Anonymity, Signaling, and Collusion in Limit Order Books

Álvaro Cartea[a,b], Patrick Chang[a,b], Rob Graumans[a,c]

[a] *Oxford-Man Institute, University of Oxford*
[b] *Mathematical Institute, University of Oxford*
[c] *Autoriteit Financiële Markten*

Abstract

A key feature in the design of a limit order book is the anonymity of limit orders. However, we analyze data with trader identification and find that market makers break the anonymity of limit orders. Market makers use limit orders with large volumes to signal themselves to other market makers to avoid trading with each other and to snipe retail limit orders. We explain the behavior of market makers with a model that considers competitive and collusive equilibria. The model shows that the behavior of market makers is consistent with that in a collusive equilibrium where market makers use signals to avoid sniping each other's limit orders. Signaling enables market makers to share the profitable benign flow from retail limit orders, and simultaneously limit competition from retail limit orders, which enables market makers to receive additional benign flow from impatient investors who would have otherwise traded with a retail investor's limit order.

Keywords: High-frequency trading, Collusion, Signaling

Disclaimer: The views expressed in this article are those of the authors, and not necessarily those of the Autoriteit Financiële Markten. The authors are solely responsible for errors or omissions.

1. Introduction

Anonymity is important for strategic interactions in financial markets. In limit order books, anonymity enables informed investors to blend in with uninformed investors (see Kyle, 1985), it facilitates investors to hide their trading intentions to prevent back-running and predatory trading

We would like to thank Elena Asparouhova, Alan Beggs, Jonathan Brogaard, Fayçal Drissi, Mauro Cesa, Winston Wei Dou, Björn Hagströmer, Terry Hendershott, Peter Hoffmann, David L. Kelly, Joep Knook, Roger Lee, Albert Menkveld, Joshua Mollner, Tom Noe, Lars Nordén, Roel Oomen, José Penalva, Maarten Pieter Schinkel, Stefan Schlamp, Jan Svitak, Chaojun Wang, Basil Williams, Noah Williams, and seminar participants at Stockholm Business School, Chinese University of Hong Kong, University of Miami, David Eccles School of Business, Barclays, Citi, Cubist, Deutsche Bank, Optiver, Deutsche Börse, AFM and ACM, ESMA, DG Competition, Financial Economics and Microstructure Workshop, Algo-Trading & DeFi Workshop, International Fintech Research Conference, the Market Microstructure, Quantitative Trading, High Frequency and Large Data Conference at the University of Chicago, and the 24th Annual SAET Conference for helpful comments. We also thank Alexandre Roubaud and Sander van Nugteren from BlackRock for helpful comments. Patrick Chang and Rob Graumans acknowledge financial support from the Oxford-Man Institute.

Preprint submitted to TBA Latest version on SSRN. This version: November 18, 2025. First version: January 3, 2025.

(see Yang and Zhu, 2019; Brunnermeier and Pedersen, 2005), and it enables market makers to conceal their inventory positions (see Biais, 1993). Importantly, anonymity improves market quality (see Comerton-Forde et al., 2005). Therefore, why would market makers reveal themselves to each other? What happens when market makers break the pre-trade anonymity of limit orders and reveal themselves to each other?

We use proprietary data from Euronext Amsterdam to study the anonymity of the limit order book in the ETF market. We find that market makers in ETFs knowingly or unknowingly use the volume of their limit orders to signal themselves to each other. Market makers submit limit orders with volumes that are very large compared with the size of limit orders sent by other market participants and compared with the size of transactions. Specifically, the median value of limit orders posted by market makers is approximately 226 times larger than the median value of limit orders posted by retail traders, and it is approximately 164 times larger than the median value of transactions.[1] Extant theory shows that providing excess liquidity and breaking anonymity entails obvious costs. There is no such thing as a free lunch, thus: what are the offsetting benefits for market makers?

For market makers, one benefit from providing excess liquidity at the best bid and at the best ask is that investors who wish to trade are forced to send a marketable limit order and cross the spread, or are forced to send a limit order inside the quoted spread to gain priority through a price improvement (see Li et al., 2021).

Another benefit from breaking anonymity is that market makers can infer who sent which limit order in an otherwise anonymous market, which affects their decision to trade with an order. In ETFs, only 1.37% of all transactions are between market makers, whereas in shares, 24.02% of all transactions are between market makers.[2] The lack of transactions between market makers cannot be fully attributed to the breaking of anonymity because there are many factors that affect the decision to trade with an order (e.g., duration and staleness).

To paint a more complete picture and to highlight the effect of breaking anonymity, we study the behavior of market participants who snipe the limit orders that improve the quoted spread. A spread-improving limit order is sniped if it enters inside the quoted spread, and all or part of it is taken out by an aggressive order within 1 millisecond.[3] In these cases, spread-improving limit

[1] For shares in Euronext, the median value of limit orders posted by market makers is 1.48 times the median value of limit orders posted by retail traders, and is approximately 1.47 times the median transaction value.

[2] Differences in quoting behavior cannot explain the disparity in transactions between market makers. In the most traded ETFs, the top five market makers are at the touch 53% of the time, while in the most traded shares, the top five market makers are at the touch 60% of the time.

[3] We focus on new spread-improving limit orders with a resting time of at most 1 millisecond to isolate the effect of the volume and the price of a limit order on the decision of high-frequency traders to trade with the order, and to eliminate any effect which might be due to the resting time of the order, such as orders becoming stale.

orders that are close to crossing the quoted spread are very likely to be sniped (62.58% probability) if the limit orders originate from a retail trader, and very unlikely to be sniped (0.08% probability) if the limit orders originate from a market maker. When close to crossing the quoted spread, a spread-improving limit order sent by a retail investor is over 2,000 times more likely to be sniped (in terms of group-wise odds) than a spread-improving limit order sent by a market maker. This behavior shows that when snipers target small limit orders, they consistently trade against retail limit orders. Snipers can trade consistently with retail investors because market makers break the anonymity of limit orders. However, why would market makers voluntary help the snipers figure out which limit orders are from retail traders?

In this market, market makers are the snipers who benefit from breaking the anonymity of limit orders, and they account for 82.24% of this sniping activity. Market makers successfully win the race to snipe retail limit orders with 90.68% probability. Their success stems from a speed advantage; their median reaction time to snipe limit orders that improve the spread is 91 microseconds. This is faster than the reaction time of any non-retail market participant. Thus, market makers in the ETF market act as makers and as takers of liquidity, and they behave as the usual high-frequency traders (HFTs) in the literature.[4,5]

We argue that the arrival of new information, flickering quotes, queue priority, and quoting requirements cannot explain the behavior of the HFTs (i.e., the market makers who are also snipers). To explain the behavior of the HFTs, we propose a model that considers competitive and collusive equilibria, and we show that their behavior is consistent with that in a collusive equilibrium.[6] Our model highlights the economic forces that provide incentives for HFTs to reveal themselves to each other. The model also provides the conditions that determine if HFTs willingly reveal themselves to each other, and analyzes how HFTs mutually benefit and collectively enforce this behavior.

In our model, there are patient and impatient investors, and there are $N \geq 3$ risk-neutral HFTs who play an infinitely repeated trading game. The trading game is a generalization of the two-period trading game in Budish et al. (2024), and it retains the main elements of latency arbitrage and adverse selection. In each trading game, our model considers the possibility that one of the HFTs receives short-lived private information, or that a patient investor arrives and sends a limit order that improves the quoted spread. Informed HFTs trading alongside patient investors intro-

[4] We call these participants market makers because that is how they officially identify themselves to the exchange.

[5] In response to the first draft of this paper, various industry participants publicly confirmed that the limit order book is indeed not anonymous. Furthermore, a former trader confirmed that firms factor this information into their trading strategies, and they suggest that this practice continues “because it does not breach any rules and provides valuable information to market makers”. See Clancy and Cesa (2025) for the various quotes from market participants.

[6] In this paper, collusion refers to the economic theory of collusion where firms use strategies that embody a reward-punishment scheme to obtain supracompetitive profits (see Harrington, 2018).

duces strategic ambiguity.[7] Ambiguity arises because a limit order that is sent inside the spread could be benign flow from a patient investor or it could be toxic flow from an informed HFT. In the latter case, the informed HFT pretends to be a patient investor and posts a toxic limit order to fool a sniper into trading with the toxic limit order, i.e., the sniper adversely selects herself when trading with the informed HFT.

This ambiguity creates a decision problem for the HFTs. Do HFTs snipe an incoming spread-improving limit order without knowing who sent it and potentially face the risk of adversely selecting themselves, or do HFTs wait until they can verify that the limit order was sent by a patient investor before they trade? The choice of HFTs to snipe or to wait-and-verify affects the equilibrium quoted spread that impatient investors (i.e., liquidity traders) face.

The intuition to determine if HFTs immediately snipe incoming limit orders that improve the quoted spread is straightforward. In the competitive equilibrium, if the benign flow from immediately sniping patient investors is sufficiently profitable, then HFTs immediately snipe any limit order that improves the quoted spread and bear the costs of being fooled by informed HFTs pretending to be patient investors. The consequence is that the impatient investors always trade at the spread set by the HFTs, i.e., due to their speed disadvantage, impatient investors cannot trade with spread-improving limit orders posted by patient investors. On the other hand, if the cost of being fooled by informed HFTs is too high, then HFTs wait until they can verify that the order was sent by a patient investor before they trade. The consequence is that HFTs cannot remove competition from the limit orders of patient investors, and impatient investors can occasionally trade with a spread-improving limit order posted by a patient investor. In this case, the HFTs collectively earn less because the HFT who provides liquidity has fewer opportunities to provide liquidity to an impatient investor, and the HFTs who act as latency arbitrageurs have fewer opportunities to latency arbitrage the HFT providing liquidity.

With a competitive baseline established, we study a collusive equilibrium to explain the observed behavior of HFTs. The collusive strategy is a reversion strategy with a cooperation phase and a punishment phase (see Green and Porter, 1984). In the cooperation phase, HFTs signal to each other to avoid trading with each other. HFTs cooperate until they observe a deviation or suspect a deviation from cooperation, both of which trigger a punishment phase that reverts to competitive play for T iterations of the trading game, after which, play returns to the cooperation phase.

In the collusive equilibrium, the primary gains from cooperation depend on the underlying competitive equilibrium used in the reversion phase of the strategy. When HFTs snipe spread-improving limit orders in the competitive equilibrium, the primary gains from cooperation arise

[7] In our model, patient investors are a proxy for the retail limit orders we see in the data.

from a wider spread relative to the competitive equilibrium (see Dutta and Madhavan, 1997). When HFTs wait-and-verify before they trade in the competitive equilibrium, the primary gains from cooperation arise from signaling because it enables HFTs to identify benign limit orders from patient investors and to identify toxic limit orders from informed HFTs. Breaking the anonymity of the limit orders enables HFTs to safely profit by sniping limit orders from patient investors, and simultaneously limit competition from patient investors, which as a byproduct, enables the HFTs to receive additional benign flow from impatient investors who would have otherwise matched with a patient investor's limit order at better prices than those quoted by HFTs. Both of which are sources of excess profits that would otherwise be unavailable in the competitive equilibrium where HFTs wait-and-verify before they trade. The resulting equilibrium in this case is one where: (i) quoted spreads are on average wider than that in the competitive equilibrium because HFTs can safely snipe limit orders sent by patient investors, and (ii) the trading costs for impatient investors are higher because they are forced to trade at the spread set by HFTs (they cannot trade with spread-improving limit orders from a patient investor unless play is in the punishment phase).

Finally, the collusive equilibrium does not always exist. Factors that affect the existence of the collusive equilibrium include the number of HFTs, arrival of private information, and the mispricing of limit orders from patient investors. Theoretically, as the number N of HFTs increases, the possibility of colluding decreases. Furthermore, when there is less private information, the possibility of colluding decreases. Lastly, the collusive equilibrium exists only when the limit orders sent by patient investors are sufficiently mispriced.

2. Related Literature

Our work is closely related to Christie and Schultz (1994) who document an absence of odd-eighth quotes for Nasdaq stocks, and who offer implicit collusion as the most likely explanation. Here, we document a phenomenon where market makers break the anonymity of limit orders by signaling themselves to each other, and similarly, we offer collusion as a possible explanation. Dutta and Madhavan (1997) provide a model of dealer markets and rationalize the results of Christie and Schultz (1994) through a collusive arrangement. Our paper provides a model of the limit order book and rationalizes signaling through a collusive arrangement. Our approach is similar to Bryzgalova et al. (2025) who present a model where arbitrageurs choose to specialize in some markets, and who present evidence consistent with their theory from the options market.

Our study shares similarities with the bid signaling that occurred in the Federal Communications Commission (FCC) spectrum auctions, where rivals used "code bidding" to send messages to their rivals to coordinate on which licenses to bid and which to avoid (see Cramton and Schwartz, 2000, 2001).[8] Here, our theory shows that HFTs signal their limit orders with large volumes to distinguish their limit orders from the limit orders sent by patient investors. Signaling enables HFTs to share the profitable benign flow of patient investors, and simultaneously limit competition from patient investors, which as a byproduct, enables the HFTs to receive additional benign flow from impatient investors who would have otherwise matched with a patient investor's limit order.

Fundamental to our result is that signaling breaks the anonymity of limit orders. This artificial form of pre-trade transparency creates a different playing field for a subset of market participants. Therefore, our work is related to the literature that analyzes the effect of transparency on market dynamics (see De Jong and Rindi, 2009). Transparency comes in various dimensions, from post-trade transparency (see e.g., Madhavan, 1995; Friederich and Payne, 2014; Meling, 2021) to pre-trade quotation transparency (see e.g., Biais, 1993; Madhavan et al., 2005). Our paper is closer to the literature below on the pre-trade anonymity of limit orders.

Simaan et al. (2003) argue that the anonymity of limit orders reduces the probability of collusion among quote setters because it limits their ability to monitor and punish deviations from cooperation. Empirically, they show that the spread of Nasdaq dealer quotes posted through anonymous electronic communication networks are tighter than those of Nasdaq dealer quotes posted through the transparent dealer quotation system. In our model, monitoring is imperfect because signaling is voluntary. However, this does not affect the existence of a collusive equilibrium because our problem is similar to one of "secret price cutting" (see Green and Porter, 1984).

Foucault et al. (2007) show that limit order anonymity can result in tighter bid-ask spreads because a non-anonymous environment can lead to free riding, to which market makers respond by quoting a wider spread. Our model has a similar flavor. When HFTs signal themselves to each other, they partially reveal their private information. A collusive equilibrium exists when the gains outweigh the costs from revealing yourself.

Comerton-Forde et al. (2005) and Comerton-Forde and Tang (2009) use natural experiments to study the impact of pre-trade anonymity and find that limit order anonymity improves liquidity. On the other hand, Comerton-Forde et al. (2011) study how brokers strategically disclose their identity to reduce their execution costs.

The main difference between our work and all the aforementioned papers on transparency is that the degree of transparency is imposed by the design of the market. In contrast, our setting is an artificial form of pre-trade transparency that is achieved only if HFTs decide to break their anonymity. The goal of our model is to understand the incentives for HFTs to reveal themselves.

Our model connects to the early literature on market microstructure that studies liquidity and asymmetric information (e.g., Glosten and Milgrom, 1985; Kyle, 1985; Glosten, 1994). Similar to

[8] See also Klemperer (2002) for other examples of signaling in auction markets.

our work is Easley and O'Hara (1987), where the authors show that order flow signals the investor's type (i.e., informed or uninformed). Our model looks at the size of limit orders as a method to signal your type. In our model, the competitive equilibrium is a pooling equilibrium where one cannot differentiate between limit orders sent by informed HFTs or limit orders sent by patient investors, while the collusive equilibrium is a separating equilibrium where HFTs voluntarily distinguish themselves from patient investors.

Our framework builds upon the literature on HFTs, see for example Biais et al. (2011), Cartea and Penalva (2012), Aït-Sahalia and Saglam (2013), Foucault et al. (2013), Hoffmann (2014), Jovanovic and Menkveld (2016), Menkveld and Zoican (2017), Foucault et al. (2017), Bernales (2017); see also Menkveld (2016) for a review. Our model retains the main elements of latency arbitrage (see e.g., Budish et al., 2015) and adverse selection (see e.g., Baldauf and Mollner, 2020; Budish et al., 2024). Similar to Li et al. (2021), our model introduces investors who send limit orders inside the quoted spread, but the main difference is that by allowing HFTs to be informed, we introduce an element of ambiguity as to *who* sent the limit order inside the quoted spread.

The empirical literature on HFTs is diverse. From understanding how HFTs affect market quality (see e.g., Brogaard, 2010; Hendershott et al., 2011; Hagströmer and Nordén, 2013; Brogaard et al., 2015; Brogaard and Garriott, 2019) to how HFTs increase price discovery (see e.g., Brogaard et al., 2014, 2019). Our work is closer to the literature that documents the behavior of HFTs (Hendershott and Riordan, 2009; Hagströmer et al., 2014; Kirilenko et al., 2017; Aquilina et al., 2021). Here, we document a phenomenon where HFTs knowingly or unknowingly signal themselves to each other.

Finally, our work contributes to the recent literature on collusion (see e.g., Calvano et al., 2020; Harrington, 2022; den Boer et al., 2022; Abada and Lambin, 2023). In financial markets, Cartea et al. (2022a,b) and Colliard et al. (2022) study how market makers who use reinforcement learning algorithms can end up quoting supracompetitive prices, while Dou et al. (2024) study how informed traders who use reinforcement learning algorithms can learn collusive strategies. Similar to Dutta and Madhavan (1997) and Bryzgalova et al. (2025), we study collusion through the usual repeated game approach.

3. Data and Institutional Details

This section discusses the dataset we use in our study, and it describes the characteristics of the market and its participants. The data are from Euronext Amsterdam and are collected by the Dutch Authority for the Financial Markets.

3.1. Data

Market. Our data consist of all the messages from Euronext Amsterdam's Cash Market. Each message includes the security identification, date and timestamp to the nanosecond, order type, order volume, and a buy/sell indicator. Importantly, our data contain the identification and dealing capacity of each member trading in the exchange.

Members. A member is any individual, corporation, partnership, association, trust, or entity who has been admitted to Euronext Securities Membership or Euronext Derivatives Membership, and whose membership has not been terminated. Only members can trade on Euronext Amsterdam (Euronext, 2023). Members include retail-brokers, investment banks, and high-frequency traders (see Euronext, 2024, for a list of members on Euronext Amsterdam).

Dealing Capacity. A member can trade on her own account or on behalf of her clients, which in turn, determines her dealing capacity. In our dataset, the dealing capacity is provided by the exchange. The dealing capacities are: (i) Client, (ii) House, (iii) Retail Liquidity Provider, (iv) Retail, (v) Liquidity Provider, and (vi) Related Party. A member can trade under one or multiple dealing capacities, but uses only one dealing capacity per message sent to the exchange.

Members with dealing capacity Retail are retail-brokers who handle retail orders. These orders are entered by retail clients, and the broker sends the orders to the exchange on behalf of the client. Examples of retail-brokers are Robinhood, Interactive Brokers, etc.

Members with dealing capacity Liquidity Provider are those who have a contractual agreement with Euronext to meet quoting obligations. We distinguish between market makers who signed the exchange's Market Making Agreement (i.e., any investment firm that pursues a market making strategy must enter into a binding agreement with the exchange according to RTS 8 MiFID II Article 1), and market makers under the exchange's Market Making Scheme. The requirements for market makers under the Market Making Scheme are stricter than those for market makers who only signed the Market Making Agreement. Market makers under the Market Making Scheme receive fee reductions for passive executions if they meet the quoting obligations.

Members with dealing capacity Client are those who send orders on behalf of larger institutional clients, and members with dealing capacity House are those who trade on their own account, but do not qualify as a Liquidity Provider. Finally, we remove orders and transactions by members who do not trade with dealing capacity Retail, Liquidity Provider, Client, and House.[9]

To streamline the terminology, we refer to members who send orders under the dealing capacity of Liquidity Provider as market makers, and members who send orders under the dealing capacity

[9] In January 2022, this removes 0.8% of transactions in ETFs, and 1.7% of transactions in shares.

of Retail as retail. Finally, we refer to members who send orders under the dealing capacity of Client or House as others, because our main focus is on market makers and retail.

3.2. ETF Market

	Order count	Trade count	Value traded in millions €	Value traded by retail in millions €	Number of instruments	Retail trade count of total %
Asset class						
Exchange traded funds (ETFs)	425,146,572	390,628	2,634	912	367	70
Common/ordinary shares	256,659,629	8,318,418	59,701	6,510	143	10
Structured instruments (without protection)	49,836,657	9,319	63	19	109	59
Structured instruments	36,838,474	4,384	14	5	87	67
Bonds	18,287,803	6,364	*	*	*	*
Asset-backed securities	8,806,692	1,059	13	2	17	59
Depositary receipts on equities	7,178,987	173,759	937	210	4	16
Medium-term notes	1,931,134	94	282	144	141	57
Money market instruments	667,202	0	0	0	4	0
Others (miscellaneous)	231,594	37	0	0	3	89
Mortgage-backed securities	65,888	0	0	0	2	0
Standard investment funds/mutual funds	6,533	51	0	0	3	100
Purchase rights	977	182	0	0	11	9
Municipal bonds	150	0	0	0	1	0
Warrants	87	5	0	0	7	20
Preferred/preference shares	60	12	0	0	1	100

Table 1: Trading activity by asset class on Euronext Amsterdam Cash Markets in January 2022.

Our focus is on exchange traded funds (ETFs), which is the second largest asset class in Euronext Amsterdam by value traded. Table 1 provides statistics of trading activity for the asset classes traded on the exchange in January 2022.[10] There are 390,628 transactions in ETFs and more than 8.3 million transactions in ordinary shares, and there are approximately 425 million order entries in ETFs and approximately 256 million order entries in shares.[11] Therefore, the order-to-trade ratio for ETFs is more than 1,000 and approximately 32 for shares.

The value traded in ETFs is approximately €2.6 billion and approximately €59.7 billion for shares. The participation of retail traders in ETFs is much higher than that in shares. In ETFs, 70% of the transactions involve a retail buyer or a retail seller (or both). In shares, only 10% of the transactions involve a retail buyer or a retail seller (or both). Finally, in January 2022, 367 ETFs and the shares of 143 firms were traded in Euronext Amsterdam.

In the ETF market, Table 2 shows that transactions are concentrated among the top ETFs. The top ten ETFs (of the 367 ETFs traded in Euronext Amsterdam) account for 61% of all transactions in ETFs and for 33% of all value traded in ETFs on Euronext Amsterdam, while the top three ETFs account for 41% of all transactions in ETFs and for 19% of the value traded. Furthermore,

[10] The price of bonds and of medium-term notes are in percentages. The value traded depends on the bond's nominal value, so we cannot calculate the value traded accurately because the nominal value is different for each bond.

[11] These statistics include amendments, which are modifications to outstanding orders that count as order entries.

	Trade count	Value traded €	Trade count %	Value traded %	Spread in ticks
Vanguard S&P 500 UCITS ETF	68,298	196,613,759	17	7	31
Vanguard FTSE All-World UCITS ETF	50,346	104,931,516	13	4	3
iShares Core MSCI World UCITS ETF USD	41,407	219,954,479	11	8	6
iShares China CNY Bond UCITS ETF USD	22,243	76,213,631	6	3	71
iShares Core S&P 500 UCITS ETF USD (Dist)	16,073	60,324,319	4	2	12
iShares AEX	11,523	62,319,456	3	2	5
iShares Core S&P 500 UCITS ETF USD (Acc)	6,742	109,725,392	2	4	122
Vanguard FTSE All-World High Dividend Yield	6,528	28,927,160	2	1	56
VanEck Vectors Sustainable World Equal Weight	6,354	15,819,200	2	1	4
iShares China CNY Bond UCITS ETF USD Hedged	5,731	18,639,365	1	1	112
Top 10	235,245	893,468,277	61	33	-
Top 20	277,316	1,248,787,082	71	47	-
Top 50	332,583	1,760,204,504	85	67	-
Top 100	364,242	2,155,727,660	93	82	-
All	390,628	2,634,932,345	100	100	-

Table 2: Trade count, value traded, and median quoted spread in ticks in ETFs on Euronext Amsterdam in January 2022.

the median quoted spread of six of the ten most traded ETFs is larger than ten ticks. In fact, the majority of ETFs are small-tick instruments, we return to this point in Section 5.

3.3. *Other Exchanges*

Exchange	Average spread (in %)	Volume traded	Trade count	Average trade size
London	0.054	7,704,828	28,915	266
Euronext Amsterdam	0.056	2,476,160	62,862	39
Bloomberg MTF EU - RTS1	-	1,697,380	-	-
Xetra	0.056	1,501,361	3,907	384
Bloomberg MTF - RTS1	-	1,235,045	-	-
Tradegate	0.055	558,759	4,977	112
Borsa Italiana	0.074	451,821	1,809	250
SIX	0.086	398,301	1,192	334

Table 3: Average spread (as a percentage of the midprice), volume traded, trade count, and average trade size for Vanguard S&P 500 UCITS ETF on some exchanges in Europe in January 2022. Source: Bloomberg.

Table 3 shows the average spread (as a percentage of the midprice), traded volume, and number of transactions on other European exchanges in Vanguard S&P 500 UCITS ETF, which is the most traded ETF on Euronext Amsterdam. The table shows that the spread in Euronext Amsterdam is similar to the spread in other European exchanges, and that Euronext Amsterdam is the second largest European venue in terms of volume traded and the largest European venue in terms of trade count, while it is the smallest among European venues in terms of the average size of transactions.

3.4. *Concentration*

Table 4 presents the market share of the top one, top three, and top five members based on their dealing capacity. We measure market share in two ways. One, Table 4a shows the percentage of

	Liquidity Provider			House			Client			Retail
	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5	all
Instruments										
Decile 1 (least traded)	72	84	84	13	13	13	1	1	1	0
Decile 2	63	74	74	18	22	22	1	1	1	0
Decile 3	61	70	70	21	25	25	2	2	2	0
Decile 4	57	69	69	22	26	26	1	2	2	0
Decile 5	50	60	60	29	32	32	2	2	2	3
Decile 6	48	57	58	29	31	32	4	4	4	4
Decile 7	42	56	56	28	34	34	4	5	5	2
Decile 8	39	46	47	34	39	40	5	6	6	5
Decile 9	34	46	48	29	33	33	6	8	8	9
Decile 10 (most traded)	38	51	53	14	18	19	9	14	14	10
Simple average	52	63	64	23	26	26	3	4	4	3

(a) Concentration measured as percentage of time at the best bid or best offer.

	Liquidity Provider			House			Client			Retail
	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5	Top 1	Top 3	Top 5	all
Instruments										
Decile 1 (least traded)	61	68	68	5	5	5	10	14	14	11
Decile 2	44	54	54	14	18	18	14	15	15	10
Decile 3	43	53	53	15	21	21	11	14	14	8
Decile 4	37	48	48	19	24	24	9	13	13	12
Decile 5	37	46	46	16	24	25	10	15	16	10
Decile 6	35	46	46	14	19	20	10	15	16	15
Decile 7	27	37	38	20	26	27	11	15	16	15
Decile 8	31	40	41	17	23	24	10	15	16	16
Decile 9	24	37	39	19	25	26	7	11	12	19
Decile 10 (most traded)	28	44	45	9	12	13	16	22	23	14
Simple average	28	49	50	9	12	12	13	17	17	15

(b) Concentration measured as percentage of passive transactions.

Table 4: Concentration of the top one, top three, and top five members per dealing capacity in ETFs on Euronext Amsterdam in January 2022. Instruments are grouped and sorted from least traded (Decile 1) to most traded (Decile 10), and percentages are averages per decile.

time that limit orders of a member (when acting under a certain dealing capacity) rest at the best bid or at the best offer in the limit order book. Two, Table 4b shows the percentage of transactions in which a member (when acting under a certain dealing capacity) was on the passive side of the transaction, i.e., the member provided liquidity.[12]

In each ETF, there are three dominant market makers (who might differ per ETF). The fourth and fifth largest market makers have little to no meaningful market share under either measure of market share. Similarly, in each ETF, there is one dominant member (who might differ per ETF) trading with dealing capacity House. Overall, the three dominant market makers and the dominant house have a combined market share that ranges from a minimum of 53% in the most traded ETFs

[12] The concentration per dealing capacity does not add up to 100 because the denominator includes time spent by (respectively, liquidity providing transactions by) all dealing capacities (including Retail Liquidity Provider and Related Party).

to a maximum of 97% in the least traded ETFs, depending on the measure of market share.[13]

4. Signaling and Sniping Orders

In this section, we show that the size of the limit orders sent by market makers is very large compared with the size of those sent by retail traders, and very large relative to the size of the transactions. Furthermore, we show that market makers respond differently to incoming small limit orders (which are mostly sent by retail investors) compared with how they respond to incoming large limit orders (which are mostly sent by other market makers). Specifically, when limit orders arrive inside the quoted spread at a price level beyond the midprice, market makers snipe the smaller retail limit orders much more frequently than they snipe the larger limit orders from other market makers.

4.1. Order Size

Table 5 shows the transaction size and the size of limit orders entered by retail traders, market makers, and others in the ETF and equity market.[14]

		ETFs (in €)				Ordinary shares (in €)		
	Retail	Market makers	Others	Transaction	Retail	Market makers	Others	Transaction
mean	4,155	210,911	121,373	6,745	17,877	8,025	16,640	7,186
25th percentile	160	61,019	8,754	202	1,333	3,263	3,567	1,751
median	515	116,144	30,257	709	3,999	5,919	7,254	4,033
75th percentile	1,595	261,797	108,319	2,986	10,537	8,726	21,108	6,643

Table 5: Statistics of limit orders and transactions in ETFs (left) and shares (right) in euros on Euronext Amsterdam in January 2022.

For all ETFs, the median value of limit orders entered by market makers and retail traders is €116,144 and €515, respectively; while the median transaction value is €709. Therefore, the median value quoted by market makers is approximately 226 times larger than the median value quoted by retail traders, and approximately 164 times larger than the median transaction value.

In contrast, for all shares, the median value of limit orders entered by market makers and retail traders is €5,919 and €3,999, respectively; while the median transaction value is €4,033. Here,

[13] When considering dominant members, we ignore members who trade as Client or Retail because these members trade on behalf of multiple clients, so their market share is less likely to translate into market power. It is more appropriate to consider these members (particularly those trading as Client) as the execution algorithms of Li et al. (2021). These execution algorithms are investors with an inelastic need to buy or to sell the security, and who can choose between limit and marketable limit orders to minimize transaction costs.

[14] We take all limit orders with validity type “day” entered by retail, market makers, and others in January 2022 between 9:15 and 17:15. We focus on limit orders because we are interested in the size of visible orders resting in the limit order book. Thus, we exclude all liquidity taking orders and iceberg orders. Iceberg orders are 6% of all orders entered in ETFs. We use 371,850,665 orders in ETFs, i.e., 87% of all orders entered in ETFs in January 2022.

the size of orders in shares does not provide sufficient information to infer accurately whether the order was sent by retail or by a market maker. Moreover, market makers provide liquidity in quantities that match what the market demands.

Although the comparison between the equity market and the ETF market is not perfect, it provides a useful point of reference. According to theory, strategic market participants would want to blend in their orders with those of other market participants to avoid leaking information (see Kyle, 1985), and market makers do not provide excess liquidity (i.e., liquidity that supplies a quantity greater than the market demand) to minimize the costs of being adversely selected and latency arbitraged (see Budish et al., 2024). In contrast, in the ETF market, we observe that market makers quote volumes that are much larger than those of retail, and they provide liquidity in quantities well above the immediacy demanded. This raises the question: why do market makers quote such large volumes?

4.2. Sniping Behavior

We study a form of sniping that highlights the benefits of posting limit orders with very large volumes. Here, we look at sniping limit orders that improve the quoted spread. Specifically, a spread-improving limit order is sniped if: (i) the spread-improving order is new, i.e., not an amendment of an existing limit order, (ii) the order it trades with initiates the transaction, (iii) the order initiating the trade is an immediate-or-cancel order, and (iv) the spread-improving order is picked up within 1 millisecond from its time of entry.[15,16,17] We focus on spread-improving limit orders with a resting time of at most 1 millisecond to isolate the effect of the volume and the price of a limit order on the decision of high-frequency traders to trade with the order, and eliminate any effect which might be due to the resting time of the order, such as orders becoming stale.

Here, the spread improvement of a limit order is given by

$$\text{Spread improvement} = \begin{cases} (p-b)/(a-b)\,, & \text{limit order improves the best bid,} \\ (a-p)/(a-b)\,, & \text{limit order improves the best offer,} \end{cases}$$

where p is the limit price of the limit order, b is the price of the best bid before the limit order ar-

[15] For new orders, the exchange assigns new identifiers, while amendments keep the original identifier. This restriction ensures that snipers cannot use the lifetime of the limit order as a feature to infer who submitted the spread-improving limit order.

[16] The spread-improving limit order does not have to be completely executed. Partial executions also count as a snipe.

[17] Aquilina et al. (2021) use 500 microseconds as the upper bound on the information horizon because a longer timeframe might allow the sniper to have seen new data from other symbols or from other exchanges which inform her decision to snipe. In our study, the number of snipes does not increase substantially for time intervals in excess of one millisecond.

rived, and a is the price of the best offer before the limit order arrived. A spread improvement near zero corresponds to a minor improvement over the best bid or over the best offer, an improvement beyond 0.5 corresponds to an improvement of the best bid or the best offer beyond the midprice, and an improvement near 1 is an improvement that nearly crosses the quoted spread.

	Number orders			Number orders sniped			Percentage sniped		
Spread improvement	Retail	Market maker	Other	Retail	Market maker	Other	Retail	Market maker	Other
0.00-0.10	1,204	9,738,427	4,446,682	0	25	30	0.00	0.00	0.00
0.10-0.20	2,212	2,221,371	992,735	13	24	53	0.59	0.00	0.01
0.20-0.30	2,584	866,450	194,078	46	34	39	1.78	0.00	0.02
0.30-0.40	2,994	479,829	85,841	94	37	42	3.14	0.01	0.05
0.40-0.50	2,876	308,894	43,735	199	65	73	6.92	0.02	0.17
0.50-0.60	3,347	302,304	34,781	341	112	106	10.19	0.04	0.30
0.60-0.70	3,243	251,414	19,976	663	136	167	20.44	0.05	0.84
0.70-0.80	3,123	163,264	8,724	985	91	242	31.54	0.06	2.77
0.80-0.90	2,750	84,662	6,957	1,263	62	315	45.93	0.07	4.53
0.90-1.00	1,582	13,617	3,826	990	11	223	62.58	0.08	5.83

Table 6: Number of spread-improving orders, number of sniped orders, and percentage of sniped orders based on level of spread improvement and dealing capacity. All spread-improving orders in all ETFs, January 2022.

Table 6 reports the number of spread-improving limit orders, and the number and percentage of orders sniped as a function of the level of spread improvement and of dealing capacity. Most spread-improving limit orders are sent by market makers and others. Specifically, 71% of the spread-improving orders are by market makers, 29% by others, and approximately 1% by retail. For retail, the distribution of spread-improving orders is relatively uniform, i.e., each level of spread improvement is equally likely. On the other hand, the level of most of the spread improvements by market makers and others is small, and the number of spread improvements decreases as the level of spread improvement increases.

The table shows that spread-improving limit orders sent by retail traders are sniped much more often than spread-improving limit orders sent by market makers and others. For example, limit orders that nearly cross the quoted spread (i.e., spread improvements between 0.9 and 1) are very likely to be sniped (62.58% probability) if they originate from a retail trader, and very unlikely to be sniped (0.08% probability) if they originate from a market maker. In these cases, a spread-improving limit order sent by a retail investor is 2,068 times more likely to be sniped than a spread-improving limit order sent by a market maker. A similar calculation shows that orders from retail traders are 27 times more likely to be sniped than those sent by others, while orders from others are 76 times more likely to be sniped than those sent by market makers.[18]

Figure 1 shows boxplots of the size of sniped and not sniped limit orders as a function of

[18] These values are the group-wise odds.

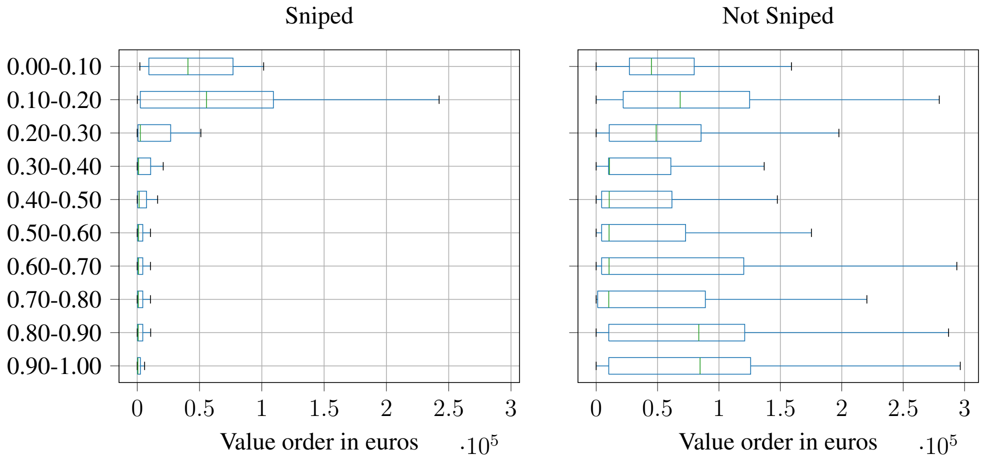


Figure 1: Size of sniped and not sniped limit orders in euros per level of spread improvement for all spread-improving orders in all ETFs in January 2022.

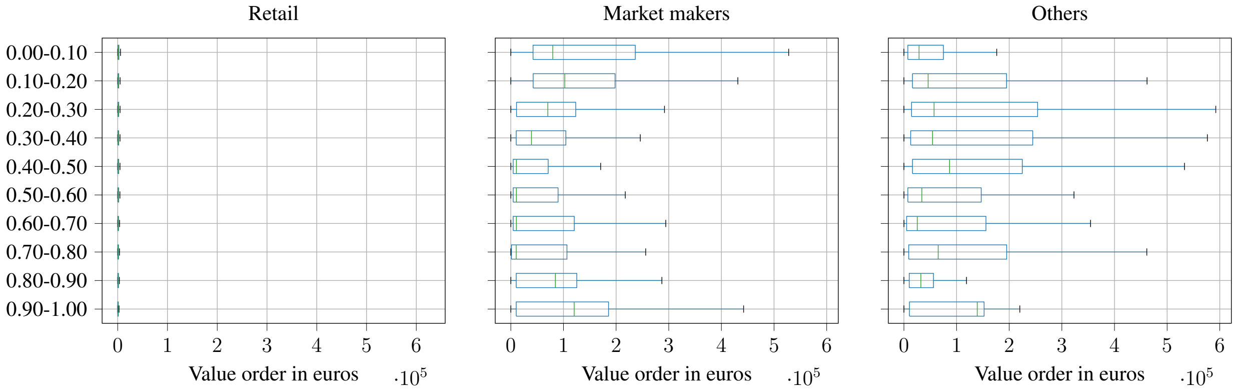


Figure 2: Size of spread-improving order in euros per level of spread improvement for all spread-improving orders in all ETFs in January 2022.

the level of spread improvement.[19] For limit orders with a level of spread improvement less than 0.2, the size of an order provides little information to determine if the limit order will be sniped. However, for limit orders with a level of spread improvement greater than 0.5, the size of sniped limit orders is smaller than those that are not sniped. Specifically, small orders are sniped more often than large orders with the same level of spread improvement. Figure 2 shows the distribution of spread-improving limit orders by dealing capacity. The figure shows that snipers consistently pick out retail limit orders by targeting small limit orders, and snipers consistently avoid trading with market makers and others because snipers avoid trading with large limit orders.

Next, we study who are the snipers in the ETF market. Only market makers or others can be

[19] The boxplots do not show datapoints that lie below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$, where $Q1$ is the first quartile (25th percentile), $Q3$ is the third quartile (75th percentile), and IQR is the interquartile range ($Q3 - Q1$).

	Sniper	
Sniped	Other	Market maker
Other	390	900
Retail	428	4,166
Market maker	333	264

Table 7: Number of spread-improving orders that are sniped by market makers and others for all ETFs in January 2022.

snipers because retail traders do not have the infrastructure or the technological capacity to compete for these spread-improving limit orders. Table 7 illustrates who are the snipers, and whose orders they mainly snipe. Market makers account for 82.24% of this sniping activity, while others account for the remaining 17.76% of this sniping activity. Market makers predominantly snipe retail traders and others, and less than 5% of their snipes are on another market maker's spread-improving limit order. On the other hand, snipers who are others are relatively uniform in whose orders they snipe.

	Reaction time in milliseconds
mean	0.141
std	0.112
min	0.024
25%	0.081
50%	0.091
75%	0.174
max	0.996

Table 8: Reaction time of market makers. Time between entry of a spread-improving order and the snipe by a market maker in all ETFs in January 2022.

Finally, Table 8 shows the reaction time of market makers when sniping a spread-improving order. The median reaction time of market makers is 91 microseconds. This speed advantage enables market makers to win the race to snipe retail limit orders with 90.68% probability.

5. Possible Explanations

In this section, we look at possible explanations for the behavior described in Section 4. We argue that while some explanations might justify part of the behavior we observe in the data, these potential explanations cannot fully explain the behavior we observe in Euronext's ETF market.

5.1. Public Signals and Inventory

Most market microstructure theories use the arrival of new information or a market maker's inventory position to explain the quoting behavior of market makers (see O'Hara, 1998). Information can be public or private.[20] If all the spread-improvement limit orders sent by market makers are the result of public information, then it is not surprising that snipers avoid trading with these spread-improving limit orders because these are loss-leading trades.

However, if all the spread-improving limit orders sent by market makers are the result of public information, then it does not explain the lack of latency arbitrage we observe in the data. Consider the framework of Budish et al. (2015), where risk-neutral market makers adjust their quotes in response to a public innovation in the fundamental value of the security. In this setting, any jump in fundamental value that exceeds the half-spread triggers a latency arbitrage race. Table 6 shows there were 815,261 instances where market makers (who are HFTs) improve the best quote beyond the midprice. Theoretically, if all these improvements were because of new public information, then all of them should trigger a latency arbitrage race. In the case of two HFTs, half of the 815,261 instances should result in a transaction because the HFT providing liquidity loses the race to cancel her stale quote. However, Table 9 shows there were only 4,677 transactions between market makers; this includes the 264 snipes from Table 7, latency arbitrage or adverse selection, and trades for liquidity needs (inventory). Thus, the assumption that all new orders sent by market makers that exceed the midpoint price are due to public information is inconsistent with the lack of transactions between market makers in the data.

			Aggressor		
		Retail	Market Maker	Other	
Passive	Retail	28,088	19,459	12,522	60,069
	Market Maker	133,708	4,677	33,111	171,496
	Other	37,452	13,683	57,533	108,668
		199,248	37,819	103,166	340,233

Table 9: Transactions between dealing capacities grouped by aggressive side (columns) and passive side (rows), all ETFs in January 2022.

Next, suppose a market maker is risk averse, then theory shows that market makers will skew their quotes if they have either a long or short inventory (see Biais, 1993).[21] If a spread-improving limit order crosses the fundamental value because of inventory constraints, and not because of su-

[20] Brogaard et al. (2019) analyze the contribution to price discovery of market and limit orders by HFTs and non-HFTs. They show that market orders have a larger individual price impact, but that limit orders are far more numerous. This results in price discovery occurring predominantly through limit orders, i.e., HFTs' limit orders contribute more than twice as much to price discovery than their market orders. In contrast, non-HFTs' market orders contribute more to price discovery than their limit orders. They use a vector autoregression (VAR) to systematically quantify the contribution to price discovery of market and limit orders by HFTs and non-HFTs, similar to Hasbrouck (1991a), Hasbrouck (1991b), and Hasbrouck (1995).

[21] In ETFs, inventory is less binding than shares if a market maker is an Authorized Participant (AP), because an AP can unwind inventory through the primary market. We cannot verify if the market makers we study are APs because those are agreements between the market maker and the ETF issuer.

			Aggressor		
		Retail	Market Maker	Other	
Passive	Retail	28,155	134,047	135,961	298,163
	Market Maker	123,692	1,917,114	1,995,043	4,035,849
	Other	113,477	2,203,337	1,329,187	3,646,001
		265,324	4,254,498	3,460,191	7,980,013

Table 10: Transactions between dealing capacities grouped by aggressive side (columns) and passive side (rows), all shares in January 2022.

perior information, then a sniper, who is less risk averse or is less constrained by inventory, should trade against the order because there is an arbitrage opportunity. Yet, the data show that snipers are unwilling to trade against the majority of spread-improving orders sent by market makers. Hence, not all spread improvements can be the result of inventory constraints.

Overall, there are few transactions between market makers in ETFs. Table 9 shows that 1.37% of all transactions are between market makers in ETFs, which corresponds to an average of 0.61 transactions between market makers per ETF, per day. On the other hand, Table 10 shows that 24.02% of all transactions are between market makers in ordinary shares, which corresponds to an average of 638.40 transactions between market makers per share, per day.

5.2. *Flickering Quotes*

Hasbrouck and Saar (2009) find that flickering quotes, i.e., limit orders entered inside the spread which are canceled quickly after entry, are the result of strategies that search for latent liquidity by executing against hidden orders or by quickly attracting a new marketable order (i.e., a snipe). In Euronext Amsterdam, the only use of flickering quotes would be to attract a snipe because there are no hidden orders in the exchange.[22] If a market maker wants to invite a snipe, then the dominant strategy is to blend among the retail limit orders so that her spread-improving limit order is executed.

5.3. *Queue Priority*

In tick-constrained instruments, queue priority is important (see Parlour, 1998). In these cases, there may be incentives for market makers to quote larger volumes to maintain queue priority.[23] Table 11 shows that the majority of ETFs are small-tick instruments where queue priority is less relevant because market makers can regain priority through minor price improvements. This is also reflected in Table 6 where the majority (67.49%) of spread improvements by market makers are minor improvements between 0 and 10 percent of the quoted spread.

[22] In Euronext Amsterdam, there are iceberg orders that are not hidden because part of the volume is visible.

[23] When market makers are HFTs, as in our case, theory still shows that HFTs do not provide excess liquidity (see Li et al., 2021).

Quoted spread in ticks	Number ETFs	Trade count %	Value traded %
2 to 5	21	24.89	19.79
6 to 10	51	20.02	23.68
11 to 20	76	12.30	15.90
21 to 50	59	22.33	16.03
51 to 100	59	12.40	11.25
101+	101	8.05	13.33

Table 11: Number of ETFs, trade count, and value traded grouped by median quoted spread in ticks on Euronext Amsterdam in January 2022.

5.4. *Quoting Requirements*

In 2022, market makers in ETFs were required to quote at least €100,000 within a spread of 2–6% around the midprice (depending on the ETF) for at least 80% of trading hours to qualify for the Market Maker Scheme. However, the volume requirement is a *cumulative* requirement, so there is no reason for market makers to quote in excess of the cumulative requirement on an order-to-order basis (see Figure 2). Furthermore, existing theory shows that, in equilibrium, market makers set quotes with convex price schedules where the marginal price is increasing per unit of the trade, i.e., limit orders are split across different price levels, because of the cost of adverse selection (see Glosten, 1994; Biais et al., 2000; Back and Baruch, 2013) and latency arbitrage (see Budish et al., 2015). Therefore, quoting requirements or existing theory cannot explain the quoting behavior of market makers.

6. Theory

In this section, we propose a model of the limit order book that considers competitive and collusive equilibria. Our model highlights the economic forces that underlie the incentives of market makers, i.e., HFTs, to reveal themselves to each other. Our model also answers the following questions: under what conditions would an HFT reveal herself to others? How do HFTs mutually benefit from revealing themselves, and how can they collectively enforce this behavior?

The model consists of impatient investors who take liquidity, patient investors who provide liquidity, and strategic HFTs who play a repeated trading game. The trading game is a generalization of the trading game in Budish et al. (2024), and it retains the main elements of latency arbitrage and adverse selection. In our model, one of the HFTs may receive short-lived private information in each trading game (see Foucault et al., 2016, for a model where HFTs trade on short-lived information), which differs from the usual high-frequency trading models with informed traders (see Baldauf and Mollner, 2020; Budish et al., 2024). The primary generalization in our model is the arrival of patient investors who send limit orders that improve the quoted spread.[24] This introduces

[24] Our patient investors are different from the slow trading firms in Budish et al. (2024) that have no intrinsic need

strategic ambiguity because market participants do not know *who* sent the spread-improving limit order. A limit order posted inside the spread could be benign flow from patient investors, or it could be toxic flow from a privately informed HFT who is pretending to be a patient investor.

6.1. Setup

Fundamental Value. Consider security x whose fundamental value is given by y, and at the end of each trading game, x can be liquidated at this fundamental value with zero cost, i.e., no frictions or fees. The value y evolves across trading games as a discrete-time jump process. In each trading game, if there is an innovation in y, then y jumps up or down with equal probability, and J is the distribution of the size of the jumps. The price grid is continuous to abstract from the queuing dynamics when non-zero tick sizes are a binding constraint.

Impatient Investor. In each trading game, an impatient investor arrives stochastically with an inelastic need to buy or to sell one unit of x. Impatient investors arrive with probability p_{LT}, and are equally likely to buy or to sell a unit of x. On arrival, impatient investors use a marketable limit order to transact immediately.[25] Impatient investors are the usual "liquidity traders" in Glosten and Milgrom (1985) or "noise traders" in Kyle (1985). These participants include mutual funds, pension funds, hedge funds, retail traders, etc. In the Euronext dataset, the majority of impatient investors are retail traders.

Patient Investor. In each trading game, a patient investor arrives stochastically with an elastic need to buy or to sell one unit of x. Patient investors arrive with probability p_{LO}, and are equally likely to post a buy or a sell limit order for one unit of x. Patient investors care only about buying or selling x at their reservation price, and they do so by submitting their limit orders inside the quoted spread. Patient investors are similar to the execution algorithms in Li et al. (2021). However, the key differences are that execution algorithms have an inelastic demand and are strategic, while patient investors have an elastic demand and are not strategic.[26] Patient investors, as seen in the dataset, are the retail traders who send limit orders that improve, but do not cross, the quoted spread. For analytical tractability, patient investors cancel their unexecuted limit orders at the end of the trading game.

If a patient investor arrives to buy one unit of the security, then he sends a buy limit order with a limit price of $y + \delta$, so he is content with buying at any price below or equal to $y + \delta$. Similarly,

to buy or to sell. Our patient investors, like impatient investors, have a need to buy or to sell one unit of the asset, but are unwilling to cross the half-spread.

[25] Marketable limit orders are limit orders with a bid price weakly greater than the best ask (if buying), or an ask price weakly lower than the best bid (if selling).

[26] Although patient investors are not strategic, Section 6.4.3 analyzes how the behavior of patient investors affects the equilibrium outcomes.

if a patient investor arrives to sell one unit of the security, then he sends a sell limit order with a limit price of $y - \delta$, so he is content with selling at any price above or equal to $y - \delta$. The value $\delta \in (0, s/2)$ is the improvement in the quoted spread by a patient investor. We focus on the case where δ is positive, so the reservation price is mispriced because the limit order improves the best bid or best offer beyond the fundamental value of the security. The improvement value δ is bounded above by the half-spread $s/2$ set by the HFTs, otherwise the limit order gets executed as a marketable limit order, in which case patient investors are subsumed as impatient investors.

HFTs. There are $N \geq 3$ HFTs who make and take liquidity, and who are present throughout all iterations of the trading game. HFTs are risk neutral and they have no intrinsic need to buy or to sell x. They buy x at prices lower than y and sell x at prices higher than y to maximize profits, and they discount future payoffs with the common discount factor $\rho \in [0, 1)$.

Public and Private Information. The probability that there is a jump in y that is public information and seen by all market participants at the same time is p_{public}, while the probability that there is a jump in y seen only by HFT i is p^i_{private}. HFTs are equally likely to receive private information, so $p^i_{\text{private}} = p_{\text{private}}/N$, where p_{private} is the probability that there is private information.

If an HFT observes private information, she can act on that information in the current trading game. Regardless of her actions, any private information becomes public at the end of the trading game. Given that ETFs are indices, it is unlikely that informed traders can acquire private information about all the constituents of an ETF. Therefore, the usual assumption that informed traders possess long-lived private information is unlikely to hold in an ETF market. Here, information is “news” as in Foucault et al. (2016), where every quote update or trade in any exchange is a source of information for the fundamental value y. Therefore, short-lived private information of HFTs is a consequence of market fragmentation. HFTs are present in multiple, but not all, exchanges and integrate information across markets (see Baron et al., 2019; Brogaard et al., 2019), but importantly, HFTs are not always connected to the same set of exchanges (we verified this with MiFID II data). If an HFT is not connected to a venue, then she can infer that there is private information from the behavior of an HFT connected to that venue, or wait until the information becomes public once she receives the data from a third-party data vendor who processes and disseminates the data.

Latency. HFTs operate with no latency. There is no delay in sending or receiving updates from the exchange. When multiple messages reach the exchange at the same time, there is a random tie-break to decide which message is processed first. In contrast, investors are slow, so when they race against HFTs to send a message to the exchange, the HFTs always win.

6.2. Timing of the Trading Game

Our trading game consists of four periods, and the four-period trading game is repeated infinitely many times. At the start of each trading game, there is a publicly observed state (y, ω), which consists of the fundamental value y, and of the outstanding bids and asks in the limit order book ω. In the first time the game is played, the initial fundamental value is y_0, and the order book is empty. In all subsequent instances of the trading game, the state is determined at the conclusion of the previous trading game, and ω consists of all outstanding limit orders.

The four-period trading game proceeds sequentially as follows:

1. **Period 1:** HFTs observe the state (y, ω) and send instructions to the exchange. These instructions are messages to submit limit orders, cancel existing limit orders, submit marketable limit orders, and submit immediate-or-cancel orders. Messages are processed by the exchange and ω is updated.

2. **Period 2:** Nature moves and selects one of two possibilities:

   (i) With probability p_{LO}, a patient investor, who is equally likely to buy or to sell a unit of x, arrives and sends a limit order inside the quoted spread at his reservation price.

   (ii) With probability $p_{\text{private}} = 1 - p_{LO}$, nature releases private information to one HFT. The informed HFT has an opportunity to send instructions to the exchange.

3. **Period 3:** If a limit order arrived in the second period, then the HFTs have an opportunity to snipe the limit order. Afterwards, HFTs also have an opportunity to send instructions to the exchange.

4. **Period 4:** Depending on nature's draws in period two, nature either does nothing or moves again. If nature selected an informed HFT in the second period, then nature does nothing. Otherwise, nature moves and selects one of three possibilities:

   (i) With probability p_{LT}, an impatient investor arrives and he is equally likely to buy or to sell a unit of x.

   (ii) With probability p_{public}, there is a publicly observable jump in y. HFTs participate in latency arbitrage as in Budish et al. (2015).

   (iii) With probability $p_N = 1 - p_{LT} - p_{\text{public}}$, there is no event.

   Finally, at the end of each trading game, any outstanding limit order from a patient investor is canceled.

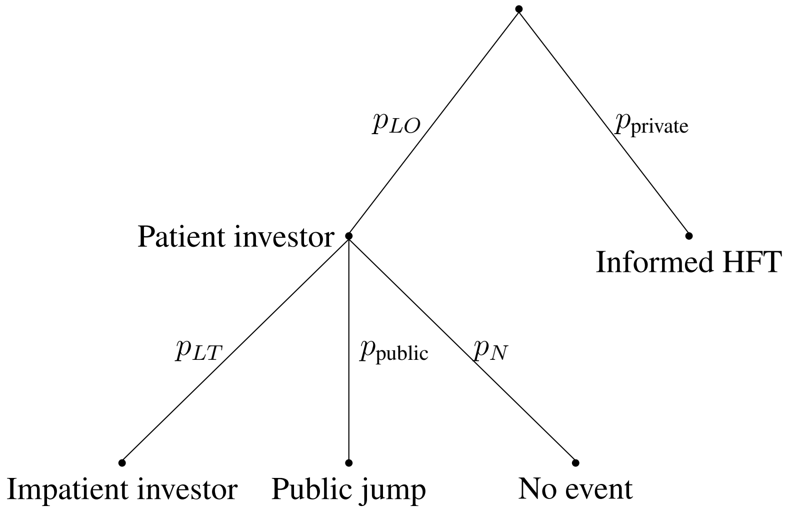


Figure 3: Nature's actions in each trading game.

Figure 3 illustrates nature's actions and their associated probabilities. Within a single trading game, nature can call upon both a patient investor and an impatient investor. This captures the incentives of HFTs to snipe limit orders from patient investors who improve the quoted spread. If HFTs do not snipe limit orders from patient investors, then these limit orders preclude HFTs from earning the half-spread on one side of the book when an impatient investor arrives in period four.

6.3. Competitive Equilibrium

For the competitive equilibrium in our infinitely repeated trading game, we restrict our attention to pure strategies where market participants condition their actions on the state (y, ω) and update their beliefs based on Bayes' rule whenever possible. The relevant beliefs are about *who* arrives in period two. We analyze each trading game in isolation, and then show that repeated play of the equilibrium for a single trading game remains an equilibrium for the infinitely repeated trading game.

6.3.1. Solution Concept

With the restriction to pure strategies, one solution concept would be a pure strategy weak perfect Bayesian equilibrium (WPBE). However, a pure strategy WPBE does not exist. As in Budish et al. (2024), excess liquidity (i.e., liquidity that supplies a quantity greater than the one unit required by the impatient investor) exposes the liquidity provider to adverse selection and latency arbitrage, but without the benefits of liquidity provision. Thus, there are no other liquidity providers willing to provide excess liquidity to constrain the spread, so the liquidity provider has a profitable deviation by widening the spread.

To restore the existence of equilibrium, we adapt the order book equilibrium from Budish et al.

(2024) to our setting. As in Budish et al. (2024), the solution concept strictly weakens a WPBE by allowing for profitable unilateral deviations to exist, provided that these deviations are rendered unprofitable by one of two specific reactions by rivals: withdrawal of liquidity (canceling limit orders), or safe profitable price improvements (of limit orders).

The solution concept captures the spirit of competitive liquidity provision as discussed and used in Glosten and Milgrom (1985). The solution concept ensures that even if excess liquidity is not provided in equilibrium, the presence of other potential liquidity providers will discipline equilibrium price levels.[27] For a more detailed discussion, see Budish et al. (2024).

6.3.2. Equilibrium Behavior of HFTs

In the first period, the weakly dominant strategy is for one of the HFTs to maintain two-sided quotes. One quote to buy one unit of x at price $y - s_a/2 - \varepsilon/2$, and the other quote to sell one unit of x at price $y + s_a/2 + \varepsilon/2$, where $s_a/2 \geq 0$ is the HFT's half-spread in period three that is solved through an equilibrium indifference condition. The small and positive constant $\varepsilon \to 0$ ensures that there are incentives for someone to undercut the quotes in period three, so the HFT can withdraw liquidity in period three (without raising suspicion) in the event she becomes informed in period two. Individually, each HFT strictly prefers to become an informed trader upon receiving private information in the second period, so none of the HFTs set quotes with $\varepsilon = 0$ in period one.

Informed HFT. In the second period, if an HFT receives private information, then the weakly dominant strategy is to pretend to be a patient investor (when profitable) by submitting a toxic limit order for one unit of x in the same direction as the jump. Specifically, if $y' > y + \delta$, then the informed HFT sends a buy limit order with a price of $y + \delta$; similarly, if $y' < y - \delta$, then the informed HFT sends a sell limit order with a price of $y - \delta$. On the other hand, if the jump size does not exceed the improvement level δ, then the informed HFT does not submit a limit order because it is not profitable to do so.[28]

In the fourth period, if the value y' is not profitable to trade (i.e., $|y' - y| \leq s_a/2$), then the informed HFT does nothing. However, if the value y' is profitable to trade (i.e., $|y' - y| > s_a/2$), then the informed HFT adversely selects the market maker and earns $|y' - y| - s_a/2$.[29]

[27] Baldauf and Mollner (2020) refer to these potential liquidity providers as enforcers.

[28] In equilibrium, the uninformed HFTs know that if the limit order originates from an informed HFT, then the jump size exceeds the improvement level δ. However, for the uninformed HFT who provides liquidity in period three, the cost of adverse selection remains the same if the informed HFTs send or do not send toxic orders. Therefore, the weakly dominant strategy of the informed HFT is always to send a toxic limit order whenever it is profitable to do so.

[29] Although the informed HFT receives private information in the second period, the dominant strategy in equilibrium is to wait until the fourth period to adversely select the market maker. Adversely selecting the market maker in the second period will make other HFTs aware of an informed HFT, which will inhibit the informed HFT's ability to profit from providing liquidity with a toxic limit order by pretending to be a patient investor. Additionally, as the quotes will tighten by ε at the end of period three, it is more profitable to adversely select the market maker in the

Uninformed HFTs. In the second period, if HFT i does not receive private information, then Nature either called upon a patient investor with probability p_{LO}, or released private information to another HFT with probability $p_{\text{private}}^{-i} = \sum_{j \neq i} p_{\text{private}}^{j} = \frac{N-1}{N} p_{\text{private}}$. There are two cases to consider as an uninformed HFT in period two: (i) no limit order arrives, or (ii) a limit order arrives.

In the first case when no limit order arrives with probability $1-\alpha$, it means that Nature released private information to another HFT, but the jump size did not exceed the improvement level δ. If the jump size does not exceed the improvement level δ, then the jump size cannot exceed $s_a/2$. In this case, all HFTs do nothing and earn nothing.

In the second case when a limit order arrives with probability $\alpha = p_{LO} + p_{\text{private}}^{-i} \, \mathbb{P}(J > \delta)$, it means that either a patient investor submitted a benign limit order, or an informed HFT submitted a toxic limit order. Applying Bayes' rule to this information set, where a limit order arrives, the probability of the limit order originating from a patient investor is p_{LO}/α, while the probability of the limit order originating from an informed HFT is $p_{\text{private}}^{-i} \, \mathbb{P}(J > \delta) / \alpha$. Furthermore, as informed HFTs will submit a toxic limit order in the same direction as the jump, uninformed HFTs learn the direction of the private jump conditional on the limit order originating from an informed HFT.

If a limit order arrives in the second period, then, in the third period, the uninformed HFTs have an opportunity to snipe the limit order, or they can wait until the fourth period when there is a public innovation in y to ensure that the limit order is from a patient investor so that it is safe and profitable to trade against. At the end of period three, the HFT who provided liquidity in period one cancels her quotes, and the uninformed HFTs endogenously sort themselves into one of two roles on each side of the book. On each side of the book, one uninformed HFT (who may or may not be the same HFT providing liquidity in period one) takes the role of a market maker and the remaining uninformed HFTs take the role of a latency arbitrageur.

First, consider the case where at least one uninformed HFT chooses to snipe in the third period whenever a limit order arrives in the second period. If the limit order is from a patient investor, then the uninformed HFTs (who choose to snipe) have an equal probability of executing against the mispriced limit order to earn an expected profit of δ. However, if the limit order is from an informed HFT, then the informed HFT earns a profit of $|y' - y| - \delta$, while the uninformed HFT who sniped the informed HFT's toxic limit order will incur a loss of $|y' - y| - \delta$.

With the limit order from period two cleared out, the trading game becomes the two-period trading game in Budish et al. (2024), but with event probabilities given by the updated beliefs. Uninformed HFTs set asymmetric quotes around y because only one side of the book faces the risk of adverse selection. We use $s_a/2$ to denote the half-spread for the side of the book that faces the risk of adverse selection, and $s_{na}/2$ to denote the half-spread for the side of the book that does

fourth period.

not face the risk of adverse selection; clearly, in equilibrium $s_a > s_{na}$. Therefore, if a buy limit order arrived inside the quoted spread in period two, then at the end of period three, after the limit order is sniped, one uninformed HFT provides one unit of liquidity to sell at $y + s_a/2$, and the same (or a different) uninformed HFT provides one unit of liquidity to buy at $y - s_{na}/2$. Similarly, if a sell limit order arrived inside the quoted spread in period two, then at the end of period three, after the limit order is sniped, one uninformed HFT provides one unit of liquidity to buy at $y - s_a/2$, and the same (or a different) uninformed HFT provides one unit of liquidity to sell at $y + s_{na}/2$.

In period four, if the public jump to the value y' is not profitable to trade, then nothing happens. However, if the value y' is profitable to trade, then a latency arbitrage race occurs. In the latency arbitrage race, the market maker races to cancel her unprofitable stale quote. The market maker will successfully cancel her quote with probability $1/N$ and lose nothing, and she will unsuccessfully cancel her quote with probability $(N-1)/N$ and lose $|y' - y| - s_a/2$ or $|y' - y| - s_{na}/2$. Simultaneously, the latency arbitrageur races to pick up the market maker's stale quote. The latency arbitrageur will win the race with probability $1/N$ and earn $|y' - y| - s_a/2$ or $|y' - y| - s_{na}/2$, and she will lose the race with probability $(N-1)/N$ and earn nothing. The losses and gains depends on the direction of the public jump, which occur with equal probability.

Next, consider the case when all HFTs wait until there is a public innovation in y in the fourth period to decide whether to trade against the limit order posted in period two. In this case, because the limit order from period two is not cleared out in period three, the uninformed HFT will only provide liquidity on one side of the book. Providing liquidity on the same side as the limit order from period two means that her order is exposed to the risk of latency arbitrage without the benefits of liquidity provision. Therefore, if a buy limit order arrives in period two and remains unexecuted by the end of period three, then the market maker will provide a quote to sell one unit of x at price $y + s_a/2$; similarly, if a sell limit order arrives in period two and remains unexecuted by the end of period three, then the market maker will provide a quote to buy one unit of x at price $y - s_a/2$.

In period four, if the public jump to the value y' is such that the market maker's quote is stale, then a latency arbitrage race occurs between the HFTs. On the other hand, if the value y' is such that the limit order from period two is stale, then all the HFTs attempt to latency arbitrage the patient investor. All HFTs have an equal probability of trading against the limit order from the patient investor to earn $|y' - y| + \delta$. Finally, if the value y' is such that no quotes are stale, then nothing happens and the limit order from the patient investor is canceled.

In what follows, we characterize the HFT's bid-ask spread in period three, and provide the conditions to determine whether an HFT snipes the limit order in period three or waits until there is an innovation in y in period four.

6.3.3. Equilibrium Analysis

To analyze the equilibrium, we first solve the equilibrium half-spreads set by uninformed HFTs at the end of period three, and then we solve whether uninformed HFTs snipe in the third period or not in equilibrium. First, we introduce some notation. Let

$$L(x) = \mathbb{P}\left(J > x\right) \mathbb{E}\left[J - x \mid J > x\right] \quad \text{ and } \quad \tilde{L}(x) = \mathbb{P}\left(J < x\right) \mathbb{E}\left[x - J \mid J < x\right]$$

denote the expected payoffs of arbitrages associated with jumps in the fundamental value. Furthermore, let

$$L_\delta(x) = \mathbb{P}\left(J_\delta > x\right) \mathbb{E}\left[J_\delta - x \mid J_\delta > x\right]$$

denote the expected payoffs of arbitrages associated with jumps in the fundamental value that are greater than δ. The expected payoff $L_\delta(x)$ accounts for Bayesian updating by uninformed HFTs, where J_δ is the distribution of the size of the jump J given that the size of the jump is greater than δ, i.e., $J|J > \delta$. Finally, let d^i denote the binary decision variable of HFT i sniping in the third period. If HFT i snipes in the third period, then $d^i = 1$, otherwise $d^i = 0$.

In equilibrium, the half-spread set by uninformed HFTs at the end of period three leaves them indifferent between the role of a market maker or that of a latency arbitrageur on either side of the book given their updated beliefs.

Uninformed HFTs do not snipe in period three. HFTs have fewer opportunities to provide liquidity (HFTs only provide liquidity on one side of the book). Therefore, latency arbitrageurs have fewer opportunities to latency arbitrage another HFT. However, because the uninformed HFTs do not snipe in period three, then, conditional on a public innovation of y in the appropriate direction, all the uninformed HFTs (including the market maker) have the opportunity to latency arbitrage the outstanding limit order from the patient investor in period four for an expected profit of $(L(-\delta) + \tilde{L}(\delta))/2$.[30]

At least one uninformed HFT snipes in period three. HFTs provide liquidity on both sides of the book, which allows latency arbitrageurs to latency arbitrage another HFT on either side of the book. If an uninformed HFT decides to snipe in period three, then she takes a share of the limit orders from patient investors at an expected payoff of δ, but simultaneously, she is also exposed to

[30] To see this, consider a spread-improving limit order on the bid resting at $y + \delta$. There is an arbitrage opportunity if the innovation is anywhere below $y + \delta$. The term $L(-\delta)/2$ corresponds to the expected payoff when the new fundamental value y' is below y. The term $\tilde{L}(\delta)/2$ corresponds to the expected payoff when the new fundamental value y' is between y and $y + \delta$.

trading with toxic limit orders from informed HFTs, who pretend to be a patient investor, at a loss of $L(\delta)$.

Therefore, conditional on HFT i being uninformed and observing a limit order arrive in period two, her expected payoff, in period three, when acting as a market maker is

$$\begin{aligned}
\frac{1}{\alpha}\Bigg[&\underbrace{\mathbb{1}_{\{\sum_{j=1}^{N} d^j=0\}}\, p_{LO}\,\frac{p_{\text{public}}}{2\,N}\,\Big(L(-\delta)+\tilde{L}(\delta)\Big)}_{\text{latency arbitraging patient investors in period four}} + d^i\underbrace{\left(\frac{p_{LO}}{\sum_j d^j}\,\delta-\sum_{j\neq i}\frac{p^j_{\text{private}}}{\sum_{k\neq j} d^k}\,L(\delta)\right)}_{\text{sniping in period three}}\\
&+\underbrace{\frac{1}{2}\,p_{LO}\,\left(p_{LT}\,\frac{s_a}{2}-p_{\text{public}}\,\frac{N-1}{N}\,L\left(s_a/2\right)\right)}_{\text{liquidity provision less latency arbitrage costs in period four}}-\underbrace{\sum_{j\neq i} p^j_{\text{private}}\,\mathbb{P}\left(J>\delta\right)\,L_\delta\left(s_a/2\right)}_{\text{adverse selection costs in period four}}\\
&+\underbrace{\mathbb{1}_{\{\sum_{j=1}^{N} d^j>0\}}\,\frac{1}{2}\,p_{LO}\,\left(p_{LT}\,\frac{s_{na}}{2}-p_{\text{public}}\,\frac{N-1}{N}\,L\left(s_{na}/2\right)\right)}_{\text{liquidity provision less latency arbitrage costs in period four}}\Bigg].
\end{aligned}\tag{1}$$

All the terms are pre-multiplied by $1/\alpha$ to account for the updated beliefs about who arrived in period two. The second term (sniping in period three) includes a double sum because when the informed HFT pretends to be a patient investor, the cost of adversely selecting yourself is shared equally among all uninformed HFTs who choose to snipe in period three. This cost excludes the informed HFT because she does not snipe her own order. The fourth term (adverse selection cost) accounts for the updated information about the jump size, because when informed HFTs send toxic orders, it means that the size of the jump exceeds the improvement level δ. This term can be rewritten as $\sum_{j\neq i} p^j_{\text{private}}\, L\left(s_a/2\right)$ because $\mathbb{P}\left(J>\delta\right)\, L_\delta\left(s_a/2\right) = L\left(s_a/2\right)$, which demonstrates that the cost of adverse selection remains the same if the informed HFTs send or do not send toxic orders. Finally, the last term is zero if uninformed HFTs do not snipe in period three because the limit order from period two is not cleared out, in which case, HFTs cannot benefit from providing liquidity.

Next, conditional on HFT i being uninformed and observing a limit order arrive in period two, her expected payoff, in period three, when acting as a latency arbitrageur is

$$\begin{aligned}
\frac{1}{\alpha}\Bigg[&\underbrace{\mathbb{1}_{\{\sum_{j=1}^{N} d^j=0\}}\, p_{LO}\,\frac{p_{\text{public}}}{2\,N}\,\Big(L(-\delta)+\tilde{L}(\delta)\Big)}_{\text{latency arbitraging patient investors in period four}} + d^i\underbrace{\left(\frac{p_{LO}}{\sum_j d^j}\,\delta-\sum_{j\neq i}\frac{p^j_{\text{private}}}{\sum_{k\neq j} d^k}\,L(\delta)\right)}_{\text{sniping in period three}}\\
&+\underbrace{\frac{1}{2}\,p_{LO}\,p_{\text{public}}\,\frac{1}{N}\,L\left(s_a/2\right)}_{\text{latency arbitrage gains in period four}}+\underbrace{\mathbb{1}_{\{\sum_{j=1}^{N} d^j>0\}}\,\frac{1}{2}\,p_{LO}\,p_{\text{public}}\,\frac{1}{N}\,L\left(s_{na}/2\right)}_{\text{latency arbitrage gains in period four}}\Bigg],
\end{aligned}\tag{2}$$

where $1/\alpha$ accounts for the updated beliefs.

When (1) and (2) are equal, uninformed HFTs are indifferent between taking the role of a latency arbitrageur or that of a market maker. Thus, equating (1) and (2) and solving for the equilibrium quotes on each side of the book separately leads to the following two equilibrium indifference conditions

$$\frac{1}{2}\, p_{LO}\, p_{LT}\, s_a/2 = \frac{1}{2}\, p_{LO}\, p_{\text{public}}\, L\left(s_a/2\right) + \frac{N-1}{N}\, p_{\text{private}}\, L\left(s_a/2\right)\ , \tag{3a}$$

$$\frac{1}{2}\, p_{LO}\, p_{LT}\, s_{na}/2 = \frac{1}{2}\, p_{LO}\, p_{\text{public}}\, L\left(s_{na}/2\right)\ . \tag{3b}$$

For the side of the book that faces the risk of adverse selection, (3a) uniquely pins down the equilibrium half-spread $s_a^*/2$ set by uninformed HFTs. For the side of the book that does not face the risk of adverse selection, if the limit order from period two is cleared out, then (3b) uniquely pins down the equilibrium half-spread $s_{na}^*/2$ set by uninformed HFTs.

Both s_a^* and s_{na}^* are positive and unique because in both equations the left-hand sides are strictly increasing in s and equal to zero when $s = 0$, while the right-hand sides are strictly decreasing in s and are positive for $s = 0$. The difference between these indifference conditions is the cost of adverse selection. To compensate for this difference, market makers require a wider half-spread, i.e., $s_a^* \geq s_{na}^*$, with equality when $p_{\text{private}} = 0$.

Next, we show when do uninformed HFTs snipe in the third period. This requires an equilibrium profile $\mathbf{d} = (d^1, ..., d^N)$ such that there are no profitable deviations in d^i. We restrict our attention to symmetric equilibria when either $\mathbf{d} = \mathbf{1} = (1, ..., 1)$, all HFTs snipe in period three, or $\mathbf{d} = \mathbf{0} = (0, ..., 0)$, HFTs do not snipe in period three. The following lemma provides the existence conditions for each equilibrium.

Lemma 1 *Let s_a^* and s_{na}^* be the unique solutions to (3a) and (3b), respectively. If*

$$p_{LO}\, \frac{p_{public}}{2\,N}\, \left(L(-\delta) + \tilde{L}(\delta)\right) > p_{LO}\, \frac{p_{public}}{2\,N}\, L\left(s_{na}^*/2\right) + p_{LO}\, \delta - \frac{N-1}{N}\, p_{private}\, L(\delta) \tag{D0}$$

holds, then HFTs do not snipe in period three, so $\mathbf{d} = \mathbf{0}$. On the other hand, if

$$p_{LO}\, \delta/N > p_{private}\, L(\delta)/N \tag{D1}$$

holds, then HFTs snipe in period three, so $\mathbf{d} = \mathbf{1}$.

To show the result, consider a deviation from the symmetric profile, and write the inequalities to ensure that the deviation is not profitable. If the expected cost of being fooled by informed HFTs is too high, then condition (D0) holds and in equilibrium it is optimal for all HFTs not to snipe in period three. On the other hand, when the expected profit from sniping patient investors outweighs the expected loss from being fooled by informed HFTs, then condition (D1) holds and in equilibrium it is optimal for all HFTs to snipe in period three.

The following proposition summarizes the equilibrium behavior.

Proposition 1 *Let s_a^* and s_{na}^* be the unique solutions to (3a) and (3b), respectively. If the no-snipe condition (D0) or the snipe condition (D1) holds, then there exists a symmetric equilibrium (with respect to $\mathbf{d}$) where the following occurs in every iteration of the trading game given state (y, ω):*

- *At the end of period 1, there is one unit of liquidity provided at $y - s_a^*/2 - \varepsilon/2$ and one unit of liquidity provided at $y + s_a^*/2 + \varepsilon/2$.*

- *In period 2: a patient investor, upon arrival, submits a limit order for a unit of security x at his reservation price; or an informed HFT, when profitable, submits a limit order at the patient investor's reservation price in the direction of her privately observed jump.*

- *In period 3:*

  - *If (D0) holds, then uninformed HFTs do not snipe the limit order from period 2. If a buy limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to sell at $y + s_a^*/2$. If a sell limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to buy at $y - s_a^*/2$. The quoted spread following period 3 is $s_a^*/2 - \delta$.*
  - *If (D1) holds, then at the beginning of period 3, uninformed HFTs race to snipe the limit order from period 2. If a buy limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to buy at $y - s_{na}^*/2$ and an uninformed HFT provides one unit of liquidity to sell at $y + s_a^*/2$. If a sell limit order arrived in period 2, then at the end of period 3, an uninformed HFT provides one unit of liquidity to buy at $y - s_a^*/2$ and an uninformed HFT provides one unit of liquidity to sell at $y + s_{na}^*/2$. The quoted spread following period 3 is $(s_a^* + s_{na}^*)/2$.*

- *In period 4: an impatient investor, upon arrival, purchases or sells one unit of x at the best bid or the best offer; an informed HFT, when profitable, purchases or sells one unit of x at the best bid or the best offer; HFTs race to latency arbitrage stale quotes when a publicly observed jump is profitable, if the stale quote belongs to the market maker, then she attempts to cancel her stale quote.*

Repeated play of the equilibrium for a single trading game is an equilibrium for the infinitely repeated trading game. This follows because there is no queuing motive to rest orders that carry over to subsequent games, and because information resets at the end of each trading game.

There are two important behavioral predictions from Proposition 1. First, when profitable, informed HFTs pretend to be a patient investor. This strategic behavior is reminiscent of Kyle (1985) where informed traders camouflage themselves among noise traders. This is unsurprising because strategic players use the anonymity of limit orders to their advantage. Second, when providing liquidity, HFTs do not provide excess liquidity, which is a common prediction across various models (see e.g., Li et al., 2021; Budish et al., 2024). Interestingly, in Euronext Amsterdam the empirical behavior of the HFTs in the ETF market is inconsistent with these basic predictions from the competitive equilibrium.

6.4. Collusive Equilibrium

The collusive strategy we consider is a reversion strategy, one in which observed deviations or suspected deviations from cooperation lead to a reversion to competitive play that lasts for T iterations of the trading game. The collusive strategy of interest recovers the key features we observe from the behavior of HFTs in the data. In our collusive strategy, we do not consider the possibility of cooperating to widen the spread beyond s_a^*.[31] In the collusive equilibrium, the primary gains from cooperation depend on the competitive equilibrium used in the reversion phase of the strategy.

The interesting case is when uninformed HFTs do not snipe in the competitive equilibrium (i.e., condition (D0) holds) because the costs of sniping toxic limit orders from privately informed HFTs is too high. In this case, signaling to avoid trading with each other enables the HFTs to share the benign flow of patient investors, and simultaneously, limits competition from patient investors. The byproduct of limiting competition is that it enables the HFTs to receive additional benign flow from impatient investors who would have otherwise matched with a patient investor's limit order at better prices than those quoted by HFTs.

On the other hand, when uninformed HFTs snipe in the competitive equilibrium (i.e., condition (D1) holds), the primary gains from cooperation arise from a wider spread relative to the competitive equilibrium (see Dutta and Madhavan, 1997). In this case, there are no additional gains from sharing the benign flow of patient investors and from limiting competition from patient investors because the competitive equilibrium is one where limit orders from period two are always cleared out.

6.4.1. Equilibrium Behavior of HFTs

There are two phases in the collusive strategy: a cooperation phase and a punishment phase. HFTs cooperate until they observe a deviation or suspect a deviation from cooperation, both of

[31] In practice, the spread is constrained by other exchanges. Table 3 shows that the spread in Euronext is similar to the spread in other European exchanges. If HFTs cooperate to widen the spread, then brokers may divert benign flow away to other exchanges with tighter spreads.

which trigger a punishment phase that lasts for T iterations of the trading game, after which, play returns to the cooperation phase. The collusive strategy is as follows.

In the cooperation phase,

- HFTs signal all their limit orders with a quantity $Q \gg 1$ so that their limit orders are easily differentiated from a patient investor's limit order which are for one unit of the security. HFTs take turns to provide liquidity. Each HFT i takes the role of the market maker for a proportion $1/N$ of the iterations played. HFTs do not latency arbitrage or adversely select market makers, and a market maker is compliant if she respects the agreement to take turns to provide liquidity at $y - s_a^*/2$ and at $y + s_a^*/2$, where s_a^* is the unique solution to (3a).

- HFTs do not pretend to be a patient investor, and in the third period HFTs race to snipe limit orders that arrived in the second period without a signal, i.e., they race to snipe patient investors.

In the punishment phase, HFTs revert to play from the competitive equilibrium, and play according to the behavior described in Section 6.3.2.

Coordinating Play. In the collusive equilibrium, there is a queuing motive for resting orders that carry over to subsequent iterations of the game, because when cooperating, HFTs forgo latency arbitraging and adversely selecting market makers, and instead share the role of the market maker. Therefore, HFTs need to coordinate their proportion of time as the market maker. This coordination uses public and private innovations in y as a correlation device.

At the start of the infinitely repeated trading game, HFTs race to provide quotes around y_0, and the probability of each HFT winning the race is $1/N$. The winner of the race remains as the incumbent market maker until the next innovation in y.[32] Whenever there is an innovation in y (public or private), HFTs race to become the new market maker, and the incumbent market maker providing stale quotes complies by canceling her stale quotes. Losers of the race cancel their limit orders because the volume Q of the limit order is too large for losers of the race to receive flow from impatient investors. When the jump in y is public information, HFTs have an equal probability of winning the race to become the new market maker, and when the jump in y is private information, the privately informed HFT becomes the new market maker.

Consider a publicly observed jump from y to y'. If the jump is such that $|y' - y| < s_a^*$, then HFTs race to provide new quotes around y', and the incumbent market maker providing stale quotes around y complies by canceling her outstanding orders. On the other hand, if the jump is

[32] This is different from the competitive equilibrium where the HFTs providing liquidity in period 1 and period 3 are not necessarily the same. Here, the same HFT continues to provide liquidity until there there is an innovation in y.

such that $|y' - y| \geq s_a^*$, then the incumbent market maker complies by canceling her stale quotes around y, and HFTs race to provide new quotes around y'. For large public innovations in y, the incumbent market maker is provided with an opportunity to first cancel her stale quotes around y so that new quotes from the race around y' do not trade with the stale quotes of the incumbent market maker. This ensures that HFTs do not latency arbitrage market makers.

Next, consider a privately observed jump from y to y'. The informed HFT always wins the race because only she knows the value y' until the end of the trading game. If the jump is such that $|y' - y| < s_a^*$, then the informed HFT provides new quotes around y', and the incumbent market maker providing stale quotes around y complies by canceling her outstanding orders. On the other hand, if the jump is such that $|y' - y| \geq s_a^*$, then the informed HFT submits a signaled limit order at the patient investor's reservation price according to the direction of her privately observed jump. The incumbent market maker providing stale quotes around y complies by canceling her outstanding orders, and the informed HFT cancels her signaled limit order and provides new quotes around y'. For large private innovations in y, the signaled limit order at the patient investor's reservation price serves as a warning to the incumbent market maker that there is a privately informed HFT. This provides the incumbent market maker with an opportunity to comply and cancel her stale quotes so that HFTs do not adversely select market makers.

By coordinating the role of the market maker through innovations in y, each HFT i becomes the incumbent market maker for a proportion $1/N$ of the iterations played.

Monitoring. To sustain cooperation, HFTs must be able to monitor deviations from cooperative play so that profitable deviations from cooperation can be deterred with the threat of punishment, i.e., reversion to competitive play. If HFTs cannot reliably monitor deviations, then the threat of punishment is not credible.

In the collusive strategy, there are three punishment triggers:

1. a market maker is latency arbitraged or adversely selected,
2. a market maker is not compliant, and
3. HFTs snipe a non-signaled limit order from period two and lose money.

In the model, monitoring is perfect for the first two triggers. If a market maker trades a quantity greater than one unit, then it is easy to infer that a market maker was latency arbitraged or adversely selected by another HFT because patient investors only demand one unit of x. On the other hand, if a market maker trades a quantity of one unit of x, then HFTs cannot infer who (HFT or impatient investor) traded with the market maker from the quantity alone. However, in equilibrium, HFTs can perfectly infer that a market maker was latency arbitraged or adversely selected by another HFT because these trades are associated with an innovation in y, while trades initiated by an impatient investor are not associated with an innovation in y, see Figure 3. Similarly, if a market maker is not compliant because they do not respect the agreement to take turns to provide liquidity, or because they undercut the agreed upon spread s_a^*, then the non-compliance is immediately observed by all HFTs.

On the other hand, monitoring is imperfect for the third trigger. There are two possibilities that lead to an HFT losing money after sniping a non-signaled limit order that was posted period two. First, the limit order is from a patient investor and the public innovation in y went in the wrong direction (the sniper was unlucky), or there was an informed HFT pretending to be a patient investor. When private information becomes public at the end of the trading game, this looks like public information from the perspective of uninformed HFTs, so uninformed HFTs cannot differentiate if the jump is public or it was a private jump that became public.[33] Hence, the problem is one related to "secret price cutting" (see Green and Porter, 1984), because HFTs cannot perfectly infer whether the trigger was a result of bad luck or because there was a cheater.

In what follows, we analyze and characterize conditions for the collusive strategy to be an equilibrium of the infinitely repeated trading game.

6.4.2. *Equilibrium Analysis*

In the collusive strategy, if play is in the punishment phase, then the expected payoff of HFT i per trading game depends on whether condition (D0) or (D1) holds. If condition (D0) holds so that HFTs do not snipe in period three, then the expected payoff of HFT i in the competitive equilibrium, per trading game, is

$$\Pi^{D0}_{\text{comp}} = \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\left(s_a^*/2\right) + \frac{1}{N}\, p_{\text{private}}\, L\left(s_a^*/2\right) + p_{LO}\, \frac{p_{\text{public}}}{2\,N}\, \Big(L(-\delta) + \tilde{L}(\delta)\Big)\ ,$$

where s_a^* is the unique solution to (3a), and all trades occur in period four. The first term on the right-hand side is the expected payoff from latency arbitraging the market maker. The second term is the expected payoff from adversely selecting the market maker, and the final term is the expected payoff from latency arbitraging limit orders posted by patient investors.

Similarly, if condition (D1) holds so that HFTs snipe in period three, then the expected payoff of HFT i in the competitive equilibrium, per trading game, is

$$\begin{aligned}\Pi^{D1}_{\text{comp}} = {} & \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\left(s_a^*/2\right) + \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\left(s_{na}^*/2\right) + \frac{1}{N}\, p_{\text{private}}\, L\left(s_a^*/2\right) \\ & + p_{LO}\, \frac{1}{N}\, \delta + \frac{1}{N}\, p_{\text{private}}\, L(\delta) - \frac{1}{N}\, p_{\text{private}}\, L(\delta)\ ,\end{aligned}$$

[33] If one assumes HFTs can determine if a jump was public or private, then the problem becomes one with perfect monitoring.

where s_a^* and s_{na}^* are the unique solutions to (3a) and (3b), respectively, and trades occur in both period three and period four. The first two terms on the right-hand side are the expected payoffs from latency arbitraging the market maker in period four. The third term is the expected payoff from adversely selecting the market maker in period four. The fourth term is the expected payoff from sniping limit orders posted by patient investors in period three. The last two terms are the expected profits and losses from sniping toxic limit orders in period three.

On the other hand, if play is in the cooperative phase, then the expected payoff of HFT i when cooperating, per trading game, is

$$\Pi_{\text{coop}} = \frac{1}{N} p_{LO}\, p_{LT}\, s_a^*/2 + p_{LO} \frac{1}{N} \delta \,,$$

which consists of the expected payoff from sharing the role of the market maker plus the expected payoff from sniping limit orders in period three that were posted by patient investors in period two. A necessary condition for collusion is that the difference between the expected payoff from cooperation is greater than that from competitive play. This difference is given by

$$c = \begin{cases} \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D0} & \text{if no-snipe condition (D0) holds}\,, \\ \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D1} & \text{if snipe condition (D1) holds}\,. \end{cases}$$

If condition (D0) holds so that HFTs do not snipe in period three in the competitive equilibrium, then the gains from cooperation are given by

$$c = \underbrace{\frac{p_{LO}}{N} \delta - p_{LO} \frac{p_{\text{public}}}{2\,N} \left(L(-\delta) + \tilde{L}(\delta) \right)}_{\text{gains from benign flow of patient investors}} \\ + \underbrace{\frac{1}{N} p_{LO}\, p_{LT}\, s_a^*/2 - \frac{1}{2\,N} p_{LO}\, p_{\text{public}}\, L\left(s_a^*/2\right) - \frac{1}{N} p_{\text{private}}\, L\left(s_a^*/2\right)}_{\text{gains from benign flow of impatient investors}} \,.$$

The tradeoff is clear. The first term on the right-hand side is the profitable benign flow from sniping patient investors in period three when cooperating, and the second term is the opportunity cost from no longer latency arbitraging patient investors in period four when there is a public innovation in y because patient investors are sniped in period three. The third term on the right-hand side is the benign flow from impatient investors when sharing the role of the market maker, the fourth term is the opportunity cost from no longer latency arbitraging the market maker when cooperating, and the final term is the opportunity cost from no longer adversely selecting the market maker when cooperating. The sum of the last three terms is positive, and the additional profits arise from the benign flow from impatient investors that would have otherwise matched with a patient investor's limit order.

On the other hand, if condition (D1) holds so that HFTs do snipe in period three in the competitive equilibrium, then the gains from cooperation are given by

$$c = \frac{1}{N}\, p_{LO}\, p_{LT}\, s_a^*/2 - \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\,(s_a^*/2) - \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\,(s_{na}^*/2) - \frac{1}{N}\, p_{\text{private}}\, L\,(s_a^*/2)\ .$$

The first term on the right-hand side is the benign flow from impatient investors when sharing the role of the market maker. The second and third terms are the opportunity costs from no longer latency arbitraging the market maker when cooperating, and the final term is the opportunity cost from no longer adversely selecting the market maker when cooperating. In this case, there are no additional gains from sharing the benign flow of patient investors or from limiting competition from patient investors because these profits are already included in Π^{D1}_{comp}. Here, the gains from cooperation arise from a wider spread of s_a^* instead of $(s_a^* + s_{na}^*)/2$, see Proposition 1.

In both cases, when $c > 0$, the profits from cooperation are higher than the profits from competitive play, which is a necessary condition to collude. However, when cooperating, HFTs have an incentive to cheat. As discussed above, there are three profitable deviations from cooperation. One, HFTs latency arbitrage or adversely select market makers. Two, HFTs are not compliant and do not respect the agreement to take turns to provide liquidity. Three, HFTs pretend to be a patient investor to fool other HFTs into sniping their informed toxic limit order.

With the first two deviations, the upper bound on the gain from deviation that one can achieve in expectation, per trading game, is

$$k = Q\, L(s_a^*/2)\,. \tag{4}$$

Even though the incentive to deviate is strong, the incentive is easily deterred because these deviations are perfectly observed. If HFTs are sufficiently patient (i.e., the discount factor ρ is sufficiently close to one) and if the punishment phase is sufficiently long (i.e., the value of T is large enough), then the loss in future payoffs from entering into a punishment phase outweighs the immediate gain from deviation. Therefore, in equilibrium, these deviations never occur and punishment phases are never triggered by these deviations.

Next, the expected gain, per trading game, from the third deviation is

$$g = \frac{p_{\text{private}}}{N}\, L(\delta)\,. \tag{5}$$

This deviation is harder to disincentivize because HFTs cannot perfectly infer if the punishment phase was triggered as a result of bad luck or because there was a cheater providing toxic liquidity. To disincentivize informed HFTs pretending to be a patient investor, the trigger condition must be sufficiently informative. An informative trigger condition is one that allows the HFTs to monitor each other with sufficient accuracy so that cheating can be punished by reverting to competitive play. If the trigger condition is sufficiently informative, if HFTs are sufficiently patient, and if the punishment phase T is sufficiently long, then cheating by pretending to be a patient investor can be deterred with intertemporal incentives.

In the collusive strategy, if play is in the cooperative phase, then the probability of triggering a punishment phase in the next trading game is the probability of sniping a patient investor and losing money, which is given by

$$1 - q_C = p_{LO}\, p_{\text{public}} \frac{\mathbb{P}(J > \delta)}{2}\,.$$

On the other hand, if play is in the cooperative phase and one HFT deviates by pretending to be a patient investor, then the probability of triggering a punishment phase in the next trading game is the probability of sniping a patient investor and losing money plus the probability of the deviator pretending to be a patient investor, which is given by

$$1 - q_D = p_{\text{private}} \frac{\mathbb{P}(J > \delta)}{N} + p_{LO}\, p_{\text{public}} \frac{\mathbb{P}(J > \delta)}{2}\,.$$

Therefore, the informativeness of the trigger condition is given by

$$q_C - q_D = p_{\text{private}} \frac{\mathbb{P}(J > \delta)}{N}\,.$$

To deter deviations by pretending to be a patient investor, the trigger condition must be sufficiently informative (i.e., $q_C - q_D$ must be sufficiently large) so that cheating can be deterred.

A key property of the collusive strategy is that, in equilibrium, reverting to competitive play is necessary even though HFTs do not pretend to be patient investors, i.e., reversion to competition occurs even though HFTs do not cheat. If the collusive strategy did not specify such a negative repercussion, then HFTs would have an incentive to cheat and the strategy profile would no longer be an equilibrium.

The following proposition provides the conditions for the collusive equilibrium to exist and summarizes the equilibrium behavior.

Proposition 2 *If the no-snipe condition* (D0) *or the snipe condition* (D1) *holds, and if the two inequalities*

$$\begin{aligned} \rho\left[c\,(q_C - q_D)\left(1 - \rho^T\right) + g\,q_C + \rho^T\,(1 - q_C)\,g\right] &\geq g\,, \\ \rho\left[c\,q_C\left(1 - \rho^T\right) + k\,q_C + \rho^T\,(1 - q_C)\,k\right] &\geq k \end{aligned} \tag{ICC}$$

hold, where the gains from deviation k and g are in (4) *and* (5)*, respectively, then the following* *reversion strategy profile is a collusive equilibrium:*

- *in punishment phases, HFTs play according to the competitive behavior described in Proposition 1, and*
- *in cooperation phases:*
  - *HFTs signal all their limit orders with a quantity $Q \gg 1$. HFTs take turns to provide liquidity. Each HFT i takes the role of the market maker for a proportion $1/N$ of the iterations played. HFTs do not latency arbitrage or adversely select market makers, and a market maker is compliant if she respects the agreement to take turns to provide liquidity at $y - s_a^*/2$ and at $y + s_a^*/2$, where s_a^* is the unique solution to (3a).*
  - *HFTs do not pretend to be a patient investor, and in period 3 HFTs race to snipe limit orders that arrived in period 2 without a signal.*

Play begins and remains in the cooperation phase until any one of the following is observed:

- *a market maker is latency arbitraged or adversely selected,*
- *a market maker is not compliant, or*
- *an HFT snipes a limit order without a signal from period 2 and loses money.*

Upon observing any of the triggers above, a punishment phase proceeds for T iterations of the trading game, after which, play returns to the cooperation phase.

The result follows from calculating the continuation values to check that there are no profitable deviations with the one-shot deviation principle. When play is in the punishment phase, there are no profitable deviations because myopic play is an equilibrium and no deviations can influence the continuation values. On the other hand, when play is in the cooperation phase, we check that the value obtained from each myopic deviation is less than the continuation value of continued cooperation. If condition (ICC) holds, then the myopic deviations are not profitable.

The (ICC) conditions hold if HFTs are sufficiently patient, if the punishment phase is sufficiently long, if the gains c from cooperation are large enough, and if the trigger condition is sufficiently informative (i.e., $q_C - q_D$ must be sufficiently large). The conditions ensure that deviations from cooperation are deterred through intertemporal incentives.

In the collusive equilibrium, the incentives to cooperate differ depending on the underlying competitive equilibrium. If uninformed HFTs snipe in the competitive equilibrium (i.e., condition (D1) holds), then the gains from cooperation arise from a wider spread of s_a^*. In this case, if the cooperation phase uses the quoted spread $(s_a^* + s_{na}^*)/2$ from the competitive equilibrium instead of s_a^*, then the collusive strategy with signaling is no longer an equilibrium when condition (D1) holds, because there are no incentives to cooperate because $c < 0$. Furthermore, in this case, signaling is not strictly necessary because if the incentives to cooperate arise from widening the spread, then there are other collusive strategies that can also result in cooperation without the need to signal (see e.g., Dutta and Madhavan, 1997). Nonetheless, the collusive strategy with signaling remains an equilibrium if condition (ICC) holds.

In the other case when uninformed HFTs do not snipe in the competitive equilibrium (i.e., condition (D0) holds), signaling to identify benign limit orders from patient investors and to identify toxic limit orders from informed HFTs has a distinct role in the incentives to cooperate. Specifically, signaling allows HFTs to safely profit from sniping limit orders posted by patient investors, and simultaneously limit competition from patient investors, which as a byproduct, enables the HFTs to receive additional benign flow from impatient investors who would have otherwise matched with a patient investor's limit order at better prices than those quoted by HFTs. These are sources of excess profits that would otherwise be unavailable in the competitive equilibrium where HFTs do not snipe. To see this, write the gains from cooperation when condition (D0) holds as

$$c = \left(\Pi_{\text{coop}} - \Pi_{\text{comp}}^{D1}\right) + \left(\Pi_{\text{comp}}^{D1} - \Pi_{\text{comp}}^{D0}\right) .$$

The first term on the right-hand side is the same as the gains from cooperation from a wider spread when condition (D1) holds, while the second term is given by

$$\begin{aligned}\Pi_{\text{comp}}^{D1} - \Pi_{\text{comp}}^{D0} &= \frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, L\left(s_{na}^*/2\right) + \frac{p_{LO}}{N}\,\delta - p_{LO}\,\frac{p_{\text{public}}}{2\,N}\left(L(-\delta) + \tilde{L}(\delta)\right)\\ &= \underbrace{\frac{1}{2\,N}\, p_{LO}\, p_{\text{public}}\, s_{na}^*/2}_{\text{additional flow from impatient investors}} + \underbrace{\frac{p_{LO}}{N}\,\delta - p_{LO}\,\frac{p_{\text{public}}}{2\,N}\left(L(-\delta) + \tilde{L}(\delta)\right)}_{\text{gains from benign flow of patient investors}},\end{aligned}$$

where the equality follows from (3b). The gains from $\Pi_{\text{comp}}^{D1} - \Pi_{\text{comp}}^{D0}$ arise solely from the ability to determine which limit orders in period two are safe to snipe and which are not. These additional profits illustrate how HFTs benefit from breaking the anonymity of the limit orders when the costs of sniping toxic limit orders from informed HFTs is too high. Furthermore, these gains alone are sufficient for the collusive strategy with signaling to be an equilibrium. That is, gains from a wider spread, i.e., $\Pi_{\text{coop}} - \Pi_{\text{comp}}^{D1}$, are not necessary to collude when condition (D0) holds.

When the costs of sniping toxic limit orders from informed HFTs is too high (i.e., condition (D0) holds), signaling to break the anonymity of the limit orders results in an equilibrium where: (i) quoted spreads are on average wider than that in the competitive equilibrium where the type of limit order (toxic or benign) cannot be identified, and (ii) the trading costs for impatient investors are higher because they are forced to trade at the spread set by HFTs (they cannot trade with limit orders from a patient investor unless play is in the punishment phase).

6.4.3. Parameter Analysis

To understand the robustness of the collusive equilibrium, we analyze how the values of the model parameters affect the existence of the collusive equilibrium. We provide both theoretical and numerical results to gain insights into what breaks the collusive equilibrium.

Theoretical Analysis. We provide a theoretical result to show how the number N of HFTs affects the existence of the collusive equilibrium, and a result to highlight how private information affects the existence of the collusive equilibrium.

Corollary 1 *Consider the set of model parameterizations such that the conditions in Proposition 2 hold. The size of this set decreases as the number N of HFTs increases.*

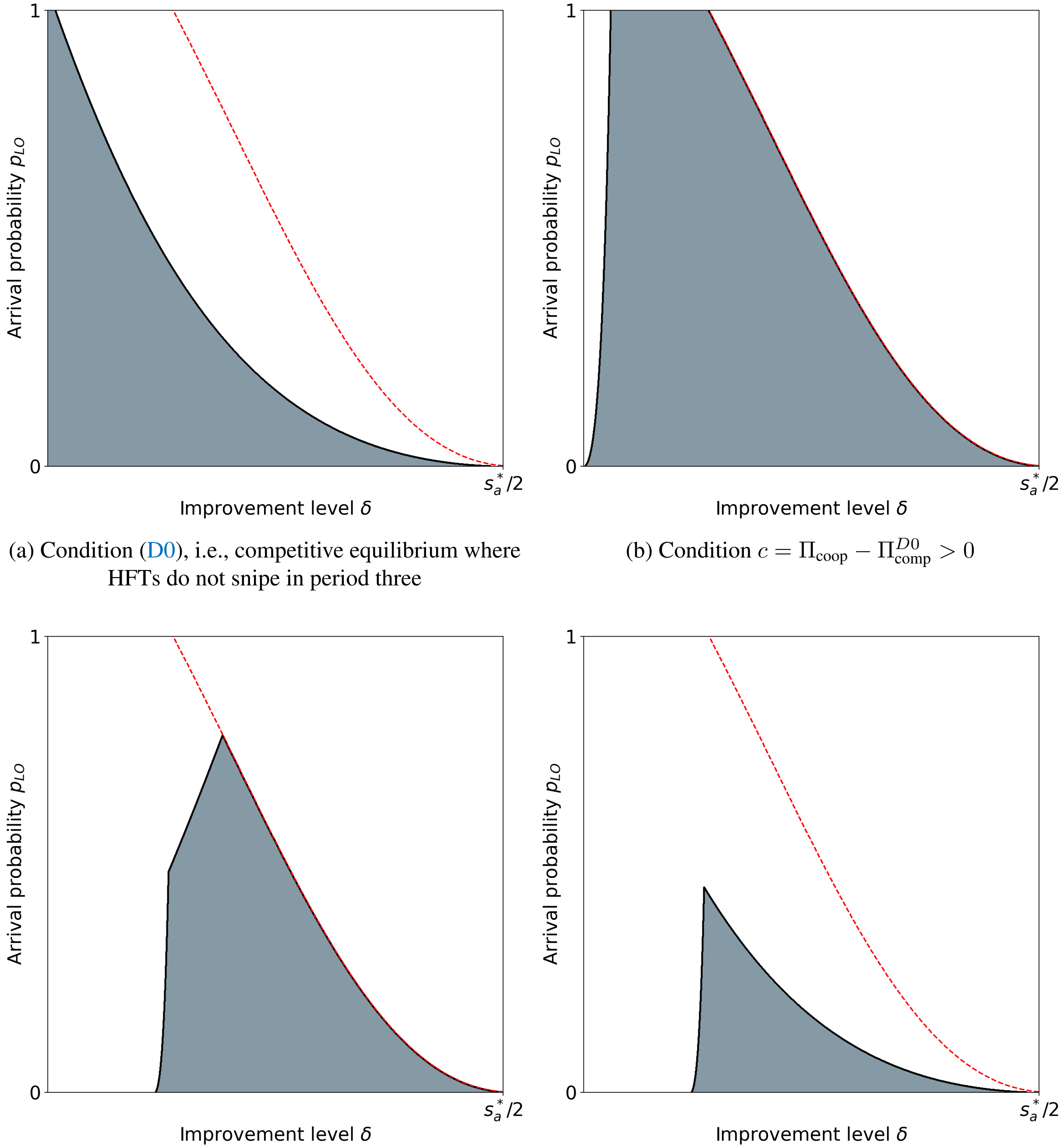
Put simply, as the number N of HFTs increases, the possibility of colluding with play described in Proposition 2 decreases. There are two forces driving this result. First, the gains c from cooperation decrease as the number of HFTs increases. Therefore, each HFT gets a smaller share of the excess profits, which decreases their incentives to cooperate. Second, the informativeness of the trigger condition $q_C - q_D$ decreases as N increases. Therefore, deviations are harder to deter, which makes it more difficult to sustain a collusive arrangement.

Corollary 2 *Consider the set of model parameterizations such that the conditions in Proposition 2 hold. The size of this set decreases as the arrival probability of private information $p_{private}$ decreases.*

In other words, when the arrival probability of private information p_{private} decreases, the possibility of colluding with play described in Proposition 2 decreases. This result follows because the trigger condition becomes uninformative, i.e., $q_C - q_D = 0$. Therefore, the collusive equilibrium unravels because cheating cannot be deterred. This result points to a potential policy response: reduce short-lived private information that arises from market fragmentation. For example, this can potentially be addressed through frequent batch auctions (see Budish et al., 2015).

Numerical Analysis. We use a toy example to analyze numerically the driver behind each of the conditions in Proposition 2 as a function of the improvement value δ of limit orders from patient investors and the arrival probability of a patient investor p_{LO}. An implicit condition in (ICC) is that there are gains from cooperation, i.e., $c > 0$ holds. We include this condition to highlight the incentives to cooperate. If all of the conditions hold, then a collusive equilibrium exists.

Figures 4 and 5 shades the regions of (δ, p_{LO}) that satisfy each of the conditions in Proposition 2 when conditions (D0) and (D1) hold, respectively. In Figure 4a, condition (D0) holds when the



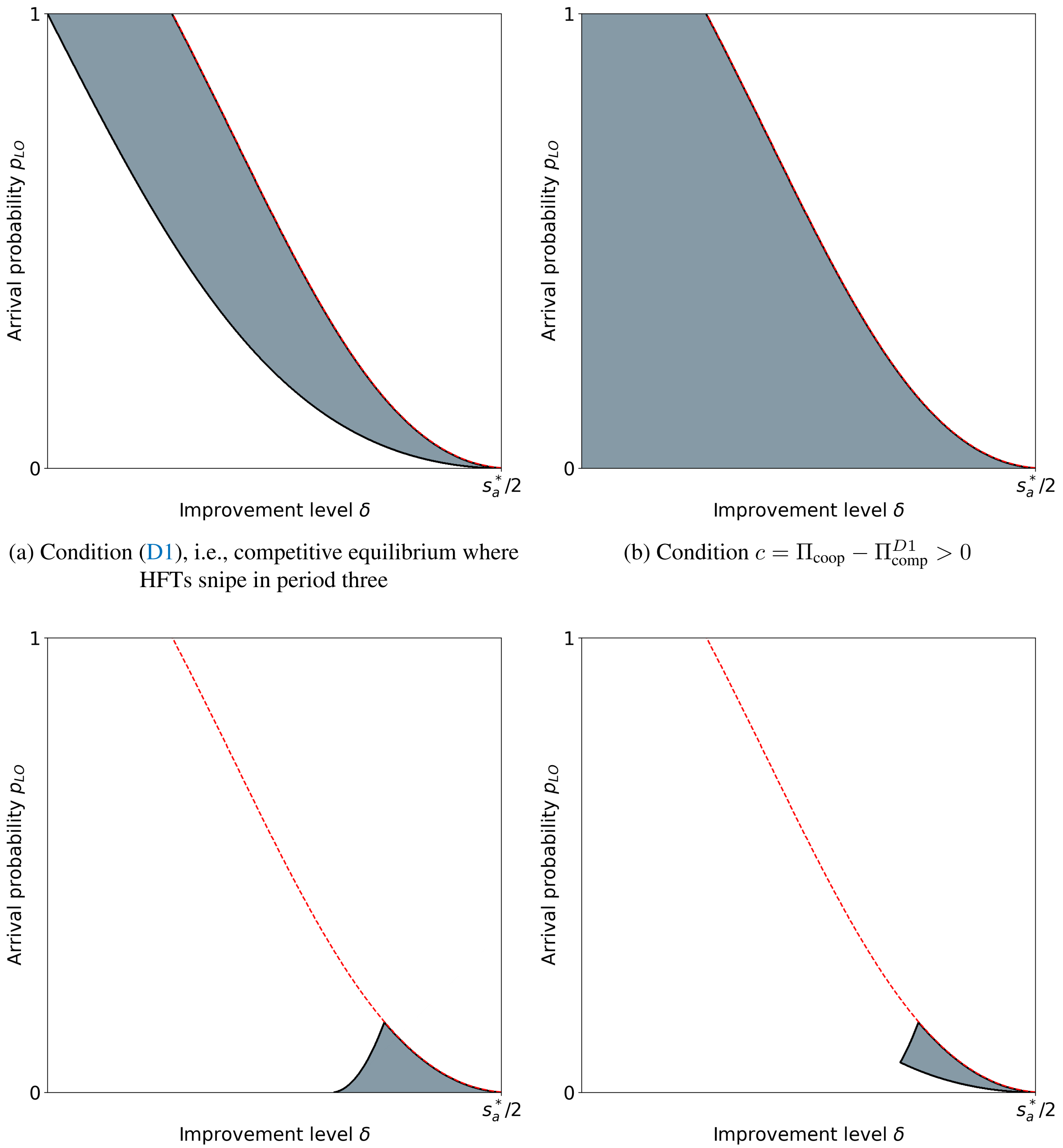
(a) Condition (D0), i.e., competitive equilibrium where HFTs do not snipe in period three

(b) Condition $c = \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D0} > 0$

(c) Condition (ICC) when $c = \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D0}$

(d) All conditions when condition (D0) holds

Figure 4: Improvement value δ of limit orders from patient investors and the arrival probability of a patient investor p_{LO} that satisfy each of the conditions in Proposition 2 when condition (D0) holds. If all of the conditions hold, then collusive equilibrium exists. Shaded regions are where the conditions are satisfied. Dashed red line is the solution to (3a) as a function of the arrival probability of a patient investor p_{LO}, which is the upper bound on the improvement level δ. Model parameters: $N = 3$, $p_{LT} = p_{\text{public}} = 0.4$, $Q = 3$, and $J \sim U(0, 10)$.



(a) Condition (D1), i.e., competitive equilibrium where HFTs snipe in period three

(b) Condition $c = \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D1} > 0$

(c) Condition (ICC) when $c = \Pi_{\text{coop}} - \Pi_{\text{comp}}^{D1}$

(d) All conditions when condition (D1) holds

Figure 5: Improvement value δ of limit orders from patient investors and the arrival probability of a patient investor p_{LO} that satisfy each of the conditions in Proposition 2 when condition (D1) holds. If all of the conditions hold, then collusive equilibrium exists. Shaded regions are where the conditions are satisfied. Dashed red line is the solution to (3a) as a function of the arrival probability of a patient investor p_{LO}, which is the upper bound on the improvement level δ. Model parameters: $N = 3$, $p_{LT} = p_{\text{public}} = 0.4$, $Q = 3$, and $J \sim U(0, 10)$.

expected payoff from sniping a patient investor in period 3 (i.e., $\delta\, p_{LO}$) is not too large, while in Figure 5a, condition (D1) holds when the expected payoff from sniping a patient investor in period 3 is large enough. In Figure 4b, there are gains from cooperation when the limit orders of patient investors are sufficiently mispriced, while in Figure 5b, the incentives to cooperate through a wider spread always exists. In Figures 4c and 5c, condition (ICC) holds when the limit orders of patient investors are sufficiently mispriced and when the arrival probability of patient investors is low.[34]

Finally, Figures 4d and 5d outline the regions of (δ, p_{LO}) that satisfy all conditions in Proposition 2. The union of the shaded regions in Figures 4d and 5d are regions of (δ, p_{LO}) where the collusive equilibrium exists. We see that the regions where the collusive equilibrium exists are relatively small. The collusive equilibrium only exists when the limit orders posted by patient investors are sufficiently mispriced, and when the arrival probability of a patient investor is low. The requirement for sufficiently mispriced limit orders posted by patient investors is consistent with Table 6, which shows that for limit orders sent by retail investors, the sniping probability increases as the level of mispricing increases.

7. Discussion

In the cooperation phase of the collusive equilibrium, the way in which HFTs can coordinate play is not unique. Similar to Folk theorems, our goal is not to specify the exact strategies used by HFTs, but rather to highlight the economic forces that underlie the incentives of HFTs to reveal themselves and how they can collectively enforce a collusive arrangement. Therefore, our model does not explain why HFTs would choose one collusive strategy profile over another.

Similarly, our model cannot explain how such a collusive arrangement might be initiated, whether it is tacit or explicit, because our model only analyzes the incentive structures in the collusive arrangement. Existing theory has little to say about how firms initiate a collusive arrangement without communication (see Green et al., 2014). An exception is Cartea et al. (2023) who show that a collusive arrangement can be initiated through learning without communication.

This raises a question on the role of quoting requirements. We argued that the quoting requirements on Euronext Amsterdam are not binding because it is a cumulative requirement. However, other exchanges such as Xetra, where many of the same ETFs are traded, have more stringent requirements that apply on an order-to-order basis (see Xetra, 2024). It is conceivable that HFTs active on exchanges with more stringent quoting requirements applied the same quoting behavior across exchanges, which allowed them to stumble into this outcome in Euronext Amsterdam.

Finally, we highlight that in the model, HFTs are invariant to the signaling mechanism. The only requirements are that limit orders posted by HFTs are easily differentiated from limit orders

[34] We say condition (ICC) holds whenever there exists a combination of (ρ, T) such that the inequality is satisfied.

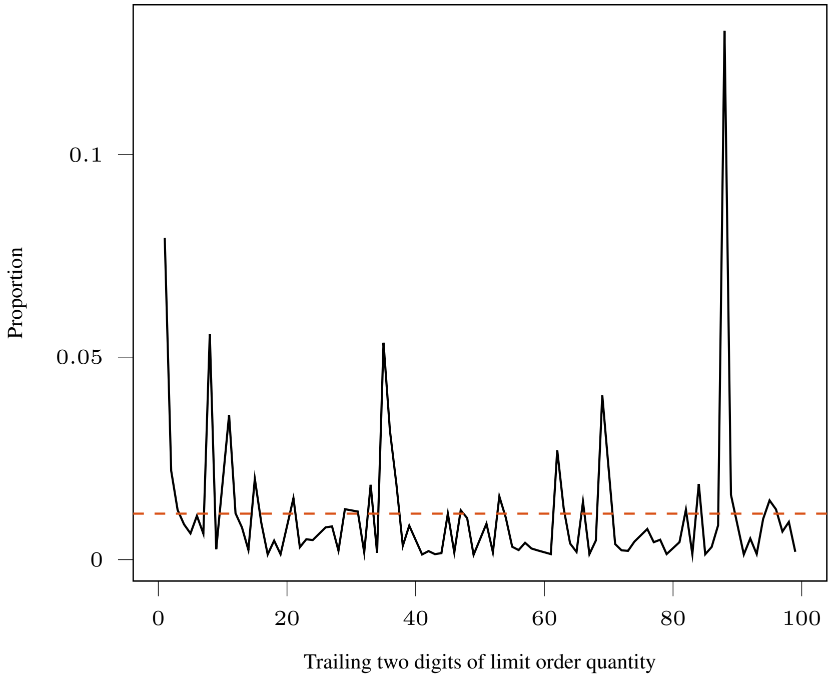


Figure 6: Proportion of limit orders sent by market makers as a function of the trailing two digits of order quantity (e.g., "87" in order quantity of "1087"). Data: Vanguard S&P 500 UCITS ETF from 2016-2021 with volumes larger than 100, excluding multiples of 10 or 25. Dashed red line is the uniform distribution.

posted by a patient investor, and that play in the cooperation phase is coordinated to accommodate for the signaling mechanism. Empirically, we find that in some ETFs, market makers also have distinct quoting patterns in the trailing digits of their limit orders. Figure 6 shows that some combinations of digits are preferred to others. This behavior is similar to that described in Cramton and Schwartz (2000, 2001), which warrants future research.

References

Abada, Ibrahim and Xavier Lambin (2023): "Artificial Intelligence: Can Seemingly Collusive Outcomes Be Avoided?" *Management Science*, 0.

Aït-Sahalia, Yacine and Mehmet Saglam (2013): "High Frequency Traders: Taking Advantage of Speed," Tech. rep., National Bureau of Economic Research.

Aquilina, Matteo, Eric Budish, and Peter O'Neill (2021): "Quantifying the High-Frequency Trading "Arms Race"," *The Quarterly Journal of Economics*, 137, 493–564.

Back, Kerry and Shmuel Baruch (2013): "Strategic Liquidity Provision in Limit Order Markets," *Econometrica*, 81, 363–392.

Baldauf, Markus and Joshua Mollner (2020): "High-Frequency Trading and Market Performance," *The Journal of Finance*, 75, 1495–1526.

Baron, Matthew, Jonathan Brogaard, Björn Hagströmer, and Andrei Kirilenko (2019):

"Risk and Return in High-Frequency Trading," *Journal of Financial and Quantitative Analysis*, 54, 993–1024.

BERNALES, ALEJANDRO (2017): "Algorithmic and High Frequency Trading in Dynamic Limit Order Markets," *Available at SSRN 2352409*.

BIAIS, BRUNO (1993): "Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets," *The Journal of Finance*, 48, 157–185.

BIAIS, BRUNO, THIERRY FOUCAULT, AND SOPHIE MOINAS (2011): "Equilibrium High Frequency Trading," in *Proceedings from the fifth annual Paul Woolley Centre conference, London School of Economics*.

BIAIS, BRUNO, DAVID MARTIMORT, AND JEAN-CHARLES ROCHET (2000): "Competing Mechanisms in a Common Value Environment," *Econometrica*, 68, 799–837.

BROGAARD, JONATHAN (2010): "High Frequency Trading and its Impact on Market Quality," *Northwestern University Kellogg School of Management Working Paper*, 66, 10.

BROGAARD, JONATHAN AND COREY GARRIOTT (2019): "High-Frequency Trading Competition," *Journal of Financial and Quantitative Analysis*, 54, 1469–1497.

BROGAARD, JONATHAN, BJÖRN HAGSTRÖMER, LARS NORDÉN, AND RYAN RIORDAN (2015): "Trading Fast and Slow: Colocation and Liquidity," *The Review of Financial Studies*, 28, 3407–3443.

BROGAARD, JONATHAN, TERRENCE HENDERSHOTT, AND RYAN RIORDAN (2014): "High-Frequency Trading and Price Discovery," *The Review of Financial Studies*, 27, 2267–2306.

——— (2019): "Price Discovery without Trading: Evidence from Limit Orders," *The Journal of Finance*, 74, 1621–1658.

BRUNNERMEIER, MARKUS K. AND LASSE HEJE PEDERSEN (2005): "Predatory Trading," *The Journal of Finance*, 60, 1825–1863.

BRYZGALOVA, SVETLANA, ANNA PAVLOVA, AND TAISIYA SIKORSKAYA (2025): "Strategic Arbitrage in Segmented Markets," *Journal of Financial Economics*, 166, 104008.

BUDISH, ERIC, PETER CRAMTON, AND JOHN SHIM (2015): "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, 130, 1547–1621.

BUDISH, ERIC, ROBIN LEE, AND JOHN J. SHIM (2024): "A Theory of Stock Exchange Competition and Innovation: Will the Market Fix the Market?" *Journal of Political Economy*, 132, 1209 – 1246.

CALVANO, EMILIO, GIACOMO CALZOLARI, VINCENZO DENICOLÓ, AND SERGIO PASTORELLO (2020): "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, 110, 3267–97.

CARTEA, ÁLVARO, PATRICK CHANG, MATEUSZ MROCZKA, AND ROEL OOMEN (2022a): "AI-Driven Liquidity Provision in OTC Financial Markets," *Quantitative Finance*, 22, 2171–2204.

CARTEA, ÁLVARO, PATRICK CHANG, AND JOSÉ PENALVA (2022b): "Algorithmic Collusion in Electronic Markets: The Impact of Tick Size," *Available at SSRN 4105954*.

CARTEA, ÁLVARO, PATRICK CHANG, JOSÉ PENALVA, AND HARRISON WALDON (2023): "Algorithmic Collusion and a Folk Theorem from Learning with Bounded Rationality," *Available at SSRN 4293831*.

CARTEA, ÁLVARO AND JOSÉ PENALVA (2012): "Where is the Value in High Frequency Trading?" *The*

Quarterly Journal of Finance, 2, 1250014.

CHRISTIE, WILLIAM G. AND PAUL H. SCHULTZ (1994): "Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *The Journal of Finance*, 49, 1813–1840.

CLANCY, LUKE AND MAURO CESA (2025): "Crossed Signals: Row Over Collusion Pits Scholars Against Traders," https://www.risk.net/markets/7961037/crossed-signals-row-over-collusion-pits-scholars-against-traders, accessed: 6 February 2025.

COLLIARD, JEAN-EDOUARD, THIERRY FOUCAULT, AND STEFANO LOVO (2022): "Algorithmic Pricing and Liquidity in Securities Markets," *HEC Paris Research Paper*.

COMERTON-FORDE, CAROLE, ALEX FRINO, AND VITO MOLLICA (2005): "The Impact of Limit Order Anonymity on Liquidity: Evidence from Paris, Tokyo and Korea," *Journal of Economics and Business*, 57, 528–540.

COMERTON-FORDE, CAROLE, TĀLIS J PUTNIŅŠ, AND KAR MEI TANG (2011): "Why do Traders Choose to Trade Anonymously?" *Journal of Financial and Quantitative Analysis*, 46, 1025–1049.

COMERTON-FORDE, CAROLE AND KAR MEI TANG (2009): "Anonymity, Liquidity and Fragmentation," *Journal of Financial Markets*, 12, 337–367.

CRAMTON, PETER AND JESSE A SCHWARTZ (2000): "Collusive bidding: Lessons from the FCC spectrum auctions," *Journal of Regulatory Economics*, 17, 229–252.

——— (2001): "Collusive Bidding in the FCC Spectrum Auctions," *Contributions in Economic Analysis & Policy*, 1, 1538–0645.1078.

DE JONG, FRANK AND BARBARA RINDI (2009): *The Microstructure of Financial Markets*, Cambridge University Press.

DEN BOER, ARNOUD V., JANUSZ M. MEYLAHN, AND MAARTEN PIETER SCHINKEL (2022): "Artificial collusion: Examining supracompetitive pricing by Q-learning algorithms," *Amsterdam Law School Research Paper*.

DOU, WINSTON WEI, ITAY GOLDSTEIN, AND YAN JI (2024): "AI-Powered Trading, Algorithmic Collusion, and Price Efficiency," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

DUTTA, PRAJIT K. AND ANANTH MADHAVAN (1997): "Competition and Collusion in Dealer Markets," *The Journal of Finance*, 52, 245–276.

EASLEY, DAVID AND MAUREEN O'HARA (1987): "Price, Trade Size, and Information in Securities Markets," *Journal of Financial Economics*, 19, 69–90.

EURONEXT (2023): "Rule book," https://www.euronext.com/en/media/1905, accessed: 7/March/2023.

——— (2024): "Member list," https://connect2.euronext.com/en/membership/resources/member-list, accessed: 16/August/2024.

FOUCAULT, THIERRY, JOHAN HOMBERT, AND IOANID ROŞU (2016): "News Trading and Speed," *The Journal of Finance*, 71, 335–381.

FOUCAULT, THIERRY, OHAD KADAN, AND EUGENE KANDEL (2013): "Liquidity Cycles and Make/Take Fees in Electronic Markets," *The Journal of Finance*, 68, 299–341.

FOUCAULT, THIERRY, ROMAN KOZHAN, AND WING WAH THAM (2017): "Toxic Arbitrage," *The Review of Financial Studies*, 30, 1053–1094.

FOUCAULT, THIERRY, SOPHIE MOINAS, AND ERIK THEISSEN (2007): "Does Anonymity Matter in Electronic Limit Order Markets?" *The Review of Financial Studies*, 20, 1707–1747.

FRIEDERICH, SYLVAIN AND RICHARD PAYNE (2014): "Trading Anonymity and Order Anticipation," *Journal of Financial Markets*, 21, 1–24.

GLOSTEN, LAWRENCE R. (1994): "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, 49, 1127–1161.

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.

GREEN, EDWARD J., ROBERT C. MARSHALL, AND LESLIE M. MARX (2014): "Tacit Collusion in Oligopoly," in *The Oxford Handbook of International Antitrust Economics, Volume 2*, ed. by Roger D. Blair and D. Daniel Sokol, Oxford University Press, 464–497.

GREEN, EDWARD J. AND ROBERT H. PORTER (1984): "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, 52, 87–100.

HAGSTRÖMER, BJÖRN AND LARS NORDÉN (2013): "The Diversity of High-Frequency Traders," *Journal of Financial Markets*, 16, 741–770.

HAGSTRÖMER, BJÖRN, LARS NORDÉN, AND DONG ZHANG (2014): "How Aggressive Are High-Frequency Traders?" *Financial Review*, 49, 395–419.

HARRINGTON, JOSEPH E. (2018): "Developing Competition Law for Collusion by Autonomous Artificial Agents," *Journal of Competition Law & Economics*, 14, 331 – 363.

——— (2022): "The Effect of Outsourcing Pricing Algorithms on Market Competition," *Management Science*, 68, 6889–6906.

HASBROUCK, JOEL (1991a): "Measuring the Information Content of Stock Trades," *The Journal of Finance*, 46, 179–207.

——— (1991b): "The Summary Informativeness of Stock Trades: An Econometric Analysis," *The Review of Financial Studies*, 4, 571–595.

——— (1995): "One Security, Many Markets: Determining the Contributions to Price Discovery," *The Journal of Finance*, 50, 1175–1199.

HASBROUCK, JOEL AND GIDEON SAAR (2009): "Technology and Liquidity Provision: The Blurring of Traditional Definitions," *Journal of Financial Markets*, 12, 143–172.

HENDERSHOTT, TERRENCE, CHARLES M. JONES, AND ALBERT J. MENKVELD (2011): "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance*, 66, 1–33.

HENDERSHOTT, TERRENCE AND RYAN RIORDAN (2009): "Algorithmic Trading and Information," *Manuscript, University of California, Berkeley*.

HOFFMANN, PETER (2014): "A Dynamic Limit Order Market with Fast and Slow Traders," *Journal of*

Financial Economics, 113, 156–169.

Jovanovic, Boyan and Albert J Menkveld (2016): "Middlemen in Limit Order Markets," *Available at SSRN 1624329*.

Kirilenko, Andrei, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017): "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance*, 72, 967–998.

Klemperer, Paul (2002): "What Really Matters in Auction Design," *The Journal of Economic Perspectives*, 16, 169–189.

Kyle, Albert S (1985): "Continuous Auctions and Insider Trading," *Econometrica*, 1315–1335.

Li, Sida, Xin Wang, and Mao Ye (2021): "Who Provides Liquidity, and When?" *Journal of Financial Economics*, 141, 968–980.

Madhavan, Ananth (1995): "Consolidation, Fragmentation, and the Disclosure of Trading Information," *The Review of Financial Studies*, 8, 579–603.

Madhavan, Ananth, David Porter, and Daniel Weaver (2005): "Should Securities Markets be Transparent?" *Journal of Financial Markets*, 8, 265–287.

Meling, Tom Grimstvedt (2021): "Anonymous Trading in Equities," *The Journal of Finance*, 76, 707–754.

Menkveld, Albert J (2016): "The Economics of High-Frequency Trading: Taking Stock," *Annual Review of Financial Economics*, 8, 1–24.

Menkveld, Albert J and Marius A Zoican (2017): "Need for speed? Exchange Latency and liquidity," *The Review of Financial Studies*, 30, 1188–1228.

O'Hara, Maureen (1998): *Market Microstructure Theory*, John Wiley & Sons.

Parlour, Christine A. (1998): "Price Dynamics in Limit Order Markets," *The Review of Financial Studies*, 11, 789–816.

Simaan, Yusif, Daniel G. Weaver, and David K. Whitcomb (2003): "Market Maker Quotation Behavior and Pretrade Transparency," *The Journal of Finance*, 58, 1247–1267.

Xetra (2024): "Designated Sponsor requirement," `https://www.xetra.com/xetra-en/trading/trading-models/liquidity-through-designated-sponsors/designated-sponsor-requirements`, accessed: 17/November/2024.

Yang, Liyan and Haoxiang Zhu (2019): "Back-Running: Seeking and Hiding Fundamental Information in Order Flows," *The Review of Financial Studies*, 33, 1484–1533.