement for a stock  $A$ . This means that the exact same stock can play the role of an above-median stock  $Z$  relative to one M&A announcement while playing the role of a below-median stock  $Z$  relative to another. Figure 5 gives an example from



**Figure 6: No Pre-Trend.** *x-axis: days before stock A is announced as M&A target. y-axis: log ETF rebalancing volume each day for stock Zs that are unrelated to stock A. Solid-black line (left axis) corresponds to subset of stock Zs with few neighbors. Dashed-red line (right axis) corresponds to subset with many neighbors.*

our dataset of this sort of thing happening for Oracle Corp.

By including fixed effects for each stock  $Z$  in our regression specification, we can estimate how the ETF rebalancing activity for the exact same stock changes when it happens to have many neighbors. This design feature means that any empirical results we find can't be explained by ETFs always trading some stocks differently than others. Any confounding variable has to explain why ETFs suddenly change their rebalancing behavior for an ever-changing subset of stock  $Z$ s in the days immediately following an M&A announcement about a completely unrelated stock  $A$ .

*Rebalancing Volume.* Table 4 provides the estimated coefficients for the regression described in Equation (22). The first column shows that ETF rebalancing volume for all unrelated stock  $Z$ s tends to rise by 5.0% on average in the wake of an M&A announcement for stock  $A$ . But, the third column shows that this growth is concentrated among unrelated stock  $Z$ s that have many neighboring stocks in the network defined by ETF rebalancing rules. Consistent with our economic story, we find that ETF rebalancing volume is 3.7% higher for the above-median group of stock  $Z$ s than for the below-median group in the five days immediately following each M&A announcement.

There are three important points to emphasize about this result. The first is that we include stock- $Z$  fixed effects in our regressions. So, because the same stock  $Z$  can have many neighbors relative to one M&A target and few neighbors relative to another, these results can't be due to persistent differences in how ETFs tend to rebalance their positions in particular stocks. The second is that there is no pre-trend. Figure 6 shows that in the run-up to each M&A announcement, the difference between the amount of ETF rebalancing activity in stock  $Z$ s with many neighbors and the amount of ETF rebalancing activity in stock  $Z$ s with few neighbors remains constant. Finally, the second and fourth columns of Table 4 confirm that the sudden spike in log ETF rebalancing volume for stock  $Z$ s with many neighbors isn't due to a general run-up in trading volume. When we include lagged trading volume in our regression specification, our point estimates are largely unchanged.



*Order Imbalance.* Table 4 gives evidence of long rebalancing cascades; whereas, Table 5 gives evidence that these cascades have an unpredictable effect on demand. This table reports the estimated coefficients for the regression described in Equation (23). The main takeaway from this table comes from comparing the coefficients in the first and third columns. While the first column shows that there is a statistically significant  $\beta = 0.0075\%$ pt increase in ETF order imbalance for unrelated stock  $Z$ s in the days immediately after an M&A announcement, the third column shows that there is no measurable difference between the ETF order imbalance of stock  $Z$ s with many neighbors and stock  $Z$ s with few neighbors. What’s more, the size of the statistically insignificant point estimate for this difference,  $\delta_{\text{etfOrdImbal}} = 0.0024\%$ pt, is more than 2 orders-of-magnitude smaller than the corresponding difference in ETF rebalancing volume,  $\delta_{\ln(\text{etfRebal})} = 3.70\%$ . Taken together, this evidence suggests that, while it’s possible to predict which stock  $Z$ s are likely to be affected by a ETF rebalancing cascade, it’s much harder to predict how these stock  $Z$ s will be affected by the resulting demand shock.

*Aggregate Tests.* Among empirical economists, it’s taken almost as an article of faith that empirical tests should be run using the most micro-level data possible. So, at this point, you might be surprised that we didn’t try to trace out the precise buy-sell-buy-sell sequences of each ETF cascade in our sample. But, there is a good reason why we didn’t do this. This empirical approach would fundamentally ignore the central message of our theoretical analysis: it is computationally intractable to make predictions about the fine-grained structure of rebalancing cascades. Instead, we need to run our tests using well-chosen macro-level variables. Even if it isn’t practical to track the precise buy-sell-buy-sell sequence of ETF rebalancings, it’s relatively easy to proxy for the total number of thresholds that a stock is close to. By analogy, even if it isn’t possible to keep track of the location and momentum of every single gas molecule in a  $1\text{m}^3$  box, it’s easy to measure macro-level variables like the pressure and temperature inside the container.

## 4 Market Reaction

We’ve just seen evidence that ETF rebalancing cascades exist and that the resulting demand shocks are unpredictable from the standpoint of an econometrician. But, maybe these demand shocks look less random to traders? In this section, we provide evidence that market participants treat the erratic demand coming from ETF

rebalancing cascades as noise by studying cross-sectional variation in liquidity.

## 4.1 Liquidity

If market participants treat the erratic demand coming from ETF rebalancing cascades as noise, then we should find that stock  $Z$ s with more neighbors also have higher levels of liquidity.

*Variable Definitions.* We calculate the liquidity of each stock  $Z$  on a daily basis in two different ways. First, we compute an intraday variant of the Amihud (2002) measure:

$$\text{amihud}_{Z,t} = \frac{1}{390} \cdot \sum_{m \in t} \frac{|R_{Z,m}|}{\$V_{Z,m}} \quad (24)$$

Above,  $m \in \{1, \dots, 390\}$  indexes the 390 minutes in each trading day,  $\text{ret}_{Z,m}$  denotes the return of stock  $Z$  in minute  $m$ , and  $\$V_{Z,m}$  denotes the number of dollars of stock  $Z$  traded in minute  $m$ . We scale this variable so that it's reported as the percent-change in stock  $Z$ 's price per million dollars of traded volume. In addition, we also compute the average bid-ask spread of stock  $Z$  during the trading day:

$$\text{baSpread}_{Z,t} = \frac{1}{P_{Z,t}} \cdot \left( \frac{1}{390} \cdot \sum_{m \in t} [P_{Z,m}^{\text{bid}} - P_{Z,m}^{\text{ask}}] \right) \quad (25)$$

We scale this second variable so that it's reported in basis points as a fraction of stock  $Z$ 's closing price on day  $t$ .

*Around Initial Shocks.* Market participants can see whether each unrelated stock  $Z$  has many neighbors or few neighbors. And, because the stock  $Z$ s with many neighbors are more likely to be hit by an ETF rebalancing cascade, market participants should realize that they are more likely to see erratic non-fundamental demand shocks for these stocks as a result. So, the stock  $Z$ s with many neighbors should have higher liquidity. The second and fourth columns of Table 6 confirm this prediction by showing that  $\gamma < 0$  when estimating the regression below for  $y \in \{\text{amihud}, \text{baSpread}\}$ :

$$\begin{aligned} y_{Z,t} = & \alpha + \beta \cdot \text{afterAncmt}_{A,t} \\ & + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z,t} \\ & + \delta \cdot \{\text{manyNbrs}_{A \rightarrow Z,t} \times \text{afterAncmt}_{A,t}\} \\ & + \dots + \varepsilon_{A \rightarrow Z,t} \end{aligned} \quad (26)$$

Note that both the Amihud (2002) measure and the bid-ask spread are inversely related to liquidity. So,  $\gamma < 0$  implies that stock  $Z$ s who happen to have more

neighbors also have more liquidity.

These same columns also show that there is no change in the relative liquidity of the stock  $Z$ s with many neighbors and the stock  $Z$ s with few neighbors in the days immediately after the initial M&A announcement for stock  $A$ . This result is in stark contrast to the earlier findings in Tables 4 and 5. But, this result is also exactly what we’d expect to find in a market where traders knew which stock  $Z$ s were most susceptible to the erratic non-fundamental demand coming from ETF rebalancing cascades. This knowledge should be priced into each stock  $Z$ ’s bid-ask spread ahead of time. The market as a whole might contain more asymmetric information after an important M&A announcement and thus be less liquid in general. But, if market participants already understood which stocks were more likely to be hit by ETF rebalancing cascades and considered this demand to be noise, then there should be no differential effect for stock  $Z$ s with many neighbors vs. those with few neighbors following the M&A announcement for stock  $A$ .

## 4.2 Panel Regressions

If market participants recognize ex ante that stocks with more neighbors are more likely to be hit by ETF rebalancing cascades, then you might expect that stocks with many neighbors actually have higher liquidity than stocks with few neighbors unconditionally. And, this is exactly what we find in the data.

*Unconditional Results.* The third and fifth columns of Table 7 show the estimation results for the regression below where  $y \in \{\text{amihud}, \text{baSpread}\}$ :

$$y_{s,t} = \alpha + \beta \cdot \text{\#nbrs}_{s,t} + \dots + \varepsilon_{s,t} \quad (27)$$

The key difference between this regression and the earlier regressions is that the data we use to estimate this regression do not only include the unrelated stock  $Z$ s around each initial M&A announcement. They include all stocks in our data sample. Again, we estimate  $\beta < 0$  for both liquidity measures, suggesting that market participants are treating the erratic demand coming from ETF rebalancing cascades as noise.

*Implication for Traders.* A natural next question is: ‘What should a trader do with this information?’ The answer isn’t to directly buy or sell stocks with many or few neighbors. Instead, these results suggest a way of amplifying the returns to any existing cross-sectional trading strategy. For example, suppose that you would like to construct a classic momentum portfolio that is long the 30% of stocks with the highest returns over the previous 6 months and short the 30% of stocks with

the lowest returns over the previous 6 months. Our results suggest that you could implement this strategy more efficiently by focusing each leg of this strategy on the stocks with the most neighbors. This position will have the same average return, but it will have lower implementation costs due to the liquidity provided by ETF rebalancing cascades.

*Neighbors vs. Holdings.* Are our results just due to the fact that ETFs like holding liquid stocks? No. And, we can use the regression specification in Equation (27) one more time to further emphasize this point. Our results are due to the overlapping network of ETF rebalancing decisions and not due to which stocks ETFs choose to hold. Specifically, we re-estimate the regression where  $y \in \{\text{amihud}, \text{baSpread}\}$ , but this time including fixed effects for the number of ETFs that hold a particular stock:

$$y_{s,t} = \alpha + \beta \cdot \text{\#nbrs}_{s,t} + \gamma \cdot \mathbf{1}_{\{\text{\#etf}=i\}} + \dots + \varepsilon_{s,t} \quad (28)$$

By doing this, we are able to estimate the amount of additional liquidity that is associated with having more neighbors while controlling for the effect of being held by more ETFs. Again, we find that  $\beta < 0$ , suggesting that market participants are reacting to a stock’s susceptibility to ETF rebalancing cascades rather than just the number of ETFs that hold it.

## 5 Conclusion

*“To generate randomness, we humans flip coins, roll dice, shuffle cards, or spin a roulette wheel. All these operations follow direct physical laws, yet casinos are in no risk of losing money. The complex interaction of a roulette ball with the wheel makes it computationally impossible to predict the outcome of any one spin, and each result is indistinguishable from random.”*  
—Fortnow (2017)

This paper proposes an analogous explanation for seemingly random demand shocks in financial markets. A stock’s demand might appear random, not because individual investors are behaving randomly, but because it’s too computationally complex to predict how a wide variety of simple, deterministic, trading rules will interact with one another. First, we show theoretically how computational complexity can generate noise by modeling a particular kind of trading-rule interaction: index-fund rebalancing cascades. Then, we give empirical evidence that index-fund rebalancing cascades actually generate noise in real-world financial markets using data on the

end-of-day holdings of exchange-traded funds (ETFs).

By showing precisely why it's computationally intractable to predict ETF rebalancing cascades, we make it possible for researchers to identify other situations where the same logic holds. For example, our theoretical model also applies to any other group of funds following a wide variety of threshold-based rebalancing rules. Think about quantitative hedge funds following strategies of the form 'Buy the top 30% and sell the bottom 30% of stocks when sorting on  $X$ ' (Khandani and Lo, 2007). Or, consider pension funds with strict portfolio mandates of the form '15% of our assets will be held in alternative investments' (Pennacchi and Rastad, 2011).

## References

- Aaronson, S. (2013). *Quantum computing since Democritus*. Cambridge University Press.
- Admati, A. (1985). A noisy rational-expectations equilibrium for multi-asset securities markets. *Econometrica*.
- Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*.
- Andrade, G., M. Mitchell, and E. Stafford (2001). New evidence and perspectives on mergers. *Journal of Economic Perspectives*.
- Anton, M. and C. Polk (2014). Connected stocks. *Journal of Finance*.
- Arora, S., B. Barak, M. Brunnermeier, and R. Ge (2011). Computational complexity and information asymmetry in financial products. *Communications of the ACM*.
- Atkeson, A., A. Eisfeldt, and P. Weill (2015). Entry and exit in OTC derivatives markets. *Econometrica*.
- Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics*.
- Barber, B. and T. Odean (2000). Trading is hazardous to your wealth: The common-stock investment performance of individual investors. *Journal of Finance*.
- Barberis, N., A. Shleifer, and J. Wurgler (2005). Comovement. *Journal of Financial Economics*.
- Barberis, N. and R. Thaler (2003). A survey of behavioral finance. In *Handbook of the Economics of Finance*.
- Ben-David, I., F. Franzoni, and R. Moussawi (2017). Do ETFs increase volatility? *Journal of Finance*.
- Ben-David, I., F. Franzoni, and R. Moussawi (2017). Exchange-traded funds. *Annual Review of Financial Economics*.
- Bernheim, D. and A. Rangel (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*.
- Black, F. (1985). Noise. *Journal of Finance*.
- Bollobás, B. (2001). *Random Graphs*. Cambridge University Press.
- Brown, D., S. Davies, and M. Ringgenberg (2016). ETF arbitrage and return predictability. Working Paper.

- Chang, Y., H. Hong, and I. Liskovich (2014). Regression discontinuity and the price effects of stock-market indexing. *Review of Financial Studies*.
- Cook, S. (1971). The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*.
- Diamond, D. and P. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *Journal of political economy*.
- Duffie, D., S. Malamud, and G. Manso (2009). Information percolation with equilibrium search dynamics. *Econometrica*.
- Erdos, P. and A. Rényi (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*.
- Fortnow, L. (2017). *Golden Ticket: P, NP, and the Search for the Impossible*. Princeton University Press.
- French, K. (2008). The cost of active investing. *Journal of Finance*.
- Gabaix, X. (2014). A sparsity-based model of bounded rationality. *Quarterly Journal of Economics*.
- Garey, M. and D. Johnson (2002). *Computers and intractability*. WH Freeman New York.
- Gill, J. (1977). Computational complexity of probabilistic turing machines. *SIAM Journal on Computing*.
- Greenwood, R. and D. Thesmar (2011). Stock-price fragility. *Journal of Financial Economics*.
- Gromb, D. and D. Vayanos (2010). Limits of arbitrage. *Annual Review of Financial Economics*.
- Grossman, S. and J. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economics Review*.
- Hellwig, M. (1980). On the aggregation of information in competitive markets. *Journal of economic theory*.
- Investment Company Institute (2015). *Investment Company Fact Book*. Investment Company Institute. [http://www.icifactbook.org/pdf/2015\\_factbook.pdf](http://www.icifactbook.org/pdf/2015_factbook.pdf).
- Israeli, D., C. Lee, and S. Sridharan (2017). Is there a dark side to exchange-traded funds? an information perspective. *Review of Accounting Studies*.
- Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

- Karp, R. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations*. Springer US.
- Keynes, J. (1921). *A Treatise on Probability*. Macmillan and Company.
- Khandani, A. and A. Lo (2007). What happened to the quants in August 2007? *Journal of Investment Management*.
- Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica*.
- Madhavan, A. (2016). *Exchange-Traded Funds and the New Dynamics of Investing*. Oxford University Press.
- Milgrom, P. and N. Stokey (1982). Information, trade, and common knowledge. *Journal of Economic Theory*.
- Moore, C. and S. Mertens (2011). *The nature of computation*. OUP Oxford.
- Newman, M., S. Strogatz, and D. Watts (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*.
- Ozsoylev, H. and J. Walden (2011). Asset pricing in large information networks. *Journal of Economic Theory*.
- Pennacchi, G. and M. Rastad (2011). Portfolio allocation for public pension funds. *Journal of Pension Economics and Finance*.
- Shleifer, A. (1986). Do demand curves for stocks slope down? *Journal of Finance*.
- Shleifer, A. and L. Summers (1990). The noise-trader approach to finance. *Journal of Economic Perspectives*.
- Shleifer, A. and R. Vishny (1997). The limits of arbitrage. *Journal of Finance*.
- Stambaugh, R. (2014). Presidential address: Investment noise and trends. *Journal of Finance*.
- Vayanos, D. and P. Woolley (2013). An institutional theory of momentum and reversal. *Review of Financial Studies*.
- Wigderson, A. (1992). The complexity of graph connectivity. In *International Symposium on Mathematical Foundations of Computer Science*, pp. 112–132.
- Wurgler, J. (2010). On the economic consequences of index-linked investing. In *Challenges to Business in the Twenty-First Century: The Way Forward*.



## A Proofs

**Definition** (Binary String). Let  $\{0, 1\}^* = \cup_{n=0,1,2,\dots} \{0, 1\}^n$  denote the set of all binary strings.

**Definition** (Problem Solving). Let  $\text{Prob} \in \{0, 1\}^*$  denote a decision problem. An algorithm  $F : \{0, 1\}^* \mapsto \{0, 1\}$  solves  $\text{Prob}$  (a.k.a., decides membership in  $\text{Prob}$ ) if for every instance  $i \in \{0, 1\}^*$

$$i \in \text{Prob} \quad \Leftrightarrow \quad F(i) = 1$$

**Problem A** ( $\text{stCon}$ ).

- *Instance:* A directed graph  $G$ , and two vertices  $(s, t)$ .
- *Question:* Is there a path from  $s$  to  $t$ ?

**Theorem A** (Wigderson, 1992).  $\text{stCon}$  is solvable in polynomial time.

**Definition** (Reduction). Let  $\text{Prob}_1$  and  $\text{Prob}_2$  denote two decision problems. We say that  $\text{Prob}_2$  is (Karp, 1972) reducible to  $\text{Prob}_1$  if there exists a polynomial-time algorithm  $F : \{0, 1\}^* \mapsto \{0, 1\}^*$  such that

$$i \in \text{Prob}_2 \quad \Leftrightarrow \quad F(i) \in \text{Prob}_1$$

**Proof** (Proposition 2.2a). If  $\hat{S}$  contains a single stock, then  $\text{lf}$  and  $\text{stCon}$  are the same problem—there is a trivial reduction from  $\text{lf}$  to  $\text{stCon}$ . Both involve finding a path from one node in a directed network to another. What’s more, each  $K$ -path to stock  $Z$  is evaluated separately. For example, in the market described by Figure 2, the path described in Equation (11) exists with or without the path described by Equation (12). This means that if  $(Z, M, T, \{s\}) \in \text{lf}$  and  $(Z, M, T, \{s'\}) \in \text{lf}$ , then  $(Z, M, T, \{s, s'\}) \in \text{lf}$ . Thus, we don’t need to check every single subset  $\hat{S} \subseteq S$  separately. To see which subsets of stocks are connected to stock  $Z$ , we can just check which stocks are connected to stock  $Z$ . This is reducible to solving  $(S - 1)$  separate instances of  $\text{stCon}$ , which is doable in polynomial time because  $\text{stCon}$  itself is solvable in polynomial time (Wigderson, 1992).  $\square$

**Remark** (Time Complexity). Let  $\text{Prob}_1$  and  $\text{Prob}_2$  denote decision problems with instances of size  $S$ .  $\text{Prob}_1$  is solvable in polynomial time if there’s a solution algorithm that runs in  $O[S^k]$  steps for some  $k > 0$ . Whereas,  $\text{Prob}_2$  requires exponential time if every solution algorithm requires  $2^{\ell \cdot S}$  steps on at least one instance for some  $\ell > 0$ .

Decision problems with polynomial-time solutions are considered tractable while those that require exponential time are not. However, a polynomial-time solution for  $\text{Prob}_1$  could require a  $k = 10000$ , and an exponential-time solution for  $\text{Prob}_2$  could use an  $\ell = 0.00001$ . For these values of  $k$  and  $\ell$ ,  $\text{Prob}_2$  would be easier to solve than  $\text{Prob}_1$  on reasonable instance sizes.

“If cases like this regularly arose in practice, then it would’ve turned out that we were using the wrong abstraction. But so far, it seems like we’re using the

right abstraction. Of the big problems solvable in polynomial time—matching, linear programming, primality testing, etc. . . —most of them really do have practical algorithms. And of the big problems that we think take exponential time—theorem-proving, circuit minimization, etc. . . —most of them really don’t have practical algorithms. (Aaronson, 2013)” In short, when seen in this context, your first guess for both  $k$  and  $\ell$  should be something like 1, 2, or 3.

**Remark** (Random Networks). To make predictions about the likelihood of being affected by an index-fund rebalancing cascade, we assume a data-generating process for the market structure. A standard way to do this is to use a random-networks model (Jackson, 2010). The particular random-networks model we use dates back to Erdos and Rényi (1960). We chose this model because it is the simplest. Our main economic insight is about complexity not networks. Proposition 2.2b can be extended to other models with power-law and exponential edge distributions. See Newman et al. (2001) for more details.

**Remark** (Percolation Threshold). The largest connected component of a directed graph is the largest set of nodes that are each connected to one another by a path. There’s a sharp phase transition in the size of the largest connected component in an Erdős-Rényi random-networks model (Bollobás, 2001). When  $\kappa < 1$ , the size of the largest connected component remains finite as  $S \rightarrow \infty$ ; whereas, when  $\kappa > 1$ , the largest connected component is infinitely large as  $S \rightarrow \infty$ . i.e., the largest connected component includes a finite fraction of infinitely many nodes. When  $\kappa > 1$ , the largest connected component is called the ‘Giant Component’. For our purposes, this percolation threshold implies that the probability of stock  $Z$  being affected by an index-fund rebalancing cascade starting somewhere else in the market is vanishingly small when  $\kappa < 1$ .

**Remark** (Connectivity Threshold). There’s a similar phase transition in the existence of small connected components for the Erdős-Rényi random-networks model (Bollobás, 2001). When  $\kappa < \log(S)$ , the typical random network will contain many small connected components; whereas, when  $\kappa > \log(S)$ , the typical random network will contain only the giant component and nodes without any edges whatsoever. For our purposes, this connectivity threshold implies that the probability stock  $Z$  isn’t affected by an index-fund rebalancing cascade starting somewhere else in the market is vanishingly small when  $\kappa > \log(S)$ .

**Proof** (Equation 13). Suppose  $\mathbf{M}$  contains  $S$  stocks and was generated using connectivity parameter  $\kappa > 0$ . If  $(s, s') \in \mathbf{S}^2$ , then stock  $s'$  will be a positive neighbor to stock  $s$  with probability  $\kappa/s$ . Because the outcome is determined independently for each stock  $s' \in \mathbf{S}$ , the probability that stock  $s$  has exactly  $n$  positive neighbors is

$$\Pr(N_s^+ = n | S) = \binom{S}{n} \cdot (\kappa/s)^n \cdot (1 - \kappa/s)^{S-n}$$

This is the probability of  $n$  successes in  $S$  independent Bernoulli trials, which implies

$$N_s^+ \sim \text{Binomial}(\kappa/s, S)$$

So, given the additional restriction that  $\kappa = O[\log(S)]$ , we know that as  $S \rightarrow \infty$

$$N_s^+ \sim \text{Poisson}(\kappa, S)$$

since the Binomial distribution converges to the Poisson distribution as  $S \rightarrow \infty$  for small values of  $\kappa$ .  $\square$

**Proof** (Proposition 2.2b). Let  $C_s \in \{\text{True}, \text{False}\}$  be an indicator variable for whether a stock  $s$  is connected to the giant component of the random graph induced by  $\mathbf{M}$ . We can write

$$\begin{aligned} \Pr[(Z, \mathbf{M}, T, \{s\}) \in \text{If} \mid N_Z = n] &= \Pr[(C_s = \text{True}) \wedge (C_Z = \text{True}) \mid N_Z = n] \\ &= \Pr[C_s = \text{True}] \cdot \Pr[C_Z = \text{True} \mid N_Z = n] \end{aligned}$$

The second line implies that  $E[\hat{S}_{\max}(Z, \mathbf{M}, T)]$  will be increasing in  $N_Z$  if and only if  $E[C_Z \mid N_Z = n]$  is increasing in  $n$  since the path connecting each stock  $s \in \mathbf{S}$  to stock  $Z$  can be evaluated independently. Bayes' rule implies

$$E[C_Z \mid N_Z = n] = \left( \frac{\Pr[N_Z = n \mid C_Z = \text{True}]}{\Pr[N_Z = n]} \right) \times E[C_Z]$$

And,  $\Pr[N_Z = n \mid C_Z = \text{True}] / \Pr[N_Z = n]$  is increasing in  $n$ . So, we can conclude that  $E[C \mid N = n]$  is increasing in  $n$ .  $\square$

**Definition** (Complexity Class NP). Let **Prob** denote a decision problem, and let  $|i|$  denote the size of instance  $i$ . We say that **Prob**  $\in$  **NP** if there exists a polynomial-time Turing machine  $\mathbf{M}$  such that

$$i \in \text{Prob} \iff \exists w \in \{0, 1\}^{\text{Poly}(|i|)} \text{ s.t. } \mathbf{M}(i, w) = 1$$

The string  $w$  is known as the ‘witness’ or ‘proof’ that  $i \in \text{Prob}$ .

**Definition** (Hardness). Let **CC** denote an arbitrary complexity class, such as **NP**. We say that **Prob** is hard with respect to **CC** if every decision problem in **CC** can be reduced to **Prob**.

**Definition** (Completeness). Let **CC** denote an arbitrary complexity class. We say that **Prob** is complete with respect to **CC** if both i) **Prob**  $\in$  **CC** and ii) **Prob** is **CC** hard.

**Problem B** (3Sat).

- *Instance: A Boolean formula defined over  $N$  input variables*

$$F : \{\text{True}, \text{False}\}^N \mapsto \{0, 1\}$$

*where some clauses contain 3 variables.*

- *Question: Is there an assignment  $\mathbf{x} \in \{\text{True}, \text{False}\}^N$  such that  $F(\mathbf{x}) = 1$ ?*

**Theorem B** (Cook, 1971). 3Sat is an NP-complete problem.

**Corollary.** Let **Prob** denote any decision problem. If **Prob** is reducible to 3Sat, then **Prob** is NP complete.

**Proof** (Proposition 2.3a). We show that **How** is **NP** complete by reducing it to **3Sat**. There are two steps to the proof.

STEP 1: First, create variables to track of the state of the rebalancing cascade:

- For each possible value of  $(x_{s,t}, \Delta x_{s,t})$ ,

$$k \in \{(0,0), (1,1), (1,0), (0, -1), (-1, -1), (-1,0), (0,1)\}$$

define for each stock  $s \in \mathcal{S}$

$$\alpha(k)_{s,t} = 1_{\{(x_{s,t}, \Delta x_{s,t})=k\}}$$

- For each pair of stocks  $(s, s') \in \mathcal{S}^2$  such that  $s \neq s'$  define

$$\beta_{s,s',t}^+ = 1_{\{s' \in \text{Out}_{s,t}^+\}}$$

$$\beta_{s,s',t}^- = 1_{\{s' \in \text{Out}_{s,t}^-\}}$$

- For each pair of stocks  $(s, s') \in \mathcal{S}^2$  such that  $s \neq s'$  define

$$\gamma_{s',s,t+1}^+ = 1_{\{s \in \text{In}_{s',t+1}^+\}}$$

$$\gamma_{s',s,t+1}^- = 1_{\{s \in \text{In}_{s',t+1}^-\}}$$

- For each stock  $s \in \mathcal{S}$  define

$$\delta_{s,t+1}^+ = 1_{\{u_{s,t+1}=1\}}$$

$$\delta_{s,t+1}^- = 1_{\{u_{s,t+1}=-1\}}$$

Total number of new variables is polynomial in  $S$ .

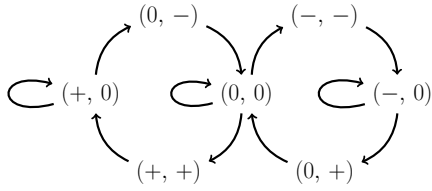
STEP 2: Encode constraints on variables in conjunctive-normal form clauses. There are two kinds of constraints to consider.

- First, there are constraints that impose variable consistency. e.g., we can't have both  $\alpha(0,0)_{s,t} = 1$  and  $\alpha(1,1)_{s,t} = 1$  at the same time:

$$(\overline{\alpha(0,0)_{s,t}} \vee \overline{\alpha(1,1)_{s,t}})$$

- Second, there are constraints that encode the rebalancing cascade updating rules. e.g., if stock  $s$  has one negative neighbor,  $s'$ , and one positive neighbor,  $s''$ , then the rebalancing-cascade rules are encoded in four different clauses:

$\delta_s^+$	$\lambda_{s',s}^+$	$\lambda_{s'',s}^-$	Violated Clause
0	0	0	✓
0	0	1	✓
0	1	0	⊗ $(\delta_s^+ \vee \bar{\lambda}_{s',s}^+ \vee \lambda_{s'',s}^-)$
0	1	1	✓
1	0	0	⊗ $(\bar{\delta}_s^+ \vee \lambda_{s',s}^+ \vee \lambda_{s'',s}^-)$
1	0	1	⊗ $(\bar{\delta}_s^+ \vee \lambda_{s',s}^+ \vee \bar{\lambda}_{s'',s}^-)$
1	1	0	✓
1	1	1	⊗ $(\bar{\delta}_s^+ \vee \bar{\lambda}_{s',s}^+ \vee \bar{\lambda}_{s'',s}^-)$



**Figure 7: State Diagram.** All possible ways that a single stock could move between the 7 possible values of  $(x_{s,t}, \Delta x_{s,t})$  in successive rounds of an index-fund rebalancing cascade. Arrows denote transitions. Loops denote unchanged values in successive rounds.

Again, the total number of new clauses is polynomial in  $S$ .

Whenever stock  $s$  has both positive and negative neighbors, some of these clauses involve 3 variables. Thus, we have a polynomial reduction of **How** to **3Sat**.  $\square$

**Definition** (Complexity Class PP). Let **Prob** denote a decision problem, and let  $r \in \{0, 1\}^*$  denote an arbitrarily long sequence of random bits. We say that **Prob**  $\in$  PP if there exists a polynomial-time randomized algorithm  $F$  such that

$$i \in \text{Prob} \iff \Pr_r[F(i, r) = 1 \mid i \notin \text{Prob}] > 1/2$$

**Problem C** (Majority).

- *Instance:* A Boolean formula defined over  $N$  input variables

$$F : \{\text{True}, \text{False}\}^N \mapsto \{0, 1\}$$

- *Question:* Is  $\sum_{\mathbf{x} \in \{\text{True}, \text{False}\}^N} F(\mathbf{x}) > 2^{N-1}$ ?

**Theorem C** (Gill, 1977).  $\text{NP} \subseteq \text{PP}$ , and **Majority** is a PP-complete problem.

**Corollary.** Let **Prob** denote any decision problem. If **Prob** is reducible to **Majority**, then **Prob** is PP hard.

**Proof** (Proposition 2.3b). The proof of Proposition 2.3a showed how to reduce instances of **How** into Boolean formulas. So, since **Majority** is defined in terms of Boolean functions, the same reduction converts instances of **MajorityHow** into instances of **Majority**. Hence, because **MajorityHow** is a PP-complete problem (Gill, 1977), the corollary above implies that **MajorityHow** is an NP-hard problem.  $\square$

**Problem D** (2Sat).

- *Instance:* A Boolean formula defined over  $N$  input variables

$$F : \{\text{True}, \text{False}\}^N \mapsto \{0, 1\}$$

where no clause contains more than 2 variables.

- *Question:* Is there an assignment  $\mathbf{x} \in \{\text{True}, \text{False}\}^N$  such that  $F(\mathbf{x}) = 1$ ?

**Theorem D** (Cook, 1971). **2Sat** is solvable in polynomial time.

**Proof** (Proposition 2.4a). If there is no alternation, then stocks only have positive neighbors. So, a stock  $Z$  will be affected by an initial shock to the stocks in  $\mathbf{A}$  if and only if there is a path from stock  $s \in \mathbf{A}$  connecting to stock  $Z$ . Without alternation, there is no way for two different paths in an index-fund rebalancing cascade to interfere with one another. And, since within a single path, each stock has only  $O$  (stock  $A$ ) or 1 (all other stocks) incoming links at any point in time, there would be no need to create clauses with more than two variables in the proof of Proposition 2.3a. Thus, without alternation, **How** is reducible to 2Sat. And, this reduction implies it's solvable in polynomial time (Cook, 1971).  $\square$

**Proof** (Proposition 2.4b). If there are no loops, then there is either a single path from any stock  $s$  to stock  $Z$  or no such path. After all, if there is more than one path, then these two paths would define a closed loop. As a result, no stock can have more than 1 incoming link. And so, the rebalancing cascade rules can be encoded using clauses with no more than 2 variables as in the proof of Proposition 2.4a. Thus, without loops, **How** is reducible to 2Sat. And, this reduction implies that it's solvable in polynomial time (Cook, 1971).  $\square$

**Problem E (SmoothHow)**. Suppose that the updating rule in Equation (6) was changed to the following for some  $\theta \in (0, 1)$ :

$$u_{s,t+1} = \frac{1}{|\ln_{s,t+1}^+| + |\ln_{s,t+1}^-|} \cdot \left( \sum_{s' \in \ln_{s,t+1}^-} x_{s',t} - \sum_{s'' \in \ln_{s,t+1}^+} x_{s'',t} \right)$$

$$x_{s,t+1} = \theta \cdot (x_{s,t} + u_{s,t+1})$$

- *Instance*: A choice for stock  $Z$ ; a market structure  $\mathbf{M}$ ; a time  $T = \text{Poly}[S]$ ; a positive constant  $\epsilon > 0$ ; and, the power set  $\hat{\mathbf{A}} \subseteq 2^{\mathbf{S}}$  of all  $\epsilon$ -small sets  $\mathbf{A} \subseteq \mathbf{S}$ .
- *Question*: Does there exist a  $\mathbf{A} \in \hat{\mathbf{A}}$  such that  $\text{Effect}_{\mathbf{M},T}(\mathbf{A}, Z) < 0$ ?

**Proposition 2.4c** (Necessity of Thresholds, Restated). Let  $i$  denote an instance of **SmoothHow**. There's a polynomial-time algorithm,  $\mathbf{F}$ , such that for any  $\delta > 0$

$$\sum_{|i|=N, i \notin \text{Prob}} \mathbf{1}_{\{\mathbf{F}(i)=1\}} < \delta$$

**Proof** (Proposition 2.4c). Because  $\theta < 1$ , the effect of a long direct path connecting to stock  $Z$  (i.e., where each stock in the path has exactly one incoming neighbor) will decay at an exponential rate. A direct path from stock  $A$  to stock  $Z$  that involves  $(K - 1)$  intermediary stocks will have an affect on stock  $Z$  proportional to  $\theta^K$ . And, the effects of any indirect paths (i.e., where each stock in the path has more than one incoming neighbor) will decay even fast due to averaging. Having more than one incoming neighbor presents that possibility that a stock will be hit by both a positive and a negative shock at the same time. So, to get an approximate solution to **SmoothHow**, just compute the effect of all direct paths connecting to stock  $Z$  of length  $K = \text{Poly}[S]$ . If there exists a path with a negative effect, then answer 'Yes'; otherwise, answer 'No'.  $\square$

## B Tables

### Summary Statistics, ETFs

a) Time Series

	Avg	Sd	1%	25%	50%	75%	99%
<b>#etf</b>	1049	153	863	911	1012	1206	1337
<b>#bmark</b>	876	128	731	768	832	1006	1117
<b>mktCap</b> [\$1b]	1291	584	2	984	1420	1683	2283

b) Cross-Section

	Avg	Sd	1%	25%	50%	75%	99%
<b>#stock</b>	241	505	1	29	75	243	2335

**Table 1:** Summary statistics for the ETF market. Data is from ETF Global. Sample period runs from January 2010 to December 2015. Panel a) reports monthly aggregates for the entire ETF market. **#etf**: number of ETFs in the sample each month. **#bmark**: number of different benchmarks reported by these ETFs. **mktCap**: total market capitalization of the ETF industry each month in billions of dollars. Panel b) reports cross-sectional statistics for fund-month observations. **#stock**: number of stocks held by an ETF in a given month.

## Summary Statistics, Trading

### a) Announcements

	Avg	Sd	Min	25%	50%	75%	Max
$\frac{\#ancmt}{month}$	14.73	5.19	4	11	14	19	28
$\frac{\#stockZ}{ancmt}$	1077.02	1151.21	162	456	641	912	4509

### b) Stocks

	Avg	Sd	Min	25%	50%	75%	Max
<b>#etfOwned</b>	17.39	16.38	1.00	2.87	13.29	27.17	60.08
$\ln(etfRebalVlm)$	7.62	2.81	0.65	5.81	8.24	9.68	12.58
<b>etfOrdImbal</b>	-0.03	0.14	-0.72	-0.05	-0.02	0.01	0.22
<b>amihud</b> [%/\$1m]	0.85	3.90	0.00	0.00	0.01	0.07	18.09
<b>baSpread</b> [%]	0.01	0.01	0.00	0.00	0.00	0.01	0.05

**Table 2:** Summary statistics for trading data. Announcement data is from Thompson Financial. Stock-market data is from CRSP and TAQ. Sample period runs from January 2010 to December 2015. Panel a) reports summary statistics for the M&A announcement data.  $\frac{\#ancmt}{month}$ : number of announcements per month.  $\frac{\#stockZ}{ancmt}$ : number of stocks that are unrelated to the stock named as the target of each M&A announcement. Panel b) reports cross-sectional summary statistics at the stock-month level. **#etfOwned**: number of ETFs that own each stock.  $\ln(etfRebalVlm)$ : log ETF rebalancing volume. **etfOrdImbal**: ETF order imbalance, which ranges from  $[-1, 1]$ . **amihud**: Amihud (2002) liquidity measure quoted in units of percent per \$1m of trading volume. **baSpread**: bid-ask spread quoted as a percent of the closing price.



# Rebalancing Volume, Stock A

		ln(etfRebalVlm <sub>A,t</sub> ) [%]				
		Actual Announcements				Placebo
$1_{\{t=t_A+1\}}$	22.07* (11.66)	-10.23 (11.78)	22.14* (11.71)	-10.15 (11.82)	-16.47 (17.41)	
$1_{\{t=t_A\}}$	165.62*** (10.25)	167.78*** (10.28)	165.69*** (10.30)	166.87*** (10.33)	-11.47 (15.42)	
$1_{\{t=t_A-1\}}$	60.66*** (10.52)	63.94*** (10.35)	60.93*** ( 9.82)	64.19*** ( 9.56)	-10.53 (19.38)	
$1_{\{t=t_A-2\}}$			5.16 ( 8.33)	6.81 ( 8.28)	10.98 (17.90)	
$1_{\{t=t_A-3\}}$			-2.45 ( 9.38)	-2.45 ( 9.35)	-13.28 (16.43)	
$1_{\{t=t_A-4\}}$			13.70 ( 9.71)	14.19 ( 9.72)	13.21 (21.36)	
$1_{\{t=t_A-5\}}$			3.59 ( 9.15)	5.81 ( 9.14)	7.45 (15.92)	
ln(vlm <sub>A,t-1</sub> )		12.29*** ( 0.41)		12.29*** ( 0.41)	12.29*** ( 0.41)	
ln(vlm <sub>A,t-2</sub> )		3.85*** ( 0.29)		3.85*** ( 0.29)	3.85*** ( 0.29)	
ln(vlm <sub>A,t-3</sub> )		3.09*** ( 0.34)		3.09*** ( 0.34)	3.09*** ( 0.34)	
Month-Year FE	Y	Y	Y	Y	Y	
Stock-Specific FE	Y	Y	Y	Y	Y	
$R^2$	57.0%	57.1%	57.0%	57.1%	27.0%	
Observations	5,197,276				5,197,276	

**Table 3:** *Effect of initial M&A announcements on ETF rebalancing volume for stock A. For the first 4 columns, the data is panel containing each M&A target in the window  $t \in \{t_A - 20, \dots, t_A + 5\}$ . The fifth column uses a new dataset of randomly selected M&A announcement dates for the same target companies.  $t_A$  denotes date of M&A announcement for stock A. ln(vlm<sub>A</sub>) is log trading volume for stock A. Table reports results for the regression:  $\ln(\text{etfRebal}_{A,t}) = \alpha + \beta \cdot 1_{\{t=t_A-1\}} + \gamma \cdot 1_{\{t=t_A\}} + \delta \cdot 1_{\{t=t_A+1\}} + \dots + \varepsilon_{A,t}$ .*

## Rebalancing Volume, Stock $Z$

	ln(etfRebalVlm $_{Z,t}$ ) [%]			
afterAncmt $_{A,t}$	4.98 <sup>***</sup> (0.28)	4.85 <sup>***</sup> (0.28)	3.27 <sup>***</sup> (0.39)	3.02 <sup>***</sup> (0.39)
manyNbrs $_{A \rightarrow Z,t}$			95.58 <sup>***</sup> (3.25)	94.05 <sup>***</sup> (3.23)
afterAncmt $_{A,t} \times$ manyNbrs $_{A \rightarrow Z,t}$			3.70 <sup>***</sup> (0.55)	3.02 <sup>***</sup> (0.55)
ln(vlm $_{Z,t}$ )		32.38 <sup>***</sup> (1.72)		31.45 <sup>***</sup> (1.70)
Announcement FE	Y	Y	Y	Y
Stock-Specific FE	Y	Y	Y	Y
$R^2$	52.6%	53.0%	53.3%	53.6%
Observations	14,736,786		14,736,786	

**Table 4:** *Effect of initial M&A announcements on ETF rebalancing volume for stock  $Z$ . Data is panel containing each stock  $Z$  that is unrelated to the target stock  $A$  in the window  $t \in \{t_A - 20, \dots, t_A + 5\}$ . For stock  $Z$  to be unrelated to a particular stock  $A$ , it has to be twice removed in the network of ETF holdings. It can't have been recently held by any ETF that also recently held stock  $A$ . And, it can't have been held by any ETF that also held a stock that was held by another ETF that held stock  $A$ . i.e., the chain has to be  $A - B - C - Z$  or longer.  $\text{afterAncmt}_A$  is an indicator variable for the 5 days following the announcement of stock  $A$  as an M&A target.  $\text{manyNbrs}_{A \rightarrow Z}$  is an indicator variable for stock  $Z$  having an above-median number of neighbors relative to stock  $A$ 's M&A announcement.  $\ln(\text{vlm}_Z)$  is log trading volume for stock  $Z$ . Table reports results for the regression:  $\ln(\text{etfRebalVlm}_{Z,t}) = \alpha + \beta \cdot \text{afterAncmt}_{A,t} + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z,t} + \delta \cdot \{\text{afterAncmt}_{A,t} \times \text{manyNbrs}_{A \rightarrow Z,t}\} + \dots + \varepsilon_{A \rightarrow Z,t}$ .*

## Order Imbalance, Stock $Z$

	etfOrdImbal $_{Z,t}$ [bps]			
afterAncmt $_{A,t}$	0.75 <sup>***</sup> (0.11)	0.74 <sup>***</sup> (0.11)	0.63 <sup>***</sup> (0.16)	0.63 <sup>***</sup> (0.16)
manyNbrs $_{A \rightarrow Z,t}$			-0.92 <sup>***</sup> (0.10)	-0.87 <sup>***</sup> (0.10)
afterAncmt $_{A,t} \times$ manyNbrs $_{A \rightarrow Z,t}$			0.24 (0.21)	0.22 (0.21)
ln(vlm $_{Z,t}$ )		-1.25 <sup>***</sup> (0.08)		-1.24 <sup>***</sup> (0.08)
Announcement FE	Y	Y	Y	Y
Stock-Specific FE	Y	Y	Y	Y
$R^2$	1.4%	1.4%	1.4%	1.4%
Observations	13,755,851		13,755,851	

**Table 5:** *Effect of initial M&A announcements on ETF order imbalance for stock  $Z$ . Data is panel containing each stock  $Z$  that is unrelated to the target stock  $A$  in the window  $t \in \{t_A - 20, \dots, t_A + 5\}$ . For stock  $Z$  to be unrelated to a particular stock  $A$ , it has to be twice removed in the network of ETF holdings. It can't have been recently held by any ETF that also recently held stock  $A$ . And, it can't have been held by any ETF that also held a stock that was held by another ETF that held stock  $A$ . i.e., the chain has to be  $A - B - C - Z$  or longer. **afterAncmt $_A$**  is an indicator variable for the 5 days following the announcement of stock  $A$  as an M&A target. **manyNbrs $_{A \rightarrow Z}$**  is an indicator variable for stock  $Z$  having an above-median number of neighbors relative to stock  $A$ 's M&A announcement. **ln(vlm $_Z$ )** is log trading volume for stock  $Z$ . Table reports results for the regression:  $\text{etfOrdImbal}_{Z,t} = \alpha + \beta \cdot \text{afterAncmt}_{A,t} + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z,t} + \delta \cdot \{\text{afterAncmt}_{A,t} \times \text{manyNbrs}_{A \rightarrow Z,t}\} + \dots + \varepsilon_{A \rightarrow Z,t}$ .*

## Liquidity Measures, Stock $Z$

	amihud $_{Z,t}$ [%/\$1m]		baSpread $_{Z,t}$ [bps]	
afterAncmt $_{A,t}$	0.80 (0.63)	0.84 (1.09)	0.17*** (0.06)	0.18* (0.10)
manyNbrs $_{A \rightarrow Z,t}$		-4.61*** (1.51)		-5.31*** (0.40)
afterAncmt $_{A,t} \times$ manyNbrs $_{A \rightarrow Z,t}$		-0.09 (1.41)		-0.04 (0.12)
Announcement FE	Y	Y	Y	Y
Stock-Specific FE	Y	Y	Y	Y
$R^2$	6.7%	6.7%	52.8%	52.9%
Observations	14,736,786		14,736,786	

**Table 6:** *Effect of initial M&A announcements on liquidity for stock  $Z$ . Data is panel containing each stock  $Z$  that is unrelated to the target stock  $A$  in the window  $t \in \{t_A - 20, \dots, t_A + 5\}$ . For stock  $Z$  to be unrelated to a particular stock  $A$ , it has to be twice removed in the network of ETF holdings. It can't have been recently held by any ETF that also recently held stock  $A$ . And, it can't have been held by any ETF that also held a stock that was held by another ETF that held stock  $A$ . i.e., the chain has to be  $A - B - C - Z$  or longer. amihud $_Z$  is Amihud (2002) illiquidity measure in units of % per million dollars. baSpread $_Z$  is bid-ask spread as a fraction of closing price. afterAncmt $_A$  is an indicator variable for the 5 days following the announcement of stock  $A$  as an M&A target. manyNbrs $_{A \rightarrow Z}$  is an indicator variable for stock  $Z$  having an above-median number of neighbors relative to stock  $A$ 's M&A announcement. For  $y \in \{\text{amihud}, \text{baSpread}\}$ , table reports results for the regression:  $y_{Z,t} = \alpha + \beta \cdot \text{afterAncmt}_{A,t} + \gamma \cdot \text{manyNbrs}_{A \rightarrow Z,t} + \delta \cdot \{\text{afterAncmt}_{A,t} \times \text{manyNbrs}_{A \rightarrow Z,t}\} + \dots + \varepsilon_{A \rightarrow Z,t}$ .*

## Unconditional Panel Regression

	ln(etfRebalVlm <sub>s,t</sub> ) [%]		amihud <sub>s,t</sub> [%/\$100m]		baSpread <sub>s,t</sub> [bps×100]	
#nhbrs <sub>s,t</sub>	0.12*** (0.01)	0.06*** (0.01)	-1.47*** (0.31)	-0.63** (0.27)	-0.88*** (0.08)	-0.25*** (0.06)
Month-Year FE	Y		Y		Y	
#etf-Specific FE	Y		Y		Y	
Stock-Specific FE	Y		Y		Y	
<i>R</i> <sup>2</sup>	60.6%	43.4%	6.5%	6.5%	53.6%	53.7%
Observations	4,915,505		4,966,292		4,986,730	

**Table 7:** *Unconditional relationship between number of neighbors and a stock’s ETF rebalancing volume and liquidity. Data consists of a panel containing all stock-month observations in our sample. #nhbrs<sub>s</sub> is the number of neighboring stocks to stock s in the ETF rebalancing-rule network. amihud<sub>s</sub> is Amihud (2002) illiquidity measure in units of % per 100 million dollars. baSpread<sub>s</sub> is bid-ask spread as a fraction of closing price times 100. #etf-specific fixed effects denote indicator variables for the number of ETFs that hold a given stock in a given month. For  $y \in \{\ln(\text{etfRebalVlm}), \text{amihud}, \text{baSpread}\}$ , table reports results for the regressions  $y_{s,t} = \alpha + \beta \cdot \text{\#nhbrs}_{s,t} + \dots + \varepsilon_{s,t}$ .*

# Information, Liquidity, and Dynamic Limit Order Markets\*

Roberto Ricc<sup>†</sup>    Barbara Rindi<sup>‡</sup>    Duane J. Seppi<sup>§</sup>

May 16, 2018

## Abstract

This paper describes price discovery and liquidity provision in a dynamic limit order market with asymmetric information and non-Markovian learning. In particular, investors condition on information in both the current limit order book and also, unlike in previous research, on the prior trading history when deciding whether to provide or take liquidity. Numerical examples show that the information content of the prior order history can be substantial. In addition, the information content of arriving orders can be non-monotone in both the direction and aggressiveness of arriving orders.

JEL classification: G10, G20, G24, D40

Keywords: Limit order markets, asymmetric information, liquidity, market microstructure

---

\*We thank Sandra Fortini, Thierry Foucault, Paolo Giacomazzi, Burton Hollifield, Phillip Illeditsch, Stefan Lewellen, Marco Ottaviani, Tom Ruchti, and seminar participants at Carnegie Mellon University for helpful comments. We are grateful to Fabio Sist for his significant contribution to the computer code for our model.

<sup>†</sup>Bocconi University. Phone: +39-02-5836-2715. E-mail: roberto.ricco@phd.unibocconi.it

<sup>‡</sup>Bocconi University and IGIER. Phone: +39-02-5836-5328. E-mail: barbara.rindi@unibocconi.it

<sup>§</sup>Tepper School of Business, Carnegie Mellon University. Phone: 412-268-2298. E-mail: ds64@andrew.cmu.edu

The aggregation of private information and the dynamics of liquidity supply and demand are closely intertwined in financial markets. In dealer markets, informed and uninformed investors trade via market orders and, thus, take liquidity, while dealers provide liquidity and try to extract information from the arriving order flow (e.g., as in Kyle (1985) and Glosten and Milgrom (1985)). However, in limit order markets — the dominant form of securities market organization today — the relation between who has information and who is trying to learn it and who supplies and demands liquidity is not well understood theoretically.<sup>1</sup> Recent empirical research highlights the role of informed traders not only as liquidity takers but also as liquidity suppliers. O’Hara (2015) argues that fast informed traders use market and limit orders interchangeably and often prefer limit orders to marketable orders. Fleming, Mizrach, and Nguyen (2017) and Brogaard, Hendershott, and Riordan (2016) find that limit orders play a significant empirical role in price discovery.<sup>2</sup>

Our paper presents the first rational expectations model of a dynamic limit order market with asymmetric information and history-dependent Bayesian learning. In particular, learning is not constrained to be Markovian. The model represents a trading day with market opening and closing effects. Our model lets us investigate the information content of different types of market and limit orders, the dynamics of who provides and demands liquidity, and the non-Markovian information content of the trading history. In addition, we study how changes in the amount of adverse selection — in terms of both asset-value volatility and the arrival probability of informed investors — affect equilibrium trading strategies, liquidity, price discovery, and welfare. We have three main results:

- Increased adverse selection does not always worsen market liquidity as in Kyle (1985). Liquidity can improve if informed traders with better information trade more aggressively by submitting more limit-orders at the inside quotes rather than using market orders.

---

<sup>1</sup>See Jain (2005) for a discussion of the prevalence of limit order markets. See Parlour and Seppi (2008) for a survey of theoretical models of limit order markets. See Rindi (2008) for a model of informed traders as liquidity providers.

<sup>2</sup>Gencay, Mahmoodzadeh, Rojcek, and Tseng (2016) investigate brief episodes of high-intensity/extreme behavior of quotation process in the U.S. equity market (bursts in liquidity provision that happen several hundreds of time a day for actively traded stocks) and find that liquidity suppliers during these bursts significantly impact prices by posting limit orders.

- The relation between limit and market orders and their information content depends on the size of private information shocks relative to the tick size. Indeed, the information content of orders can even be opposite the order direction and aggressiveness.
- The learning dynamics are non-Markovian in that the order history has information in addition to the current state of the limit order book. In particular, the incremental information content of arriving limit and market orders is history-dependent.

Dynamic limit order markets with uninformed investors are studied in a large literature. This includes Foucault (1999), Parlour (1998), Foucault, Kadan, and Kandel (2005), and Goettler, Parlour, and Rajan (2005). There is some previous theoretical research that allows informed traders to supply liquidity. Kumar and Seppi (1994) is a static model in which optimizing informed and uninformed investors use profiles of multiple limit and market orders to trade. Kaniel and Liu (2006) extend the Glosten and Milgrom (1985) dealership market to allow informed traders to post limit orders. Aït-Sahalia and Saglam (2013) also allow informed traders to post limit orders, but they do not allow them to choose between limit and market orders. Moreover, the limit orders posted by their informed traders are always at the best bid and ask prices. Goettler, Parlour, and Rajan (2009) allow informed and uninformed traders to post limit or market orders, but their model is stationary and assumes Markovian learning. Roşu (2016b) studies a steady-state limit order market equilibrium in continuous-time with Markovian learning and additional information-processing restrictions. These last two papers are closest to ours. Our model differs from them in two ways: First, they assume Markovian learning in order to study dynamic trading strategies with order cancellation, whereas we simplify the strategy space (by not allowing dynamic order cancellations and submissions) in order to investigate non-Markovian learning (i.e., our model has a larger state space with full order histories). Second, we model a non-stationary trading day with opening and closing effects. Market opens and closes are important daily events in the dynamics of liquidity in financial markets. Bloomfield, O’Hara, and Saar (2005) show in an experimental asset market setting that informed traders sometimes provide more liquidity than uninformed traders. Our model provides equilibrium examples of liquidity provision by informed investors.

A growing literature investigates the relation between information and trading speed (e.g., Biais,



Foucault, and Moinas (2015); Foucault, Hombert, and Roşu (2016); and Roşu (2016a)). However, these models assume Kyle or Glosten-Milgrom market structures and, thus, cannot consider the roles of informed and uninformed traders as endogenous liquidity providers and demanders. We argue that understanding price discovery dynamics in limit order markets is an essential precursor to understanding speedbumps and cross-market competition given the real-world prevalence of limit order markets.

## 1 Model

We consider a limit order market in which a risky asset is traded at five times  $t_j \in \{t_1, t_2, t_3, t_4, t_5\}$  over a trading day. The fundamental value of the asset after time  $t_5$  at the end of the day is

$$\tilde{v} = v_0 + \Delta = \begin{cases} \bar{v} = v_0 + \delta & \text{with } Pr(\bar{v}) = \frac{1}{3} \\ v_0 & \text{with } Pr(v_0) = \frac{1}{3} \\ \underline{v} = v_0 - \delta & \text{with } Pr(\underline{v}) = \frac{1}{3} \end{cases} \quad (1)$$

where  $v_0$  is the ex ante expected asset value, and  $\Delta$  is a symmetrically distributed value shock. The limit order market allows for trading through two types of orders: Limit orders are price-contingent orders that are collected in a limit order book. Market orders are executed immediately at the best available price in the limit order book. The limit order book has a price grid with four prices,  $P_i \in \{A_2, A_1, B_1, B_2\}$ , two each on the ask and bid sides of the market. The tick size is equal to  $\kappa > 0$ , and the ask prices are  $A_1 = v_0 + \frac{\kappa}{2}$ ,  $A_2 = v_0 + \kappa$ ; and by symmetry the bid prices are  $B_1 = v_0 - \frac{\kappa}{2}$ ,  $B_2 = v_0 - \kappa$ . Order execution in the limit order book follows time and price priority.

Investors arrive sequentially over time to trade in the market. At each time  $t_j$  one investor arrives. Investors are risk-neutral and asymmetrically informed. A trader is informed with probability  $\alpha$  and uninformed with probability  $1 - \alpha$ . Informed investors know the realized value shock  $\Delta$  perfectly. Uninformed investors do not know  $\Delta$ , so they use Bayes' Rule and their knowledge of the equilibrium to learn about  $\Delta$  from the observable market dynamics over time. An investor arriving at time  $t_j$  may also have a personal private-value trading motive, which — we assume

for tractability — causes them to adjust their valuation of  $v_0$  to  $\beta_{t_j}v_0$  where the factor  $\beta_{t_j}$  may be greater than or less than 1. Non-informational private-value motives include preference shocks, hedging needs, and taxation. The absence of a non-informational trading motive would lead to the Milgrom and Stokey (1982) no-trade result. The factor  $\beta_{t_j}$  at time  $t_j$  is drawn from a truncated normal distribution,  $Tr[\mathcal{N}(\mu, \sigma^2)]$ , with support over the interval  $[0, 2]$ . The mean is  $\mu = 1$ , which corresponds to a neutral private valuation. Traders with neutral private factors tend to provide liquidity symmetrically on both the buy and sell sides of the market, while traders with extreme private valuations provide one-sided liquidity or actively take liquidity. The parameter  $\sigma$  determines the dispersion of a trader’s private-value factor  $\beta_{t_j}$ , as shown in Figure 1, and, thus, the probability of large private gains-from-trade due to extreme investor private valuations.

The sequence of arriving investors is independently and identically distributed in terms of whether they are informed or uninformed and in terms of their individual private-value factors  $\beta_{t_j}$ . In one specification of our model, only uninformed investors have private valuations, while in a second richer specification both informed and uninformed investors have private valuations. A generic informed investor is denoted as  $I$ , where we denote the informed investor as  $I_{\bar{v}}$  if the value shock is positive ( $\Delta = \delta$ ), as  $I_{\underline{v}}$  if the shock is negative ( $\Delta = -\delta$ ), and as  $I_{v_0}$  if the shock is zero ( $\Delta = 0$ ). Informed investors arriving at different times during the day all have the identical asset-value information (i.e., there is only one realized  $\Delta$ ). Uninformed investors are denoted as  $U$ .

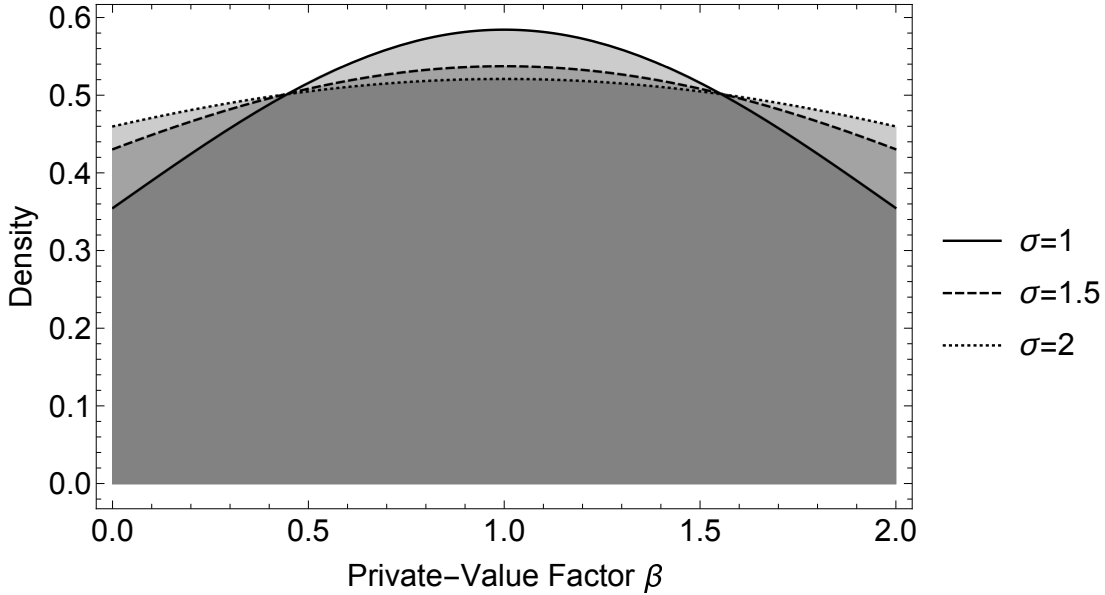
An investor arriving at time  $t_j$  can take one of seven possible actions  $x_{t_j}$ : One possibility is to submit a buy or sell market order  $MOA_{i,t_j}$  or  $MOB_{i,t_j}$  to buy or sell immediately at the best available ask or bid respectively in the limit order book at time  $t_j$ . A subscript  $i = 1$  indicates that the best standing quote at time  $t_j$  is at the inside prices  $A_1$  or  $B_1$ , and  $i = 2$  means the best quote is at the outside prices  $A_2$  or  $B_2$ . Alternatively, the investor can submit one of four possible limit orders  $LOA_i$  and  $LOB_i$  on the ask or bid side of the book, respectively. A subscript  $i = 1$  denotes an aggressive limit order posted at the inside quote, and  $i = 2$  is a less aggressive limit order at the outside quotes.<sup>3</sup> Yet another alternative is to choose to do nothing ( $NT$ ).

For tractability, we make a few simplifying assumptions. Limit orders cannot be modified or

---

<sup>3</sup>For tractability, it is assumed that investors cannot post buy limit orders at  $A_1$  and sell limit orders at  $B_1$ . This is one way in which the investor action space is simplified in our model.

**Figure 1: Distribution of Traders' Private-Value Factors -  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This figure shows the truncated Normal probability density Function (PDF) of trader private-value factors  $\beta_{t_j}$  with a mean  $\mu = 1$  and three different values of dispersion  $\sigma$ .



canceled after submission. Thus, each arriving investor has one and only one opportunity to submit an order. There is also no quantity decision. Orders are to buy or sell one share. Lastly, investors can only submit one order. Taken together, these assumptions let us express the traders' action set as  $X_{t_j} = \{MOB_{i,t_j}, LOA_1, LOA_2, NT, LOB_2, LOB_1, MOA_{i,t_j}\}$ , where each of the orders denotes an order for one share.<sup>4</sup>

In addition to the arriving informed and uninformed traders, there is a market-making *trading crowd* that submits limit orders to provide liquidity. By assumption, the crowd just posts single limit orders at the outside prices  $A_2$  and  $B_2$ . The market opens with an initial book submitted by the crowd at time  $t_0$ . After the order-submission by arriving informed and uninformed investors at each time  $t_j$ , the crowd replenishes the book at the outside prices, as needed, when either side of the book is empty. Otherwise, if there are limit orders on both sides of the book, the crowd does not submit any further limit orders. For tractability, we assume public limit orders by the arriving informed and uninformed investors have priority over limit orders from the crowd. The focus of our

<sup>4</sup>The action space  $X_{t_j}$  of orders that can be submitted at time  $t_j$  includes market orders at the standing best bid or offer at time  $t_j$ . Our notation  $MOB_{i,t_j}$  and  $MOA_{i,t_j}$  reflects the fact that the bid or offer at time  $t_j$  is not a fixed number but rather depends on the incoming state of the limit order book. There is no time script in the limit order notation  $LOA_1, \dots$  because these are just limit orders at particular fixed prices in the price grid.

model is on market dynamics involving information and liquidity given the behavior of optimizing informed and uninformed investors. The crowd is simply a modeling device to insure it is always possible for arriving investors to trade with market orders if they so choose.

Market dynamics over the trading day are intentionally non-stationary in our model in order to capture market opening and closing effects. When the market opens at  $t_1$ , the only standing limit orders in the book are those at prices  $A_2$  and  $B_2$  from the trading crowd.<sup>5</sup> At the end of the day all unexecuted limit orders are cancelled. The state of the limit order book at a generic time  $t_j$  during the day is

$$L_{t_j} = [q_{t_j}^{A_2}, q_{t_j}^{A_1}, q_{t_j}^{B_1}, q_{t_j}^{B_2}] \quad (2)$$

where  $q_{t_j}^{A_i}$  and  $q_{t_j}^{B_i}$  indicate the total depths at prices  $A_i$  and  $B_i$  at time  $t_j$ . The limit order book changes over time due to the arrival of new limit orders (which augment the depth of the book) and market orders (which remove depth from the book) from arriving informed and uninformed investors and due to the submission of limit orders from the crowd. The resulting dynamics are:

$$L_{t_j} = L_{t_{j-1}} + Q_{t_j} + C_{t_j} \quad j = 1, \dots, 5 \quad (3)$$

---

<sup>5</sup>In practice, daily opening limit order books include uncanceled orders from the previous day and new limit orders from opening auctions. For simplicity, we abstract from these interesting features of markets.

where  $Q_{t_j}$  is the change in the limit order book due an arriving investor's action  $x_{t_j} \in X_{t_j}$  at  $t_j$ :<sup>6</sup>

$$Q_{t_j} = [Q_{t_j}^{A_2}, Q_{t_j}^{A_1}, Q_{t_j}^{B_1}, Q_{t_j}^{B_2}] = \begin{cases} [-1, 0, 0, 0] & \text{if } x_{t_j} = MOA_2 \\ [0, -1, 0, 0] & \text{if } x_{t_j} = MOA_1 \\ [+1, 0, 0, 0] & \text{if } x_{t_j} = LOA_2 \\ [0, +1, 0, 0] & \text{if } x_{t_j} = LOA_1 \\ [0, 0, 0, 0] & \text{if } x_{t_j} = NT \\ [0, 0, +1, 0] & \text{if } x_{t_j} = LOB_1 \\ [0, 0, 0, +1] & \text{if } x_{t_j} = LOB_2 \\ [0, 0, -1, 0] & \text{if } x_{t_j} = MOB_1 \\ [0, 0, 0, -1] & \text{if } x_{t_j} = MOB_2 \end{cases} \quad (4)$$

where “+1” with a limit order denotes the arrival of an additional order at a particular limit price and “-1” with a market order denotes execution of an earlier BBO limit order and where  $C_{t_j}$  is the change in the limit order book due to any limit orders submitted by the crowd

$$C_{t_j} = \begin{cases} [1, 0, 0, 0] & \text{if } q_{t_{j-1}}^{A_2} + Q_{t_j}^{A_2} = 0 \\ [0, 0, 0, 1] & \text{if } q_{t_{j-1}}^{B_2} + Q_{t_j}^{B_2} = 0. \\ [0, 0, 0, 0] & \text{otherwise.} \end{cases} \quad (5)$$

A potentially important source of information at time  $t_j$  is the observed history of orders at prior times  $t_1, \dots, t_{j-1}$ . In particular, when traders arrive in the market, they observe the history of market activity up through the current standing limit order book at the time they arrive. However, since orders from the crowd have no incremental information beyond that in the arriving investor orders, we exclude them from the notation for the portion of the order-flow history used for informational updating of investor beliefs, which we denote by  $\mathcal{L}_{t_{j-1}} = \{Q_{t_1}, \dots, Q_{t_{j-1}}\}$ .

Investors trade using optimal order-submission strategies given their information and any private-value motive. If an uninformed investor arrives at time  $t_j$ , then his order  $x_{t_j}$  is chosen to maximize

---

<sup>6</sup>There are nine alternatives in (4) because we allow separately for cases in which the best bid and ask for market sells and buys at time  $t_j$  are at the inside and outside quotes.

his expected terminal payoff

$$\begin{aligned}
\max_{x \in X_{t_j}} w^U(x | \beta_{t_j}, \mathcal{L}_{t_{j-1}}) &= E[(\beta_{t_j} v_0 + \Delta - p(x)) f(x) | \beta_{t_j}, \mathcal{L}_{t_{j-1}}] \\
&= \begin{cases} [\beta_{t_j} v_0 + E[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x] - p(x)] Pr(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}}) & \text{if } x \text{ is a buy order} \\ [p(x) - (\beta_{t_j} v_0 + E[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x])] Pr(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}}) & \text{if } x \text{ is a sell order} \end{cases}
\end{aligned} \tag{6}$$

where  $p(x)$  is the price at which order  $x$  trades, and  $f(x)$  denotes the amount of the submitted order that is actually “filled.” If  $x$  is a market order, then  $p(x)$  is the best standing quote on the other side of the market at time  $t_j$ , and  $f(x) = 1$  for a market buy and  $f(x) = -1$  for a market sell (i.e., all of the order is executed). If  $x$  is a non-marketable limit order, then the execution price  $p(x)$  is its limit price, but the fill amount  $f(x)$  is random variable equal to zero if the limit order is never executed and equal to 1 if a limit buy is filled and  $-1$  if a limit sell is filled. If the investor does not trade — either because no order is submitted ( $NT$ ) or because a limit order is not filled — then  $f(x)$  is zero. In the second line of (6), the expression  $\theta_{t_j}^x$  denotes the set of future trading states in which an order  $x$  submitted at time  $t_j$  is executed.<sup>7</sup> This conditioning matters for limit orders because the sequence of subsequent orders in the market, which may or may not result in the execution of a limit order submitted at time  $t_j$ , is correlated with the asset value shock  $\Delta$ . For example, future market buy orders are more likely if the  $\Delta$  shock is positive (since the average  $I_{\bar{v}}$  investors will want to buy but not the average  $I_{\underline{v}}$  investor). Uninformed investors rationally take the relation between future orders and  $\Delta$  into account when forming their expectation  $E[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$  of what the asset will be worth in states in which their limit orders are executed. The second line of (6) also makes clear that uninformed investors use the prior order history  $\mathcal{L}_{t_{j-1}}$  in two ways: It affects their beliefs about limit order execution probabilities  $Pr(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}})$  and their execution-state-contingent asset-value expectations  $E[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$ .

---

<sup>7</sup>A market orders  $x_{t_j}$  is executed immediately at time  $t_j$  and so is executed for sure.

An informed investor who arrives at  $t_j$  chooses an order  $x_{t_j}$  to maximize her expected payoff

$$\begin{aligned} \max_{x \in X_{t_j}} w^I(x|v, \beta_{t_j}, \mathcal{L}_{t_{j-1}}) &= E[(\beta_{t_j} v_0 + \Delta - p(x)) f(x) | \beta_{t_j}, \mathcal{L}_{t_{j-1}}] \\ &= \begin{cases} [\beta_{t_j} v_0 + \Delta - p(x)] Pr(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}}) & \text{if } x \text{ is a buy order} \\ [p(x) - (\beta_{t_j} v_0 + \Delta)] Pr(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}}) & \text{if } x \text{ is a sell order} \end{cases} \end{aligned} \quad (7)$$

The only uncertainty for informed investors is about whether any limit orders they submit will be executed. Their belief about order-execution probabilities  $Pr(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}})$  are conditioned on both the trading history up through the current book and on their knowledge about the ending asset value. Thus, informed traders condition on  $\mathcal{L}_{t_{j-1}}$ , not to learn about the value shock  $\Delta$  (which they already know) or about future investor private-value factors  $\beta_{t_j}$  (which are i.i.d. over time), but rather because they understand that the trading history is an input in the trading behavior of future uninformed investors with whom they might trade in the future. Our analysis considers two model specifications for the informed investors. In the first, informed investors have no private-value motive, so that their  $\beta$  factors are equal to 1. In the second specification, their  $\beta$  factors are random and are independently drawn from the same truncated normal distribution  $Tr[\mathcal{N}(\mu, \sigma^2)]$  as the uninformed investors.

The optimization problem in (6) defines sets of actions  $x_{t_j} \in X_{t_j}$  that are optimal for the uninformed investor at different times  $t_j$  given different private-value factors  $\beta_{t_j}$  and order histories  $\mathcal{L}_{t_{j-1}}$ . These optimal orders can be unique, or there may be multiple orders which make the uninformed investor equally well-off. The *optimal order-submission strategy* for the uninformed investor is a probability function  $\varphi_{t_j}^U(x | \beta_{t_j}, \mathcal{L}_{t_{j-1}})$  that is zero if the order  $x$  is suboptimal and equals a mixing probability over optimal orders. If an optimal order  $x$  is unique, then  $\varphi_{t_j}(x | \beta_{t_j}, \mathcal{L}_{t_{j-1}}) = 1$ . Similarly, the optimization problem in (7) can be used to define an optimal order-submission strategy  $\varphi_{t_j}^I(x | \beta_{t_j}, v, \mathcal{L}_{t_{j-1}})$  for informed investors at time  $t_j$  given their factor  $\beta_{t_j}$ , their knowledge about the asset value  $v$ , and the order history  $\mathcal{L}_{t_{j-1}}$ .

## 1.1 Equilibrium

An equilibrium is a set of mutually consistent optimal strategy functions and beliefs for uninformed and informed investors for each time  $t_j$ , given each order history  $\mathcal{L}_{t_j-1}$ , private-value factor  $\beta_{t_j}$ , and (for informed traders) private information  $v$ . This section explains what “mutually consistent” means and then gives a formal definition of an equilibrium.

A central feature of our model is asymmetric information. The presence of informed traders means that, by observing orders over time, uninformed traders can infer information about the asset value  $v$  and use it in their order-submission strategies. More precisely, uninformed traders rationally learn from the trading history about the probability that  $v$  will go up, stay constant, or go down. However, investors cannot learn about the private values ( $\beta$ ) or information status ( $I$  or  $U$ ) of future traders since, by assumption, these are both i.i.d over time. Informed traders do not need to learn about  $v$  since they know it directly. However, they do condition their orders on  $v$  (both because  $v$  is the final stock value and also because  $v$  tells them what types of informed traders will arrive in the future along with the uninformed traders). Informed investors also condition on the order-flow history  $\mathcal{L}_{t-1}$ , since  $\mathcal{L}_{t-1}$  affects the trading behavior of future investors.<sup>8</sup>

The underlying *economic state* in our model is the realization of the asset value  $v$  and a realized sequence of investors who arrive in the market. The investor who arrives at time  $t_j$  is described by two characteristics: their status as being informed or uninformed,  $I$  or  $U$ , and their private-value factor  $\beta_{t_j}$ . The underlying economic state is exogenously chosen over time by Nature. More formally, it follows an exogenous stochastic process described by the model parameters  $\delta$ ,  $\alpha$ ,  $\mu$ , and  $\sigma$ . A sequence of arriving investors together with a pair of strategy functions — which we denote here as  $\Phi = \{\varphi_{t_j}^U(x|\beta_{t_j}, \mathcal{L}_{t_j-1}), \varphi_{t_j}^I(x|\beta_{t_j}, v, \mathcal{L}_{t_j-1})\}$  — induce a sequence of trading actions  $x_{t_j}$  which — together with the predictable actions of the trading crowd — results in a sequence of observable changes in the state  $L_{t_j}$  of the limit order book. Thus, the stochastic process generating paths of order histories is induced by the economic state process and the strategy functions. Given the order-path process, several probabilistic quantities can compute directly: First, we can com-

---

<sup>8</sup>The order history  $\mathcal{L}_{t-1}$  is an input in the uninformed-investor learning problem and, thus, is an input in their order-submission strategy. In addition, since future informed investors know that  $\mathcal{L}_{t-1}$  can affect uninformed investor trading behavior, it also enters the order-submission strategies of future informed investors.



pute the unconditional probabilities of different paths  $Pr(\mathcal{L}_{t_j})$  and the conditional probabilities  $Pr(Q_{t_j}|\mathcal{L}_{t_{j-1}})$  of particular order book changes  $Q_{t_j}$  due to arriving investors given a prior history  $\mathcal{L}_{t_{j-1}}$ . Certain paths of orders are *possible* (i.e., have positive probability  $Pr(\mathcal{L}_{t_j})$ ) given the strategy functions  $\{\varphi_{t_j}^U(x|\beta, \mathcal{L}_{t_{j-1}}), \varphi_{t_j}^I(x|\beta, v, \mathcal{L}_{t_{j-1}})\}$ , and certain paths of orders are not possible (i.e., for which  $Pr(\mathcal{L}_{t_j}) = 0$ ). Second, the endogenous order-path process also determines the order-execution probabilities  $Pr(\theta_{t_j}^x|v, \mathcal{L}_{t_{j-1}})$  and  $Pr(\theta_{t_j}^x|\mathcal{L}_{t_{j-1}})$  for informed and uninformed investors for various orders  $x$  submitted at time  $t_j$ . Computing each of these probabilities is simply a matter of listing all of the possible underlying economic states, mechanically applying the order-submission rules, identifying the relevant outcomes path-by-path, and then taking expectations across paths.

Let  $\ell$  denote the set of all feasible histories  $\{\mathcal{L}_{t_j} : j = 1, \dots, 4\}$  of physically available orders of lengths up to four trading periods. A four-period long history is the longest history a order-submission strategy can depend on in our model. In this context, *feasible* paths are simply sequences of actions from the action choice sets  $X_{t_j}$  over time without regard to whether they are *possible* in the sense that they occur with positive probability given the strategy functions  $\Phi$ . Let  $\ell^{in, \Phi}$  denote the subset of all possible trading paths in  $\ell$  that have positive probability,  $Pr(\mathcal{L}_{t_j}) > 0$ , given a pair of order strategies  $\Phi$ . Let  $\ell^{off, \Phi}$  denote the complementary set of trading paths that are *feasible* but *not possible* given  $\Phi$ . This notation will be useful when discussing “off equilibrium” beliefs. In our analysis, strategy functions  $\Phi$  are defined for all feasible paths in  $\ell$ . In particular, this includes all of the possible paths in  $\ell^{in, \Phi}$  given  $\Phi$  and also the paths in  $\ell^{off, \Phi}$ . As a result, the probabilities  $Pr(Q_{t_j}|\mathcal{L}_{t_{j-1}})$ ,  $Pr(\theta_{t_j}^x|v, \mathcal{L}_{t_{j-1}})$  and  $Pr(\theta_{t_j}^x|\mathcal{L}_{t_{j-1}})$  are always well-defined, because the continuation trading process going forward — even after an unexpected order-arrival event (i.e., a path  $\mathcal{L}_{t_{j-1}} \in \ell^{off, \Phi}$ ) — is still well-defined.

The stochastic process for order paths and its relation to the underlying economic state also determine the uninformed-investor expectations  $E[v|\mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$  of the terminal asset value given the previous order history ( $\mathcal{L}_{t_{j-1}}$ ) and conditional on future limit-order execution ( $\theta_{t_j}^x$ ). These expectations are determined as follows:

- Step 1: The conditional probabilities  $\pi_{t_j}^v = Pr(v|\mathcal{L}_{t_j})$  of a particular final asset value  $v = \bar{v}, v_0$  or  $\underline{v}$  given a possible trading history  $\mathcal{L}_{t_j} \in \ell^{in, \Phi}$  up through time  $t_j$  is given by Bayes’ Rule.

At time  $t_1$ , this probability is

$$\begin{aligned}
\pi_{t_1}^v &= \frac{Pr(v, \mathcal{L}_{t_1})}{Pr(\mathcal{L}_{t_1})} = \frac{Pr(\mathcal{L}_{t_1}|v)Pr(v)}{Pr(\mathcal{L}_{t_1})} = \frac{Pr(Q_{t_1}|v)Pr(v)}{Pr(Q_{t_1})} \\
&= \frac{Pr(Q_{t_1}|v, I)Pr(I) + Pr(Q_{t_1}|U)Pr(U)}{Pr(Q_{t_1})} Pr(v) \\
&= \frac{E^\beta[\varphi_{t_1}^I(x_{t_1}|\beta_{t_1}^I, v)|v]\alpha + E^\beta[\varphi_{t_1}^U(x_{t_1}|\beta_{t_1}^U)](1-\alpha)}{Pr(Q_{t_1})} \pi_{t_0}^v
\end{aligned} \tag{8}$$

where the prior is the unconditional probability  $\pi_{t_0}^v = Pr(v)$ ,  $x_{t_1}$  is the order at time  $t_1$  that leads to the order book change  $Q_{t_1}$ , and  $\beta_{t_1}^I$  and  $\beta_{t_1}^U$  are independently distributed private-value  $\beta$  realizations for informed and uninformed investors at time  $t_1$ .<sup>9</sup> At times  $t_j > t_1$ , the history-conditional probabilities are given recursively by<sup>10</sup>

$$\begin{aligned}
\pi_{t_j}^v &= \frac{Pr(v, \mathcal{L}_{t_j})}{Pr(\mathcal{L}_{t_j})} = \frac{Pr(v, Q_{t_j}, \mathcal{L}_{t_{j-1}})}{Pr(Q_{t_j}, \mathcal{L}_{t_{j-1}})} \\
&= \frac{\left( \begin{aligned} &Pr(Q_{t_j}|v, \mathcal{L}_{t_{j-1}}, I)Pr(I|\mathcal{L}_{t_{j-1}})Pr(v|\mathcal{L}_{t_{j-1}}) \\ &+ Pr(Q_{t_j}|v, \mathcal{L}_{t_{j-1}}, U)Pr(U|\mathcal{L}_{t_{j-1}})Pr(v|\mathcal{L}_{t_{j-1}}) \end{aligned} \right)}{Pr(Q_{t_j}|\mathcal{L}_{t_{j-1}})} \\
&= \frac{E^\beta[\varphi_{t_j}^I(x_{t_j}|\beta_{t_j}^I, v, \mathcal{L}_{t_{j-1}})|v, \mathcal{L}_{t_{j-1}}] \alpha + E^\beta[\varphi_{t_j}^U(x_{t_j}|\beta_{t_j}^U, \mathcal{L}_{t_{j-1}})|\mathcal{L}_{t_{j-1}}] (1-\alpha)}{Pr(Q_{t_j}|\mathcal{L}_{t_{j-1}})} \pi_{t_{j-1}}^v
\end{aligned} \tag{9}$$

Given these probabilities, the expected asset value conditional on the order history is

$$E[\tilde{v}|\mathcal{L}_{t_{j-1}}] = \pi_{t_{j-1}}^{\bar{v}} \bar{v} + \pi_{t_{j-1}}^{v_0} v_0 + \pi_{t_{j-1}}^v v \tag{10}$$

- Step 2: The conditional probabilities  $\pi_{t_j}^v$  given a “feasible but not possible in equilibrium” order history  $\mathcal{L}_{t_j} \in \ell^{off, \Phi}$  in which a limit order book change  $Q_{t_j}$  that is inconsistent with the strategies  $\Phi$  at time  $t_j$  are set as follows:

<sup>9</sup>A trader’s information status ( $I$  or  $U$ ) is independent of the asset value  $v$ , so  $P(I|v) = Pr(I)$  and  $Pr(U|v) = Pr(U)$ . Furthermore, uninformed traders have no private information about  $v$ , so the probability  $Pr(Q_{t_1}|U)$  with which they take a trading action  $Q_{t_1}$  does not depend on  $v$ .

<sup>10</sup>A trader’s information status is again independent of  $v$ , and it is also independent of the past trading history  $\mathcal{L}_{t_1}$ . While the probability with which an uninformed trader takes a trading action  $Q_{t_1}$  may depend on the past order history  $\mathcal{L}_{t_j}$ , it does not depend directly on  $v$  which uninformed traders do not know.

1. If the priors are fully revealing in that  $\pi_{t_{j-1}}^v = 1$  for some  $v$ , then  $\pi_{t_j}^v = \pi_{t_{j-1}}^v$  for all  $v$ .
  2. If the priors are not fully revealing at time  $t_j$ , then  $\pi_{t_j}^v = 0$  for any  $v$  for which  $\pi_{t_{j-1}}^v = 0$  and the probabilities  $\pi_{t_j}^v$  for the remaining  $v$ 's can be any non-negative numbers such that  $\pi_{t_j}^{\bar{v}} + \pi_{t_j}^{v_0} + \pi_{t_j}^v = 1$ .
  3. Thereafter, until any next unexpected trading event, the subsequent probabilities  $\pi_{t_{j'}}^v$  for  $j' > j$  are updated according to Bayes' Rule as in (9).
- Step 3: The execution-contingent conditional probabilities  $\hat{\pi}_{t_j}^v = Pr(v|\mathcal{L}_{t_{j-1}}, \theta_{t_j}^x)$  of a final asset value  $v$  conditional on a prior path  $\mathcal{L}_{t_{j-1}}$  and on execution of a limit order  $x$  submitted at time  $t_j$  is

$$\begin{aligned}\hat{\pi}_{t_j}^v &= \frac{Pr(\mathcal{L}_{t_{j-1}})Pr(v|\mathcal{L}_{t_{j-1}})\Pr(\theta_{t_{j-1}}^x|v, \mathcal{L}_{t_{j-1}})}{Pr(\theta_{t_j}^x, \mathcal{L}_{t_{j-1}})} \\ &= \frac{Pr(\theta_{t_j}^x|v, \mathcal{L}_{t_{j-1}})}{Pr(\theta_{t_j}^x|\mathcal{L}_{t_{j-1}})}\pi_{t_{j-1}}^v\end{aligned}\tag{11}$$

This holds when adjusting for a future execution contingency both when the probabilities  $\pi_{t_{j-1}}^v$  given the prior history  $\mathcal{L}_{t_{j-1}}$  are for possible paths in  $\ell^{in, \Phi}$  (from (8) and (9) in Step 1) and also for feasible but not possible paths in  $\ell^{off, \Phi}$  (from Step 2). These execution-contingent probabilities  $\hat{\pi}_{t_j}^v$  are used to compute the execution-contingent conditional expected value

$$E[\tilde{v}|\mathcal{L}_{t_{j-1}}, \theta_{t_j}^x] = \hat{\pi}_{t_j}^{\bar{v}} \bar{v} + \hat{\pi}_{t_j}^{v_0} v_0 + \hat{\pi}_{t_j}^v v\tag{12}$$

used by uninformed traders to compute expected payoffs for limit orders. In particular, the probabilities in (12) are the execution-contingent probabilities  $\hat{\pi}_{t_j}^v$  from (11) rather than the probabilities  $\pi_{t_j}^v$  from (9) that just condition on the prior trading history but not on the future states in which the limit order is executed.

Given these updating dynamics, we can now define an equilibrium.

**Definition.** A *Perfect Bayesian Nash Equilibrium* of the trading game in our model is a collection  $\{\varphi_{t_j}^{U,*}(x|\beta_{t_j}, \mathcal{L}_{t_{j-1}}), \varphi_{t_j}^{I,*}(x|\beta_{t_j}, v, \mathcal{L}_{t_{j-1}}), Pr^*(\theta_{t_j}^x|v, \mathcal{L}_{t_{j-1}}), Pr^*(\theta_{t_j}^x|\mathcal{L}_{t_{j-1}}), E^*[\tilde{v}|\mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]\}$  of

order-submission strategies, execution-probability functions, and execution-contingent conditional expected asset-value functions such that:

- The equilibrium execution probabilities  $Pr^*(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}})$  and  $Pr^*(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}})$  are consistent with the equilibrium order-submission strategies  $\{\varphi_{t_{j+1}}^{U,*}(x|\beta_{t_{j+1}}, \mathcal{L}_{t_j}), \dots, \varphi_{t_5}^{U,*}(x|\beta_{t_5}, \mathcal{L}_{t_4})\}$  and  $\{\varphi_{t_{j+1}}^{I,*}(x|\beta_{t_{j+1}}, v, \mathcal{L}_{t_j}), \dots, \varphi_{t_5}^{I,*}(x|\beta_{t_5}, v, \mathcal{L}_{t_4})\}$  after time  $t_j$ .
- The execution-contingent conditional expected asset values  $E^*[\tilde{v} | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$  agree with Bayesian updating equations (8), (9), (11), and (12) in Steps 1 and 3 when the order  $x$  is consistent with the equilibrium strategies  $\varphi_{t_j}^{U,*}(x|\beta_{t_j}, \mathcal{L}_{t_{j-1}})$  and  $\varphi_{t_j}^{I,*}(x|\beta_{t_j}, v, \mathcal{L}_{t_{j-1}})$  at date  $t_j$  and, when  $x$  is an off-equilibrium action inconsistent with the equilibrium strategies, with the off-equilibrium updating in Step 2.
- The positive-probability supports of the equilibrium strategy functions  $\varphi_{t_j}^{U,*}(x|\beta_{t_j}, \mathcal{L}_{t_{j-1}})$  and  $\varphi_{t_j}^{I,*}(x|\beta_{t_j}, v, \mathcal{L}_{t_{j-1}})$  (i.e., the orders with positive probability in equilibrium) are subsets of the sets of optimal orders for uninformed and informed investors computed from their optimization problems (6) and (7) and the equilibrium execution probabilities and outcome-contingent conditional asset-value expectation functions  $Pr^*(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}})$ ,  $Pr^*(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}})$ , and  $E^*[\tilde{v} | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$ .

Appendix A explains the algorithm used to compute the equilibria in our model. To help with intuition, the next section walks through the order-submission and Bayesian updating mechanics for a particular path in the extensive form of the trading game.

Our equilibrium concept differs from the Markov Perfect Bayesian Equilibrium used in Goettler et al. (2009). Beliefs and strategies in our model are path-dependent; that is to say, traders use Bayes Rule given the full prior order history when they arrive in the market. In contrast, Goettler et al. (2009) restricts Bayesian updating to the current state of the limit order book and does not allow for conditioning on the previous order history. Roşu (2016b) also assumes a Markov Perfect Bayesian Equilibrium. The quantitative importance of the order history is considered when we discuss our results in Section 2.

## 1.2 Illustration of order-submission mechanics and Bayesian updating

This section uses an excerpt of the extensive form of the trading game in our model to illustrate order-submission and trading dynamics and the associated Bayesian updating process. The particular trading history path in Figure 2 is from the equilibrium for the model specification in which informed and uninformed investors both have random private-value motives. The parameter values are  $\kappa = 0.10$ ,  $\sigma = 1.5$ ,  $\alpha = 0.8$ , and  $\delta = 0.16$ , which is a market with a relatively high informed-investor arrival probability and large value shocks. In this example, Nature has chosen an economic state in which there is good news ( $\bar{v}$ ) about the asset, and the realized sequence of arriving traders over time is  $\{I, U, U, I, I\}$ . At each node shown here, Figure 2 reports the total book  $L_{t_j}$  of limit orders from both arriving investors and the crowd. Trading starts at  $t_1$  with a book  $[1, 0, 0, 1]$  consisting of no orders from informed and uninformed investors (since none have arrived yet) plus the additional limit orders from the trading crowd (i.e., 1 each at the outside prices  $A_2$  and  $B_2$ ). For simplicity, our discussion here only reports a few nodes of the trading game with their associated equilibrium strategies. For example, we do not include  $NT$  at the end of  $t_1$ , since, as we show later in Section 2.2,  $NT$  is not an equilibrium action at  $t_1$  for these parameters.

Investors in our equilibrium choose from a discrete number of possible orders given their respective information and any private-value trading motives. Along the particular equilibrium path considered here, the optimal strategies do not involve any randomization. Optimal orders are unique given the inputs. However, orders are random due to randomness in the private factor  $\beta$ . Figure 2 shows below each order type at each time the probabilities with which the different orders are submitted by the trader who arrived. For example, if an informed trader  $I_{\bar{v}}$  arrives at  $t_1$ , she chooses a limit order  $LOA_2$  to sell at  $A_2$  with probability 0.118. Each of these unique optimal orders is associated with a different range of  $\beta$  types (for both informed and uninformed investors) and value signals (for informed investors). Figure 3 illustrates where the order-submission probabilities come from by superimposing the upper envelope of the expected payoffs for the different optimal orders at time  $t_1$  on the truncated Normal  $\beta$  distribution. It shows how different  $\beta$  ranges correspond to a discrete set of optimal orders delimited by the  $\beta$  thresholds. At each trading time, as the trading game progresses along this path, traders submit orders (or do not trade) following their equilibrium

order-submission strategies. The equilibrium execution probabilities of their orders depend on the order-submission decisions of future traders, which, in turn, depend on their trading strategies and the input information (i.e., their  $\beta$  realizations, any private knowledge about  $v$ , and the order history path when they arrive). At time  $t_1$ , the initial trader has rational-expectation beliefs that the execution probability of her  $LOA_2$  order posted at  $t_1$  is 0.644.<sup>11</sup> This equilibrium execution probability depends on all of the possible future trading paths proceeding from submission time  $t_1$  up through time  $t_5$ . For example, one possibility is that the  $LOA_2$  order will be hit by an investor arriving at time  $t_2$  who submits a market order. Another possibility (which is what happens along this particular path) is that an uninformed trader will arrive at  $t_2$  and post a limit order  $LOA_1$  to sell at  $A_1$ , thereby undercutting the earlier  $LOA_2$  order — so that the book at the end of  $t_2$  is  $[2, 1, 0, 1]$ . In this scenario, the initial  $LOA_2$  order from  $t_1$  will only be executed provided that the  $LOA_1$  order submitted at  $t_2$  is executed first. For example, the probability of a market order  $MOA_1$  hitting the limit order at  $A_1$  at  $t_3$  is 0.365, and then the probability of another market order hitting the initial limit sell at  $A_2$  is 0.423 at  $t_4$  and 0.505 at  $t_5$ .<sup>12</sup> Therefore, there is a chance that the  $LOA_2$  order from  $t_1$  will still be executed even after it is undercut by the order  $LOA_1$  at  $t_2$ .

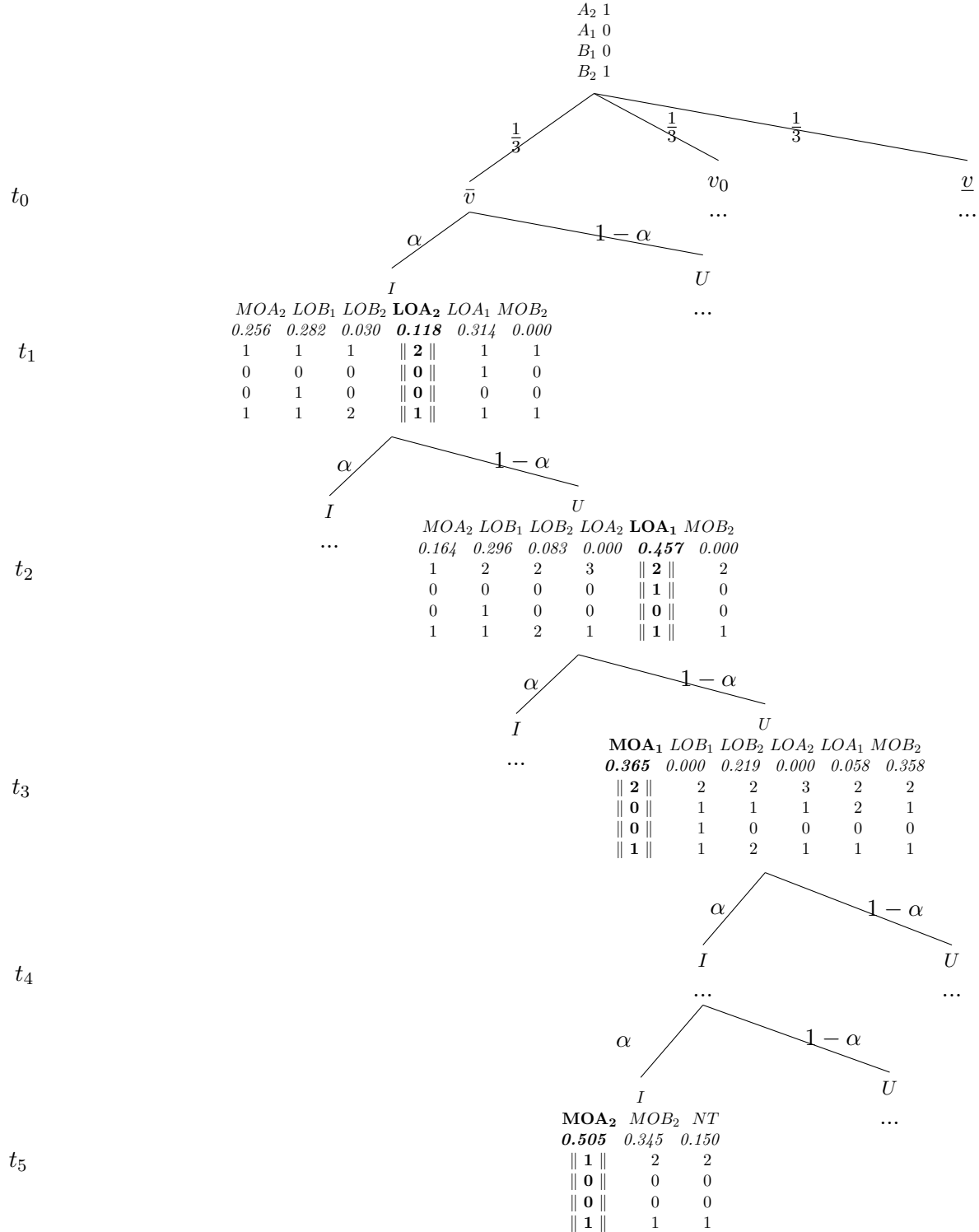
The path in Figure 2 also illustrates Bayesian updating in the model. After the investor at  $t_1$  has been observed submitting a limit order  $LOA_2$ , the uninformed trader who arrives in this example at time  $t_2$  — who just knows the submitted order at time  $t_1$  but not the identity or information of the trader at time  $t_1$  — updates his equilibrium conditional valuation to be  $E[\tilde{v}|LOA_2] = 1.056$  and his execution-contingent expectation given his limit order  $LOA_1$  at time  $t_2$  to be  $E[\tilde{v}|LOA_2, \theta_{t_2}^{LOA_1}] = 1.089$ . In subsequent periods, later investors observe additional realized orders and then further update their beliefs.

---

<sup>11</sup>Some of the numerical values discussed here are from equilibrium calculations reported in more detail in Tables 3 and 4 and Table B2 in Appendix B. Others are unreported calculations available from the authors upon request.

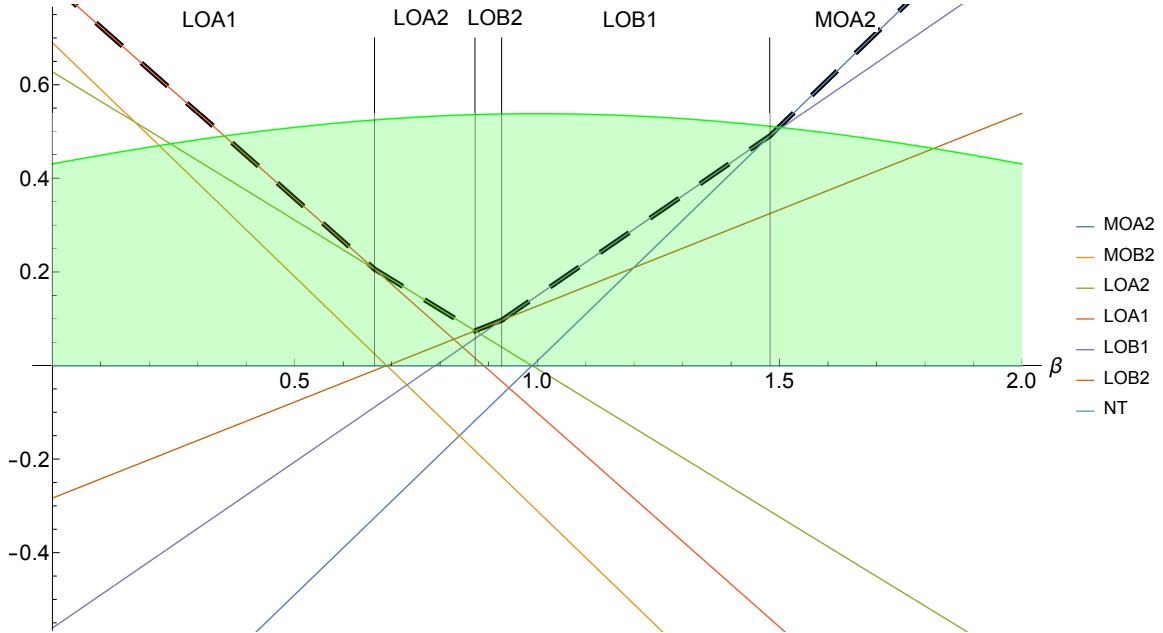
<sup>12</sup>Due to space constraints, we do not include the  $t_4$  node in Figure 2.

**Figure 2: Excerpt of the Extensive Form of the Trading Game.** This figure shows one possible trading path of the trading game with parameters  $\alpha = 0.8$ ,  $\delta = 0.16$ ,  $\mu = 1$ ,  $\sigma = 1.5$ ,  $\kappa = 0.10$ , and 5 time periods. Before trading starts at time  $t_1$ , the incoming book  $[1, 0, 0, 1]$  from time  $t_0$  consists of just the initial limit orders from the crowd at  $A_2$  and  $B_2$ . Nature selects a realized final value  $v = \{\bar{v}, v_0, \underline{v}\}$  with probabilities  $\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ . At each trading period nature also selects an informed trader ( $I$ ) with probability  $\alpha$  and an uninformed trader ( $U$ ) with probability  $1 - \alpha$ . Arriving traders choose the optimal order at each period which may potentially include limit orders  $LOA_i$  ( $LOB_i$ ) or market orders at the best ask,  $MOA_{i,t}$ , or at the best bid,  $MOB_{i,t}$ . Below each optimal trading strategy we report in italics its equilibrium order-submission probability. Boldfaced equilibrium strategies and associated states of the book (within double vertical bar) indicate the states of the book that we consider at each node of the chosen trading path.



**Figure 3:  $\beta$  Distribution and Upper Envelope for Informed Investor  $I_{\bar{v}}$  at time  $t_1$ .**

This figure shows the private-value factor  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$  distribution superimposed on the plot of the expected payoffs the informed investor  $I_{\bar{v}}$  with good news at time  $t_1$  for each equilibrium order type  $MOA_2, MOB_2, LOA_2, LOA_1, LOB_1, LOB_2, NT$ , (solid colored lines) when the total book (including crowd limit orders) opens  $L_{t_0} = [1 \ 0 \ 0 \ 1]$ . The dashed line shows the investor’s upper envelope for the optimal orders. The vertical black lines show the  $\beta$ -thresholds at which two adjacent optimal strategies yield the same expected payoffs. For example  $LOA_1$  is the optimal strategy for values of  $\beta$  between 0 and the first vertical black line;  $LOA_2$  is instead the optimal strategy for the values of beta between the first and the second vertical lines; and so forth. The parameters are  $\alpha = 0.8, \delta = 0.16, \mu = 1, \sigma = 1.5$ , and  $\kappa = 0.10$ .



## 2 Results

Our analysis investigates how liquidity supply and demand decisions of informed and uninformed traders and the learning process of uninformed traders affect market liquidity, price discovery, and investor welfare. This section presents numerical results for our model. We first consider a model specification in which only uninformed investors have a random private-value trading motive. In a second specification, we generalize the analysis and show the robustness of our findings and extend them. The tick size  $\kappa$  is fixed at 0.10, and the private-value dispersion  $\sigma$  is 1.5 throughout.

We focus on two time windows. The first is when the market opens at time  $t_1$ . The second is over the middle of the trading day from times  $t_2$  through  $t_4$ . We look at these two windows because our model is non-stationary over the trading day. Much like actual trading days, our



model has start-up effects at the beginning of the day and terminal horizon effects at the market close. When the market opens at time  $t_1$ , there are time-dependent incentives to provide, rather than to take, liquidity: The incoming book is thin (with limit orders only from the crowd), and there is the maximum time for future investors to arrive to hit limit orders from  $t_1$ . There are also time-dependent disincentives to post limit orders. Information asymmetries are maximal at time  $t_1$ , since there has been no learning from the trading process. Over the day, information is revealed (lessening adverse selection costs), but the book can also become fuller (i.e., there is competition in liquidity provision from earlier limit orders which have time priority at their respective limit prices), and the remaining time for market orders to arrive and execute limit orders becomes shorter. Comparing these two time windows shows how market dynamics change over the day. The market close at  $t_5$  is also important, but trading then is straightforward. At the end of the day, investors only submit market orders (or do not trade), because the execution probability for new limit orders submitted at  $t_5$  is zero given our assumption that unfilled limit orders are canceled once the market closes.

We use our model to investigate three questions: First, who provides and takes liquidity, and how does the amount of adverse selection affect investor decisions to take and provide liquidity? Second, how does market liquidity vary with different amounts of adverse selection? Third, how does the information content of different types of orders depend on an order's direction, aggressiveness, and on the prior order history?

The amount of adverse selection can change in two ways: The proportion of informed traders can change, and the magnitude of asset value shocks can change. We present comparative statics using four different combinations of parameters with high and low informed-investor arrival probabilities ( $\alpha = 0.8$  and  $0.2$ ) and high and low value-shock volatilities ( $\delta = 0.16$  and  $0.02$ ). We call markets with  $\delta = 0.02$  *low-volatility* markets and markets with  $\delta = 0.16$  *high-volatility* markets, because the arriving information is small relative to the tick size  $\kappa = 0.10$  in the former parameterization and larger relative to the tick size in the later. In high-volatility markets, the final asset value  $v$  given good or bad news is beyond the outside quotes  $A_2$  or  $B_2$ , and so even market orders at the outside prices are profitable for informed traders. However, in low-volatility markets,  $v$  is always within

the inside quotes  $A_1$  and  $B_1$ , and so market orders are never profitable for informed investors.

## 2.1 Uninformed traders with random private-value motives

In our first model specification, only uninformed traders have random private values. Informed traders have fixed neutral private-value factors  $\beta = 1$ . Thus, as in Kyle (1985), there is a clear differentiation between investors who speculate on private information and those who trade for purely non-informational reasons. Unlike Kyle (1985), informed and uninformed investors can trade using limit or market orders rather than being restricted to just market orders.

### 2.1.1 Trading strategies

We begin by investigating who supplies and takes liquidity and how these decisions change with the amount of adverse selection. Table 1 reports results about trading early in the day at time  $t_1$  using a  $2 \times 2$  format. Each of the four cells corresponds to a different combination of parameters. Comparing cells horizontally shows the effect of a change in the value-shock size  $\delta$  while holding the arrival probability  $\alpha$  for informed traders fixed. Comparing cells vertically shows the effect of a change in the informed-investor arrival probability while holding the value-shock size fixed. In each cell corresponding to a set of parameters, there are four columns reporting conditional results for informed investors with good news, neutral news, and bad news about the asset ( $I_{\bar{v}}$ ,  $I_{v_0}$ ,  $I_{\underline{v}}$ ) and for an uninformed investor ( $U$ ) and a fifth column with the unconditional market results ( $Uncond$ ). The table reports the order-submission probabilities and several market-quality metrics. Specifically, we report expected bid-ask spreads conditioning on the three informed-investor types  $E[Spread | I_v]$  and on the uninformed trader  $E[Spread | U]$ , the unconditional expected market spread  $E[Spread]$ , and expected depths at the inside prices ( $A_1$  and  $B_1$ ) and total depths ( $A_1 + A_2$  and  $B_1 + B_2$ ) on each side of the market. As we shall see, our results are symmetric for the directionally informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  and on the buy and sell sides of the market. In addition, we report the probability-weighted contributions to the different investors' welfare (i.e., expected gains-from-trade) from limit and market orders respectively, and their total expected welfare.<sup>13</sup> Table B1 in Appendix B

<sup>13</sup>Let  $W(\beta_{t_1})$  and  $W(v, \beta_{t_1})$  denote the value functions when (6) and (7) are evaluated at time  $t_1$  using the optimal strategies for the uninformed and informed investors respectively. The total welfare gain is  $E[W(\beta_{t_1})]$  for

provides additional results about conditional and unconditional future execution probabilities for the different orders ( $P^{EX}(x_{t_1})$ ) and also the uninformed investor's updated expected asset value  $E[v|x_{t_1}]$  given different types of buy orders  $x_{t_1}$  at time  $t_1$ .

Table 2 shows average results for times  $t_2$  through  $t_4$  during the day using a similar  $2 \times 2$  format. The averages are across time and trading histories. Comparing results for time  $t_1$  with the trading averages for  $t_2$  through  $t_4$  shows intraday changes in properties of the trading process. There is no table for time  $t_5$ , because only market orders are used at the market close.

**Result 1** Changes in adverse selection due to the value-shock size  $\delta$  affect trading strategies differently than changes in the informed-investor arrival probability  $\alpha$ .

The fact that different forms of adverse selection affect investors' trading decisions differently can be shown theoretically from first principles. Suppose the informed-investor arrival probability  $\alpha$  is close to zero. If the value-shock volatility  $\delta$  is close to zero, then directionally informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  with good or bad news never use market orders, since the final asset value  $v$  is always between the inside bid and ask prices. However, if  $\delta$  is sufficiently large, then investors with good and bad news will start to use market orders given the guaranteed execution. Thus, the set of orders used by directionally informed investors can change in these small  $\alpha$  scenarios when  $\delta$  changes. In contrast, consider a market in which  $\delta$  is close to zero. Now informed investors with good or bad news never use market orders for any informed-investor arrival probability  $\alpha$ . Thus, the set of orders used by directionally informed investors never changes to include market orders in these small  $\delta$  scenarios when  $\alpha$  changes.

---

the uninformed investor where the expectation is taken over  $\beta_{t_1}$  and  $E[W(v, \beta_{t_1})]$  for the informed investor where the expectation is taken over  $v$  and  $\beta_{t_1}$ .

**Table 1: Trading Strategies, Liquidity, and Welfare at Time  $t_1$  in an Equilibrium with Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different informed-investor arrival probabilities  $\alpha$  (0.8 and 0.2) and two different value-shock volatilities  $\delta$  (0.16 and 0.02). The private-value factor parameters are  $\mu = 1$  and  $\sigma = 1.5$ , and the tick size is  $\kappa = 0.10$ . Each cell corresponding to a set of parameters reports the equilibrium order-submission probabilities, the expected bid-ask spreads and expected depths at the inside prices ( $A_1$  and  $B_1$ ) and total depths on each side of the market after order submissions at time  $t_1$ , and expected welfare of the market participants. The first four columns in each parameter cell are for informed traders with positive, neutral and negative signals,  $(I_{\bar{v}}, I_{v_0}, I_v)$  and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports unconditional results for the market.

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	$LOA_2$	0	0.500	0.650	0.143	0.335	0	0.500	1.000	0.052	0.410
	$LOA_1$	0	0	0.350	0	0.093	0	0	0	0.079	0.016
	$LOB_1$	0.350	0	0	0	0.093	0	0	0	0.079	0.016
	$LOB_2$	0.650	0.500	0	0.143	0.335	1.000	0.500	0	0.052	0.410
	$MOA_2$	0	0	0	0.357	0.071	0	0	0	0.369	0.074
	$MOA_1$	0	0	0	0	0	0	0	0	0	0
	$MOB_1$	0	0	0	0	0	0	0	0	0	0
	$MOB_2$	0	0	0	0.357	0.071	0	0	0	0.369	0.074
	$NT$	0	0	0	0	0	0	0	0	0	0
	E[Spread $ \cdot$ ]	0.265	0.300	0.265	0.300	0.281	0.300	0.300	0.300	0.284	0.297
	E[Depth $A_2+A_1$ $ \cdot$ ]	1.000	1.500	2.000	1.143	1.429	1.000	1.500	2.000	1.131	1.426
	E[Depth $A_1$ $ \cdot$ ]	0	0	0.350	0	0.093	0	0	0	0.079	0.016
	E[Depth $B_1$ $ \cdot$ ]	0.350	0	0	0	0.093	0	0	0	0.079	0.016
	E[Depth $B_1+B_2$ $ \cdot$ ]	2.000	1.500	1.000	1.143	1.429	2.000	1.500	1.000	1.131	1.426
	E[Welfare LO $ \cdot$ ]	0.034	0.053	0.034	0.018		0.029	0.069	0.029	0.015	
E[Welfare MO $ \cdot$ ]	0	0	0	0.337		0	0	0	0.339		
E[Welfare $ \cdot$ ]	0.034	0.053	0.034	0.355		0.029	0.069	0.029	0.354		
$\alpha = 0.2$	$LOA_2$	0	0.500	0.110	0.063	0.091	0	0.500	1.000	0.063	0.150
	$LOA_1$	0	0	0.890	0.374	0.358	0	0	0	0.397	0.318
	$LOB_1$	0.890	0	0	0.374	0.358	0	0	0	0.397	0.318
	$LOB_2$	0.110	0.500	0	0.063	0.091	1.000	0.500	0	0.063	0.150
	$MOA_2$	0	0	0	0.064	0.051	0	0	0	0.040	0.032
	$MOA_1$	0	0	0	0	0	0	0	0	0	0
	$MOB_1$	0	0	0	0	0	0	0	0	0	0
	$MOB_2$	0	0	0	0.064	0.051	0	0	0	0.040	0.032
	$NT$	0	0	0	0	0	0	0	0	0	0
	E[Spread $ \cdot$ ]	0.211	0.300	0.211	0.225	0.228	0.300	0.300	0.300	0.221	0.236
	E[Depth $A_2+A_1$ $ \cdot$ ]	1.000	1.500	2.000	1.436	1.449	1.000	1.500	2.000	1.460	1.468
	E[Depth $A_1$ $ \cdot$ ]	0	0	0.890	0.374	0.358	0	0	0	0.397	0.318
	E[Depth $B_1$ $ \cdot$ ]	0.890	0	0	0.374	0.358	0	0	0	0.397	0.318
	E[Depth $B_1+B_2$ $ \cdot$ ]	2.000	1.500	1.000	1.436	1.449	2.000	1.500	1.000	1.460	1.468
	E[Welfare LO $ \cdot$ ]	0.273	0.146	0.273	0.316		0.081	0.150	0.081	0.360	
E[Welfare MO $ \cdot$ ]	0	0	0	0.099		0	0	0	0.064		
E[Welfare $ \cdot$ ]	0.273	0.146	0.273	0.415		0.081	0.150	0.081	0.424		

**Table 2: Averages for Trading Strategies, Liquidity, and Welfare across Times  $t_2$  through  $t_4$  for Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different informed-investor arrival probabilities  $\alpha$  (0.8 and 0.2) and for two different asset-value volatilities  $\delta$  (0.16 and 0.02). The private-value factor parameters are  $\mu = 1$  and  $\sigma = 1.5$ , and the tick size is  $\kappa = 0.10$ . Each cell corresponding to a set of parameters reports the equilibrium order-submission probabilities, the expected bid-ask spreads and expected depths at the inside prices ( $A_1$  and  $B_1$ ) and total depths on each side of the market after order submissions at times  $t_2$  through  $t_4$ , and expected welfare for the market participants. The first four columns in each parameter cell are for informed traders with positive, neutral and negative signals, ( $I_{\bar{v}}, I_{v_0}, I_v$ ) and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports unconditional results for the market.

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	$LOA_2$	0	0.191	0.051	0.157	0.096	0.399	0.255	0.108	0.026	0.209
	$LOA_1$	0	0.258	0.257	0.023	0.142	0.192	0.239	0.288	0.064	0.205
	$LOB_1$	0.257	0.258	0	0.023	0.142	0.288	0.239	0.192	0.064	0.205
	$LOB_2$	0.051	0.191	0	0.157	0.096	0.108	0.255	0.399	0.026	0.209
	$MOA_2$	0.493	0	0	0.286	0.189	0	0	0	0.347	0.069
	$MOA_1$	0.001	0	0	0.031	0.006	0	0	0	0.058	0.012
	$MOB_1$	0	0	0.001	0.031	0.006	0	0	0	0.058	0.012
	$MOB_2$	0	0	0.493	0.286	0.189	0	0	0	0.347	0.069
	$NT$	0.198	0.061	0.198	0.007	0.124	0.013	0.010	0.013	0.011	0.012
	E[Spread $\cdot$ ]	0.217	0.212	0.217	0.251	0.223	0.227	0.228	0.227	0.278	0.237
	E[Depth $A_2+A_1 \cdot$ ]	1.047	2.276	2.480	1.755	1.899	2.165	2.300	2.433	1.608	2.161
	E[Depth $A_1 \cdot$ ]	0	0.438	0.829	0.243	0.387	0.226	0.362	0.506	0.131	0.318
	E[Depth $B_1 \cdot$ ]	0.829	0.438	0	0.243	0.387	0.506	0.362	0.226	0.131	0.318
	E[Depth $B_1+B_2 \cdot$ ]	2.480	2.276	1.047	1.755	1.899	2.433	2.300	2.165	1.608	2.161
	E[Welfare LO $\cdot$ ]	0.010	0.020	0.010	0.106		0.014	0.013	0.014	0.005	
	E[Welfare MO $\cdot$ ]	0.009	0	0.009	0.298		0	0	0	0.354	
	E[Welfare $\cdot$ ]	0.019	0.020	0.019	0.405		0.014	0.013	0.014	0.359	
	$\alpha = 0.2$	$LOA_2$	0	0.358	0.508	0.102	0.139	0.375	0.389	0.443	0.093
$LOA_1$		0	0.122	0.258	0.056	0.070	0.044	0.096	0.116	0.066	0.070
$LOB_1$		0.258	0.122	0	0.056	0.070	0.116	0.096	0.044	0.066	0.070
$LOB_2$		0.508	0.358	0	0.102	0.139	0.443	0.389	0.375	0.093	0.155
$MOA_2$		0.130	0	0	0.219	0.184	0	0	0	0.218	0.175
$MOA_1$		0.088	0	0	0.119	0.101	0	0	0	0.120	0.096
$MOB_1$		0	0	0.088	0.119	0.101	0	0	0	0.120	0.096
$MOB_2$		0	0	0.130	0.219	0.184	0	0	0	0.218	0.175
$NT$		0.016	0.035	0.016	0.006	0.010	0.022	0.030	0.022	0.005	0.009
E[Spread $\cdot$ ]		0.205	0.190	0.205	0.280	0.264	0.221	0.217	0.221	0.300	0.284
E[Depth $A_2+A_1 \cdot$ ]		1.305	2.089	2.512	1.583	1.660	1.932	2.091	2.257	1.576	1.680
E[Depth $A_1 \cdot$ ]		0.194	0.451	0.740	0.301	0.333	0.346	0.414	0.442	0.262	0.290
E[Depth $B_1 \cdot$ ]		0.740	0.451	0.194	0.301	0.333	0.442	0.414	0.346	0.262	0.290
E[Depth $B_1+B_2 \cdot$ ]		2.512	2.089	1.305	1.583	1.660	2.257	2.091	1.932	1.576	1.680
E[Welfare LO $\cdot$ ]		0.119	0.086	0.119	0.052		0.060	0.064	0.060	0.050	
E[Welfare MO $\cdot$ ]		0.018	0	0.018	0.343		0	0	0	0.342	
E[Welfare $\cdot$ ]		0.137	0.086	0.137	0.394		0.060	0.064	0.060	0.392	

Our numerical analysis illustrates this first result and also other facets of how adverse selection affects investor trading strategies. Consider the directionally informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  with good or bad news. First, hold the informed-investor arrival probability  $\alpha$  fixed and increase the amount of adverse selection by increasing the value-shock volatility  $\delta$ . In a low-volatility market in which value shocks  $\Delta$  are small relative to the tick size, informed traders with good and bad news are unwilling to pay a large tick size to trade on their information and instead act as liquidity providers who supply liquidity asymmetrically depending on the direction of their information. This can be seen in Table 1 where in both of the two parameter cells on the right (with  $\alpha = 0.8$  and  $0.2$  and a small  $\delta = 0.02$ ) informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  at time  $t_1$  use limit orders at the outside quotes  $A_2$  and  $B_2$  exclusively. In contrast, in a high-volatility market where value shocks are large relative to the tick size, informed investors with good or bad news trade more aggressively. This can be seen in the left two parameterization cells (with  $\alpha = 0.8$  and  $0.2$  and a large  $\delta = 0.16$ ) where now informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  use limit orders at both the inside quotes  $A_1$  and  $B_1$  as well at the outside quotes with positive probability at time  $t_1$ . Now compare this to a change in the amount of adverse selection due to a change in the informed-investor arrival probability  $\alpha$  while holding the value-shock size  $\delta$  fixed. In this case, changing the amount of adverse selection does not affect which orders informed investors with good and bad news use at time  $t_1$ . This can be seen by comparing the lower two parameter cells (with  $\delta = 0.02$  and  $0.16$  and a small  $\alpha$ ) with the upper two parameter cells (with the same  $\delta$ s and a larger  $\alpha$ ).

The average order-submission probabilities at times  $t_2$  through  $t_4$  in Table 2 are qualitatively similar to those for time  $t_1$ . In low-volatility markets, informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  with good and bad news tend to supply liquidity via limit orders following strategies in which order-submission probabilities are somewhat skewed on the two sides of the market in the direction of their small amount of private information. In contrast, in high-volatility markets, informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  switch from providing liquidity on both sides of the market at times  $t_2$  to  $t_4$  to using a mix of taking liquidity via market orders and supplying liquidity via limit orders on the same side of the market as their information. Thus, once again, the trading strategies for informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  are qualitatively similar holding  $\delta$  fixed and changing  $\alpha$ , but their trading strategies change

qualitatively when  $\alpha$  is held fixed and  $\delta$  is changed.

Next, consider informed investors  $I_0$  who know that the value shock  $\Delta$  is 0 and, thus, that the unconditional prior  $v_0$  is correct. Tables 1 and 2 show that their liquidity provision trading strategies are qualitatively the same at time  $t_1$  and on average over times  $t_2$  through  $t_4$ . In contrast, uninformed investors  $U$  become less willing to provide liquidity via limit orders at the inside quotes as the adverse selection problem they face using limit orders worsens. Rather, they increasingly take liquidity via market orders or supply liquidity by less aggressive limit orders at the outside quotes. This reduction in liquidity provision at the inside quotes by uninformed investors happens at time  $t_1$  (Table 1) and at times  $t_2$  through  $t_4$  (Table 2) both when the value shocks become larger and when the arrival probability of informed investors increases.

Two equilibrium effects are noteworthy in this context. First, while the uninformed  $U$  investors reduce their liquidity provision at the inside quotes as adverse selection increases, the  $I_0$  informed investors increase their liquidity provision at the inside quotes. This is because  $I_0$  informed investors have an advantage in liquidity provision over the uninformed  $U$  investors in that there is no adverse selection risk for them. These results are qualitatively consistent with the intuition of Bloomfield, O'Hara and Saar (BOS, 2005). Informed traders are more likely to use limit orders than market orders when the value-shock volatility is low (and, thus, the profitability from trading on information asymmetries is low), and to use market orders when the reverse is true.

Second, uninformed  $U$  investors are unwilling to use aggressive limit orders at the inside quotes when the adverse selection risk is sufficiently high as in the upper left parametrization ( $\alpha = 0.8$  and  $\delta = 0.16$ ). This explains the fact that informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$  use aggressive limit orders at the inside quotes with a higher probability at time  $t_1$  in the lower left parametrization (0.890 with  $\alpha = 0.2$  and  $\delta = 0.16$ ) than in the upper left parameterization (0.350). At first glance this might seem odd since competition from future informed investors (and the possibility of being undercut by later limit orders) is greater when the informed-investor arrival probability is large ( $\alpha = 0.8$ ) than when  $\alpha$  is smaller. However, in equilibrium there is camouflage from the uninformed  $U$  investors limit orders at the inside quotes in the lower left parametrization, whereas limit orders at the inside quotes are fully revealing in the upper left parametrization. As a result, Table B1

in Appendix B shows that the execution probabilities for the fully revealing limit orders at prices which are revealed to be far from the asset’s actual value are much lower (0.078) relative to the non-fully revealing limit orders (0.717).

### 2.1.2 Market quality

Market liquidity changes when the amount of adverse selection in a market changes. The standard intuition, as in Kyle (1985), is that liquidity deteriorates given more adverse selection. For example, Roşu (2016b) also finds worse liquidity (a wider bid-ask spread) given higher value volatility. However, we find that the standard intuition is not always true.

**Result 2** Liquidity need not always deteriorate when adverse selection increases.

Markets can become more liquid given greater value-shock volatility if, given the tick size, high volatility makes the value shock  $\Delta$  large relative the price grid. In addition, different measures of market liquidity — expected spreads, inside depth, and total depth — can respond differently to changes in adverse selection.

The impact of adverse selection on market liquidity follows directly from the trading strategy effects discussed above. Two intuitions are useful in understanding our market liquidity results. First, different investors affect liquidity differently. Informed traders with neutral news ( $I_{v_0}$ ) are natural liquidity providers. Their impact on liquidity comes from whether they supply liquidity at the inside ( $A_1$  and  $B_1$ ) or outside ( $A_2$  and  $B_2$ ) prices. In contrast, informed traders with directional news ( $I_{\bar{v}}$  and  $I_{\underline{v}}$ ) and uninformed traders ( $U$ ) affect liquidity depending on whether they opportunistically take or supply liquidity. Second, the most aggressive way to trade (both on directional information and private values) is via market orders, which takes liquidity. However, the next most aggressive way to trade is via limit orders at the inside prices. Thus, changes in market conditions (i.e.,  $\delta$  and  $\alpha$ ) that make informed investors trade more aggressively (i.e., that reduce their use of limit orders at the outside prices  $A_2$  and  $B_2$ ) can potentially improve liquidity if their stronger trading interest migrates to aggressive limit orders at the inside quotes ( $A_1$  and  $B_1$ ) rather than to market orders.



Our analysis shows that the standard intuition that adverse selection reduces market liquidity depends on the relative magnitudes of asset-value shocks and the tick size. In Table 1, the expected spread narrows at time  $t_1$  (markets become more liquid) when the value-shock volatility  $\delta$  increases (comparing parameterizations horizontally so that  $\alpha$  is kept fixed). Liquidity improves in these cases because the informed traders  $I_{\bar{v}}$  and  $I_{\underline{v}}$  submit limit orders at the inside quotes in these high-volatility markets, whereas they only use limit orders at the outside quotes in low-volatility markets. In contrast, the expected spread at time  $t_1$  widens when the informed-investor arrival probability  $\alpha$  increases holding the value-shock size  $\delta$  constant, as predicted by the standard intuition. The evidence against the standard intuition is even stronger in Table 2. At times  $t_2$  through  $t_4$ , the expected spread narrows both when information becomes more volatile ( $\delta$  is larger) and when there are more informed traders (when  $\alpha$  is larger). The qualitative results for the expected depth at the inside quotes goes in the same direction as the results for the expected spread. This is because both results are driven by limit-order submissions at the inside quotes. The results for adverse selection and total depth at both the inside and outside quotes are mixed. For example, total depth at time  $t_1$  increases in Table 1 when value-shock volatility  $\delta$  increases when the informed-investor arrival probability  $\alpha$  is high (comparing horizontally the two parametrizations on the top), but decreases in  $\delta$  when the informed  $\alpha$  is low. In contrast, average total depth at times  $t_2$  through  $t_4$  in Table 2 is decreasing in the value-shock volatility (comparing parameterizations horizontally). This is opposite the effect on the inside depth. Thus, different metrics for liquidity give mixed results.

The main result in this section is that the relation between adverse selection and market liquidity is more subtle than the standard intuition. Increased adverse selection can improve liquidity. The ability of investors to choose endogenously whether to supply or demand liquidity and at what limit prices is what can overturn the standard intuition. Goettler et al. (2009) also have a model specification with informed traders who have no private-value trading motive and uninformed having only private-value motives. In their model, when volatility increases, informed traders reduce their provision of liquidity and increase their demand of liquidity; with the opposite holding for uninformed traders. Our results are more nuanced. Increased value-shock volatility is associated with increased liquidity supply in some cases and with decreased liquidity in others. This is because

the tick size of the price grid constrains the prices at which liquidity can be supplied and demanded.

### 2.1.3 Information content of orders

Traders in real-world markets and empirical researchers are interested in the information content of different types of arriving orders.<sup>14</sup> A necessary condition for an order to be informative is that informed investors use it. However, the magnitude of order informativeness is determined by the mix of equilibrium probabilities with which both informed and uninformed traders use an order. If uninformed traders use the same orders as informed investors, they add noise to the price discovery process, and orders become less informative. In our model, the mix of information-based and noise-based orders depends on the underlying proportion of informed investors  $\alpha$  and the value-shock volatility  $\delta$ .

We expect different market and limit orders to have different information content. A natural conjecture is that the sign of the information revision associated with an order should agree with the direction of the order (e.g., buy market and limit orders should lead to positive valuation revisions). Another natural conjecture is that the magnitude of information revisions should be greater for more aggressive orders. However, while the order-sign conjecture is true in our first model specification, the order-aggressiveness conjecture does not always hold here.

**Result 3** Order informativeness is not always increasing in the aggressiveness of an order.

This, at-first-glance surprising, result is another consequence of the impact of the tick size on how informed investors trade on their information. As a result, the relative informativeness of different market and limit orders can flip in high-volatility and low-volatility markets. The result is immediate for market orders versus (less aggressive) limit orders in low-volatility markets in which informed investors avoid market orders (see Table 1). However, this reversed ordering can also hold for aggressive limit orders at the inside quotes ( $A_1$  and  $B_1$ ) versus less aggressive limit orders at the outside quotes ( $A_2$  and  $B_2$ ).

---

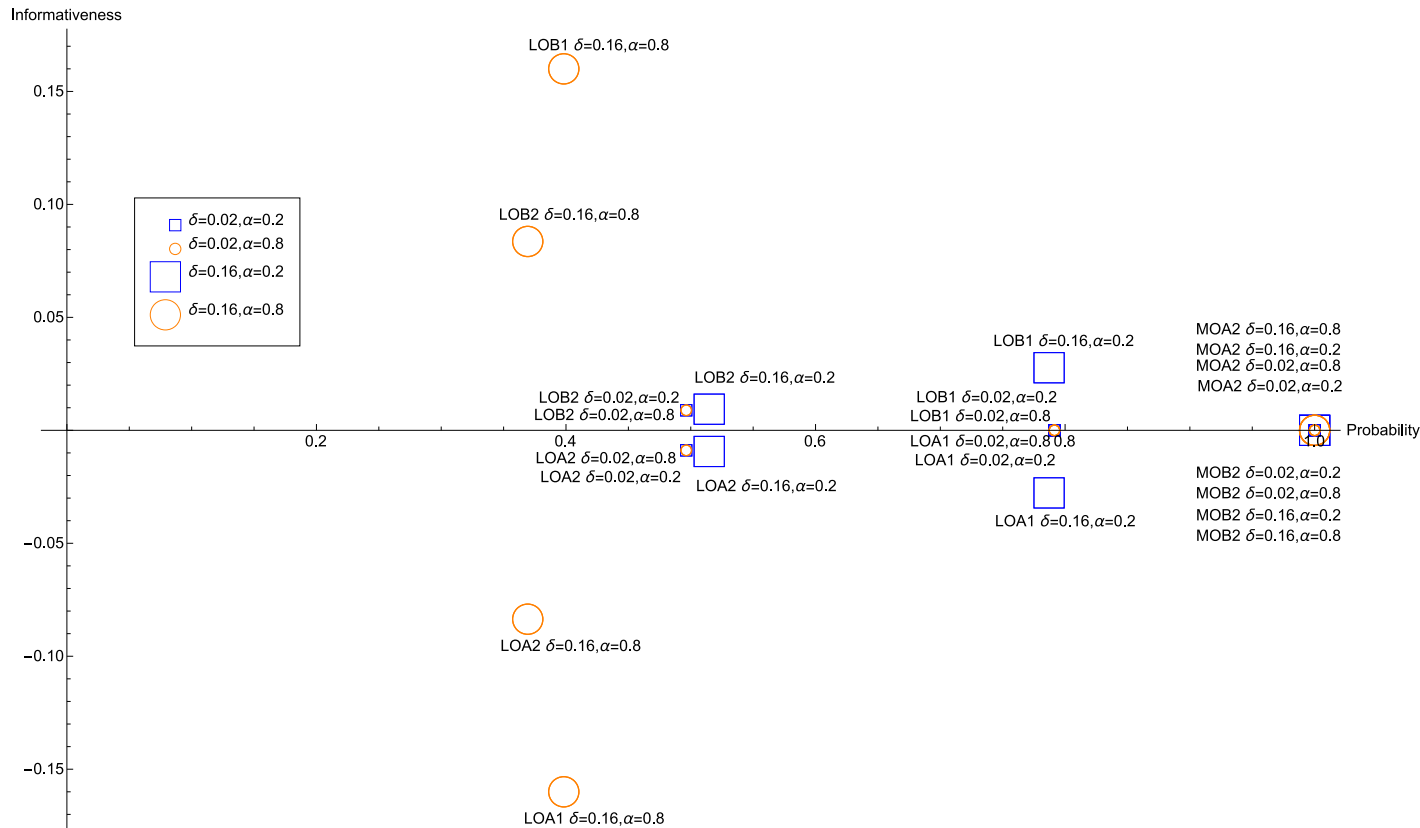
<sup>14</sup>Fleming et al. (2017) extend the VAR estimation approach of Hasbrouck (1991) to estimate the price impacts of limit orders as well as market orders. See also Brogaard et al. (2016).

Figure 4 shows the informativeness of different types of orders at time  $t_1$ . Informativeness at time  $t_1$  is measured here as the Bayesian revision  $E[v|x_{t_1}] - E[v]$  in the uninformed investor's expectation of the terminal value  $v$  after observing different types of orders  $x_{t_1}$  at time  $t_1$ . The informational revisions for the different orders are plotted against the respective order-execution probabilities on the horizontal axis. Orders with higher execution probabilities are statistically more aggressive than orders with low execution probabilities. The results for the four parameterizations are indicated using different symbols: high vs low informed-investor arrival probabilities (circles vs squares), and high vs low value-shock volatility (large vs small symbols). These are described in the figure legend. For example, in the low  $\alpha$  and high  $\delta$  scenario (large squares), the informativeness of a limit buy order at  $B_1$  at time  $t_1$  is 0.026 and the order-execution probability is 78.9 percent (see Table B1 in the Appendix B).

Consider first markets with high informed-investor arrival probabilities. The case with a high informed-investor arrival probability and high value-shock volatility is denoted with large circles. Informed investors in this case use limit orders at both the outside quotes ( $LOA_2$  and  $LOB_2$ ) and inside quotes ( $LOA_1$  and  $LOB_1$ ) at time  $t_1$ , so these are therefore the only informative orders. Since uninformed investors also use the outside limit orders, they are not fully revealing, however the inside limit orders are fully revealing. Thus, the price impacts for the inside and outside limit orders here are consistent with the order-aggressiveness conjecture. The market orders ( $MOB_2$  and  $MOA_2$ ) are also used in equilibrium, but only by uninformed investors ( $U$ ). Thus, they are not informative. While market orders would be profitable for the informed investors, the potential price improvement with the limit orders leads informed investors to use the limit orders despite the zero price impact and guaranteed execution probability of the market orders. Since both inside and outside limit orders have larger price impacts than the market orders, this case is inconsistent with the order-aggressiveness conjecture.

Next, consider the case of low value-shock volatility and high informed-investor arrival probability, denoted here with small circles. Once again, the order-aggressiveness conjecture is not true. The most informative orders are now, not the most aggressive orders, but rather the most patient limit orders posted at  $A_2$  and  $B_2$  (since these are the only orders used by informed in-

**Figure 4: Informativeness of Orders after Trading at Time  $t_1$  for the Model with Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This figure plots the *Informativeness* of the equilibrium orders at the end of  $t_1$  against the probability of order execution. Four different combinations of informed-investor arrival probabilities and value-shock volatilities are considered. The informativeness of an order is measured as  $E[v|x_{t_1}] - E[v]$ , where  $x_{t_1}$  denotes one of the different possible equilibrium orders at time  $t_1$ .



vestors). The market orders and more aggressive inside limit orders are non-informative here (since only uninformed investors with extreme  $\beta$ s use them). In this case, this — again at first glance perhaps counterintuitive — result is a consequence of the fact that the informed trader’s potential information is small relative to the tick size. Low-volatility makes market orders unprofitable for informed traders given good and bad news, and it also increases the importance of price improvement attainable through limit orders deeper in the book relative to limit orders at the inside quotes.

Similar results hold when the proportion of insiders is low ( $\alpha = 0.2$ ). When the asset-value volatility is high (large squares), the most aggressive orders ( $LOB_1$  and  $LOA_1$ ) are again the most informative ones in contrast to the market orders. However, when volatility is low (small squares), the most informative orders, as before, are the least aggressive orders ( $LOB_2$  and  $LOA_2$ ).

Therefore, the potential failures of the order-aggressiveness conjecture are robust to variation in informed-investor arrival probabilities and value-shock volatility.

#### 2.1.4 Non-Markovian learning

This section investigates the role of the order history on Bayesian learning at times later in the day. One of the main differences between our model and Goettler et al. (2009) and Roşu (2016b) is that they assume that information dynamics are Markovian and that the current limit order book is a sufficient statistic for the information content of the prior trading history. Thus, the first question we consider is whether the prior order history has information about the asset value  $v$  in excess of the information in the current limit order book.

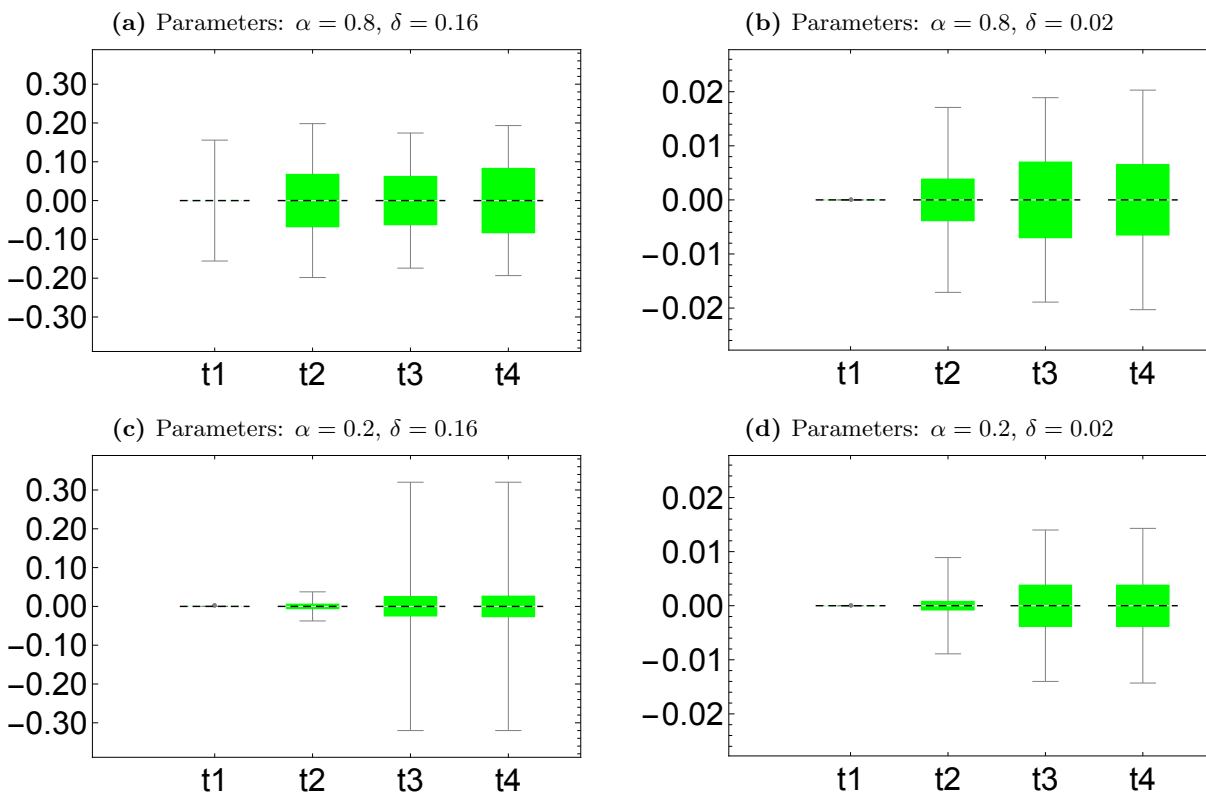
The candlestick plots in Figure 5 measure the incremental information content of order histories as the difference  $E[v | \mathcal{L}_{t_j}(L_{t_j})] - E[v | L_{t_j}]$ , which is the uninformed investors' expected asset value conditional on an order history path  $\mathcal{L}_{t_j}(L_{t_j})$  ending with a particular limit order book  $L_{t_j}$  at time  $t_j$  net of the corresponding expectation conditional on just the ending book  $L_{t_j}$ . In particular, we are interested in books  $L_{t_j}$  that can be preceded in equilibrium by more than one different prior history. If learning is Markov, then order histories  $\mathcal{L}_{t_j}(L_{t_j})$  preceding a book  $L_{t_j}$  should convey no additional information beyond  $L_{t_j}$ ; in which case the difference in expectations should be zero. The candlestick plots show the maximum and minimum values, the interquartile range, and the median of the incremental information of the prior history. The horizontal axis in the plots shows the times  $t_1$  through  $t_4$  at which different orders  $x_{t_j}$  are submitted. Time  $t_1$  is included in the plot because books at  $t_1$  can potentially be produced by different sequences of investor actions  $x_{t_1}$  and crowd responses at  $t_1$ . Each plot is for a different combination of adverse-selection parameters.

The main result from Figure 5 is that there is substantial informational variation in the Bayesian revisions conditional on different trading histories.

**Result 4** The price discovery dynamics can be significantly non-Markovian.

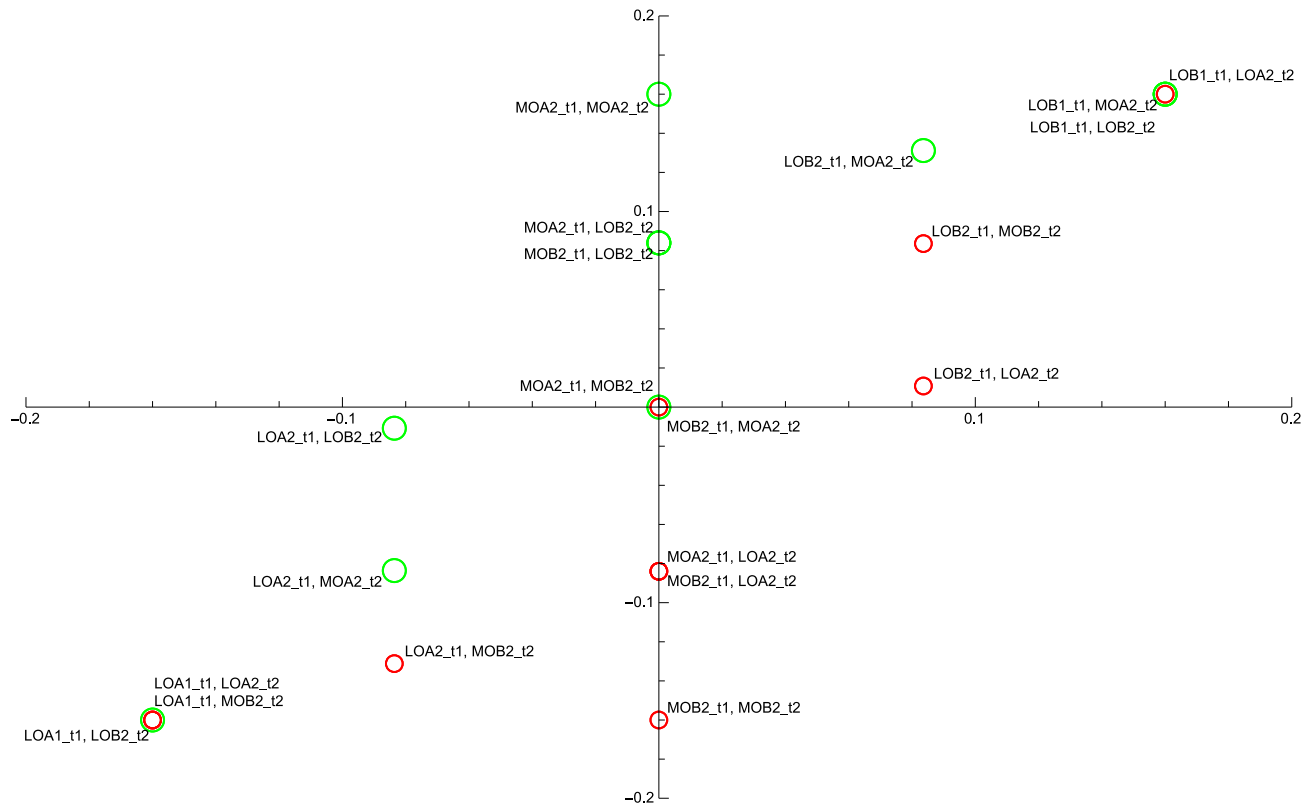
As expected, the variation in the incremental information content of the prior trading history in Figure 5 is greater when the shock volatility  $\delta$  is greater (note the differences in vertical scales).

**Figure 5: Informativeness of the Order History for the Model with Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$  for Times  $t_1$  through  $t_4$ .** This figure shows the incremental information content of the past order history in excess of the information in the current limit order book observed at the end of time  $t_j$  as measured by  $E[v|\mathcal{L}_{t_j}(L_{t_j})] - E[v|L_{t_j}]$  where  $\mathcal{L}_{t_j}(L_{t_j})$  is a history ending in the limit order book  $L_{t_j}$ . We only consider books  $L_{t_j}$  when they occur in equilibrium in the different trading periods. The candlesticks indicate for each of these two metrics the maximum, the minimum, the median and the 75<sup>th</sup> (and 25<sup>th</sup>) percentile respectively as the top (bottom) of the bar.



Given that learning is non-Markov, our next question is about how the size of the valuation revisions depends on the prior trading history. In Figure 6, the horizontal axis shows the price impact of different equilibrium orders at  $t_1$ , and the vertical axis gives the corresponding cumulative price impact of the sequence of a given action at time  $t_1$  and different subsequent equilibrium actions at time  $t_2$ . Consistent with our previous analysis, the size of the valuation revision depends crucially on the insiders' equilibrium strategies. As informed investors do not use market orders at  $t_1$  (see Table 1), market orders have a zero price impact at  $t_1$  and, thus, the points for pairs of time  $t_1$

**Figure 6: Order Informativeness for the Model with Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$  for times  $t_1$  to  $t_2$  and parameters  $\alpha = 0.8$ ,  $\delta = 0.16$ .** The horizontal axis reports  $E(v|x_{t_1}) - E(v)$  which shows how the uninformed traders' Bayesian value-forecast changes with respect to the unconditional expected value of the fundamental when uninformed traders observe at  $t_1$  an equilibrium order  $x_{t_1}$ . The vertical axis reports  $E(v|x_{t_2}, x_{t_1}) - E(v)$  which shows how the uninformed traders' Bayesian value-forecast changes with respect to the unconditional expected value of the fundamental when uninformed traders observe at  $x_{t_2}$  at  $t_2$ . We consider all the equilibrium strategies at  $t_1$  and  $t_2$  which are symmetrical. Red (green) circles show equilibrium sell (buy) orders at  $t_2$ .



and  $t_2$  price-impacts for sequences of a market order at  $t_1$  and then different orders at time  $t_2$  all line up on the vertical axis line. Interestingly, there are no observations in the second and fourth quadrants in our model, which means there are no sign reversals in the direction of the cumulative price impacts. The first and third quadrants (which are perfectly symmetrical) show the pairs of orders which have a positive and a negative price impact, respectively. The pairs with the highest price impact are driven by the insiders' equilibrium strategies at  $t_1$  and are limit orders at the inside quotes followed any other order. In fact, Table 1 shows that insiders' limit orders at the inside quotes at  $t_1$  are fully revealing. So once more, the price impact does not depend on the

aggressiveness of the orders but on the informed investors' orders choice. Overall, Figure 6 also confirms that the price impact is non-Markovian: for example the price impact of  $MOB_2$  at  $t_2$  may be either positive or negative depending on whether it is preceded by  $LOB_2$  or  $LOA_2$  at  $t_1$ .

### 2.1.5 Summary

The analysis of our first model specification has identified a number of empirically testable predictions. First, liquidity and the relative information content of different orders differ in high-volatility markets (in which value shocks are large relative to the tick size) vs. in low-volatility markets (in which value shocks are small relative to the tick size). Second, the price impact of order flow varies conditional on different trading histories and on the standing book when new orders are submitted.

## 2.2 Informed and uninformed traders both have private-value motives

Our second model specification generalizes our earlier analysis so that now informed investors also have random private-valuation factors  $\beta$  with the same truncated Normal distribution  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$  as the uninformed investors. Hence, informed traders not only speculate on their information, but they also have a private-value motive to trade. As a result, informed investors with the same signal may end up buying and selling from each other. This combination of trading motives has not been investigated in earlier models of dynamic limit order markets. We use our second model specification to show the robustness of the results in Section 2.1 and to extend them.

### 2.2.1 Trading strategies

Tables 3 and 4 report order submission probabilities and other statistics for our second model specification for time  $t_1$  by itself and for averages over times  $t_2$  through  $t_4$ . Since all investors have private-value motives to trade, all investors use all of the possible limit orders at time  $t_1$ . In particular, now informed investors also use market orders at  $t_1$ . Over times  $t_2$  through  $t_4$ , all investors again use all types of limit orders and also market orders. In particular, directionally informed investors trade sometimes opposite their asset-value information because their private-value motive adds non-informational randomness to their orders. Informed investor with neutral



**Table 3: Trading Strategies, Liquidity, and Welfare at Time  $t_1$  in an Equilibrium with Informed and Uninformed Traders both with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different informed-investor arrival probabilities  $\alpha$  (0.8 and 0.2) and two different value-shock volatilities  $\delta$  (0.16 and 0.02). The private-value factor parameters are  $\mu = 1$  and  $\sigma = 1.5$ , and the tick size is  $\kappa = 0.10$ . Each cell corresponding to a set of parameters reports the equilibrium order-submission probabilities, the expected bid-ask spreads and expected depths at the inside prices ( $A_1$  and  $B_1$ ) and total depths on each side of the market after order submissions at time  $t_1$ , and the expected welfare of the market participants. The first four columns in each parameter cell are for informed traders with positive, neutral and negative signals,  $(I_{\bar{v}}, I_{v_0}, I_v)$  and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports unconditional results for the market.

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	<i>LOA</i> <sub>2</sub>	0.118	0.054	0.031	0.064	0.067	0.054	0.048	0.042	0.048	0.048
	<i>LOA</i> <sub>1</sub>	0.314	0.446	0.282	0.426	0.363	0.438	0.452	0.466	0.452	0.452
	<i>LOB</i> <sub>1</sub>	0.282	0.446	0.314	0.426	0.363	0.466	0.452	0.438	0.452	0.452
	<i>LOB</i> <sub>2</sub>	0.031	0.054	0.118	0.064	0.067	0.042	0.048	0.054	0.048	0.048
	<i>MOA</i> <sub>2</sub>	0.256	0	0	0.009	0.070	0	0	0	0	0
	<i>MOA</i> <sub>1</sub>	0	0	0	0	0	0	0	0	0	0
	<i>MOB</i> <sub>1</sub>	0	0	0	0	0	0	0	0	0	0
	<i>MOB</i> <sub>2</sub>	0	0	0.256	0.009	0.070	0	0	0	0	0
	<i>NT</i>	0	0	0	0	0	0	0	0	0	0
	E[Spread $\cdot$ ]	0.240	0.211	0.240	0.215	0.227	0.210	0.210	0.210	0.210	0.210
	E[Depth $A_2+A_1$ $\cdot$ ]	1.432	1.500	1.312	1.491	1.430	1.492	1.500	1.508	1.500	1.500
	E[Depth $A_1$ $\cdot$ ]	0.314	0.446	0.282	0.426	0.363	0.438	0.452	0.466	0.452	0.452
	E[Depth $B_1$ $\cdot$ ]	0.282	0.446	0.314	0.426	0.363	0.466	0.452	0.438	0.452	0.452
	E[Depth $B_1+B_2$ $\cdot$ ]	1.312	1.500	1.432	1.491	1.430	1.508	1.500	1.492	1.500	1.500
	E[Welfare LO $\cdot$ ]	0.259	0.445	0.259	0.410		0.446	0.446	0.446	0.446	
	E[Welfare MO $\cdot$ ]	0.187	0	0.187	0.015		0	0	0	0	
E[Welfare $\cdot$ ]	0.446	0.445	0.446	0.425		0.446	0.446	0.446	0.446		
$\alpha = 0.2$	<i>LOA</i> <sub>2</sub>	0.063	0.051	0.042	0.051	0.051	0.049	0.048	0.046	0.048	0.048
	<i>LOA</i> <sub>1</sub>	0.356	0.449	0.476	0.449	0.445	0.441	0.452	0.464	0.452	0.452
	<i>LOB</i> <sub>1</sub>	0.476	0.449	0.356	0.449	0.445	0.464	0.452	0.441	0.452	0.452
	<i>LOB</i> <sub>2</sub>	0.042	0.051	0.063	0.051	0.051	0.046	0.048	0.049	0.048	0.048
	<i>MOA</i> <sub>2</sub>	0.063	0	0	0	0.004	0	0	0	0	0
	<i>MOA</i> <sub>1</sub>	0	0	0	0	0	0	0	0	0	0
	<i>MOB</i> <sub>1</sub>	0	0	0	0	0	0	0	0	0	0
	<i>MOB</i> <sub>2</sub>	0	0	0.063	0	0.004	0	0	0	0	0
	<i>NT</i>	0	0	0	0	0	0	0	0	0	0
	E[Spread $\cdot$ ]	0.217	0.210	0.217	0.210	0.211	0.210	0.210	0.210	0.210	0.210
	E[Depth $A_2+A_1$ $\cdot$ ]	1.419	1.500	1.518	1.500	1.496	1.490	1.500	1.510	1.500	1.500
	E[Depth $A_1$ $\cdot$ ]	0.356	0.449	0.476	0.449	0.445	0.441	0.452	0.464	0.452	0.452
	E[Depth $B_1$ $\cdot$ ]	0.476	0.449	0.356	0.449	0.445	0.464	0.452	0.441	0.452	0.452
	E[Depth $B_1+B_2$ $\cdot$ ]	1.518	1.500	1.419	1.500	1.496	1.510	1.500	1.490	1.500	1.500
	E[Welfare LO $\cdot$ ]	0.394	0.445	0.394	0.442		0.447	0.446	0.447	0.446	
	E[Welfare MO $\cdot$ ]	0.059	0	0.059	0		0	0	0	0	
E[Welfare $\cdot$ ]	0.453	0.445	0.453	0.442		0.447	0.446	0.447	0.446		

**Table 4: Averages for Trading Strategies, Liquidity, and Welfare across Times  $t_2$  through  $t_4$  for Informed and Uninformed Traders both with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different informed-investor arrival probabilities  $\alpha$  (0.8 and 0.2) and for two different asset-value volatilities  $\delta$  (0.16 and 0.02). The private-value factor parameters are  $\mu = 1$  and  $\sigma = 1.5$ , and the tick size is  $\kappa = 0.10$ . Each cell corresponding to a set of parameters reports the equilibrium order-submission probabilities, the expected bid-ask spreads and expected depths at the inside prices ( $A_1$  and  $B_1$ ) and total depths on each side of the market after order submissions at times  $t_2$  through  $t_4$ , and the expected welfare of the market participants. The first four columns in each parameter cell are for informed traders with positive, neutral and negative signals, ( $I_{\bar{v}}, I_{v_0}, I_v$ ) and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports unconditional results for the market.

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	<i>LOA</i> <sub>2</sub>	0.140	0.121	0.090	0.114	0.117	0.127	0.123	0.119	0.123	0.123
	<i>LOA</i> <sub>1</sub>	0.108	0.058	0.050	0.067	0.071	0.057	0.053	0.048	0.053	0.053
	<i>LOB</i> <sub>1</sub>	0.050	0.058	0.108	0.067	0.071	0.048	0.053	0.057	0.053	0.053
	<i>LOB</i> <sub>2</sub>	0.090	0.121	0.140	0.114	0.117	0.119	0.123	0.127	0.123	0.123
	<i>MOA</i> <sub>2</sub>	0.275	0.192	0.113	0.195	0.194	0.207	0.194	0.181	0.194	0.194
	<i>MOA</i> <sub>1</sub>	0.158	0.127	0.062	0.122	0.117	0.133	0.128	0.124	0.129	0.128
	<i>MOB</i> <sub>1</sub>	0.062	0.127	0.158	0.122	0.117	0.124	0.128	0.133	0.129	0.128
	<i>MOB</i> <sub>2</sub>	0.113	0.192	0.275	0.195	0.194	0.181	0.194	0.207	0.194	0.194
	<i>NT</i>	0.003	0.003	0.003	0.005	0.004	0.004	0.003	0.004	0.004	0.004
	E[Spread $\cdot$ ]	0.253	0.259	0.253	0.274	0.259	0.268	0.269	0.268	0.269	0.268
	E[Depth $A_2+A_1 \cdot$ ]	1.599	1.600	1.537	1.563	1.576	1.590	1.593	1.596	1.593	1.593
	E[Depth $A_1 \cdot$ ]	0.301	0.339	0.338	0.314	0.324	0.324	0.333	0.344	0.333	0.334
	E[Depth $B_1 \cdot$ ]	0.338	0.339	0.301	0.314	0.324	0.344	0.333	0.324	0.333	0.334
	E[Depth $B_1+B_2 \cdot$ ]	1.537	1.600	1.599	1.563	1.576	1.596	1.593	1.590	1.593	1.593
	E[Welfare LO $\cdot$ ]	0.089	0.071	0.089	0.072		0.067	0.067	0.067	0.067	
	E[Welfare MO $\cdot$ ]	0.328	0.332	0.328	0.331		0.336	0.336	0.336	0.336	
	E[Welfare $\cdot$ ]	0.418	0.403	0.418	0.404		0.403	0.403	0.403	0.403	
	$\alpha = 0.2$	<i>LOA</i> <sub>2</sub>	0.131	0.123	0.114	0.122	0.122	0.124	0.123	0.122	0.123
<i>LOA</i> <sub>1</sub>		0.059	0.054	0.049	0.053	0.054	0.053	0.053	0.052	0.053	0.053
<i>LOB</i> <sub>1</sub>		0.049	0.054	0.059	0.053	0.054	0.052	0.053	0.053	0.053	0.053
<i>LOB</i> <sub>2</sub>		0.114	0.123	0.131	0.122	0.122	0.122	0.123	0.124	0.123	0.123
<i>MOA</i> <sub>2</sub>		0.257	0.194	0.137	0.196	0.196	0.202	0.194	0.186	0.194	0.194
<i>MOA</i> <sub>1</sub>		0.160	0.127	0.090	0.127	0.127	0.133	0.128	0.124	0.128	0.128
<i>MOB</i> <sub>1</sub>		0.090	0.127	0.160	0.127	0.127	0.124	0.128	0.133	0.128	0.128
<i>MOB</i> <sub>2</sub>		0.137	0.194	0.257	0.196	0.196	0.186	0.194	0.202	0.194	0.194
<i>NT</i>		0.004	0.003	0.004	0.004	0.004	0.004	0.003	0.004	0.004	0.004
E[Spread $\cdot$ ]		0.266	0.267	0.266	0.269	0.269	0.269	0.269	0.269	0.269	0.269
E[Depth $A_2+A_1 \cdot$ ]		1.547	1.595	1.636	1.591	1.591	1.587	1.593	1.599	1.592	1.592
E[Depth $A_1 \cdot$ ]		0.288	0.334	0.378	0.332	0.332	0.327	0.333	0.339	0.333	0.333
E[Depth $B_1 \cdot$ ]		0.378	0.334	0.288	0.332	0.332	0.339	0.333	0.327	0.333	0.333
E[Depth $B_1+B_2 \cdot$ ]		1.636	1.595	1.547	1.591	1.591	1.599	1.593	1.587	1.592	1.592
E[Welfare LO $\cdot$ ]		0.068	0.068	0.068	0.067		0.067	0.067	0.067	0.067	
E[Welfare MO $\cdot$ ]		0.348	0.334	0.348	0.335		0.336	0.336	0.336	0.336	
E[Welfare $\cdot$ ]		0.416	0.403	0.416	0.402		0.403	0.403	0.403	0.403	

news  $I_{v_0}$  no longer just provide liquidity using limit orders. Now, due to their private-value motive, they sometimes also take liquidity using market orders.

Consider next the impact of adverse selection on trading behavior. Tables 3 and 4 show at time  $t_1$  and on average over times  $t_2$  through  $t_4$  respectively that the effects of an increase in value-shock volatility on the strategies of informed traders with good or bad news differs when we consider traders' own vs. opposite sides of the market. In particular, the "own" side of the market for an informed investor with good news is the bid (buy) side of the limit order book. The effect on the informed trader's own-side behavior is similar to the previous model specification in Section 2.1. With higher value-shock volatility, the private information about the asset value is more valuable, and both  $I_{\bar{v}}$  and  $I_v$  investors change some of their aggressive limit orders into market orders. Table 3 shows that, when  $\delta$  is increased with  $\alpha$  fixed at 0.8, the  $I_{\bar{v}}$  investors at time  $t_1$  reduce the strategy probability for  $LOB_1$  orders from 0.466 to 0.282 and increase the strategy probability for  $MOA_2$  orders from 0 to 0.256, and symmetrically  $I_v$  investors shifts from  $LOA_1$  to  $MOB_2$ .

The effects of higher volatility on uninformed traders slightly differs at  $t_1$  as opposed to times  $t_2$  through  $t_4$ . At  $t_1$  uninformed traders post slightly more aggressive orders when they demand liquidity (the strategy probabilities for  $MOA_2$  and  $MOB_2$  increase from 0 to 0.009), and more patient orders when they supply liquidity (the strategy probabilities for  $LOB_2$  and  $LOA_2$  increase slightly from 0.048 to 0.064). This change in order-submission strategies is the consequence of uninformed traders facing higher adverse selection costs. They feel safer hitting the trading crowd at  $A_2$  and  $B_2$  and offering liquidity at more profitable price levels to make up for the increased adverse selection costs. In later periods  $t_1$  through  $t_4$ , as uninformed traders learn about the fundamental value of the asset, they still take liquidity at the outside quotes (the probabilities of  $MOA_2$  and  $MOB_2$  increase slightly to 0.195 in Table 4), but move to the inside quotes to supply liquidity ( $LOA_1$  and  $LOB_1$  increase to 0.067 for times  $t_2$  through  $t_4$ ). As they learn about the future value of the asset, uninformed traders perceive less adverse selection costs and can afford to offer liquidity at more aggressive quotes. In contrast, the effects of increased value-shock volatility on the trading behavior of  $I_{v_0}$  investors are relatively modest both at time  $t_1$  and at times  $t_2$  through  $t_4$ .

The effects of an increase in the value-shock volatility is different on the opposite side than on the own side. For example, when the volatility  $\delta$  increases from 0.02 to 0.16,  $I_{\bar{v}}$  investors at time  $t_1$  switch on the own side from  $LOB_1$  limit orders to aggressive  $MOA_2$  market orders but at the same time they switch on the opposite side from aggressive limit orders to more patient limit orders. The reason why  $I_{\bar{v}}$  investors with low private-values become more patient when selling via limit orders on the opposite side is that they know that the execution probability of limit sells at  $A_2$  is higher because other  $I_{\bar{v}}$  investors in future periods will hit limit sell orders at  $A_2$  more aggressively given that  $\bar{v}$  is much bigger (see the increased order submission probabilities for  $MOA_2$  in Table 4).

### 2.2.2 Market quality

The effect of value-shock volatility on market liquidity is mixed in our second model specification. This is not surprising given the different effects of increased volatility on informed-investor trading behavior on the own and opposite sides of the market. At time  $t_1$ , holding the informed-investor arrival probability  $\alpha$  fixed, increased value-shock volatility leads to wider spreads and less total depth. However, the average effects over times  $t_2$  through  $t_4$  switches with increased asset-value volatility leading now to narrower spreads and smaller depth. This is due — in particular in the high  $\alpha$  markets — to uninformed traders perceiving greater adverse selection costs and therefore being less willing to supply liquidity. Interestingly, the effects of an increase in the arrival probability of informed investors ( $\alpha$ ) on the equilibrium strategies is qualitatively similar to that of an increase in volatility ( $\delta$ ) in this second model specification.

Lastly, our model shows how an increase in volatility and in the proportion of insiders affect the welfare of market participants. When volatility increases, directional informed investors are generally better off as their signal is stronger and hence more profitable: At  $t_1$  their welfare is unchanged with high proportion of insiders (0.446), whereas it increases in all the other scenarios, with low proportion of insiders (0.453) and in later periods with both high and low  $\alpha$  (0.418 and 0.416). At  $t_1$  uninformed traders are worse off because liquidity deteriorates with higher volatility. At later periods the result is ambiguous: there are cases in which the uninformed investors are better off and cases in which they are worse off.

### 2.2.3 Information content of orders

Figure 7 plots the Bayesian revisions for different orders at time  $t_1$  against the corresponding order-execution probabilities for our second model specification. Once again, the magnitudes and signs of the Bayesian updates depends on the mix of informed and uninformed investors who submit these orders. Consider, for example, the market with both high value-shock volatility and a high informed-investor arrival probability (large circles). The most informative orders are the market orders  $MOA_2$  and  $MOB_2$  as they are chosen much more often by informed investors than by uninformed investors (See Table 3). However, the next most aggressive orders are the inside limit orders  $LOB_1$  and  $LOA_1$ , and they are less informative than the less aggressive  $LOB_2$  and  $LOA_2$  limit orders. Even though the aggressive limit orders  $LOB_1$  and  $LOA_1$  are posted with a relatively high probability (0.282 and 0.314) by informed investors  $I_{\bar{v}}$  and  $I_{\underline{v}}$ , they are also submitted with even high probabilities by uninformed investors (0.426), and  $I_{v_0}$  informed investors with neutral (0.446). As a result, they are less informative.<sup>15</sup> Thus, this is another example in which order informativeness is not increasing in order aggressiveness.

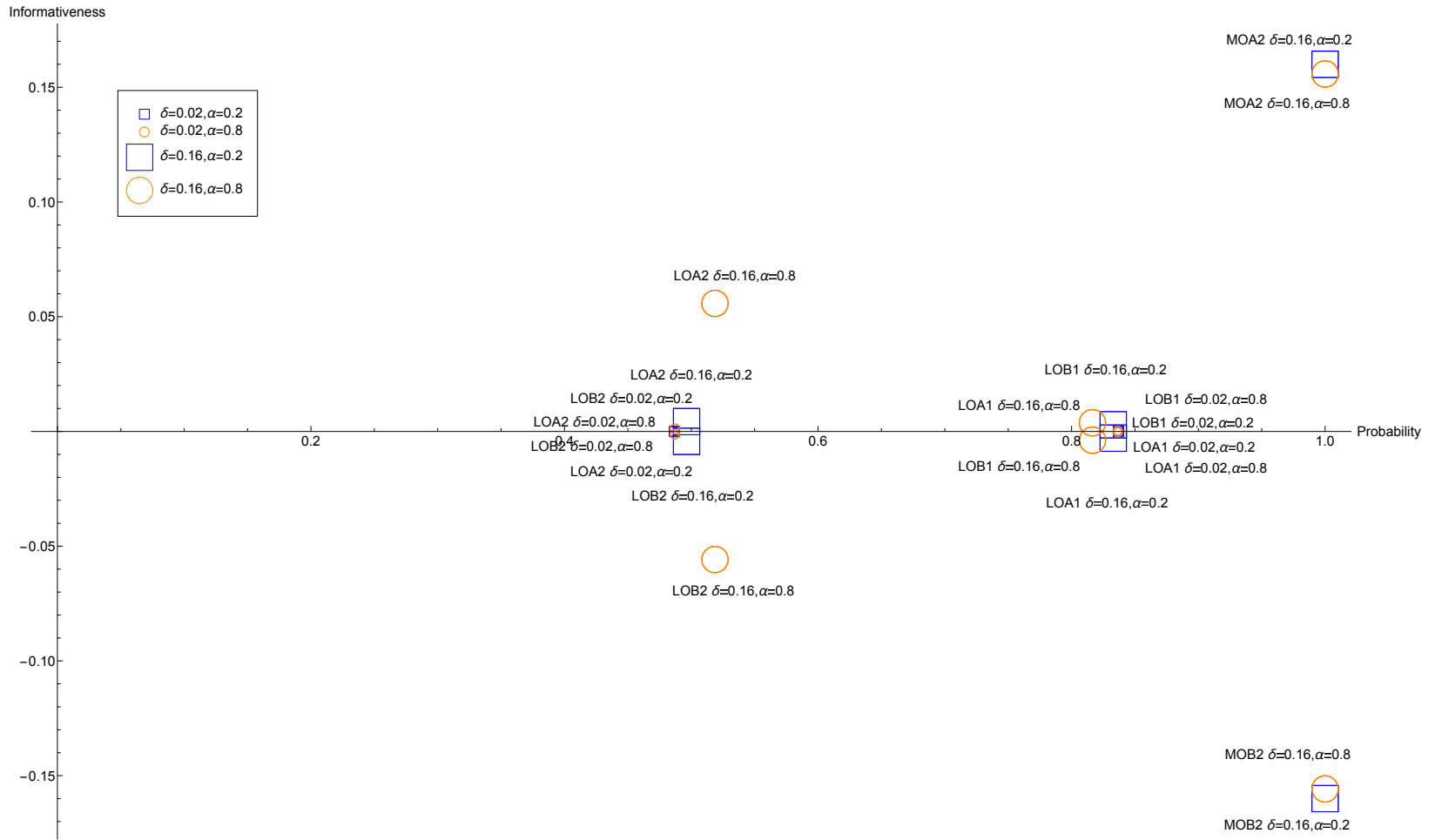
Perhaps more surprisingly, the order-sign conjecture need not hold in our second model specification:

**Result 5** The Bayesian value revision can be opposite the direction of an order.

This is to say that the direction of orders is sometimes different from the sign of their information content. For example, a limit sell  $LOA_1$  signals good news (rather than bad news as one might expect) because limit sells at  $A_1$  are used by informed investor to trade on the opposite side of their information (i.e., due to their private-value  $\beta$  factors) more frequently than these orders are used to trade on the same side of their information. In particular,  $I_{\underline{v}}$  investors usually sell using market orders  $MOB_2$  rather than using limit sells. This goes back to our previous discussion of how informed investors trade differently on the own side of their information (when their private value  $\beta$  reinforces the trading direction from their information) and on the opposite side of their information (when their  $\beta$  reverses the trading incentive from their information).

<sup>15</sup>The informativeness of limit orders  $LOA_1$  and  $LOB_1$  in Table B2 in Appendix B are 0.004 and  $-0.004$  respectively, whereas the informativeness of limit orders  $LOA_2$  and  $LOB_2$  are 0.056 and  $-0.056$  respectively.

**Figure 7: Informativeness of Orders at the End of  $t_1$  for the Model with Informed and Uninformed Traders both with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This figure plots the *Informativeness* of the equilibrium orders at the end of  $t_1$  against the probability of execution. We consider four different combinations of informed investors arrival probability. The informativeness of an order is measured as  $E[v|x_{t_1}] - E[v]$ , where  $x_{t_1}$  denotes one of the different possible orders that can arrive at time  $t_1$ .



### 2.2.4 Non-Markovian price discovery

This section continues our investigation of the importance of non-Markovian effects in information aggregation. Figure 8 shows once again the variation in the incremental information  $E[v|\mathcal{L}_{t_j}(L_{t_j})] - E[v|L_{t_j}]$  of prior order histories  $\mathcal{L}_{t_j}(L_{t_j})$  preceding different books  $L_{t_j}$ . The plots here confirm our earlier results about non-Markovian learning.

**Figure 8: History Informativeness for Informed and Uninformed Traders both with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$  for times  $t_2$  through  $t_4$ .** This Figure shows the incremental information content of the past order history in excess of the information in the current limit order book observed at the end of time  $t_j$  as measured by  $E[v|\mathcal{L}_{t_j}(L_{t_j})] - E[v|L_{t_j}]$  where  $\mathcal{L}_{t_j}(L_{t_j})$  is a history ending in the limit order book  $L_{t_j}$ . We only consider books  $L_{t_j}$  when they occur in equilibrium in the different trading periods. The candlesticks indicate for each of these two metrics the maximum, the minimum, the median and the 75<sup>th</sup> (and 25<sup>th</sup>) percentile respectively as the top (bottom) of the bar.

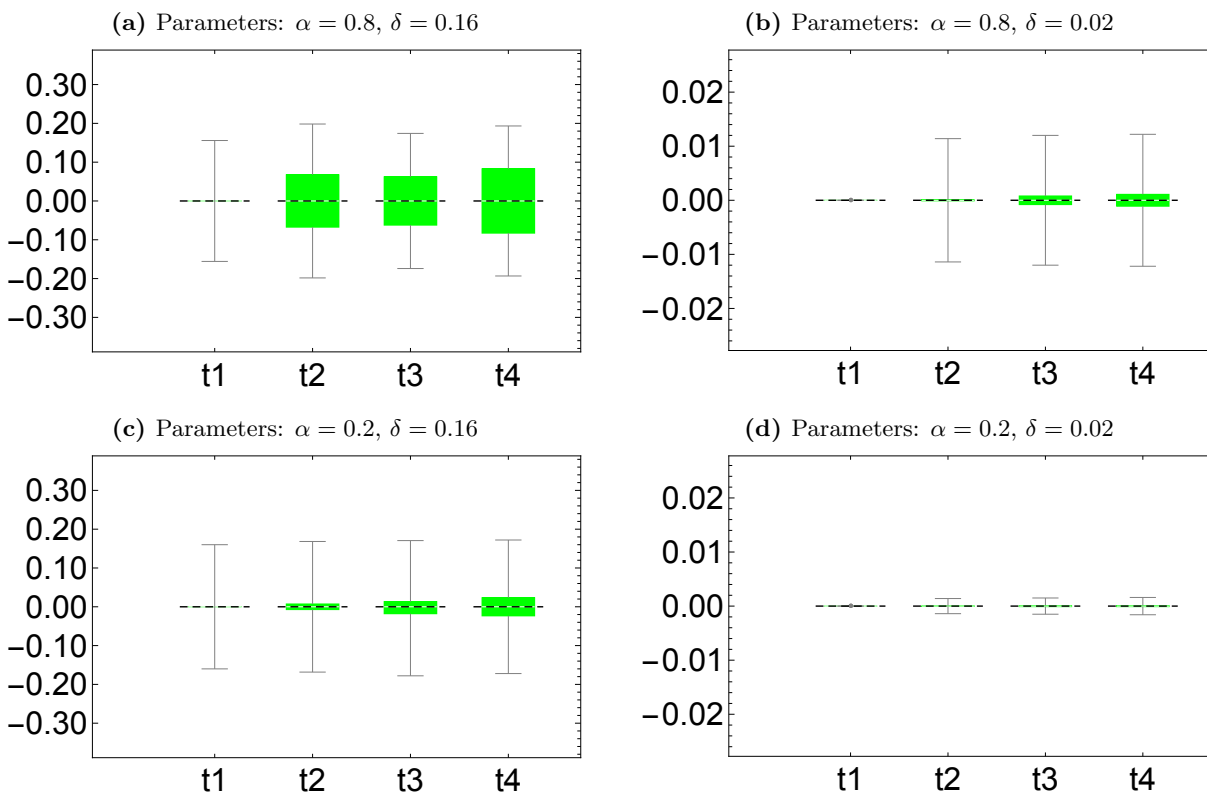


Figure 8 shows the uninformed investor's expectation of the asset value conditional on the path and various books. It also shows the expectation of these expectations across the paths, which, by

iterated expectation, is the expectation conditional on the book. Again, we see that the trading history has substantial information content above and beyond the information in the book alone. The figure also shows the standard deviation of the valuation forecast errors. Here again, the information updating dynamics are non-Markovian.

### 2.3 Summary

The results for our second model specification — with a richer specification of the informed investors’ trading motives — confirm and extend the analysis from our first model specification.

- When all market participants trade not only to speculate on their signal but also to satisfy their private-value motive, all investors use both market and limit orders in equilibrium.
- Increased adverse selection affects informed-investor trading behavior differently when they trade with their information versus (because of private-value shocks) against their information. As a result, the effect of asset-value volatility and informed investor arrival probability on market liquidity is mixed.
- The informativeness of orders can be opposite the order aggressiveness and now also the order direction. The information content of order arrivals is again history-dependent.

## 3 Robustness

Our analysis makes a number of simplifying assumptions for tractability, but we conjecture that our qualitative results are robust to relaxing these assumptions. We consider two of these assumptions here. First, our model of the trading day only has five periods. Relatedly, our analysis abstracts from limit orders being carried over from one day to the next. However, our results about the impact of adverse selection on investor trading strategies and about order informativeness are driven in large part by the relative size of information shocks and the tick size rather than by the number of rounds of trading. In addition, increasing the trading horizon leads to longer histories that are potentially even more informative. Second, arriving investors are only allowed to submit single orders that cannot be cancelled or modified subsequently. However, it seems likely that order-flow



histories will still be informative if orders at different points in time are correlated due to correlated actions of returning investors.

## 4 Conclusions

This paper has studied information aggregation and liquidity provision in dynamic limit order markets. We show a number of notable theoretical properties in our model. First, informed investors switch between endogenously demanding liquidity via market orders and supplying liquidity via limit orders. Second, the information content/price impact of orders can be non-monotone in the direction of the order and in the aggressiveness of their orders. Third, the information aggregation process is non-Markovian. In particular, the prior order history has information content beyond that in the current limit order book.

Our model suggests several directions for future research. Most importantly, our analysis provides a framework for empirical research about the changing price impacts of order flow conditional on order-flow history and time of day. There are also promising directions for future theory. First, the model can be enriched by allowing investors to trade dynamically over time (rather than just submitting an order one time) and to face quantity decisions and to use multiple orders. Second, the model could be extended to allow for trading in multiple co-existing limit order markets. This would be a realistic representation of current equity trading in the US. Third, the model could be used to study high frequency trading in limit order markets and the effect of different investors being able to process and trade on different types of information at different latencies.

## 5 Appendix A: Algorithm for computing equilibrium

The computational problem to solve for a Perfect Bayesian Nash equilibrium in our model (as defined in Section 1.1) is complex. Given investors' equilibrium beliefs, the optimal order-submission problems in (6) and (7) require computing limit-order execution probabilities and stock-value expectations that are conditional on both the past order history and on future state-contingent limit-order execution at each time  $t_j$  at each node of the trading game. For an informed trader (who

knows the asset value  $v$ ), there is no uncertainty about the payoff of a market order. In contrast, the payoff of a market order for an uninformed trader entails uncertainty about the future asset value and, therefore, computing the optimal order requires computing the expected stock value  $E[v|\mathcal{L}_{t_{j-1}}]$  conditional on the prior trading history up to time  $t_j$ . For limit orders, the expected payoff depends on the future limit-order execution probabilities,  $Pr(\theta_{t_j}^x|v, \mathcal{L}_{t_{j-1}})$  and  $Pr(\theta_{t_j}^x|\mathcal{L}_{t_{j-1}})$ , for informed and uninformed investors, which depend, in turn, on the optimal order-submission probabilities of future informed and uninformed investors. In addition, the uninformed investors' learning problem for limit orders requires uninformed investors to extract information about the expected future stock value  $E[v|\mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$  from both the past trading history and also from state-contingent future order execution given that the future states in which limit orders are executed are correlated with the stock value. Thus, optimal actions at each time  $t_j$  depend on past information and future order-flow contingencies where future orders also depend on the then-prior histories at future dates (which include the action at time  $t_j$ ) as traders dynamically update their equilibrium beliefs as the trading process unfolds. Thus, the learning problem for limit order beliefs is both backward- and forward-looking. Lastly, rational expectations (RE) involves finding a fixed point so that the equilibrium beliefs underlying the optimal order-submission strategies are consistent with the execution probabilities and value expectations that the endogenous optimal strategies produce in equilibrium.

Our numerical algorithm uses backward induction to solve for optimal order strategies given a set of asset-value beliefs for all dates and nodes in the trading game and uses an iterative recursion to solve for RE equilibrium asset-value and order-execution beliefs. The backward induction makes order-execution probabilities consistent with optimal future behavior by later-arriving investors. It also takes future state-contingent execution into account in the uninformed investors' beliefs. Given a set of history-contingent asset-value probability beliefs, we start at time  $t_5$  — when traders only use market orders which allows us to compute the execution probabilities of limit orders at  $t_4$  — and recursively solve the model for optimal order strategies back to time  $t_1$ . We then embed the optimal order strategy calculation in an iterative recursion to solve for a fixed point for the RE asset-value beliefs. For a generic round  $r$  in this recursion, the outgoing asset-value probabilities  $\pi_{t_j}^{v,r-1}$  from round  $r-1$  are used iteratively as incoming asset-value beliefs in round  $r$ . In particular,

these beliefs are used in the learning problem of the uninformed investor to extract information about the ending asset value  $v$  from the prior trading histories. They also affect the behavior of informed investors whose order-execution probability beliefs depend in part on the behavior of uninformed traders. Thus, the recursion for a generic round  $r$  involves solving by backward induction for optimal strategies for buyers

$$\max_{x \in X_{t_j}} w^{I,r}(x | v, \mathcal{L}_{t_{j-1}}) = [\beta_{t_j} v_0 + \Delta - p(x)] Pr^r(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}}) \quad (13)$$

and

$$\max_{x \in X_{t_j}} w^{U,r}(x | \mathcal{L}_{t_{j-1}}) = [\beta_{t_j} v_0 + E^r[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x] - p(x)] Pr^r(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}}) \quad (14)$$

where

$$E^r[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x] = (\hat{\pi}_{t_j}^{\bar{v},r} \bar{v} + \hat{\pi}_{t_j}^{v_0,r} v_0 + \hat{\pi}_{t_j}^{\underline{v},r} \underline{v}) - v_0 \quad (15)$$

$$\hat{\pi}_{t_j}^{v,r} = \frac{\Pr^r(\theta_{t_j}^x | v, \mathcal{L}_{t_j})}{Pr^r(\theta_{t_j}^x | \mathcal{L}_{t_j})} \pi_{t_j}^{v,r-1} \quad (16)$$

and where the calculations for sellers are symmetric. Note that at each time  $t_j$  the backward induction has already determined the future contingencies  $\theta_{t_j}^x$  for limit order executions at times  $t > t_j$ . Thus, the order-execution probabilities  $Pr^r(\theta_{t_j}^x | v, \mathcal{L}_{t_{j-1}})$  and  $Pr^r(\theta_{t_j}^x | \mathcal{L}_{t_{j-1}})$ , and the history- and execution-contingent probabilities  $\hat{\pi}_{t_j}^{v,r}$  and associated asset-value expectations  $E^r[\Delta | \mathcal{L}_{t_{j-1}}, \theta_{t_j}^x]$  are “mongrel” moments in that they are computed using the outgoing history-contingent asset value beliefs  $\pi_{t_j}^{v,r-1}$  from round  $r-1$  and then updated given the order-execution contingencies computed by backward induction in round  $r$  using the round  $r-1$  asset-value beliefs. At the end of round  $r$ , we then compute updated outgoing asset-value beliefs  $\pi_{t_j}^{v,r}$  for round  $r$ , which are used as incoming beliefs for the next round  $r+1$ . The recursion is iterated to find a RE fixed point  $\pi_{t_j}^v$  in the uninformed investor beliefs.

The fixed-point recursion is started in round  $r=1$  by setting the initial asset-value beliefs  $\pi_{t_j}^{v,0}$  of uninformed traders at each time  $t_j$  in the backward induction to be the unconditional priors

$Pr(v)$  in (1). In particular, the algorithm starts by ignoring conditioning on history in the initial round  $r = 1$ . Hence the traders' optimization problems in (14) and (13) in round  $r = 1$  simplify to:

$$\max_{x \in X_{t_j}} w^{I,r=1}(x | v, \mathcal{L}_{t_{j-1}}) = [\beta_{t_j} v_0 + \Delta - p(x)] Pr^1(\theta_{t_j}^x | v) \quad (17)$$

$$\max_{x \in X_{t_j}} w^{U,r=1}(x | \mathcal{L}_{t_{j-1}}) = [\beta_{t_j} v_0 + E^1[\Delta | \theta_{t_j}^x] - p(x)] Pr^1(\theta_{t_j}^x) \quad (18)$$

The order-execution contingencies in round  $r$  are modeled as follows: In each round  $r$  given the asset-value beliefs  $\pi_{t_j}^{v,r-1}$  in that round, we solve for investors' optimal trading strategies by backward induction. Starting at  $t_5$ , the execution probability for new limit orders is zero, and therefore optimal order-submission strategies only use market orders. Given the linearity of the expected payoffs in the private-value factor  $\beta$  in (13) and (14), the optimal orders for an informed trader at  $t_5$  are<sup>16</sup>

$$x_{t_5}^{I,r}(\beta | \mathcal{L}_{t_4}, v) = \begin{cases} MOB_{i,t_5} & \text{if } \beta \in [0, \beta^{MOB_{i,t_5}^{I,r}, NT_{t_5}^{I,r}}) \\ NT & \text{if } \beta \in [\beta^{MOB_{i,t_5}^{I,r}, NT_{t_5}^{I,r}}, \beta^{NT_{t_5}^{I,r}, MOA_{i,t_5}^{I,r}}) \\ MOA_{i,t_5} & \text{if } \beta \in [\beta^{NT_{t_5}^{I,r}, MOA_{i,t_5}^{I,r}}, 2] \end{cases} \quad (19)$$

where for each possible combination of  $MOB_{i,t_5} = MOB_1, MOB_2$  and  $MOA_{i,t_5} = MOA_1, MOA_2$

$$\begin{aligned} \beta^{MOB_{i,t_5}^{I,r}, NT_{t_5}^{I,r}} &= \frac{B_{i,t_5} - \Delta}{v} \\ \beta^{NT_{t_5}^{I,r}, MOA_{i,t_5}^{I,r}} &= \frac{A_{i,t_5} - \Delta}{v} \end{aligned} \quad (20)$$

are the critical thresholds that solve  $w^{I,r}(MOB_{i,t_5} | v, \mathcal{L}_{t_4}) = w^{I,r}(NT | v, \mathcal{L}_{t_4})$  and  $w^{I,r}(NT | v, \mathcal{L}_{t_4}) = w^{I,r}(MOA_{i,t_5} | v, \mathcal{L}_{t_4})$ , respectively. The optimal trading strategies and  $\beta$  thresholds for an unin-

<sup>16</sup>For instance, an informed trader would post a  $MOA_1$  only if the payoff is positive and thus outperforms the NT payoff of zero, i.e.  $\beta v + \Delta - A_1 > 0$  or  $\beta > \frac{A_1 - \Delta}{v}$ .

formed traders are similar but the conditioning set does not include the asset value  $v$ :

$$x_{t_5}^{U,r}(\beta|\mathcal{L}_{t_4}) = \begin{cases} MOB_{i,t_5} & \text{if } \beta \in [0, \beta^{MOB_{i,t_5}^{U,r}, NT_{t_5}^{U,r}}) \\ NT & \text{if } \beta \in [\beta^{MOB_{i,t_5}^{U,r}, NT_{t_5}^{U,r}}, \beta^{NT_{t_5}^{U,r}, MOA_{i,t_5}^{U,r}}) \\ MOA_{i,t_5} & \text{if } \beta \in [\beta^{NT_{t_5}^{U,r}, MOA_{i,t_5}^{U,r}}, 2] \end{cases} \quad (21)$$

where

$$\begin{aligned} \beta^{MOB_{i,t_5}^{U,r}, NT_{t_5}^{U,r}} &= \frac{B_{i,t_5} - E^{r-1}[\Delta|\mathcal{L}_{t_4}]}{v} \\ \beta^{NT_{t_5}^{U,r}, MOA_{i,t_5}^{U,r}} &= \frac{A_{i,t_5} - E^{r-1}[\Delta|\mathcal{L}_{t_4}]}{v} \end{aligned} \quad (22)$$

Given the  $\beta$  ranges associated with each possible action at  $t_5$ , we compute the submission probabilities associated with each optimal order at  $t_5$  using the truncated-Normal density  $\mathbf{n}(\cdot)$  for the private factor  $\beta$ .<sup>17</sup> At time  $t_4$  these are the execution probabilities for new limit orders by an informed investor at the different possible best bids and asks,  $B_{i,t_4}$  and  $A_{i,t_4}$  respectively at time  $t_5$ :

$$Pr^r(\theta_{t_4}^{LOB_i}|\mathcal{L}_{t_3}, v) = \begin{cases} \alpha \left[ \int_0^{\beta^{MOB_{i,t_5}^{I,r}, NT_{t_5}^{I,r}}} \mathbf{n}(\beta) d\beta \right] + (1 - \alpha) \left[ \int_0^{\beta^{MOB_{i,t_5}^{U,r}, NT_{t_5}^{U,r}}} \mathbf{n}(\beta) d\beta \right] \\ 0 \end{cases} \quad \text{otherwise} \quad (23)$$

$$Pr^r(\theta_{t_4}^{LOA_i}|\mathcal{L}_{t_3}, v) = \begin{cases} \alpha \left[ \int_{\beta^{NT_{t_5}^{I,r}, MOA_{i,t_5}^{I,r}}}^2 \mathbf{n}(\beta) d\beta \right] + (1 - \alpha) \left[ \int_{\beta^{NT_{t_5}^{U,r}, MOA_{i,t_5}^{U,r}}}^2 \mathbf{n}(\beta) d\beta \right] \\ 0 \end{cases} \quad \text{otherwise} \quad (24)$$

where the book is either empty at  $A_1$  and/or  $B_1$  (but may have non-crowd limit orders at the outside prices) or is empty except for just crowd orders at  $A_2$  and  $B_2$ . The analogous execution

---

<sup>17</sup>The discussion here is for the case where both informed and uninformed investors have random private factors  $\beta$ .

probabilities for an uninformed investor arriving at time  $t_4$  are:

$$Pr^r(\theta_{t_4}^{LOB_i} | \mathcal{L}_{t_3}) = \begin{cases} \alpha \left[ \sum_{v \in \{\bar{v}, v_0, \underline{v}\}} \hat{\pi}_{t_4}^{v,r} \int_0^{\beta_{t_5}^{MOB_{i,t_5}^{Iv,r}, NT_{t_5}^{Iv,r}}} \mathbf{n}(\beta) d\beta \right] + (1 - \alpha) \left[ \int_0^{\beta_{t_5}^{MOB_{i,t_5}^{Uv,r}, NT_{t_5}^{Uv,r}}} \mathbf{n}(\beta) d\beta \right] \\ 0 \end{cases} \quad \text{otherwise} \quad (25)$$

$$Pr^r(\theta_{t_4}^{LOA_i} | \mathcal{L}_{t_3}) = \begin{cases} \alpha \left[ \sum_{v \in \{\bar{v}, v_0, \underline{v}\}} \hat{\pi}_{t_4}^{v,r} \int_{\beta_{t_5}^{NT_{t_5}^{Iv,r}, MOA_{i,t_5}^{Iv,r}}}^2 \mathbf{n}(\beta) d\beta \right] + (1 - \alpha) \left[ \int_{\beta_{t_5}^{NT_{t_5}^{Uv,r}, MOA_{i,t_5}^{Uv,r}}}^2 \mathbf{n}(\beta) d\beta \right] \\ 0 \end{cases} \quad \text{otherwise} \quad (26)$$

At  $t_4$  there is only one period before the end of the trading game. Thus, the execution probability of a limit order is positive if and only if the order is posted at the best price on its own side of the market ( $A_{i,t_4}$  or  $B_{i,t_4}$ ), and if there are no non-crowd limit orders already standing in the limit order book at that price at the time the new limit order is posted.

Having obtained the execution probabilities in (23) – (26) for the different limit orders at  $t_4$ , we next derive the optimal order-submission strategies at  $t_4$ . The incoming book can be configured in many different ways at  $t_4$  depending on the different possible prior order paths  $\mathcal{L}_3$  in the trading game up through time  $t_3$ . As the payoffs of both limit and market orders are functions of  $\beta$ , we rank all the payoffs of adjacent optimal strategies in terms of  $\beta$  and equate them to determine the  $\beta$  thresholds at time  $t_4$ .<sup>18</sup> Consider, for example, an order path such that  $t_4$  has only crowd orders in the book, so that new limit and market orders are both potentially optimal orders at  $t_4$ . For an

---

<sup>18</sup>Recall that the upper envelope only includes strategies that are optimal.

informed trader, the the optimal orders are given by:

$$x_{t_4}^{I,r}(\beta | \mathcal{L}_{t_3}, v) = \begin{cases} MOB_2 & \text{if } \beta \in [0, \beta^{MOB_{2,t_4}^{I,r}, LOA_{1,t_4}^{I,r}}) \\ LOA_1 & \text{if } \beta \in [\beta^{MOB_{2,t_4}^{I,r}, LOA_{1,t_4}^{I,r}}, \beta^{LOA_{1,t_4}^{I,r}, LOA_{2,t_4}^{I,r}}) \\ LOA_2 & \text{if } \beta \in [\beta^{LOA_{1,t_4}^{I,r}, LOA_{2,t_4}^{I,r}}, \beta^{LOA_{2,t_4}^{I,r}, NT_{t_4}^{I,r}}) \\ NT & \text{if } \beta \in [\beta^{LOA_{2,t_4}^{I,r}, NT_{t_4}^{I,r}}, \beta^{NT_{t_4}^{I,r}, LOB_{2,t_4}^{I,r}}) \\ LOB_2 & \text{if } \beta \in [\beta^{NT_{t_4}^{I,r}, LOB_{2,t_4}^{I,r}}, \beta^{LOB_{2,t_4}^{I,r}, LOB_{1,t_4}^{I,r}}) \\ LOB_1 & \text{if } \beta \in [\beta^{LOB_{2,t_4}^{I,r}, LOB_{1,t_4}^{I,r}}, \beta^{LOB_{1,t_4}^{I,r}, MOA_{2,t_4}^{I,r}}) \\ MOA_2 & \text{if } \beta \in [\beta^{LOB_{1,t_4}^{I,r}, MOA_{2,t_4}^{I,r}}, 2] \end{cases} \quad (27)$$

and for an uninformed trader the optimal strategies are qualitatively similar but with different values for the  $\beta$  thresholds given the uninformed investor's different information.<sup>19</sup> As the payoffs of both limit and market orders are functions of  $\beta$ , we can rank all the payoffs of adjacent optimal strategies in terms of  $\beta$  and equate them to determine the  $\beta$  thresholds at  $t_4$ . For example, for the first  $\beta$  threshold we have:

$$\beta_{t_4}^{MOB_{2,t_4}^{I,r}, LOA_{1,t_4}^{I,r}} = \beta \in \mathbb{R} \text{ s.t. } w_{t_4}^{I,r}(MOB_2 | v, \beta, \mathcal{L}_{t_3}) = w_{t_4}^{I,r}(LOA_1 | v, \beta, \mathcal{L}_{t_3}) \quad (28)$$

and we obtain the other thresholds similarly.

The next step is to use the  $\beta$  thresholds together with the truncated Normal cumulative distribution  $\mathbb{N}(\cdot)$  for  $\beta$  to derive the probabilities of the optimal order-submission strategies at each possible node of the extensive form of the game at  $t_4$ . For example, the submission probability of  $LOA_1^{I,r}$  is:

$$Pr^r[LOA_1^{I,r} | \mathcal{L}_{t_3}, v] = \mathbb{N}(\beta^{LOA_{1,t_4}^{I,r}, LOA_{2,t_4}^{I,r}} | \mathcal{L}_{t_3}, v) - \mathbb{N}(\beta^{MOB_{2,t_4}^{I,r}, LOA_{1,t_4}^{I,r}} | \mathcal{L}_{t_3}, v) \quad (29)$$

and the submission probabilities of the equilibrium strategies can be obtained in a similar way.

Next, given the market-order submission probabilities at  $t_4$  — which together with the execution

---

<sup>19</sup>If the incoming book from  $t_3$  has non-crowd orders on any level of the book, the equilibrium strategies would be different. For example, if the book has a  $LOA_1$  limit order, then new limit orders on the ask side cannot be equilibrium orders since their execution probability would be zero.

probabilities at  $t_5$  determine the execution probabilities for new limit orders at time  $t_3$  — we can solve the optimal orders at  $t_3$  and then recursively continue to solve the model by backward induction in this fashion back to time  $t_1$ .

**Off-equilibrium beliefs:** At each time  $t_j$ , round  $r$  of the recursion needs history-contingent asset-value beliefs  $\pi_{t_j}^{v,r-1} = Pr^{r-1}(v|\mathcal{L}_{t_j})$  from round  $r - 1$  for all feasible paths that traders may use. Beliefs for paths that occur with positive probability in round  $r - 1$  are computed using Bayes' rule to update the probability  $Pr^{r-1}(v|\mathcal{L}_{t_{j-1}})$  of the time- $t_{j-1}$  sub-path  $\mathcal{L}_{t_{j-1}}$  that path  $\mathcal{L}_{t_j}$  extends. In contrast, Bayes' Rule cannot be used to update probabilities of paths that involve orders that are not used with positive probability in round  $r - 1$ . Our algorithm deals with this by setting  $Pr^{r-1}(v|\mathcal{L}_{t_j})$  to be  $Pr^{r-1}(v|\mathcal{L}_t)$  where  $\mathcal{L}_t$  is the longest positive-probability sub-path from  $t_0$  to some time  $t < t_k$  in round  $r - 1$  that is contained in path  $\mathcal{L}_{t_j}$ . For example, consider a path  $\{MOA_2, MOB_2, LOA_1\}$  at time  $t_3$  where orders  $\{MOA_2, MOB_2\}$  are used with positive probability at times  $t_1$  and  $t_2$  in round  $r - 1$ , but  $LOA_1$  is not used at time  $t_3$  after the first two orders in round  $r - 1$ . Our recursion algorithm sets the round  $r - 1$  belief uninformed traders use for path  $\{MOA_2, MOB_2, LOA_1\}$  to be their round  $r - 1$  belief for the positive-probability sub-path  $\{MOA_2, MOB_2\}$ . If instead  $MOB_2$  is not a positive-probability order at  $t_2$  in round  $r - 1$ , then we assume that uninformed traders use their belief at  $t_1$  conditional on the shorter sub-path  $\{MOA_2\}$ . Finally, if  $MOA_2$  is also not a positive-probability order at  $t_1$  in round  $r - 1$ , then we assume that traders use their unconditional prior belief  $Pr(v)$ .

**Mixed strategies:** We allow for both pure and mixed strategies in our Perfect Bayesian Nash equilibrium. When different orders have equal expected payoffs, we assume that traders randomize with equal probabilities across all such optimal orders. By construction, the expected payoffs of two different strategies are the same in correspondence of the  $\beta$  thresholds; however because we are considering single points in the support of the  $\beta$  distribution, the probability associated with any strategy that corresponds to those specific points is equal to zero. This means that mixed strategies that emerge in correspondence of the  $\beta$  thresholds, although feasible, have zero probability. Mixed strategies may also emerge in the framework in which informed traders have a fixed neutral private-



value factor  $\beta = 1$  (section 2.1). More specifically it may happen that the payoffs of two perfectly symmetrical strategies of  $I_{v_0}$  are the same, and in this case  $I_{v_0}$  randomizes between these two strategies.

In the setting of our model where informed traders have fixed neutral private-value factors  $\beta = 1$ , it may happen that both informed and uninformed traders switch their strategies back and forth from one round to the next. When this happens, to reach an equilibrium we assume that the informed traders play mixed strategies and at each subsequent round strategically reduce the probability with which they choose the most profitable strategy until the equilibrium is reached. As an example at  $t_1$  informed traders with positive news,  $I_{\bar{v}}$ , play  $LOB_2$  in round  $r = 1$ . However, in round  $r = 2$  in the subsequent periods uninformed traders do not send market orders to sell at  $B_2$  and in round  $r = 3$ , informed traders react by changing their strategy to  $LOB_1$ . However, in the subsequent periods uninformed traders do not send market orders to sell, this time at  $B_1$ . To find an equilibrium, we assume that at each round informed traders play mixed strategies and assign a greater weight to the most profitable strategy. In this case we assume they start playing  $LOB_2$  with probability 0.99 and  $LOB_1$  with probability 0.01. If these mixed strategies do not lead to an equilibrium outcome, in the subsequent round we assume that the informed traders play  $LOB_2$  with probability 0.98 and  $LOB_1$  with probability 0.02. We proceed by lowering the probability with which informed traders choose the most profitable strategy until we reach an equilibrium set of strategies.

**Convergence:** RE beliefs for a Perfect Bayesian Nash equilibrium are obtained by solving the model recursively for multiple rounds. In particular, the asset-value probabilities  $\pi_{t_j}^{v,1}$  from round  $r = 1$  from above are used as the priors to solve the model in round  $r = 2$  (i.e., the round 1 probabilities are used in place of the unconditional priors used in round 1).<sup>20</sup> The asset-value probabilities  $\pi_{t_j}^{v,2}$  from round  $r = 2$  are then used as the priors in round  $r = 3$  and so on. The recursive iteration is continued until the updating process converges to a fixed point, which are the RE beliefs. In particular, the recursive process has converged to the RE beliefs when uninformed traders no longer revise their asset-value beliefs. Operationally, we consider convergence to the RE

---

<sup>20</sup>In the second round of solutions we again solve the full 5-period model.

beliefs to have occurred when the probabilities  $\pi_{t_j}^{\bar{v},r}$ ,  $\pi_{t_j}^{v_0,r}$  and  $\pi_{t_j}^{v,r}$  in round  $r$  are “close enough” to the corresponding probabilities from round  $r - 1$ :

$$\begin{aligned}
\pi_{t_j}^{\bar{v},*} & \text{ when } \left| \pi_{t_j}^{\bar{v},r} - \pi_{t_j}^{\bar{v},r-1} \right| < 10^{-7} \\
\pi_{t_j}^{v_0,*} & \text{ when } \left| \pi_{t_j}^{v_0,r} - \pi_{t_j}^{v_0,r-1} \right| < 10^{-7} \\
\pi_{t_j}^{v,*} & \text{ when } \left| \pi_{t_j}^{v,r} - \pi_{t_j}^{v,r-1} \right| < 10^{-7}
\end{aligned} \tag{30}$$

A fixed-point solution to this recursive algorithm is an equilibrium in our model.

## 6 Appendix B: Additional numerical results

The tables in this section provide additional information on the execution probabilities of limit orders for informed investor with positive, neutral and negative signals,  $(I_{\bar{v}}, I_{v_0}, I_v)$  and for uninformed traders. The tables also report the asset value expectations of the uninformed investor at time  $t_2$  after observing all the possible buy orders submissions at time  $t_1$ . The expectations for sell orders are symmetric with respect to 1. Table B1 reports results for our first model specification in which only uninformed traders have a random private value factor. Table B2 reports results for our second model in which both the informed and uninformed traders have private-value motives.

**Table B1: Order Execution Probabilities and Asset-Value Expectation for Informed Traders with  $\beta = 1$  and Uninformed Traders with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different values of the informed-investor arrival probability  $\alpha$  (0.8 and 0.2) and for two different values of the asset-value volatility  $\delta$  (0.16 and 0.02).  $\sigma = 1.5$ . For each set of parameters, the first four columns report the equilibrium limit order probabilities of executions for informed traders with positive, neutral and negative signals,  $(I_{\bar{v}}, I_{v_0}, I_v)$  and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports the unconditional order-execution probabilities in the market. Next, the columns report conditional and unconditional future order execution probabilities and the asset-value expectations of an uninformed investor at time  $t_2$  after observing different order submissions at time  $t_1$ .

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	$P^{EX}(LOA_2 \cdot)$	0.955	0.175	0.055	0.395	0.395	0.180	0.229	0.170	0.193	0.193
	$P^{EX}(LOA_1 \cdot)$	0.989	0.125	0.078	0.397	0.397	0.323	0.323	0.323	0.323	0.323
	$P^{EX}(LOB_1 \cdot)$	0.078	0.125	0.989	0.397	0.397	0.323	0.323	0.323	0.323	0.323
	$P^{EX}(LOB_2 \cdot)$	0.055	0.175	0.955	0.395	0.395	0.170	0.229	0.180	0.193	0.193
	$E[v LOB_1 \cdot]$					1.160					1.000
	$E[v LOB_2 \cdot]$					1.083					1.013
	$E[v MOA_1 \cdot]$										
	$E[v MOA_2 \cdot]$					1.000					1.000
	$P^{EX}(LOA_2 \cdot)$	0.651	0.487	0.394	0.511	0.511	0.514	0.499	0.476	0.496	0.496
	$P^{EX}(LOA_1 \cdot)$	0.886	0.766	0.717	0.789	0.789	0.792	0.792	0.790	0.791	0.791
$P^{EX}(LOB_1 \cdot)$	0.717	0.766	0.886	0.789	0.789	0.790	0.792	0.792	0.791	0.791	
$P^{EX}(LOB_2 \cdot)$	0.394	0.487	0.651	0.511	0.511	0.476	0.499	0.514	0.496	0.496	
$\alpha = 0.2$	$E[v LOB_1 \cdot]$					1.026					1.000
	$E[v LOB_2 \cdot]$					1.013					1.009
	$E[v MOA_1 \cdot]$										
	$E[v MOA_2 \cdot]$					1.000					1.000

**Table B2: Order Execution Probabilities and Asset-Value Expectation for Informed and Uninformed Traders both with  $\beta \sim Tr[\mathcal{N}(\mu, \sigma^2)]$ .** This table reports results for two different values of the informed-investor arrival probability  $\alpha$  (0.8 and 0.2) and for two different values of the asset-value volatility  $\delta$  (0.16 and 0.02).  $\sigma = 1.5$ . For each set of parameters, the first four columns report the equilibrium limit order probabilities of executions for informed traders with positive, neutral and negative signals,  $(I_{\bar{v}}, I_{v_0}, I_v)$  and for uninformed traders ( $U$ ). The fifth column (*Uncond.*) reports the unconditional order-execution probabilities in the market. Next, the columns report conditional and unconditional future order execution probabilities and the asset-value expectations of an uninformed investor at time  $t_2$  after observing different order submissions at time  $t_1$ .

		$\delta = 0.16$					$\delta = 0.02$				
		$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>	$I_{\bar{v}}$	$I_{v_0}$	$I_v$	$U$	<i>Uncond.</i>
$\alpha = 0.8$	$P^{EX}(LOA_2 \cdot)$	0.644	0.502	0.410	0.519	0.135	0.502	0.487	0.472	0.487	0.116
	$P^{EX}(LOA_1 \cdot)$	0.913	0.834	0.702	0.817	0.392	0.849	0.837	0.824	0.836	0.470
	$P^{EX}(LOB_1 \cdot)$	0.702	0.834	0.913	0.817	0.392	0.824	0.837	0.849	0.836	0.470
	$P^{EX}(LOB_2 \cdot)$	0.410	0.502	0.644	0.519	0.135	0.472	0.487	0.502	0.487	0.116
	$E[v LOB_1 \cdot]$					0.996					1.000
	$E[v LOB_2 \cdot]$					0.944					0.999
	$E[v MOA_1 \cdot]$										
	$E[v MOA_2 \cdot]$					1.156					
	$P^{EX}(LOA_2 \cdot)$	0.525	0.494	0.470	0.496	0.402	0.490	0.487	0.483	0.487	0.394
	$P^{EX}(LOA_1 \cdot)$	0.853	0.833	0.813	0.833	0.737	0.839	0.837	0.834	0.837	0.745
$P^{EX}(LOB_1 \cdot)$	0.813	0.833	0.853	0.833	0.737	0.834	0.837	0.839	0.837	0.745	
$P^{EX}(LOB_2 \cdot)$	0.470	0.494	0.525	0.496	0.402	0.483	0.487	0.490	0.487	0.394	
$\alpha = 0.2$	$E[v LOB_1 \cdot]$					1.003					1.000
	$E[v LOB_2 \cdot]$					0.996					1.000
	$E[v MOA_1 \cdot]$										
	$E[v MOA_2 \cdot]$					1.160					

## References

- Aït-Sahalia, Yacine, and Mehmet Saglam, 2013, High frequency traders: Taking advantage of speed, Technical report, National Bureau of Economic Research.
- Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2015, Equilibrium fast trading, *Journal of Financial Economics* 116, 292–313.
- Bloomfield, Robert, Maureen O’Hara, and Gideon Saar, 2005, The “make or take” decision in an electronic market: Evidence on the evolution of liquidity, *Journal of Financial Economics* 75, 165–199.

- Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2016, Price discovery without trading: Evidence from limit orders.
- Fleming, Michael J, Bruce Mizrach, and Giang Nguyen, 2017, The microstructure of a us treasury ecn: The brokertec platform, *Journal of Financial Markets* 1–52.
- Foucault, Thierry, 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* 2, 99–134.
- Foucault, Thierry, Johan Hombert, and Ioanid Roşu, 2016, News trading and speed, *Journal of Finance* 71, 335–382.
- Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.
- Gencay, Ramazan, Soheil Mahmoodzadeh, Jakub Rojcek, and Michael C Tseng, 2016, Price impact and bursts in liquidity provision.
- Glosten, Lawrence R, and Paul R Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71–100.
- Goettler, Ronald L, Christine A Parlour, and Uday Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149–2192.
- Goettler, Ronald L, Christine A Parlour, and Uday Rajan, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67–87.
- Jain, Pankaj K, 2005, Financial market design and the equity premium: Electronic versus floor trading, *Journal of Finance* 60, 2955–2985.
- Kaniel, Ron, and Hong Liu, 2006, So what orders do informed traders use?, *The Journal of Business* 79, 1867–1913.
- Kumar, Praveen, and Duane J Seppi, 1994, Limit and market orders with optimizing traders.
- Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica* 1315–1335.

- Milgrom, Paul, and Nancy Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.
- O’Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257–270.
- Parlour, Christine A, 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789–816.
- Parlour, Christine A, and Duane J Seppi, 2008, Limit order markets: A survey, *Handbook of Financial Intermediation and Banking* 5, 63–95.
- Rindi, Barbara, 2008, Informed traders as liquidity providers: Anonymity, liquidity and price formation, *Review of Finance* 497–532.
- Roşu, Ioanid, 2016a, Fast and slow informed trading.
- Roşu, Ioanid, 2016b, Liquidity and information in order driven markets.