Institutional Brokerage Networks: Facilitating Liquidity Provision*

Munhee Han Sanghyun (Hugh) Kim Vikram K. Nanda

April 29, 2019

*Han (munhee.han@utdallas.edu), Kim (sanghyun.kim1@utdallas.edu), and Nanda (vikram.nanda@utdallas.edu) are at the University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080. We thank Steven Xiao, Kelsey Wei, Paul Irvine, Stacey Jacobsen (discussant), Junghoon Lee, Charles Trzcinka, Talis Putnins (discussant), Johan Sulaeman (discussant), Kumar Venkataraman, seminar participants at the University of Texas at Dallas and Texas Christian University, and conference participants at the 29th Annual Conference on Financial Economics and Accounting (CFEA), the 31st Australasian Finance and Banking Conference, Society for Financial Studies (SFS) Cavalcade Asia-Pacific 2018, and the 2019 Midwest Finance Association (MFA) Annual Meeting for helpful comments.

Institutional Brokerage Networks: Facilitating Liquidity Provision

Abstract

We argue institutional brokerage networks facilitate liquidity provision and mitigate price impact of large non-information motivated trades. We use commission payments to map trading networks of mutual-funds and brokers. We find central-funds outperform peripheral-funds, especially in terms of return gap. Outperformance is more pronounced when trading is primarily liquidity driven to accommodate large redemptions. The fund–centrality premium is enhanced by brokers' incentives to generate greater commissions and by trading relationships between brokers and funds. Exploiting large brokerage mergers as exogenous shocks to network structure, we show that shocks to network centrality are accompanied by predicted changes in return gap.

Keywords: Institutional brokerage networks, mutual funds, return gap, trading costs, liquidity provision

1 Introduction

Brokers play a vital role in institutional trading in equity markets. When executing large client orders, brokers can mitigate price impact by actively searching for potential counterparties across various trading venues and, on occasion, by committing their own capital and acting more as dealers. Brokers often break up their clients' large orders and then strategically reveal to other clients who may be willing to fill the orders, while concealing from those who might front-run them (see Harris (2002) for an overview). Thus, trading between institutional investors tends to be broker-intermediated, with its efficacy closely tied to the trading networks of institutional investors and their brokers. In this paper, we argue that institutional brokerage networks facilitate liquidity provision and mitigate price impact for non-information driven trades.

Using brokerage commission payments, we map trading networks of mutual funds and their brokers as affiliation networks in which mutual funds are connected through their overlapping brokerage relationships. In these networks, mutual funds that trade through brokers that are also heavily used by other funds will tend to be more central. A key finding of the paper is that central funds in institutional brokerage networks outperform peripheral funds, especially as measured by their trading performance. In order to shed light on the specific mechanisms driving the positive relation between mutual funds' brokerage network centrality and their trading performance i.e., *fund–centrality premium*, we propose a liquidity provision hypothesis.

Our notion is that centrality in brokerage networks is especially valuable when mutual funds are forced to trade for liquidity reasons. As is well-recognized, open-end mutual funds incur substantial trading costs due to the adverse market impact of their trades when they liquidate holdings in response to investor redemptions (e.g., Edelen (1999)). In market microstructure models (such as Glosten and Milgrom (1985), Kyle (1985)), risk-neutral market makers are unable to identify trading motives. In these models, market makers set market prices and expect to lose to informed traders, while breaking even with gains from uninformed, liquidity traders. Thus, as in Admati and Pfleiderer (1991), liquidity traders who are transacting large quantities for non-informational reasons have an incentive to make their trading intentions known (i.e., engage in "sunshine trading") to distinguish themselves from informed traders and

attract more traders to provide liquidity.[1] While large liquidity traders may be unable to signal their trading motives directly to market participants, our view is that they might achieve the desired outcome by relying on their brokerage network and upstairs block trading.

We contend that institutions trading for liquidity reasons may be able to credibly convey their trading motives to brokers with whom they have well-established relationships. The credibility of a mutual fund will be enhanced if misrepresentation of its trading motives is likely to be costly in terms of a loss of reputation capital and trust in the broker-institution relationship. Central funds, connected to a larger network of brokers and funds will have more at risk in terms of potential loss of reputation and, hence, are likely to have greater credibility. The fund's brokers, in turn, could certify their clients' liquidity motives and execute trades at better prices (Seppi (1990)).[2] In addition, upstairs brokers can expand the available liquidity pool using information about their clients' latent trading interests and reaching out to wider set of potential counterparties to lower trading costs (Grossman (1992)).[3] Thus, even though all funds may have similar access to the available pool of expressed liquidity, for instance, through an electronic limit order book in the downstairs market, central funds will be better positioned to tap into larger pools of unexpressed liquidity through their brokers, especially when submitting large blocks of liquidity-motivated orders.[4]

[1] A concern, however, is that strategic traders that become aware of, say, a large liquidation could engage in "predatory trading", an argument advanced in Brunnermeier and Pedersen (2005). The notion is that strategic traders would trade the asset in the same direction prior to or simultaneously with the liquidating trader, before subsequently reversing the trade, to profit from the price impact at the expense of the liquidating trader. Bessembinder et al. (2016), however, show that traders supply liquidity to rather than exploit predictable trades in resilient markets and provide empirical evidence that a larger number of individual trading accounts provide liquidity around the time of large and predictable futures "roll" trades undertaken by a large exchange-traded fund (ETF) designed to provide returns that track crude oil prices.

[2] An upstairs market is an off-exchange market where a block broker facilitates the trading process by locating counterparties to the trade, and it operates as a search-brokerage mechanism where the terms of trade are determined through negotiation. Madhavan and Cheng (1997), Smith, Turnbull, and White (2001), and Booth et al. (2002) present evidence consistent with the Seppi (1990) hypothesis that upstairs market makers effectively screen out information-motivated orders and execute large liquidity-motivated orders at a lower cost than the downstairs market in the New York Stock Exchange (NYSE), the Toronto Stock Exchange (TSE), and the Helsinki Stock Exchange (HSE), respectively.

[3] Bessembinder and Venkataraman (2004) present direct evidence in support of the Grossman (1992) prediction that upstairs brokers lower execution costs by tapping into unexpressed liquidity. The authors find that execution costs for upstairs trades on the Paris Bourse are much lower than would be expected if the trades were executed against the expressed (displayed and hidden) liquidity in the downstairs limit order book.

[4] In a related literature on inter-dealer networks in the over-the-counter (OTC) municipal bond market, Li and Schürhoff (Forthcoming) find that dealers that are more central in the networks have better access to clients and more

Liquidity traders may have their own concerns about revealing their trading intentions to brokers. In the context of outflow-driven fire sales, Barbon et al. (Forthcoming) document that institutional brokers appear to foster predatory trading by leaking their clients' order flow information about impending fire sales to other important clients. These clients then sell the stocks being liquidated – only to buy them back later at lower prices. Our view is that while brokers may occasionally disclose client trades, they are unlikely to do so against their important clients, if it puts their trading relationships in jeopardy. Institutional investors would share trading intentions only if brokerage firms had valuable reputation capital: capital that could be lost if brokers did not act in their clients' interests. A broker disclosing client information faces the risk of being readily detected due to, for instance, the visibility of the price impacts (see, e.g., Smith, Turnbull, and White (2001)). In a broader context, our contention is that brokers will tend to use information about large liquidity-motivated orders to mitigate trading costs associated with adverse selection and invite more traders to provide liquidity, especially when the brokers' reputation costs are sufficiently high. The brokers used by central funds are apt to have greater reputation capital as indicated, for instance, by their well-established relationships to many other funds (and greater costs to being seen as untrustworthy). Hence, central funds are likely to benefit from lower costs for their liquidity motivated trades.[5]

To test our liquidity provision hypothesis, we exploit a unique dataset on brokerage commissions for a comprehensive sample of mutual funds from Form N-SAR semi-annual reports filed with the Securities and Exchange Commission (SEC). Using techniques from graph theory, we map the connections between mutual funds and their brokers as affiliation networks represented by weighted bi-partite graphs.[6] The

information about which securities are available and who wants to buy or sell, which results in shorter "intermediation chains," i.e., that fewer dealers are involved before a bond is transferred to another customer.

[5] Our paper is complementary to Barbon et al. (Forthcoming) in the sense of Carlin, Lobo, and Viswanathan (2007), who present a multi-period model of trading based on liquidity needs. In their model, traders cooperate most of the time through repeated interaction, providing liquidity to one another. However, "episodically" this cooperation breaks down when the stakes are high enough, leading to predatory trading.

[6] In affiliation networks, members are connected with one another through the organizations to which they belong. One can imagine, for instance, how movie stars are connected to one another through the movies in which they have co-appeared. Affiliation networks can be represented by bi-partite graphs, which have two types of nodes with one node of one type only connected to another node of a different type. In our case, a mutual fund is directly connected to its brokers and any pair of mutual funds can be connected with each other only indirectly through their overlapping brokerage connections. The connection between two funds is stronger if the extent to which their

weight of the bi-partite graph represents the strength of connection between a given fund-broker pair and is calculated as a fraction of brokerage commissions paid to the given broker. Further, to measure mutual funds' brokerage network centrality, we reduce this bi-partite graph of funds and brokers into a mono-partite graph in which fund-to-fund links are operationalized through their overlapping broker ties. We then use degree centrality and eigenvector centrality to quantify the importance of a given fund's position in the network.

Mutual funds that trade through many brokers that many other funds also trade through tend to be central in the network. Goldstein et al. (2009) note that most institutions concentrate their order flows with a small number of brokers in order to become their important clients, whereas large institutions can easily obtain the premium status from most brokers. Consistent with this observation, we find that funds that are large or belong to large fund families tend to be more central in the network, as they can afford to trade through a large number of brokers that are themselves central in the network.[7] We also find that mutual funds' brokerage network centrality is highly persistent, reflecting the persistence in the underlying brokerage relationships.

We begin our empirical analysis by showing that mutual funds' brokerage network centrality positively predicts their trading performance. Since we do not directly observe trading activities of mutual funds, we use as our measure of trading performance the return gap, which is calculated as the difference between the reported fund return and the return on a hypothetical portfolio that invests in the previously disclosed fund holdings (Grinblatt and Titman (1989), Kacperczyk, Sialm, and Zheng (2008)). We find that mutual funds in the highest quintile of brokerage network centrality have average monthly return gaps that are about five basis points larger than mutual funds in the lowest quintile over the period from July 1994 to December 2016. The results are statistically significant, insensitive to the choice of centrality measures, and robust to risk adjustments.

The economic magnitude of the relation between brokerage network centrality and return gap is

brokerage connections overlap is larger.

[7] However, other fund characteristics do not explain much of variation in brokerage network centrality. In contrast, fixed-effects, especially fund fixed-effects, account for a large amount of variation in brokerage network centrality, suggesting that we can identify the network effects that are orthogonal to the size effects.

meaningful as well. To put the numbers in perspective, we find that the return gap differential between the highest and lowest quintile portfolios sorted on brokerage network centrality is nearly half as large as that sorted on past return gap (Kacperczyk, Sialm, and Zheng (2008)). Furthermore, in our sub-sample analysis, we find that the fund–centrality premium is economically large and statistically significant in both early (1994-2007) and later (2008-2016) periods. This suggests that even in today's fragmented market with dark pools and smart order-routing systems, upstairs trading and institutional brokerage networks remain highly relevant to large institutional investors, as reported in the *Wall Street Journal*.[8]

In order to understand the specific mechanisms driving the return–gap premium associated with mutual funds' brokerage network centrality, it is useful to recognize key factors affecting the return gap. The return gap was originally proposed by Grinblatt and Titman (1989) as a measure of total transactions costs for mutual funds. Thus, at first brush, the fund–centrality premium is pretty much in line with our hypothesis that institutional brokerage networks mitigate mutual fund trading costs. Grinblatt and Titman (1989), however, point out that the return gap may be affected by interim trades within a quarter (Puckett and Yan (2011)) and possibly window-dressing activities. Kacperczyk, Sialm, and Zheng (2008) further note that skilled fund managers can use their informational advantage to time the trades of individual stocks optimally and show that the past return gap helps predict fund performance.[9]

We also recognize that the network formation is likely endogenous.[10] In order to rule out potential

[8] "'Upstairs' Trading Draws More Big Investors," by Bradley Hope, the *Wall Street Journal*, December 8, 2013. The article quotes a trader as stating that "It's like trying to fill up your gas tank, but you have to go to 15 gas stations. By the time you get to the 15th one, they've increased the price because they've heard you were coming. Wouldn't someone rather go to two or three stations and fill up the tank in blocks?"

[9] It may seem plausible as an alternative hypothesis that central funds can acquire privileged information about company fundamentals through their strong brokerage connections and trade on it. Put it differently, under the information channel hypothesis, the positive relation between brokerage network centrality and return gap could be driven by interim trades within a quarter, rather than trading costs. As we will show in our subsequent analyses, however, the fund–centrality premium is more pronounced when funds' trading activities are largely driven by liquidity reasons, rather than information motivated.

[10] For instance, marginal benefits of brokerage networks are likely higher for better skilled ones, fund managers with superior trading skills might self-select into central positions in institutional brokerage networks. There might exist an unobservable (to the econometrician) factor that is correlated with both brokerage network centrality and return gap. For instance, Anand et al. (2012) show that institutional trading costs are closely linked to trading desks' execution skills over and above selecting better brokers. In Section 5, we provide evidence supportive of our causal interpretation that institutional brokerage networks *improve* institutional trading performance, by exploiting mergers of large brokerage houses as plausibly exogenous shocks to the network structure.

confounding factors, we use panel regressions with fund fixed-effects to control for fund characteristics and unobserved heterogeneity. Consistent with our time-series results, we continue to find robust evidence that brokerage network centrality positively predicts future return gap, even after controlling for fund characteristics, including past return gap, and fund fixed-effects.

Now we turn to testing key predictions of our liquidity provision hypothesis. The primary prediction that we can derive from our hypothesis is that the fund–centrality premium should be more pronounced when funds' trading activities are largely driven by liquidity motives and funds can credibly signal this to their brokers. We use large outflow events to identify such periods of liquidity-motivated trading. When a mutual fund is experiencing severe redemptions, the fund is forced to liquidate a large fraction of its holdings in several stocks and their selling is, to a large extent, uninformed (see, e.g., Coval and Stafford (2007), Alexander, Cici, and Gibson (2007)). In addition, such forced liquidations are likely to send a particularly strong signal to the brokers that its sell orders are driven by liquidity reasons, rather than information motivated, thus helping the brokers communicate more credibly with other institutional clients to take the other end of the trades. Consistent with this prediction, we find that the fund–centrality premium is more pronounced when funds are forced to unwind their positions to accommodate large outflows.[11]

Second, our liquidity provision hypothesis also requires an active role on the part of brokers, such as in discerning their clients' uninformed trading motives and communicating with other institutional clients. As made clear in Carlin, Lobo, and Viswanathan (2007), whether the brokers facilitate liquidity provision or foster predatory trading is likely to hinge on the incentives they face and the strength of repeated interaction with their clients. To the extent that brokers are incentivized to maximize the expected value of future commission revenue streams, central funds with greater commission revenue generating potential are most likely to benefit from liquidity provision facilitated by their brokers. Using aggregate

[11] One potential concern is that the above results could be also consistent with cross-subsidization within a fund family: when a fund is suffering severe redemptions, another fund in the same family could step in to provide liquidity. For instance, Bhattacharya, Lee, and Pool (2013) show that affiliated funds of mutual funds that invest only in other funds within the family provide an insurance pool against temporary liquidity shocks to other funds in the family. This alternative cross-subsidization hypothesis may seem plausible because we find that funds that belong to large families are more central and large fund families are likely better equipped to provide cross-subsidization. Nevertheless, we continue to find qualitatively similar results when we exclude funds that belong to large fund families.

brokerage commissions as a proxy for the broker's incentives, we find that the fund–centrality premium is more pronounced for the funds that are likely more valuable for the brokers. Furthermore, we find that the effect of brokers' incentives on the fund–centrality premium is further amplified when funds are experiencing severe investor redemptions.

Third, our hypothesis relies on the repeated nature of interaction between institutional clients and their brokers. Institutional investors must build reputation for being truthful in order to credibly signal liquidity motives for their uninformed orders to their brokers. The brokers, in turn, must develop their reputation capital for being discreet when handling their clients' orders. Thus, the signaling and certification of uninformed trading motives is likely most effective if funds have already built strong trading relationships their brokers. Consistent with this prediction, we find that the fund–centrality premium is larger for the clients that have stronger existing trading relationships with their brokers, especially when funds are forced to liquidate to accommodate large outflows.[12]

One could still argue that central funds can obtain the return–gap premium because central funds can more easily slice up large orders and spread across many brokers who can then further split their clients' orders across many counterparties. Although not mutually exclusive with this alternative hypothesis, our liquidity provision hypothesis has clear predictions about the relation between the fund–centrality premium and the information content of trading. We provide further evidence that the fund–centrality premium is mostly concentrated in the periods that can be characterized by uninformed trading activities, e.g., when funds are trading *with flows*, rather than *against flows*. In addition, the fund–centrality premium is further amplified when the orders are also likely larger, suggesting that central funds can obtain the return–gap premium when central funds submit large *uninformed* orders.

Before concluding, we provide evidence supportive of our causal interpretation that institutional brokerage networks *improve* institutional trading performance, by exploiting mergers of large brokerage houses as plausibly exogenous shocks to the network structure. Following Hong and Kacperczyk (2010),

[12] This result is also consistent with that found in a related literature on client-dealer networks. For instance, Di Maggio, Kermani, and Song (2017) show that prior trading relationships are valuable especially in turbulent times in the OTC corporate bond market

we are able to identify and match a total of 26 brokerage mergers with our N–SAR data during the period from 1995 to 2015. The shock strength, however, is a major concern for our natural experiment, given the complexity of our network structure (which typically consists of thousands of nodes connected by tens of thousands of edges). In other words, moderate-sized brokerage mergers, especially as stand-alone events (which amount to cutting a small number of edges connected to a single node), are unlikely to serve as meaningful shocks. Therefore, we focus on two waves of five largest mergers of institutional brokers that took place around 2000 and 2008.[13]

Another challenge for our natural experiment is that the treatment of a shock is *a priori* unclear. However, we can reason that funds that traded largely through the acquiring brokers but not heavily though the target brokers are most likely to benefit from exogenous shocks to the network, since the acquiring broker would retain at least some of the target broker's clients. Following this intuition, we first construct hypothetical post-merger brokerage networks as would be formed if every fund were to maintain its pre-merger brokerage relationships and the funds hiring target brokers were to simply redistribute commissions to their remaining brokers on a pro-rata basis.[14] We then estimate the expected change in brokerage network centrality for each fund by calculating the difference between its hypothetical post-merger network centrality and its actual pre-merger network centrality. We take top ten percent of funds with largest expected change as the treatment group. Using a difference-in-differences (DiD) with matching, we find that funds in the treatment group experience significant increases in both brokerage network centrality and return gap after the merger relative to a control group of funds. These findings provide plausible evidence that institutional brokerage networks have a causal impact on institutional trading performance.

The remainder of this paper is organized as follows. In the next section, we discuss our paper in the context of related literature. Section 3 introduces our data and describes how we construct networks. We report our main results in Section 4 and conduct a natural experiment in Section 5. Section 6 concludes.

[13] These five brokerage mergers include Credit Suisse First Boston (CFBS)' acquisition of Donaldson, Lufkin & Jenrette (DLJ) and UBS's acquisition of Paine Webber in 2000 and JP Morgan Chase's acquisition of Bear Stearns, Barclays' acquisition of Lehman Brothers, and Bank of America's acquisition of Merrill Lynch in 2008.

[14] In our status-quo assumption, the funds that did not trade through the target broker (candidate treated funds) do not change their brokerage relationships, as they don't need to, but nonetheless experience exogenous increases in brokerage network centrality after the merger, because *other* funds need to reconfigure their brokerage relationships.

2 Related Literature

Our paper uncovers novel network effects in equity markets by documenting the return–gap premium associated with mutual funds' brokerage network centrality.[15] We contribute to a growing literature on broker-dealer networks in financial markets by shedding light on a unique role of institutional brokers in facilitating liquidity provision through the network. Whereas there is a large literature on dealer networks in over-the-counter (OTC) markets (see Section V. D of Bessembinder, Spatt, and Venkataraman (Forthcoming) for a comprehensive survey), studies on broker networks in the stock market have been relatively scant and our paper attempts to fill this gap.

In a recent paper, Di Maggio et al. (Forthcoming) shows that central brokers can extrapolate large informed trades from order flows and selectively leak this information to their more important clients, thereby facilitating "back-running" as described by Yang and Zhu (Forthcoming). Given such rent-extraction behavior, it is thus unclear whether central brokers can obtain "best execution" for their institutional clients. Our paper shows that central funds that trade through *many* central brokers can obtain the return–gap premium by effectively leveraging their strong brokerage connections to mitigate trading costs associated with adverse selection. Our paper is consistent with a related literature on client-dealer networks in the OTC corporate bond market. Hendershott et al. (2017) shows that many insurers use only one dealer, but execution costs decrease as a non-monotone function of the network size until it reaches 20 dealers, consistent with insurers trading off the benefits of relationship trading against dealer competition.

Our paper is related to, but differs from, recent studies that document evidence of information flows or leakages from some clients to the others through the brokers. Chung and Kang (2016) shows strong return comovement among hedge funds sharing the same prime broker and argue that the prime broker provides profitable information to its hedge fund clients. As potential sources of such profitable information, Kumar et al. (2018) points to privileged information on corporate borrowers from the affiliated

[15] There has been a growing interest in studying network effects in equity markets. For instance, Ahern (2013) shows that industries that are more central in intersectoral trade networks earn higher stock returns than industries that are less central. Ozsoylev et al. (2014) estimate empirical investor networks using account-level trading data from the Istanbul Stock Exchange and find that more central individual investors earn higher returns and trade earlier than peripheral investors with respect to information events.

banking division of an investment bank with prime brokerage business and Di Maggio et al. (Forthcoming) hints at client order flow information about large informed trades by hedge funds or activist investors right before 13D filings. Our paper, however, differs substantially from these papers in that our focus is on information flows regarding large liquidity-motivated trades, rather than private information about company fundamentals.[16]

Our paper is most closely related to Barbon et al. (Forthcoming) who document that institutional brokers can foster predatory trading by leaking their clients' order flow information about impending fire sales to other important clients, such as prime brokerage hedge fund clients. The clients then sell the stocks being liquidated along with the distressed funds only to buy them back later at much lower prices, thereby exacerbating price impacts. Brokerage firms, however, value their reputation capital and institutional clients can easily monitor whether a particular broker is acting in their interests thanks to the visibility of the price impacts and the ongoing broker–client relationships (see, e.g., Smith, Turnbull, and White (2001)).

In a broader context, we show that brokers tend to use information about large liquidity-motivated orders to mitigate trading costs associated with adverse selection and invite more traders to provide liquidity, especially when the brokers' reputation costs are sufficiently high. Our paper is complementary to Barbon et al. (Forthcoming) in the sense of Carlin, Lobo, and Viswanathan (2007), who present a multi-period model of trading based on liquidity needs. In their model, traders cooperate most of the time through repeated interaction, providing liquidity to one another. However, "episodically" this cooperation breaks down when the stakes are high enough, leading to predatory trading.[17]

[16] In a similar sense, our paper differs from the papers that shows how institutional investors can gain informational advantage through their brokerage connections. Examples of such information channels include early access to sell-side research or tipping (Irvine, Lipson, and Puckett (2007)) and invitations to broker-hosted investor conferences (Green et al. (2014)).

[17] Some investment banks generate a substantial amount of fee revenues from hedge funds that use their prime brokerage services, such as securities lending, margin financing, and risk management. Consistent with high-powered incentives of prime brokerage business, Kumar et al. (2018) find strong evidence that investment banks sometimes leak privileged information about their corporate borrowers to their prime brokerage hedge fund clients who subsequently trade on and profit from it, whereas Griffin, Shu, and Topaloglu (2012) find little evidence of such information-based trading by the average brokerage house client of investment banks.

3 Data and Variable Construction

Section 3.1 describes our primary data on brokerage commissions and explains how we construct other fund-level variables. Section 3.2 explains how we construct institutional brokerage networks and centrality measures, discusses the characteristics of the network, and examines the determinants of mutual funds' brokerage network centrality.

3.1 Brokerage Commissions and Other Fund-Level Variables

Our primary data comes from the SEC Form N–SAR filings, which we combine with other data sets. We obtain data on mutual fund monthly returns, total net assets (TNA), and fund expenses from the Center for Research in Security Prices (CRSP) Survivor-Bias-Free Mutual Fund Database. The returns are net of fees, expenses, and brokerage commissions, but before any front-end or back-end loads. The stock holdings of mutual funds are from Thomson-Reuter Ownership Database (Thomson s12). We use the MFLINKS files available through Wharton Research Data Services (WRDS) to merge CRSP and Thomson data sets. For funds with multiple share classes in CRSP, we aggregate share-class-level variables at the fund-level by computing the sum of total net assets and the value-weighted average of returns and expenses.

Under the Investment Company Act of 1940, all registered investment companies are required to file Form N–SAR with the SEC on a semi-annual basis. N–SAR reports are filed at the registrant level. A registrant typically consists of a single mutual fund and thus is simply referred to as a fund in our paper, except when the distinction is likely important.[18] N–SAR filings disclose information about fund operations and financials under 133 numbered items with alphabetized sub-items. We extract all N-SAR reports filed between 1994 and 2016 available through the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system.

[18] A registrant can consist of multiple funds or be part of a fund family, although it is just a single mutual fund in about 65% of the N–SAR filings. We emphasize that a registrant does not refer to a fund family, but rather is a filing unit under which a fund family reports its funds together in a single filing. For instance, according to our N–SAR data, Fidelity reported its 466 mutual funds with about $1.5 trillion assets under management using 82 separate N–SAR filings during the first half of 2016. Many items are reported at the fund level, but some of the items such as brokerage commissions are aggregated and reported at the registrant level.

Since our focus is on U.S. domestic equity funds, we exclude N-SAR funds that are not equity-oriented (Item 66.A), international funds (Item 68.B), and the funds with percentage of TNA invested in common stocks (Item 74.F divided by Item 74.T) below 80% or above 105%. We also exclude N-SAR reports where aggregate brokerage commissions paid (Item 21) are reported as zero or missing.[19] From the CRSP–Thomson merged data set, we eliminate international, municipal, bonds and preferred, and metals funds using the investment objective code from Thomson (ioc) and screen for U.S. domestic equity funds using the investment objective code from CRSP (crsp_obj_cd). We also exclude all observations where the fund's TNA does not exceed $5 million or the number of stock holdings does not exceed 10.

After the above data screens, we automatically match N–SAR fund names (Item 1.A and a colon followed by Item 7.C) with CRSP fund names after removing share-class identifiers using the generalized Levenshtein (1966) edit distance while exploiting the typical structure of CRSP fund name (*FUND FAMILY NAME: FUND NAME; SHARE CLASS*). In the automated name matching process, we require that the monthly average net assets (TNA) during the reporting period (Item 75.B) and the corresponding TNA value constructed from CRSP and MFLINKS be within the 5% range from each other. Finally, we manually check the accuracy of the matches and remove the ones that appear inaccurate. The total number and aggregate TNA of our CRSP–Thomson–NSAR matched sample funds are reported in Table A1 in the Appendix.

Of particular interest to our study are brokerage commissions paid to the ten brokers that received the largest amount from the fund during the reporting period and the names of those brokers (Item 20). Table 1 provides an example of brokerage commission payments along with some descriptive statistics.

[Insert Table 1]

We recognize that N–SAR filings do not report all brokerage firms to which the fund paid commissions and, as a result, we are likely to miss some of the less important brokerage connections. As an example, Panel A of Table 1 reports brokerage commissions that T. Rowe Price Blue Chip Growth Fund paid to

[19] Reuter (2006) reports that in his sample, approximately 82% of the N-SAR filings that report paying no brokerage commissions are from investment companies that consist solely of bond funds, which do not pay explicit brokerage commissions on their transactions.

its top ten brokers and the aggregate commissions paid to all brokers during the first half of 2016. As is typically the case, the sum of brokerage commissions do not add up to the aggregate commissions, suggesting that the fund employed more than ten brokers.[20] In general, as shown in Panel B of Table 1, brokerage commissions are highly concentrated with a few important brokers for each fund, but the top ten brokers reported in N–SAR filings on average account for only 72.45% (or 71.62% at the median) of the aggregate brokerage commissions that the fund paid to *all* brokers. Nevertheless, partial data issues are unlikely to cause any bias in our results, since centrality calculated in the reduced network is highly correlated with full-network centrality (Ozsoylev et al. (2014)).[21]

Panel C presents a transition probability matrix of annual changes in broker rankings for each fund and shows strong persistence in brokerage relationships between a fund and its key brokers. If a broker is ranked top this year by the commission payments, the probability of the same broker staying on top for the same fund next year is close to 50%. As we move down the rankings, the persistence becomes gradually weaker. The concentration of commissions with several important brokers and the persistence in business relationships funds maintain with those brokers are generally in line with the literature on institutional brokers (e.g., Goldstein et al. (2009)).

Next, we describe how we construct other fund-level variables. We take the fund TNA directly from N–SAR (Item 74.T) and use the fund family code reported by the fund (Item 19.C) to calculate the fund family TNA. The trading volume is calculated by the sum of purchases (Item 71.A) and sales (Item 71.B). Since brokerage commissions are reported at the registrant level, we calculate the commission rate as a ratio of the aggregate commission payments (Item 21) to the sum of aggregate trading volumes across all funds reported together, following Edelen, Evans, and Kadlec (2012). This pro-rata algorithm implicitly assumes

[20] Mutual funds and institutional investors typically employ a large number of brokers not only for daily trade executions but also for various services that brokers provide, such as early access to sell-side research or tipping (Irvine, Lipson, and Puckett (2007)), favorable allocations of hot IPO stocks (Reuter (2006)), invitations to broker-hosted investor conferences (Green et al. (2014)), and marketing and retail distribution support (Edelen, Evans, and Kadlec (2012)).

[21] For example, in simulations Ozsoylev et al. (2014) show that even when a reduced network represents only 10% of the links in the full network, the correlation between true centrality and centrality calculated in the reduced network is about 0.5. In our study, the reduced network typically represents more than 70% of the weighted links in the full network that be constructed from the complete information on commissions paid to all brokers.

the commission rates to be the same for all the funds of which a registrant consists. In a similar spirit, we estimate the fund's commission payments by taking the product of the commission rate and the fund trading volume. We take an index fund indicator from N–SAR (Item 69). For each fund-quarter, size, value, and momentum percentiles are calculated as percentiles of market capitalization, book-to-market ratio, and 12-month returns skipping the most recent month, respectively, averaged across all stock holdings. For each fund-halfyear, we take the most recent quarterly observation of average size-value-momentum percentiles.

Last, following the literature (e.g., Coval and Stafford (2007)), we calculate monthly net flows for each fund share class i during month t as follows:

$$FLOW_{i,t} = TNA_{i,t} - TNA_{i,t-1} \times (1 + R_{i,t}) \tag{1}$$

where $FLOW_{i,t}$ is the dollar value of fund flow (net new issues and redemptions), $TNA_{i,t}$ is the total net asset, and $R_{i,t}$ is the monthly return. To compute the monthly fund flow for the fund, we sum monthly fund flows for all share classes belonging to the same fund as identified by MFLINKS. Monthly fund flows are summed over the half-year to calculate the semi-annual fund flow. For the percentage figures, we divide the dollar value of fund flows by the beginning-of-period TNA. The summary statistics are reported in Table 2.

[Insert Table 2]

3.2 Institutional Brokerage Networks

Using brokerage commission payments, we map trading networks of mutual funds and their brokers as affiliation networks represented by weighted bi-partite graphs. In a graph, agents can be represented by nodes and connections (ties) between agents by edges. In a bi-partite graph, nodes can be partitioned into two types and nodes of one type can only be connected to the nodes of the other type, not with the ones of the same type. Such bi-partite graphs are typically used to model affiliation networks where members form networks through organizations to which they belong. We illustrate how we construct institutional brokerage networks and calculate brokerage network centrality step-by-step using a simple example in

Figure 1. Panel A of Figure 1 presents a graphical representation of the network consisting of ten funds and four brokers.

[Insert Figure 1]

Like any graph, a bi-partite graph can be represented by an adjacency matrix, denoted G, where rows index mutual funds and columns index brokers. Each element $g_{i,k}$ of G represents the strength of connection between fund i and broker k and is defined as the brokerage commissions paid to broker k, scaled by the sum of brokerage commissions paid to the top ten brokers. If broker k does not appear as one of the top ten brokers for fund i, then $g_{i,k}$ is assumed zero. Panel B of Figure 1 shows the transpose of the adjacency matrix G representing our simple network in extended-form.

To measure a mutual fund's connections to all the other mutual funds through their overlapping brokerage connections, we reduce the bi-partite graph of mutual funds and brokers into a mono-partite graph of mutual funds only by defining its adjacency matrix A as

$$a_{i,j} = \sum_{k} \min(g_{i,k}, g_{j,k}) \text{ if } i \neq j \tag{2}$$

where i and j index funds and k indexes brokers. The strength of connection between any pair of funds is simply the percentage overlap (Jaccard distance) of brokerage connections between two funds. Panel C of Figure 1 shows the adjacency matrix A representing our simple network in reduced-form. We emphasize that the connections between funds are *indirect* and made through overlapping brokerage connections. For instance, Fund 1 and Fund 2 are connected through Broker A and Broker B, and the strength of connection between these two funds is 0.35 $(= \min(0.85, 0.20) + \min(0.15, 0.33) + \min(0, 0.22) + \min(0, 0.25))$.

We borrow techniques from graph theory and social network literature to quantify the importance of a mutual fund's position in the network. The importance of a node in a network is typically measured by its centrality and we use degree centrality (Freeman (1979)) and eigenvector centrality (Bonacich (1972, 1987)).[22] Degree centrality is defined as the sum of each row in the adjacency matrix, A, defining the

[22] Many different measures of centrality have been proposed and among the most commonly used measures

network, scaled by the number of rows minus one. Eigenvector centrality is defined as the principal eigenvector of the adjacency matrix defining the network. That is,

$$\lambda v = Av \tag{3}$$

where A is the adjacency matrix of the graph, λ is a constant (the eigenvalue), and v is the eigenvector.

Panel D of Figure 1 reports brokerage network centrality calculated for all funds in our simple network. For instance, degree centrality for Fund 1 is 0.339 ($= (0.35 + 0.30 + 0.50 + 0.35 + 0.15 + 0.35 + 0.30 + 0.60 + 0.15)/9$). As can be seen in Panel A of Figure 1, funds that are positioned in the center of the network (e.g., 2, 3, and 5) are indeed more central than funds located in the periphery (e.g., 1, 4, and 10). In general, funds that trade through many brokers that many other funds also trade through tend to be central in the network.

In order to line up with the semi-annual N–SAR reporting frequency, we construct networks every half-year at the end of June (December) for N–SAR filings with reporting period ending in January to June (July to December) from the first half of 1994 to the first half of 2016. Since brokerage commission payments are only reported at the registrant level and are not broken down by fund, we construct networks at the registrant level and all funds within the same registrant inherit the same network structure. Figure 2 shows institutional brokerage networks constructed using our N–SAR data for the first half of 2016.

[Insert Figure 2]

Now we examine what types of mutual funds are more central in institutional brokerage networks

of centrality are degree, closeness, betweenness, and eigenvector centrality. When choosing the most appropriate measure, one must be careful about the implicit assumptions underlying these centrality measures. As laid out in Borgatti (2005), closeness centrality and betweenness centrality are built upon an implicit assumption that traffic flows along the shortest paths until it reaches a pre-determined destination like the package delivery process. In institutional brokerage networks, traffic is likely to freely flow from one fund (the fund submitting a trade order) to another (a potential fund that could absorb the submitted trade order) through the broker intermediating the trade. Since this type of traffic must flow through unrestricted walks, rather than via geodesics, closeness centrality and betweenness centrality can be safely ruled out. See also Ahern (2013) for a similar discussion.

by estimating the following linear regression model:

$$Centrality_{i,t} = \gamma \times Covariates_{i,t} + \alpha_i + \theta_t + \varepsilon_{i,t} \tag{4}$$

where i indexes mutual funds and t indexes time in half-years. The dependent variable is $Centrality_{i,t}$, fund i's brokerage network centrality measured at the end of half-year t. $Covariates_{i,t}$ are a vector of fund-level characteristics that include log of fund TNA, log of family TNA, expense ratio, commission rate, trading volume, and average size-value-momentum percentiles of stock holdings, all measured at the end of half-year t. α_i denotes fund fixed-effects, θ_t denotes time fixed-effects, and standard errors are clustered at the fund level.

Table 3 presents the regression results. The dependent variable is degree centrality in columns (1) through (5) and eigenvector centrality in columns (6) through (10). Overall, we find that funds that are large or belong to large fund families tend to be more central in the network, as these funds can afford to trade through a large number of brokers that are themselves central in the network. This result suggests that brokerage relationships are costly to build and is consistent with Goldstein et al. (2009) who note that most institutions concentrate their order flows with a small number of brokers in order to become their important clients, whereas large institutions can easily obtain the premium status from most brokers.

[Insert Table 3]

As can be seen in columns (1) and (6), fund and family sizes alone can explain 19% and 28% of variation in degree centrality and eigenvector centrality, respectively. Adding other fund characteristics in columns (2) and (7) only marginally improves the explanatory power, raising adjusted R^2 to 26% and 31% for degree centrality and eigenvector centrality, respectively. In contrast, fixed-effects, especially fund fixed-effects, account for a large amount of variation in brokerage network centrality, suggesting that we can identify the network effects that are orthogonal to the size effects. Adding time and fund fixed-effects in columns (5) and (10) raises adjusted R^2 to 74% and 72% for degree centrality and eigenvector centrality, respectively. This result also implies that brokerage network centrality is highly persistent, reflecting the

persistence in the underlying brokerage relationships.

4 Brokerage Network Centrality and Trading Performance

In Section 4.1, we begin our empirical analysis by showing that mutual funds' brokerage network centrality predicts their trading performance as measured by return gap. In Section 4.2, we turn to inspecting the specific mechanisms behind the return–gap premium associated with mutual funds' brokerage network centrality (simply the fund–centrality premium or the return–gap premium).

4.1 The Fund–Centrality Premium

4.1.1 The Time-Series Evidence

Despite extensive disclosure requirements, mutual funds are only required to disclose their holdings on a quarterly basis and their trading activities are generally unobservable (Kacperczyk, Sialm, and Zheng (2008)). In order to examine how institutional brokerage networks affect mutual fund trading performance, we use the return gap as our measure of trading performance. The return gap is calculated as the difference between the reported fund return and the return on a hypothetical portfolio that invests in the previously disclosed fund holdings (Grinblatt and Titman (1989), Kacperczyk, Sialm, and Zheng (2008)):

$$Return\ Gap_{i,t} = RET_{i,t} - (HRET_{i,t} - EXP_{i,t}) \tag{5}$$

where $RET_{i,t}$, is the fund i's reported return net of expenses during month t, $EXP_{i,t}$, is the expense ratio for fund i reported prior to month t, and $HRET_{i,t}$ is the fund i's holdings return during month t, which is defined as:

$$HRET_{i,t} = \sum_{k} w_{i,k,t-1} R_{k,t} \tag{6}$$

where $w_{i,k,t-1}$ is the fund i's portfolio weight on stock k at the end of month $t-1$ and $R_{k,t}$ is the return on stock k during month t.

At the end of every June and December, we sort mutual funds into quintile portfolios, based on their brokerage network centrality. The average time-series monthly returns from July 1994 to December 2016 are reported in Table 4. The full-sample results reported in Panel A show that the average return gap increases monotonically from the portfolio of peripheral funds (the lowest quintile of brokerage network centrality) to the portfolio of central funds (the highest quintile). The difference in average return gaps between central funds and peripheral funds is about five basis points per month (t-statistic = 5.03 to 5.26). After adjusting for the Fama–French–Carhart four–factor loadings, the central–minus–peripheral portfolio delivers an average alpha of four basis points per month (t-statistic = 4.48 to 4.75).

[Insert Table 4]

The economic magnitude of the relation between brokerage network centrality and return gap is meaningful as well. To put the numbers in perspective, we find that the return gap differential between the highest and lowest quintile portfolios sorted on brokerage network centrality is nearly half as large as that sorted on past return gap (Kacperczyk, Sialm, and Zheng (2008)). Furthermore, in our sub-sample analysis, we find that the fund–centrality premium is economically large and statistically significant in both early (1994-2007) and later (2008-2016) periods reported in Panel B and Panel C, respectively. This suggests that even in today's fragmented market with dark pools and smart order-routing systems, upstairs trading and institutional brokerage networks remain highly relevant to large institutional investors, as reported in the *Wall Street Journal*.[23]

4.1.2 The Cross-Sectional Evidence

In order to understand the specific mechanisms driving the return–gap premium associated with brokerage network centrality, it is important to recognize key factors affecting the return gap. The return gap is originally proposed by Grinblatt and Titman (1989) as a measure of total transactions costs for mutual funds. Therefore, at first brush, the fund–centrality premium is very much in line with our hypothesis that institutional brokerage networks mitigate mutual fund trading costs. Grinblatt and Titman (1989),

[23] "'Upstairs' Trading Draws More Big Investors," by Bradley Hope, the *Wall Street Journal*, December 8, 2013.

however, point out that the return gap may be affected by interim trades within a quarter and possibly window-dressing activities. Kacperczyk, Sialm, and Zheng (2008) further note that skilled fund managers can use their informational advantage to time the trades of individual stocks optimally and show that the past return gap helps predict fund performance.

We also recognize that the network formation is likely endogenous. For instance, marginal benefits of institutional brokerage networks are likely higher for better skilled ones, fund managers with superior trading skills might self-select into central positions in the network. There might also exist an unobservable (to the econometrician) factor that is correlated with both mutual funds' brokerage network centrality and their trading performance. For example, Kacperczyk, Sialm, and Zheng (2008) document the persistence in return gap and propose the return gap as a measure of interim trading skills of fund managers (see also Puckett and Yan (2011)). Anand et al. (2012) show that trading costs are closely linked to trading desks' execution skills over and above selecting better brokers.

In order to mitigate these confounding factors, we use cross-sectional regressions with fund fixed-effects to control for unobserved heterogeneity along with observable fund characteristics. Specifically, we estimate the following linear regression model:

$$Return\ Gap_{i,t} = \beta \times Centrality_{i,t-1} + \gamma \times Covariates_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \tag{7}$$

where i indexes mutual funds and t indexes time in half-years. The dependent variable is $Return\ Gap_{i,t}$ which is fund i's average return gap during half-year t. $Centrality_{i,t-1}$ is fund i's brokerage network centrality measured at the end of half-year $t-1$. $Covariates_{i,t-1}$ are a vector of fund-level characteristics that include log of fund TNA, log of family TNA, expense ratio, commission rate, trading volume, and average size-value-momentum percentiles of stock holdings, all measured at the end of half-year $t-1$. Depending on the specification, the regression includes fund fixed-effects ((α_i)) and lagged return gap. All regressions include time fixed-effects (θ_t) and standard errors are clustered at the fund level.

We present the regression results in Table 5. Columns (1) and (4) report our baseline specification including fund characteristics and time-fixed effects. The coefficients on $Centrality_{i,t-1}$ are all positive and

statistically significant at 1% levels. Interestingly, the our main coefficients change little when we add lagged return gap in columns (2) and (5). In the remaining columns, our main coefficients remain positive and statistically significant even after the inclusion of fund fixed-effects, mitigating endogeneity concerns that the fund–centrality premium could be driven by some unobserved heterogeneity.

[Insert Table 5]

Later in Section 5, we further address endogeneity concerns that could arise, for instance, from reverse causality. By exploiting mergers of large brokerage houses as plausibly exogenous shocks to the network structure, we provide evidence supportive of our causal interpretation that institutional brokerage networks *improve* institutional trading performance. Next, we turn our attention to testing our our hypothesis that institutional brokerage networks facilitate liquidity provision and mitigate trading costs associated with adverse selection.

4.2 Inspecting the Mechanism

4.2.1 The Fund–Centrality Premium when Funds Experience Severe Redemptions

The primary prediction that we can derive from our hypothesis is that the fund–centrality premium should be more pronounced when funds' trading activities are largely driven by liquidity motives and funds can credibly signal this to their brokers. We use large outflow events to identify such periods of liquidity-motivated trading. When a mutual fund is experiencing severe redemptions, the fund is forced to liquidate a large fraction of its holdings in several stocks and their selling is, to a large extent, uninformed (see, e.g., Coval and Stafford (2007), Alexander, Cici, and Gibson (2007)). In addition, such forced liquidations are likely to send a particularly strong signal to the brokers that its sell orders are driven by liquidity reasons, rather than information motivated, thus helping the brokers communicate more credibly with other institutional clients to take the other end of the trades.

In order to test this prediction, we estimate the following linear regression model:

$$\begin{aligned} Return\ Gap_{i,t} = \delta \times Centrality_{i,t-1} \times \mathbb{1}(Outflow_{i,t} > 5\%) + \beta \times Centrality_{i,t-1} \\ + \rho \times \mathbb{1}(Outflow_{i,t} > 5\%) + \gamma \times Covariates_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned} \tag{8}$$

where $\mathbb{1}(Outflow_{i,t} > 5\%)$ is an indicator variable that is equal to 1 if fund i's outflow during half-year t exceeds five percent and the rest of the model is the same as in Equation (7). In some specifications, we include fund fixed-effects (α_i). All regressions include time fixed-effects (θ_t) and standard errors are clustered at the fund level.

We present the regression results in Table 6. The dependent variable is degree centrality in columns (1) and (2) and eigenvector centrality in columns (3) and (4). In the baseline specification without fund fixed-effects in columns (1) and (3), the coefficients on $Centrality_{i,t-1}$ and $Centrality_{i,t-1} \times \mathbb{1}(Outflow_{i,t} > 5\%)$ are all positive and statistically significant at 1% levels. These results suggest that central funds tend to outperform peripheral funds in terms of return gap during normal times, but the fund–centrality premium is more pronounced when funds are faced with large outflows.

[Insert Table 6]

Next, in columns (2) and (4), we add fund fixed-effects to our baseline specification to control for unobserved heterogeneity such as trading skills of fund managers and execution skills of trading desks. By exploiting within-fund variation in investor flows, we continue to find that the fund–centrality premium is more pronounced when funds are forced to liquidate due to large outflows. As a robustness check, we re-define a large outflow event as a half-year during which the fund's outflow exceeds ten percent, instead of five percent, and still obtain qualitatively similar results, reported in Panel A of Table A2 in the Appendix.

One potential concern is that the above results could be also consistent with cross-subsidization within a fund family: when a fund is suffering severe redemptions, another fund in the same family could step in to provide liquidity. For instance, Bhattacharya, Lee, and Pool (2013) show that affiliated funds of mutual funds that invest only in other funds within the family provide an insurance pool against temporary liquidity shocks to other funds in the family. This alternative cross-subsidization hypothesis

may seem plausible because we find that funds that belong to large families are more central and large fund families are likely better equipped to provide cross-subsidization. Nevertheless, we continue to find qualitatively similar results when we exclude funds that belong to large fund families, reported in Panel B of Table A2 in the Appendix.

Before we move on, we can further rule out another important alternative hypothesis. Many studies on brokerage connections have focused on various information channels.[24] Thus, it may seem plausible that central funds can acquire privileged information through their strong brokerage connections and trade on it. Our evidence, however, is at odds with this alternative information channel hypothesis: the fund–centrality premium is more pronounced when funds' trading activities are largely driven by liquidity reasons, rather than information motivated. We provide further evidence along this line in Section 4.2.4.

4.2.2 The Fund–Centrality Premium for Valuable Clients

Second, our liquidity provision hypothesis requires an active role on the part of brokers, such as in discerning their clients' uninformed trading motives and communicating with other institutional clients. As made clear in Carlin, Lobo, and Viswanathan (2007), whether the brokers facilitate liquidity provision or foster predatory trading is likely to hinge on the incentives they face and the strength of repeated interaction with their clients. To the extent that brokers are incentivized to maximize the expected value of future commission revenues, central funds with greater revenue generating potential for brokers are most likely to benefit from liquidity provision facilitated by their brokers. In addition, combined with our primary prediction, the effect of brokers' incentives on the fund–centrality premium should be further amplified when funds are forced to liquidate in order to accommodate severe redemptions.

In order to test these predictions, we first interact a proxy for brokers' incentives with brokerage

[24] Such information channels include, but not limited to, early access to sell-side research or tipping (Irvine, Lipson, and Puckett (2007)), invitations to broker-hosted investor conferences (Green et al. (2014)), and information leakages on company fundamentals, especially in the context of hedge funds and their prime brokers (Chung and Kang (2016), Kumar et al. (2018), Di Maggio et al. (Forthcoming))

network centrality and estimate the following linear regression model:

$$\begin{aligned} \textit{Return Gap}_{i,t} = \delta \times \textit{Centrality}_{i,t-1} \times \textit{Broker Incentive}_{i,t-1} + \beta \times \textit{Centrality}_{i,t-1} \\ + \rho \times \textit{Broker Incentive}_{i,t-1} + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned} \tag{9}$$

where $\textit{Broker Incentive}_{i,t-1}$ is our proxy for brokers' incentives as measured by fund i's aggregate dollar commissions during half-year $t-1$ and the rest of the model is the same as in Equation (7).

We present the regression results in Panel A of Table 7. In columns (1) and (2), $\textit{Broker Incentive}_{i,t-1}$ is an indicator variable that is equal to one if fund i's aggregate dollar commissions during half-year $t-1$ is greater than its top quartile value. Consistent with our prediction that brokers' incentives drive up the fund–centrality premium, we find a positive and statistically significant coefficient on $\textit{Centrality}_{i,t-1} \times \textit{Broker Incentive}_{i,t-1}$. In contrast, the coefficients on $\textit{Centrality}_{i,t-1}$ are small and statistically insignificant, suggesting that the fund–centrality premium is mostly accrued to central funds that are also likely valuable for brokers. As a robustness check in columns (3) and (4), we replace an indicator variable with its continuous counterpart, log of aggregate dollar commissions, for $\textit{Broker Incentive}_{i,t-1}$. We continue to obtain qualitatively similar, albeit somewhat weaker, results that essentially brokers' incentives drive up the fund–centrality premium.

[Insert Table 7]

Next, we add an indicator variable for contemporaneous large outflows as an additional interaction term in Equation (9) and run triple interaction regressions. We present the results in Panel B of Table 7. In all specifications, the coefficients on the triple interaction term, $\textit{Centrality}_{i,t-1} \times \textit{Broker Incentive}_{i,t-1} \times \mathbb{1}(\textit{Outflow}_{i,t} > 5\%)$, are positive and statistically significant at conventional levels. Overall, these results suggest that the effect of brokers' incentives on the fund–centrality premium is further amplified when funds' trading activities are largely driven by liquidity motives and funds can credibly signal this to their brokers.

4.2.3 The Fund–Centrality Premium for Relationship Clients

Third, our hypothesis relies on the repeated nature of interaction between institutional clients and their brokers. Institutional investors must build reputation for being truthful in order to credibly signal liquidity motives for their uninformed orders to their brokers. The brokers, in turn, must develop their reputation capital for being discreet when handling their clients' orders. Thus, the signaling and certification of uninformed trading motives is likely most effective if funds have already built strong trading relationships with their brokers.

In order to test this prediction, we interact a measure of existing trading relationships with brokerage network centrality and estimate the following linear regression model:

$$\begin{aligned} \textit{Return Gap}_{i,t} = \delta \times \textit{Centrality}_{i,t-1} \times \textit{Trading Relationship}_{i,t-1} + \beta \times \textit{Centrality}_{i,t-1} \\ + \rho \times \textit{Trading Relationship}_{i,t-1} + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned} \tag{10}$$

where $\textit{Trading Relationship}_{i,t-1}$, or simply, $\textit{Relationship}_{i,t-1}$ is our proxy for fund i's strength of trading relationships with its current set of brokers, as measured by taking the minimum of a fraction of fund i's commissions paid to its broker k during half-year $t-1$ (current) and that during $t-3$ (a year before) and then summing it over all brokers currently employed by the fund. Intuitively, $\textit{Relationship}_{i,t-1}$ measures the extent (Jaccard distance) to which fund i's current set of brokers overlap with the set of brokers the fund traded through a year before. The rest of the model is the same as in Equation (7).

We present the regression results in Panel A of Table 8. We find some evidence that trading relationships drive up the fund–centrality premium. In all specifications, the coefficients on $\textit{Centrality}_{i,t-1} \times \textit{Trading Relationship}_{i,t-1}$ are positive, but statistically significant only in columns (3) and (4) when we use eigenvector centrality. These somewhat weaker results, however, are not inconsistent with our liquidity provision hypothesis, which predicts that the fund–centrality premium is primarily driven by liquidity-motivated trades.

[Insert Table 8]

To test whether trading relationships drive up the fund–centrality premium especially in periods of

heavy liquidity-motivated trades, we add an indicator variable for contemporaneous large outflows as an additional interaction term in Equation (10) and run triple interaction regressions. We present the results in Panel B of Table 8. Consistent with our prediction, the coefficients on the triple interaction term, $Centrality_{i,t-1} \times Broker\ Relationship_{i,t-1} \times \mathbb{1}(Outflow_{i,t} > 5\%)$, are positive and statistically significant at conventional levels in all specifications. Our results are also consistent with those found in a related literature on client-dealer networks. For instance, Di Maggio, Kermani, and Song (2017) show that prior trading relationships are valuable especially in turbulent times in the OTC corporate bond market.

4.2.4 The Fund–Centrality Premium When Funds Submit Uninformed Large Orders

Our results thus far suggest that the return–gap premium associated with brokerage network centrality is more pronounced when funds' trading activities are largely driven by liquidity reasons and funds can credibly signal this to their brokers. In addition, we find that brokers' incentives and trading relationships further drive up the fund–centrality premium, corroborating our liquidity provision hypothesis. One could still argue that central funds can obtain the return–gap premium because central funds can more easily slice up large orders and spread across many brokers who can then further spread their clients' orders across many counterparties. Although not mutually exclusive with this alternative hypothesis, our liquidity provision hypothesis has clear predictions about the relation between the fund–centrality premium and the information content of trading. We provide further evidence that the fund–centrality premium is mostly concentrated in the periods that can be characterized by uninformed trading activities. We do so using an alternative measure, which puts emphasis on funds' trading volume in relation to fund flows, to identify such periods. In addition, we show that the fund–centrality premium is further amplified when the orders are also likely larger.

We identify periods of heavy information-motivated buying and selling activities following Alexander, Cici, and Gibson (2007). We calculate BF and SF metrics as follows:

$$BF_{i,t} = \frac{BUYS_{i,t} - FLOW_{i,t}}{TNA_{i,t-1}} \quad \& \quad SF_{i,t} = \frac{SELLS_{i,t} + FLOW_{i,t}}{TNA_{i,t-1}}$$

where $BUYS_{i,t}$ is fund i's dollar volume of stock purchases during half-year t, $SELLS_{i,t}$ is fund i's dollar volume of stock sales during half-year t, $FLOW_{i,t}$ is fund i's net investor flow (inflow minus outflow) during half-year t, and $TNA_{i,t-1}$ is fund i's total net assets at the end of half-year $t-1$. Exploiting within-fund variation in BF and SF metrics, Alexander, Cici, and Gibson (2007) show that buy (sell) portfolios with high BF (SF) tend to outperform buy (sell) portfolios with low BF (SF). Intuitively, trading against investor flows is likely motivated by superior private information, whereas trading with flows is likely driven by liquidity reasons, i.e., scaling up to accommodate inflows and scaling down to accommodate outflows (see also Coval and Stafford (2007)).

Since heavy informed buying activities do not necessarily coincide with heavy informed selling activities, we assign half-years in which both BF and SF fall below its respective top quartile value as periods of uninformed trading (or at least less informed trading). We interact an indicator variable for period of uninformed trading with brokerage network centrality and estimate the following linear regression model:

$$\begin{aligned} Return\ Gap_{i,t} = \delta \times Centrality_{i,t-1} \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3) + \beta \times Centrality_{i,t-1} \\ + \rho \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3) + \gamma \times Covariates_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned} \tag{11}$$

where $\mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3)$ is an indicator variable that is equal to 1 if both $BF_{i,t}$ and $SF_{i,t}$ fall below its respective top quartile value during half-year t and the rest of the model is the same as in Equation (7).

We present the regression results in Panel A of Table 9. We find that the coefficients on $Centrality_{i,t-1} \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3)$ are positive and statistically significant at 1% and 5% levels, whereas the coefficients on $Centrality_{i,t-1}$ are small and statistically insignificant. These results are consistent with our main results based on large outflow events and further suggest that the fund–centrality premium is associated with trading motives and mostly concentrated in periods of uninformed trading, i.e., when funds are trading *with flows*, rather than *against flows*.

[Insert Table 9]

Next, we proxy for average order sizes using average trade sizes inferred from consecutive portfolio

disclosures, adjusting for trading volume in the market as follows:

$$\overline{Trade\ Size}_{i,t} = \frac{1}{N_{i,t}} \sum_{k} \frac{\mid Shares_{i,k,t} - Shares_{i,k,t-1} \mid}{\overline{VOL}_{k,t}^{CRSP}} \tag{12}$$

where $Shares_{i,k,t}$ is the split-adjusted number of shares held in stock k by fund i at the end of half-year (or quarter) t, $\overline{VOL}_{k,t}^{CRSP}$ is the average CRSP monthly volume between portfolio disclosures, and the averages are taken over stocks for which $Shares_{i,k,t} \neq Shares_{i,k,t-1}$. To arrive at the semi-annual figure, we take the average of quarterly numbers, if two quarterly observations are available.

In order to test whether the fund–centrality premium is more pronounced when funds submit *large* uninformed orders, we add in Equation (11) an additional interaction term, $\mathbb{1}(\overline{\text{Trade Size}} > Q_3)$, which is an indicator variable that is equal to 1 if $\overline{Trade\ Size}_{i,t}$ is above its top quartile value. We present the triple interaction results in Panel B of of Table 9. In columns (1) and (3), the coefficients on the triple interaction term, $Centrality_{i,t-1} \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3) \times \mathbb{1}(\overline{Trade\ Size}_{i,t} > Q_3)$, are positive and significant at 5% and 10% levels. In contrast, the coefficients on $Centrality_{i,t-1} \times \mathbb{1}(\overline{Trade\ Size}_{i,t} > Q_3)$ are small and statistically insignificant. These results suggest that central funds can obtain the return–gap premium when central funds submit large *uninformed* orders. As a robustness check in columns (3) and (4), we replace an indicator variable with its continuous counterpart, $\overline{Trade\ Size}_{i,t}$ and continue to obtain qualitatively similar results. Overall, our results suggest that the fund–centrality premium is more pronounced when the trading orders are larger, but only when the trades are likely motivated by liquidity reasons. These results are largely consistent with our hypothesis that institutional brokerage networks facilitate liquidity provision and mitigate trading costs associated with adverse selection.

5 A Natural Experiment

We recognize that our results are not completely free from endogeneity concerns that could be derived from, for instance, reverse causality. Hence, we conduct a natural experiment to provide evidence supportive of our causal interpretation that institutional brokerage networks *improve* institutional trading

performance. To accomplish this, we exploit mergers of large brokerage houses as plausibly exogenous shocks to the network structure.

5.1 Backgrounds on Brokerage Mergers and Identification

Following Hong and Kacperczyk (2010), we identify mergers among brokerage houses by relying on information from the SDC Mergers and Acquisition database. We choose all the mergers that the acquiring broker belongs to the four-digit SIC code 6211 ("investment Commodity Firms, Dealers, and Exchanges"). Next, we manually match brokerage mergers identified in the SDC data using broker names and narrow down to the mergers in which broker names show up in at least 100 N-SAR filings.[25] This process gives rise to twenty six brokerage mergers during the period from 1995 to 2015. Table 10 lists all twenty six brokerage mergers. The table also reports average broker shares before (from18 months to 6 months) and after (from 6 months to 18 months) the merger and changes in average broker shares around the merger.

[Insert Table 10]

The shock strength, however, is a major concern for our natural experiment, given the complexity of the network structure (which typically consists of thousands of nodes connected by tens of thousands edge). Moderate-sized brokerage mergers, especially as stand-alone events (which amounts to cutting a small number of edges connected to a single node) are unlikely to have an economically meaningful impact on the entire structure of institutional brokerage networks. Therefore, we focus on two waves of five largest mergers of institutional brokerage houses that took place around 2000 and 2008, in which more than ten percent of edges were served.[26]

Figure 3 plots the changes in average broker shares around each of these mergers. A visual inspection suggests that these five mergers were likely to have a meaningful impact on institutional brokerage networks. Specifically, the average brokerage shares of the acquired brokers dramatically decreased following the

[25] Our N–SAR sample period runs from 1994 to 2016. But we exclude the first and last years to facilitate a difference-in-differences (DiD) analysis around the merger.

[26] These five brokerage mergers include CSFB's acquisition of DLJ and UBS'acquistion of PaineWebber in 2000 and JP Morgan Chase's acquisition of Bear Stearns, Barclay's acquisition of Lehman Brothers, and Bank of America's acquisition of Merrill Lynch in 2008.

merger in all cases, whereas those of the acquiring brokers increased notably after the merger except for the case of Bank of America. For instance, mutual funds on average paid about 4.02 % of its brokerage commissions to CSFB as one of the top 10 brokers, while the figure for DLJ was 4.40%. After the merger, CSFB's average broker shares increased to 6.40%. One notable exception is Bank of America's acquisition of Merrill Lynch. After the merger, the merged firm's brokerage services were carried out under the name of Merrill Lynch for a while and thus reported as such in N–SAR reports.

[Insert Figure 3]

5.2 Empirical Design and Results

Our analysis of the causal effect of mutual funds' brokerage network centrality on their trading performance exploits large brokerage mergers in a quasi-natural experiment setting to overcome potential concerns about endogeneous network formation. As stated earlier, we exploit two waves of five largest mergers of brokerage houses and the empirical methodology of our analysis is a difference-in-differences (DiD). In a standard DiD approach, the sample needs to be divided into treatment and control groups. Here comes another challenge for our natural experiment: the treatment of shock is *a priori* unclear. Nevertheless, we can reason that mutual funds that traded largely through the acquiring brokers but not heavily through the target (acquired) brokers are most likely to benefit from exogenous shocks to the network, since the acquiring broker would retain at least some of the target broker's clients.

Building on this intuition, we construct hypothetical post-merger brokerage network centrality under a fairly conservative assumption. Specifically, we assume that funds who had relationships with a target broker before the merger were to simply redistribute commissions to their existing brokers on a pro-rata basis following the merger.[27] Then, we proceed by calculating the expected change in brokerage network

[27] In particular, we re-scale each mutual fund's normalized commission payment vector $(g_{i,\cdot})$ a half-year prior to the merger event window, denoted $\tilde{g}_{i,\cdot}$, as follows:

$$\tilde{g}_{i,k} = \begin{cases} 0 & \text{if } k \in S; \\ \frac{g_{i,k}}{\sum_{k \notin S} g_{i,k}} & \text{if } k \notin S. \end{cases} \tag{13}$$

where i indexs funds, k indexs brokers, and S denotes set of acquired brokers. For instance, if a mutual fund i hired

centrality as the difference between the hypothetical post-merger centrality and the actual pre-merger centrality around the merger event. Under this assumption, the funds that did not trade through the target broker (candidate treated funds) do not change their brokerage relationships, as they don't need to, but nonetheless experience exogenous increases in brokerage network centrality after the merger, because *other* funds need to reconfigure their brokerage relationships. We form the treatment group by choosing the top ten percent of mutual funds sorted based on the expected change in brokerage network centrality.

Our empirical methodology also requires that we specify the event window around the mergers. In general, most event studies focus on a very narrow window because choosing a window that is too long may include irrelevant information with the focused events (Hong and Kacperczyk (2010)). However, a window that is too short would result in the loss of many observations containing relevant information and we thus choose a relatively longer time window than other event studies. Specifically, we examine one year before and one year after the event window of brokerage mergers. Figure 4 illustrates the event timelines for our natural experiment.

[Insert Figure 4]

If we denote the average outcome variables in the treatment (T) and control (C) groups in the pre- and post-event periods by $O_{T,1}$, $O_{T,2}$, $O_{C,1}$, and $O_{C,2}$, respectively, the partial effect of change due to the merger can be estimated as

$$DiD = (O_{T,2} - O_{T,1}) - (O_{C,2} - O_{C,1}). \tag{14}$$

A potential concern with the above estimation is that the results could be affected by fund characteristics. In other words, if the funds in the treatment and control groups have different fund characteristics, then those characteristics could potentially bias our results. To resolve this concern, we use a matching technique. As mentioned earlier, we assign top ten percent of funds with the largest expected change in brokerage network centrality as the treatment group. Among the remaining 90% of the sample, we construct the control group by matching on pre-treatment (pre-event) outcome variables and all fund characteristics

broker A, B, C, and D and its $g_i = \begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.2 \end{bmatrix}$ and C is an acquired broker, then $\tilde{g}_i = \begin{bmatrix} \frac{0.1}{0.6} & \frac{0.3}{0.6} & 0 & \frac{0.2}{0.6} \end{bmatrix}$.

used in our previous analyses except for *log(Family TNA)*[28] following *Genetic Matching* algorithm proposed by Diamond and Sekhon (2013). Matching on observable pre-event fund characteristics and pre-treatment outcome variables can remove (at least to a degree) common influences of fund characteristics that could affect return gap other than changes in brokerage network centrality.

Table 11 reports the results of our matching using *Degree Centrality* as our measure of brokerage network centrality. As seen in the table, some of the variables are remarkably different before matching, but those differences largely disappear after matching. Panel A presents the matching balance results for the brokerage mergers in 2000. Before matching, *Degree Centrality* is significantly different at the 1% level, i.e., the treated funds were more central to begin with. In addition, among covariates, *Expense Ratio*, *Size Percentile*, and *Value Percentile* are significantly different at the conventional levels. Our matching appears successful and all p-values for post-matching differences in means are above 10%, with the smallest p-value of 0.18. Panel B presents the matching balance results for the brokerage mergers in 2008. Before matching, *Degree Centrality* is also significantly different at the 1% level and several covariates including *log(Fund TNA)*, *Expense Ratio*, *Commission Rate*, and *Trade Volume* are also significantly different at the conventional levels. Again, the matching appears similarly successful.

[Insert Table 11]

To be consistent with our causal interpretation, if the brokerage mergers had indeed served as positive exogenous shocks to the brokerage network centrality of mutual funds in the treatment group, then the return gap of treated funds would have experienced significant increases relative to that of the control group of funds following the mergers.

Table 12 presents our DiD results. Panel A shows the results of our DiD analysis of the brokerage mergers in 2000. The average *Degree Centrality* of the treatment group increased from 0.206 to 0.235, while the average *Degree Centrality* of the matched control group only increased from 0.205 to 0.222. Thus, we observe a discernible increase in *Degree Centrality* of 0.013, using a DiD estimator. This effect

[28] It turns out that it is very difficult to match on fund family size and *all* the other fund characteristics including pre-event outcome variables.

is statistically significant at the 5% level. Moreover, the average *Return Gap* also substantially increased around the mergers in 2000 by 9.3 basis points per month relative to a control group of funds, significant at the 10% level. Similarly, Panel B presents the results of our DiD analysis of the brokerage mergers in 2008. We similarly observe a discernible increase in *Degree Centrality* by 0.034, using a DiD estimator, significant at the 1% level. At the same time, the average *Return Gap* of the treated funds also substantially increased by 6.8 basis points per month relative to a control group of funds, significant at the 10% level. In sum, the DiD results indicate that exogenous changes in brokerage network centrality due to large brokerage mergers are accompanied by predicted changes in return gap performance.

[Insert Table 12]

As a robustness check, we re-do our DiD analysis with *Eigenvector Centrality* instead of *Degree Centrality*. We obtain qualitatively similar results, as reported in Table A3 and Table A4. To sum up, positive changes in brokerage network centrality as a result of exogenous shocks to the brokerage network are accompanied by positive changes in return gap. These results are consistent with our causal interpretation that institutional brokerage networks *improve* institutional trading performance.

6 Conclusion

Using a unique dataset on brokerage commission payments for a comprehensive sample of mutual funds, we map trading networks of mutual funds and their brokers as affiliation networks in which mutual funds are connected through their overlapping brokerage relationships. Mutual funds that trade through many brokers that many other funds also trade through are central in the network. We find that central funds outperform peripheral ones, especially in terms of return gap. In order to shed light on the specific mechanisms behind the return–gap premium associated with brokerage network centrality (simply the fund–centrality premium), we propose a liquidity provision hypothesis.

Suppose, for instance, that a mutual fund faced with an extreme fund outflow is forced to sell large blocks of its holdings in several stocks at the same time. The sell orders would tend to be submitted to

brokers with which the fund has strong relationships and that could infer the underlying liquidity reasons for the orders. The brokers, in turn, may be likely to turn to other institutional clients with whom they have strong relationships to absorb the orders while communicating the likely liquidity motives for the trades to ease their concerns about trading against better informed traders. Thus, central funds are better positioned to tap into larger pools of unexpressed liquidity, especially when submitting large blocks of liquidity-motivated orders.

Consistent with our liquidity provision hypothesis, we find that the fund–centrality premium is more pronounced when funds' trading activities are largely driven by liquidity motives, such as to accommodate large fund outflows. We also find that the fund–centrality premium is further driven up by brokers' incentives to generate greater commission revenues and by trading relationships that funds have established with their brokers. Exploiting large brokerage mergers as plausibly exogenous shocks to the network structure, we provide evidence supportive of our causal interpretation that institutional brokerage networks improve institutional trading performance.

References

Admati, A. R., and P. Pfleiderer. 1991. Sunshine Trading and Financial Market Equilibrium. *Review of Financial Studies* 4:443–481.

Ahern, K. R. 2013. Network Centrality and the Cross Section of Stock Returns. *University of Southern California Working Paper* pp. 1–51.

Alexander, G. J., G. Cici, and S. Gibson. 2007. Does Motivation Matter when Assessing Trade Performance? An Analysis of Mutual Funds. *Review of Financial Studies* 20:125–150.

Anand, A., P. Irvine, A. Puckett, and K. Venkataraman. 2012. Performance of Institutional Trading Desks: An Analysis of Persistence in Trading Costs. *Review of Financial Studies* 25:557–598.

Barbon, A., M. Di Maggio, F. A. Franzoni, and A. Landier. Forthcoming. Brokers and Order Flow Leakage: Evidence from Fire Sales. *Journal of Finance* .

Bessembinder, H., A. Carrion, L. Tuttle, and K. Venkataraman. 2016. Liquidity, resiliency and market quality around predictable trades: Theory and evidence. *Journal of Financial Economics* 121:142–166.

Bessembinder, H., C. S. Spatt, and K. Venkataraman. Forthcoming. A Survey of the Microstructure of Fixed-Income Markets. *Journal of Financial and Quantitative Analysis* .

Bessembinder, H., and K. Venkataraman. 2004. Does an Electronic Stock Exchange Need an Upstairs Market? *Journal of Financial Economics* 73:3–36.

Bhattacharya, U., J. H. Lee, and V. K. Pool. 2013. Conflicting Family Values in Mutual Fund Families. *Journal of Finance* 68:173–200.

Bonacich, P. 1972. Factoring and Weighting Approaches to Status Scores and Clique Identification. *The Journal of Mathematical Sociology* 2:113–120.

Bonacich, P. 1987. Power and Centrality: A Family of Measures. *American Journal of Sociology* 92:1170–1182.

Booth, G. G., J.-C. Lin, T. Martikainen, and Y. Tse. 2002. Trading and Pricing in Upstairs and Downstairs Stock Markets. *Review of Financial Studies* 15:1111–1135.

Borgatti, S. P. 2005. Centrality and Network Flow. *Social Networks* 27:55–71.

Brunnermeier, M. K., and L. H. Pedersen. 2005. Predatory Trading. *Journal of Finance* 60:1825–1863.

Carlin, B. I., M. S. Lobo, and S. Viswanathan. 2007. Episodic Liquidity Crises: Cooperative and Predatory Trading. *Journal of Finance* 62:2235–2274.

Chung, J.-W., and B. U. Kang. 2016. Prime Broker-level Comovement in Hedge Fund Returns: Information or Contagion? *Review of Financial Studies* 29:3321–3353.

Coval, J., and E. Stafford. 2007. Asset Fire Sales (and Purchases) in Equity Markets. *Journal of Financial Economics* 86:479–512.

Di Maggio, M., F. A. Franzoni, A. Kermani, and C. Sommavilla. Forthcoming. The Relevance of Broker Networks for Information Diffusion in the Stock Market. *Journal of Financial Economics* .

Di Maggio, M., A. Kermani, and Z. Song. 2017. The value of trading relations in turbulent times. *Journal of Financial Economics* 124:266–284.

Diamond, A., and J. S. Sekhon. 2013. Genetic Matching for Estimating Causal Effects. *The Review of Economics and Statistics* 95:932–945.

Edelen, R. M. 1999. Investor Flows and the Assessed Performance of Open-end Mutual Funds. *Journal of Financial Economics* 53:439–466.

Edelen, R. M., R. B. Evans, and G. B. Kadlec. 2012. Disclosure and agency conflict: Evidence from mutual fund commission bundling. *Journal of Financial Economics* 103:308–326.

Freeman, L. C. 1979. Centrality in Social Networks Conceptual Clarification. *Social Networks* 1:215–239.

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

Goldstein, M. A., P. Irvine, E. Kandel, and Z. Wiener. 2009. Brokerage Commissions and Institutional Trading Patterns. *Review of Financial Studies* 22:5175–5212.

Green, T. C., R. Jame, S. Markov, and M. Subasi. 2014. Broker-hosted Investor Conferences. *Journal of Accounting and Economics* 58:142–166.

Griffin, J. M., T. Shu, and S. Topaloglu. 2012. Examining the Dark Side of Financial Markets: Do Institutions Trade on Information from Investment Bank Connections? *Review of Financial Studies* 25:2155–2188.

Grinblatt, M., and S. Titman. 1989. Mutual Fund Performance: An Analysis of Quarterly Portfolio Holding. *Journal of Business* 62:393–416.

Grossman, S. J. 1992. The Informational Role of Upstairs and Downstairs Trading. *Journal of Business* 65:509–528.

Harris, L. 2002. *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford University Press.

Hendershott, T., D. Li, D. Livdan, and N. Schürhoff. 2017. Relationship Trading in OTC Markets. *Swiss Finance Institute Research Paper No. 17-30* .

Hong, H., and M. Kacperczyk. 2010. Competition and Bias. *Quarterly Journal of Economics* 125:1683–1725.

Irvine, P., M. Lipson, and A. Puckett. 2007. Tipping. *Review of Financial Studies* 20:741–768.

Kacperczyk, M., C. Sialm, and L. Zheng. 2008. Unobserved Actions of Mutual Funds. *Review of Financial Studies* 21:2379–2416.

Kumar, N., K. Mullally, S. Ray, and Y. Tang. 2018. Prime (Information) Brokerage. *Working Paper, University of Florida* .

Kyle, A. S. 1985. Continuous Auctions and Insider Trading. *Econometrica* 53:1315–1335.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10:707–701.

Li, D., and N. Schürhoff. Forthcoming. Dealer Networks. *Journal of Finance* .

Madhavan, A., and M. Cheng. 1997. In Search of Liquidity: Block Trades in the Upstairs and and Downstairs. *Review of Financial Studies* 10:175–203.

Ozsoylev, H. N., J. Walden, M. D. Yavuz, and R. Bildik. 2014. Investor Networks in the Stock Market. *Review of Financial Studies* 27:1323–1366.

Puckett, A., and X. S. Yan. 2011. The Interim Trading Skills of Institutional Investors. *Journal of Finance* 66:601–633.

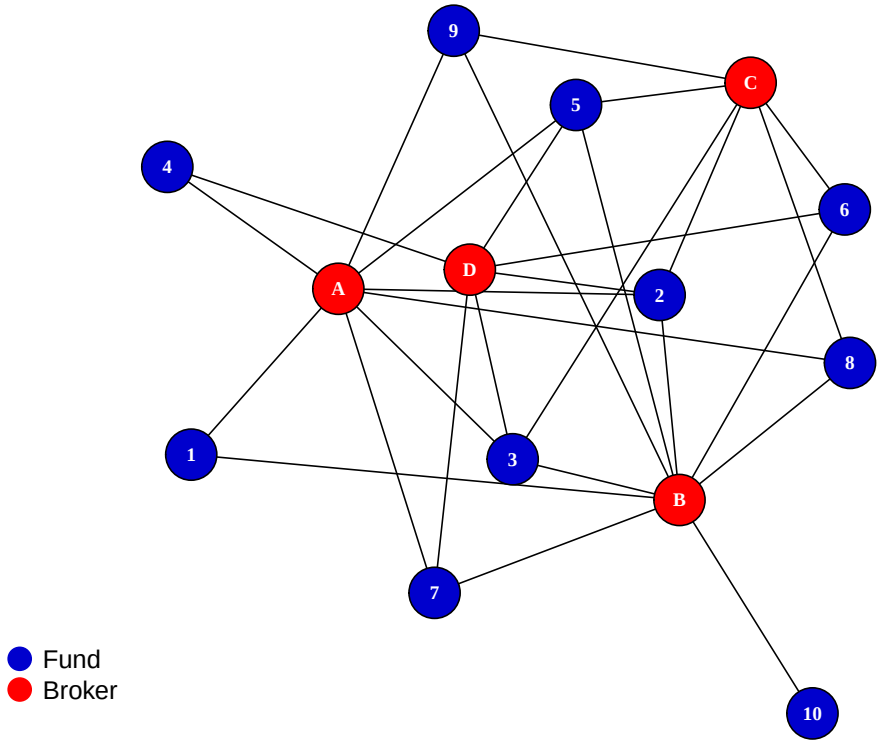
Reuter, J. 2006. Are IPO Allocations for Sale? Evidence from Mutual Funds. *Journal of Finance* 61:2289–2324.

Seppi, D. J. 1990. Equilibrium Block Trading and Asymmetric Information. *Journal of Finance* 45:73–94.

Smith, B. F., D. A. S. Turnbull, and R. W. White. 2001. Upstairs Market for Principal and Agency Trades: Analysis of Adverse Information and Price Effects. *Journal of Finance* 56:1723 –1746.

Yang, L., and H. Zhu. Forthcoming. Back-Running: Seeking and Hiding Fundamental Information in Order Flows. *Review of Financial Studies* .

Panel A: Graphical representation



Panel B: Extended-form representation

$$G^T = \begin{array}{c} \\ \text{Broker} \\ \\ \end{array} \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \text{Fund} \\ \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \left[\begin{array}{cccccccccc} 0.85 & 0.20 & 0.15 & 0.50 & 0.20 & 0.00 & 0.20 & 0.15 & 0.50 & 0.00 \\ 0.15 & 0.33 & 0.40 & 0.00 & 0.17 & 0.25 & 0.50 & 0.50 & 0.10 & 1.00 \\ 0.00 & 0.22 & 0.10 & 0.00 & 0.28 & 0.40 & 0.00 & 0.35 & 0.40 & 0.00 \\ 0.00 & 0.25 & 0.35 & 0.50 & 0.35 & 0.35 & 0.30 & 0.00 & 0.00 & 0.00 \end{array}\right] \end{array}$$

Panel C: Reduced-form representation

$$A = \begin{array}{c} \\ \text{Fund} \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \text{Fund} \\ \begin{bmatrix} 0.00 & 0.35 & 0.30 & 0.50 & 0.35 & 0.15 & 0.35 & 0.30 & 0.60 & 0.15 \\ 0.35 & 0.00 & 0.83 & 0.45 & 0.84 & 0.72 & 0.78 & 0.70 & 0.52 & 0.33 \\ 0.30 & 0.83 & 0.00 & 0.50 & 0.77 & 0.70 & 0.85 & 0.65 & 0.35 & 0.40 \\ 0.50 & 0.45 & 0.50 & 0.00 & 0.55 & 0.35 & 0.50 & 0.15 & 0.50 & 0.00 \\ 0.35 & 0.84 & 0.77 & 0.55 & 0.00 & 0.80 & 0.67 & 0.60 & 0.58 & 0.17 \\ 0.15 & 0.72 & 0.70 & 0.35 & 0.80 & 0.00 & 0.55 & 0.60 & 0.50 & 0.25 \\ 0.35 & 0.78 & 0.85 & 0.50 & 0.67 & 0.55 & 0.00 & 0.65 & 0.30 & 0.50 \\ 0.30 & 0.70 & 0.65 & 0.15 & 0.60 & 0.60 & 0.65 & 0.00 & 0.60 & 0.50 \\ 0.60 & 0.52 & 0.35 & 0.50 & 0.58 & 0.50 & 0.30 & 0.60 & 0.00 & 0.10 \\ 0.15 & 0.33 & 0.40 & 0.00 & 0.17 & 0.25 & 0.50 & 0.50 & 0.10 & 0.00 \end{bmatrix} \end{array}$$

Panel D: Brokerage network centrality

Fund	1	2	3	4	5	6	7	8	9	10
Degree Centrality	0.339	0.613	0.594	0.389	0.592	0.513	0.572	0.528	0.450	0.267
Eigenvector Centrality	0.550	1.000	0.972	0.654	0.974	0.870	0.928	0.864	0.730	0.468

Figure 1: Institutional Brokerage Networks: A Toy Example

This figure illustrates how we construct institutional brokerage networks and calculate brokerage network centrality using a simple example network consisting of ten funds and four brokers.

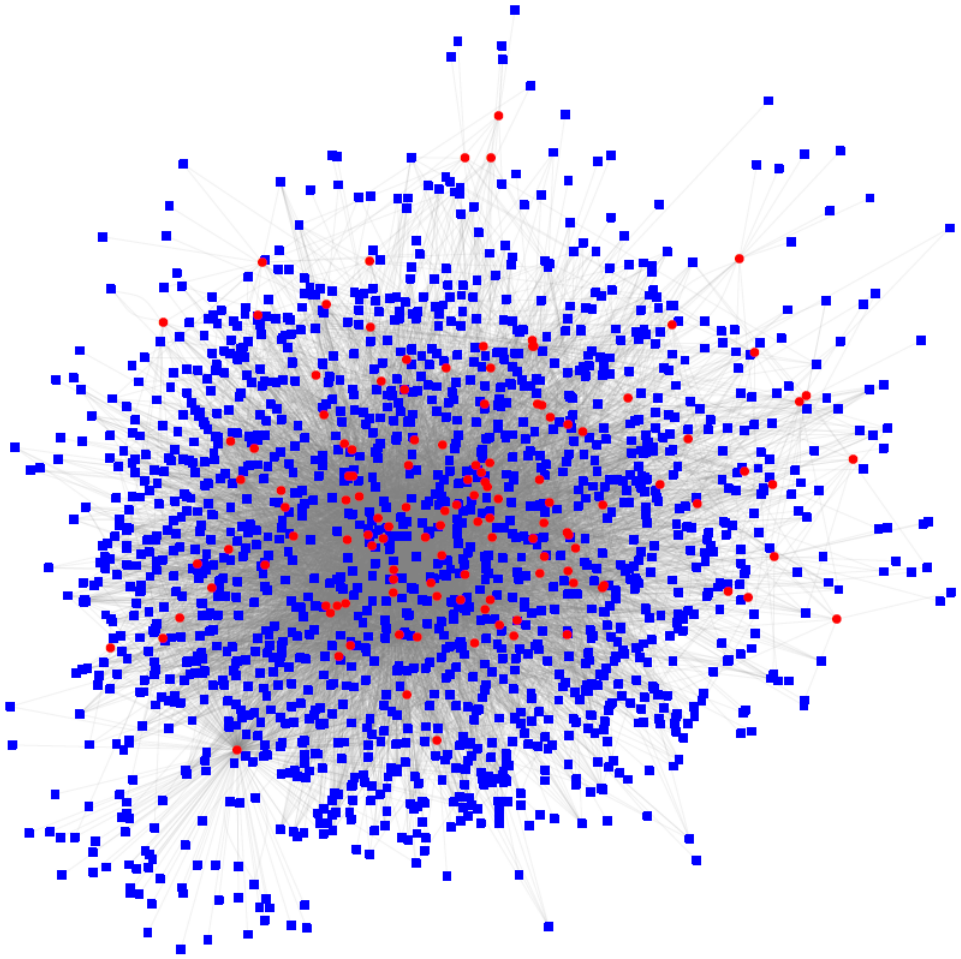
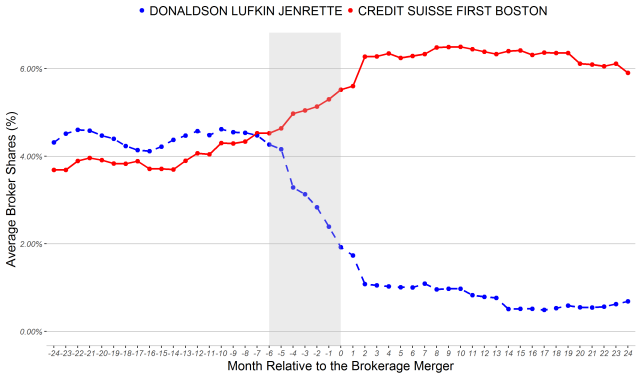
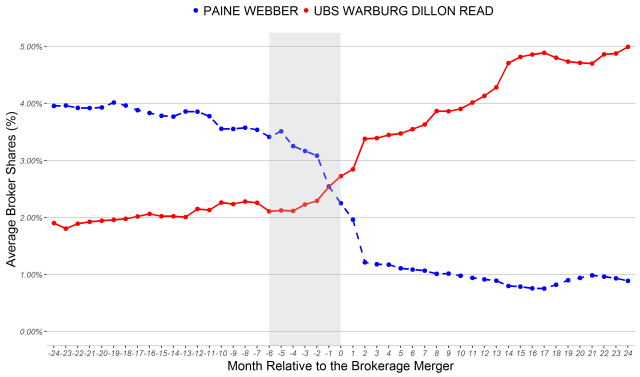


Figure 2: Institutional Brokerage Networks

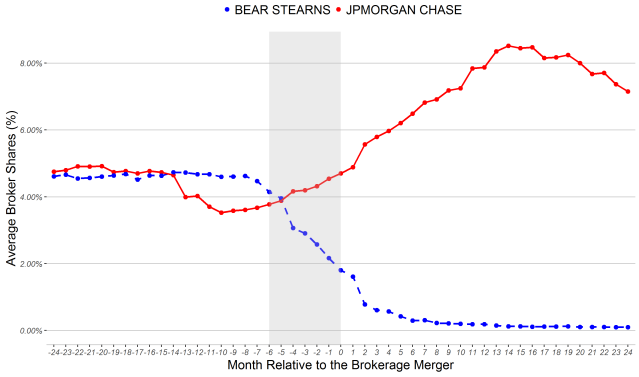
This figure shows a snapshot of institutional brokerage networks at the end of June 2016. Blue nodes represent mutual funds, red nodes represent institutional brokers, and lines represent connections between mutual funds and their brokers.



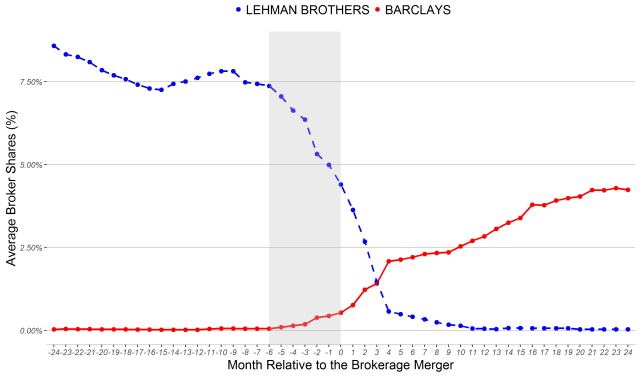
(a) CSFB's acquisition of DLJ



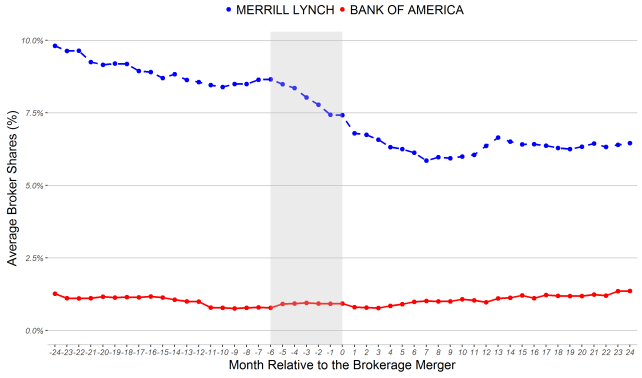
(b) UBS Warburg Dillon Read's acquisition of Paine Webber



(c) JP Morgan Chase's acquisition of Bear Stearns



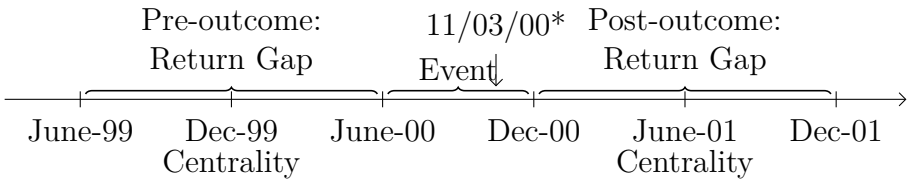
(d) Barclay's acquisition of Lehman Brothers



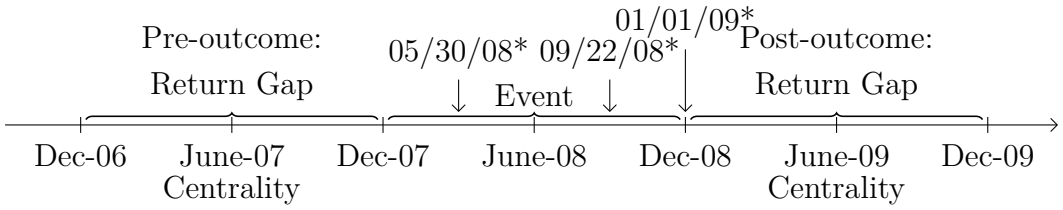
(e) Bank of America's acquisition of Merrill Lynch

Figure 3: Average Brokerage Share around Brokerage Merger

This figure shows changes in average broker shares for the acquiring brokers and target brokers around the mergers. A broker share is defined as a fraction of the commission payments to the given broker by the fund and broker shares are averaged across funds each month on a rolling basis around each of the following mergers: Credit Suisse First Boston (CSFB)'s acquisition of Donaldson Lufkin Jenrette (DLJ) in 2000 (a); UBS Warburg Dillon Read's acquisition of Paine Webber in 2000 (b); JP Morgan Chase's acquisition of Bear Stearns in 2008 (c); Barclays's acquisition of Lehman Brothers in 2008 (d); and Bank of America's acquisition of Merrill Lynch in 2008 (e).



(a) Event Timeline of Brokerage Mergers in 2000



(b) Event Timeline of Brokerage Mergers in 2008

Figure 4: Event Timeline of Brokerage Mergers

Figure 4a depicts the event timeline of the 2000 mergers: Credit Suisse First Boston (CSFB)'s acquisition of Donaldson Lufkin Jenrette (DLJ) and UBS Warburg Dillon Read's acquisition of Paine Webber in 2000. The effective date of both mergers is November 3rd, 2000. We set the second half of 2000 as the event window.
Figure 4b depicts the event timeline of the 2008 mergers: JP Morgan Chase's acquisition of Bear Stearns, Barclays's acquisition of Lehman Brothers, and Bank of America's acquisition of Merrill Lynch in 2008. The effective dates are May 30th, 2008, September 22nd, 2008, and January 1st, 2009, respectively. We set the entire year of 2008 as the event window.
* Effective date is as reported by SDC Platinum Financial Securities Data.

Table 1: Brokerage Commission Payments: Example and Descriptions

This table provides an example of and some descriptive statistics on brokerage commission payments. N-SAR filings report brokerage commissions paid to the 10 brokers that received the largest amount (Item 20) from the fund and the aggregate brokerage commission payments (Item 21). Panel A provides an example for T. Rowe Price Blue Chip Growth Fund for the period ending in June 30, 2016. Panel B reports the concentration level of brokerage commissions for the top 1, 3, 5, 7, and 10 brokers to which the fund paid the largest amount. Panel C reports the transition matrix of year-to-year changes in the broker rankings for the fund by the amount of commission payments.

Panel A: Example: T ROWE PRICE BLUE CHIP GROWTH FUND (CIK = 902259), June 30, 2016

Item 20	Name of Broker	IRS Number	Commisions ($000)
1	BANK OF AMERICA MERRILL LYNCH	13-5674085	415
2	JPMORGAN CHASE	13-4994650	292
3	MORGAN STANLEY CO INC	13-2655998	252
4	DEUTSCHE BANK SECURITIES	13-2730828	207
5	RBC CAPITAL MARKETS	41-1416330	159
6	CITIGROUP GLOBAL MARKETS INC	11-2418191	157
7	CS FIRST BOSTON	13-5659485	153
8	BAIRD ROBERT W	39-6037917	148
9	GOLDMAN SACHS	13-5108880	144
10	SANFORD C BERNSTEIN	13-2625874	115
Item 21	Aggregate Brokerage Commissions ($000)		3107

Panel B: Concentration of Brokerage Commissions

Broker Share (%)	Mean	St. Dev.	Pctl(1)	Pctl(25)	Median	Pctl(75)	Pctl(99)
Top 1 Broker	25.65	22.21	5.53	11.54	16.88	30.00	100.00
Top 1–3 Brokers	45.24	23.95	13.02	27.59	37.44	56.76	100.00
Top 1–5 Brokers	56.60	22.53	19.00	39.26	51.17	71.13	100.00
Top 1–7 Brokers	64.47	20.91	23.80	48.25	61.32	80.63	100.00
Top 1–10 Brokers	72.45	18.91	29.30	58.08	71.62	88.89	100.00

Panel C: Persistence in Brokerage Relationship (Transition Matrix)

Probability (%)	Next Year									
Current Year	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
Top 1	46.74	20.55	13.35	10.06	7.57	6.44	5.41	4.99	4.41	4.00
Top 2	17.37	23.69	17.29	13.03	10.96	8.89	7.58	6.63	6.30	5.65
Top 3	10.71	15.83	17.64	14.52	12.34	10.93	9.61	8.53	7.14	7.62
Top 4	7.17	11.24	13.33	15.12	13.31	11.82	10.10	9.59	9.42	8.47
Top 5	5.31	8.09	10.53	12.73	13.83	12.84	12.16	10.54	10.10	9.67
Top 6	4.00	6.47	8.87	10.49	11.84	13.00	12.91	12.12	10.78	10.67
Top 7	3.12	5.30	6.65	8.18	10.38	11.67	12.77	13.11	12.71	11.21
Top 8	2.41	3.60	5.00	6.56	8.02	9.93	11.72	13.73	13.70	13.53
Top 9	1.85	2.86	4.06	5.12	6.15	8.22	10.22	11.28	14.08	13.93
Top 10	1.32	2.36	3.28	4.19	5.60	6.25	7.52	9.48	11.35	15.26

Table 2: Summary Statistics

This table reports the summary statistics on degree centrality (Freeman (1979)), eigenvector centrality (Bonacich (1972, 1987)), and other fund-level characteristics over the period from the first half of 1994 through the first half of 2016. The fund TNA (Item 74.T) and an indicator for an index fund (Item 69) are directly taken from N-SAR filings and we use the family code reported by the fund (Item 19.C) to calculate the family TNA. The fund trading volume is calculated as the sum of purchases (Item 71.A) and sales (Item 71.B). Since brokerage commissions are reported at the registrant level, we calculate the commission rate as a ratio of the aggregate commission payments (Item 21) to the sum of all trading volumes across equity-oriented funds within the same registrant. We estimate the fund's commission payments as the product of the commission rate and the fund trading volume. The expense ratio is from CRSP and we calculate monthly net flows for each fund share class i during month t as follows: $FLOW_{i,t} = TNA_{i,t} - TNA_{i,t-1} \times (1 + R_{i,t})$ where $FLOW_{i,t}$ is the dollar value of fund flow (net new issues and redemptions), $TNA_{i,t}$ is the total net asset, and $R_{i,t}$ is the monthly return. To compute the monthly fund flow for the fund, we sum monthly fund flows for all share classes belonging to the same fund as identified by MFLINKS. Monthly fund flows are summed over the half-year to calculate the semi-annual fund flow. We scale the semi-annual fund flows by the beginning-of-period TNA. For each fund-quarter, size, value, and momentum percentiles are calculated as percentiles of market capitalization, book-to-market ratio, and 12-month returns skipping the most recent month, respectively, averaged across all stock holdings. For each fund-halfyear, we take the most recent quarterly observation of average size-value-momentum percentiles.

Variable	Obs.	Mean	St. Dev.	Q_1	Median	Q_3
Degree Centrality	54,331	0.16	0.08	0.10	0.16	0.21
Eigenvector Centrality	54,331	0.53	0.25	0.33	0.57	0.73
Return Gap (%)	54,331	−0.03	0.39	−0.20	−0.02	0.14
Fund TNA ($billion)	54,331	1.42	3.43	0.08	0.29	1.09
Family TNA ($billion)	54,331	122.09	277.91	2.97	20.50	79.80
Expense Ratio (%)	54,331	1.13	0.42	0.92	1.11	1.35
Commission Rate (%)	54,331	0.12	0.13	0.06	0.09	0.14
Trade Volume, as % of TNA	54,331	86.32	80.10	34.81	63.84	108.64
$\mathbb{1}$(Index Fund)	54,331	0.10	0.30	0	0	0
Size Percentile	54,331	84.97	12.60	76.87	89.52	95.03
Value Percentile	54,331	37.46	11.92	28.02	37.21	46.15
Momentum Percentile	54,331	57.69	9.56	51.55	57.07	63.46
Fund Flow, as % of TNA	54,331	1.98	22.44	−8.01	−2.25	5.71

Table 3: Determinants of Mutual Funds' Brokerage Network Centrality

This table presents the results of regressing degree centrality and eigenvector centrality on contemporaneous fund-level characteristics including log(fund TNA), log(family TNA), expense ratio, commission ratio, trading volume, and size-value-momentum percentiles. The details on the fund-level variables are reported in Table 2. The standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Dependent variable:	Degree Centrality ×100					Eigenvector Centrality ×100				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Constant	2.34***	−2.54**				−1.32	−21.41***			
	(6.18)	(−2.33)				(−1.08)	(−5.86)			
log(Fund TNA)	0.09	0.25***	0.14**	0.20***	0.21***	0.64***	0.90***	0.58***	0.85***	0.83***
	(1.53)	(4.12)	(2.53)	(3.11)	(3.13)	(3.19)	(4.50)	(2.93)	(3.72)	(3.59)
log(Family TNA)	1.36***	1.44***	1.62***	0.60***	0.59***	5.21***	5.33***	5.78***	1.92***	1.91***
	(29.57)	(32.21)	(36.92)	(8.02)	(7.95)	(34.98)	(36.49)	(38.31)	(7.53)	(7.44)
Expense Ratio (%)		1.35***			−0.22		1.52*			−1.76**
		(5.54)			(−0.89)		(1.90)			(−2.07)
Commission Rate (%)		3.94***			−0.82***		6.21***			−1.72*
		(8.76)			(−2.86)		(4.58)			(−1.78)
Trading Volume, as % of TNA		0.01***			0.002***		0.03***			0.01***
		(12.34)			(4.11)		(9.73)			(3.49)
Size Percentile		0.09***			0.02		0.27***			0.03
		(11.16)			(1.59)		(9.43)			(0.72)
Value Percentile		−0.05***			−0.01		−0.10***			−0.003
		(−6.75)			(−0.77)		(−3.68)			(−0.12)
Momentum Percentile		−0.10***			−0.01***		−0.12***			−0.03*
		(−15.80)			(−2.85)		(−6.37)			(−1.94)
Time Fixed Effects	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Fund Fixed Effects	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Observations	54,331	54,331	54,331	54,331	54,331	54,331	54,331	54,331	54,331	54,331
Adjusted R^2	0.19	0.26	0.42	0.74	0.74	0.28	0.31	0.33	0.72	0.72

Table 4: Brokerage Network Centrality and Mutual Fund Performance: Portfolio Sorts

This table reports the average time-series monthly returns from July 1994 to December 2016. Funds are sorted into quintile portfolios based on degree centrality (in columns (1) to column (6)) and eigenvector centrality (in columns (7) to (12)). The investor return is decomposed into the holdings return (net of expenses) and the return gap following Equation (5). Raw returns as well as four-factor adjusted returns are reported for average return gap, average holdings return (net of expenses), and average investor return. Panel A reports the full sample results, whereas Panel B and Panel C report the split-sample results. The heteroskedasticity robust t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Panel A: Full Sample: July 1994 to December 2016

	Raw Return (% per month)						4-Factor Alpha (% per month)					
	Peripheral	Q2	Q3	Q4	Central	C – P	Peripheral	Q2	Q3	Q4	Central	C – P
Sorted on Degree Centrality												
Return Gap	−0.06***	−0.05**	−0.04**	−0.03	−0.01	0.05***	−0.02	−0.02	−0.01	0.003	0.02	0.04***
	(−3.12)	(−2.58)	(−2.32)	(−1.54)	(−0.74)	(5.03)	(−1.41)	(−1.08)	(−0.61)	(0.20)	(1.19)	(4.48)
Holdings Return	0.88***	0.87***	0.88***	0.86***	0.87***	−0.01	0.13***	0.13***	0.13***	0.13***	0.12***	−0.005
	(3.03)	(3.04)	(3.05)	(2.99)	(2.96)	(−0.19)	(3.20)	(3.28)	(2.95)	(2.95)	(3.18)	(−0.19)
Investor Return	0.81***	0.83***	0.83***	0.84***	0.86***	0.05	0.11***	0.12***	0.12***	0.13***	0.14***	0.03
	(2.98)	(3.00)	(3.04)	(3.02)	(3.04)	(1.55)	(2.67)	(2.85)	(2.78)	(2.98)	(3.58)	(1.53)
Sorted on Eigenvector Centrality												
Return Gap	−0.07***	−0.04**	−0.04**	−0.03	−0.01	0.05***	−0.02*	−0.01	−0.01	0.002	0.02	0.04***
	(−3.27)	(−2.43)	(−2.25)	(−1.61)	(−0.71)	(5.26)	(−1.66)	(−0.85)	(−0.56)	(0.11)	(1.22)	(4.75)
Holdings Return	0.88***	0.87***	0.88***	0.87***	0.87***	−0.02	0.13***	0.12***	0.13***	0.13***	0.12***	−0.01
	(3.06)	(3.01)	(3.04)	(3.01)	(2.96)	(−0.52)	(3.39)	(3.09)	(2.98)	(2.87)	(3.26)	(−0.46)
Investor Return	0.82***	0.82***	0.83***	0.84***	0.85***	0.04	0.11***	0.11***	0.12***	0.13***	0.14***	0.03
	(2.99)	(2.99)	(3.03)	(3.03)	(3.04)	(1.34)	(2.80)	(2.74)	(2.79)	(2.93)	(3.64)	(1.44)
Sorted on Past Return Gap												
Return Gap	−0.09***	−0.05***	−0.05***	−0.02	0.02	0.11***	−0.03	−0.02	−0.03**	−0.002	0.05***	0.08***
	(−3.50)	(−3.24)	(−3.33)	(−1.47)	(0.77)	(4.57)	(−1.41)	(−1.48)	(−2.02)	(−0.16)	(2.67)	(4.16)
Holdings Return	0.88***	0.86***	0.88***	0.87***	0.88***	−0.003	0.08	0.12***	0.17***	0.16***	0.11*	0.03
	(2.88)	(3.07)	(3.21)	(3.07)	(2.79)	(−0.05)	(1.54)	(2.72)	(4.11)	(3.38)	(1.93)	(0.53)
Investor Return	0.79***	0.81***	0.83***	0.84***	0.90***	0.11*	0.05	0.10**	0.15***	0.16***	0.16***	0.11*
	(2.74)	(2.99)	(3.12)	(3.09)	(3.03)	(1.95)	(1.01)	(2.28)	(3.60)	(3.35)	(2.90)	(1.96)

Table 4–*Continued*

Panel B: Sub-sample: July 1994 to December 2007

	Raw Return (% per month)						4-Factor Alpha (% per month)					
	Peripheral	Q2	Q3	Q4	Central	C – P	Peripheral	Q2	Q3	Q4	Central	C – P
Sorted on Degree Centrality												
Return Gap	−0.04	−0.03	−0.03	−0.01	0.01	0.05***	0.01	0.02	0.02	0.03*	0.05***	0.03***
	(−1.53)	(−1.18)	(−1.11)	(−0.33)	(0.23)	(3.36)	(0.80)	(1.05)	(1.09)	(1.93)	(2.71)	(2.68)
Holdings Return	0.99***	0.98***	0.99***	0.97***	0.97***	−0.02	0.21***	0.23***	0.22***	0.22***	0.19***	−0.02
	(2.82)	(2.80)	(2.82)	(2.73)	(2.65)	(−0.51)	(3.95)	(3.81)	(3.31)	(3.36)	(3.41)	(−0.66)
Investor Return	0.94***	0.95***	0.96***	0.96***	0.97***	0.03	0.23***	0.24***	0.24***	0.25***	0.24***	0.01
	(2.89)	(2.88)	(2.90)	(2.85)	(2.82)	(0.64)	(4.31)	(4.22)	(3.76)	(4.00)	(4.45)	(0.47)
Sorted on Eigenvector Centrality												
Return Gap	−0.05*	−0.03	−0.03	−0.01	0.01	0.06***	0.01	0.02	0.02	0.03**	0.04**	0.04***
	(−1.75)	(−1.02)	(−1.01)	(−0.33)	(0.20)	(3.60)	(0.36)	(1.41)	(1.23)	(1.99)	(2.57)	(2.88)
Holdings Return	1.00***	0.97***	0.98***	0.97***	0.96***	−0.04	0.23***	0.21***	0.22***	0.21***	0.20***	−0.03
	(2.85)	(2.77)	(2.80)	(2.75)	(2.65)	(−0.89)	(4.20)	(3.57)	(3.37)	(3.22)	(3.56)	(−0.86)
Investor Return	0.95***	0.95***	0.96***	0.96***	0.97***	0.02	0.23***	0.23***	0.24***	0.25***	0.24***	0.01
	(2.91)	(2.86)	(2.89)	(2.88)	(2.81)	(0.38)	(4.50)	(4.04)	(3.81)	(3.94)	(4.52)	(0.40)
Sorted on Past Return Gap												
Return Gap	−0.08**	−0.05**	−0.04**	−0.003	0.08**	0.17***	0.01	−0.01	−0.01	0.03*	0.11***	0.11***
	(−2.33)	(−2.42)	(−2.13)	(−0.15)	(2.18)	(4.94)	(0.25)	(−0.44)	(−0.74)	(1.70)	(4.82)	(4.05)
Holdings Return	1.02***	0.99***	0.97***	0.94***	0.97**	−0.05	0.14**	0.22***	0.26***	0.25***	0.20**	0.06
	(2.66)	(2.97)	(2.96)	(2.76)	(2.39)	(−0.53)	(2.22)	(3.99)	(4.37)	(3.43)	(2.10)	(0.64)
Investor Return	0.94***	0.94***	0.93***	0.94***	1.05***	0.11	0.14**	0.22***	0.25***	0.28***	0.31***	0.16**
	(2.64)	(2.95)	(2.95)	(2.89)	(2.78)	(1.35)	(2.46)	(4.19)	(4.54)	(3.93)	(3.42)	(2.03)

Panel C: Sub-sample: January 2008 to December 2016

	Raw Return (% per month)						4-Factor Alpha (% per month)					
	Peripheral	Q2	Q3	Q4	Central	C – P	Peripheral	Q2	Q3	Q4	Central	C – P
Sorted on Degree Centrality												
Return Gap	−0.10***	−0.07***	−0.07**	−0.06**	−0.05*	0.05***	−0.06***	−0.05***	−0.04**	−0.04*	−0.02	0.04***
	(−3.23)	(−3.04)	(−2.47)	(−2.27)	(−1.72)	(4.37)	(−3.42)	(−2.75)	(−2.06)	(−1.84)	(−1.15)	(4.36)
Holdings Return	0.71	0.71	0.71	0.71	0.73	0.02	−0.07*	−0.06**	−0.06*	−0.06*	−0.04	0.03
	(1.43)	(1.45)	(1.45)	(1.46)	(1.48)	(0.55)	(−1.91)	(−2.03)	(−1.84)	(−1.73)	(−1.29)	(1.01)
Investor Return	0.61	0.64	0.65	0.65	0.68	0.07**	−0.13***	−0.11***	−0.10***	−0.10**	−0.07*	0.06***
	(1.29)	(1.34)	(1.36)	(1.37)	(1.43)	(2.15)	(−3.39)	(−3.31)	(−2.84)	(−2.56)	(−1.76)	(2.73)
Sorted on Eigenvector Centrality												
Return Gap	−0.10***	−0.07***	−0.07**	−0.06**	−0.04	0.05***	−0.06***	−0.05***	−0.04**	−0.04**	−0.02	0.04***
	(−3.19)	(−3.01)	(−2.51)	(−2.43)	(−1.61)	(4.53)	(−3.34)	(−2.75)	(−2.11)	(−2.03)	(−1.01)	(4.42)
Holdings Return	0.71	0.71	0.72	0.71	0.73	0.02	−0.07*	−0.06**	−0.06*	−0.06*	−0.04	0.03
	(1.43)	(1.45)	(1.46)	(1.45)	(1.48)	(0.56)	(−1.95)	(−2.06)	(−1.73)	(−1.75)	(−1.31)	(1.02)
Investor Return	0.61	0.64	0.65	0.65	0.68	0.07**	−0.13***	−0.11***	−0.10***	−0.10***	−0.06*	0.07***
	(1.30)	(1.34)	(1.36)	(1.36)	(1.44)	(2.14)	(−3.39)	(−3.35)	(−2.74)	(−2.72)	(−1.70)	(2.74)
Sorted on Past Return Gap												
Return Gap	−0.10***	−0.05**	−0.06***	−0.05**	−0.07*	0.03	−0.07***	−0.03*	−0.04**	−0.04**	−0.03	0.03
	(−2.77)	(−2.22)	(−2.83)	(−2.45)	(−1.82)	(0.96)	(−2.73)	(−1.70)	(−2.49)	(−2.09)	(−1.27)	(1.31)
Holdings Return	0.67	0.66	0.74	0.75	0.74	0.07	−0.12**	−0.11***	−0.005	−0.01	−0.04	0.08
	(1.33)	(1.36)	(1.56)	(1.54)	(1.48)	(1.29)	(−2.27)	(−3.18)	(−0.18)	(−0.31)	(−0.96)	(1.33)
Investor Return	0.57	0.61	0.69	0.69	0.67	0.11	−0.19***	−0.14***	−0.05*	−0.05	−0.08	0.11*
	(1.18)	(1.28)	(1.46)	(1.47)	(1.41)	(1.64)	(−3.21)	(−3.70)	(−1.72)	(−1.35)	(−1.50)	(1.77)

Table 5: Brokerage Network Centrality and Return Gap: Panel Regressions

This table examines whether our previous results documenting the fund–centrality premium based on portfolio sorts continue to hold after controlling for fund characteristics, including lagged return gap, and fund fixed-effects. Specifically, this table presents the results of our baseline linear regression model:

$$\textit{Return Gap}_{i,t} = \beta \times \textit{Centrality}_{i,t-1} + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t}$$

where i indexes mutual funds and t indexes time in half-years. The dependent variable is $\textit{Return Gap}_{i,t}$ which is fund i's average return gap during half-year t. The independent variable of interest is $\textit{Centrality}_{i,t-1}$, which is fund i's brokerage network centrality (degree centrality or eigenvector centrality) measured at the end of half-year $t-1$. $\textit{Covariates}_{i,t-1}$ are a vector of fund-level variables that are measured at the end of time $t-1$ and include log(fund TNA), log(family TNA), expense ratio, commission rate, trading volume, and average size-value-momentum percentiles of the stocks in the fund's portfolio. More details on fund-level variables are provided in Table 2. In some specifications, the regression includes lagged return gap and fund fixed-effects (α_i) and all regressions include time fixed-effects (θ_t). Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Dependent variable:	Return Gap (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
Degree Centrality	0.15***	0.13***	0.09**			
	(4.62)	(4.44)	(1.97)			
Eigenvector Centrality				0.04***	0.04***	0.03**
				(4.70)	(4.52)	(2.01)
Past Return Gap (%)		0.08***	0.01		0.08***	0.01
		(11.07)	(0.89)		(11.08)	(0.89)
log(Fund TNA)	−0.01***	−0.01***	−0.03***	−0.01***	−0.01***	−0.03***
	(−6.91)	(−6.87)	(−9.72)	(−6.92)	(−6.89)	(−9.72)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
	(6.25)	(6.24)	(2.61)	(6.16)	(6.14)	(2.61)
Expense Ratio (%)	−0.01	−0.01	0.02	−0.01	−0.01	0.02
	(−1.09)	(−1.10)	(1.20)	(−1.05)	(−1.07)	(1.22)
Commission Rate (%)	−0.04***	−0.04***	−0.07***	−0.04***	−0.04***	−0.07***
	(−2.97)	(−3.00)	(−3.98)	(−2.99)	(−3.02)	(−3.99)
Trading Volume, as % of TNA	0.0000	0.0000	0.0001*	0.0000	0.0000	0.0001*
	(0.75)	(0.78)	(1.92)	(0.76)	(0.79)	(1.93)
Size Percentile	−0.001***	−0.001***	0.001	−0.001***	−0.001***	0.001
	(−4.11)	(−3.99)	(1.35)	(−4.13)	(−4.01)	(1.36)
Value Percentile	−0.002***	−0.002***	−0.001***	−0.002***	−0.002***	−0.001***
	(−10.66)	(−10.65)	(−2.73)	(−10.67)	(−10.66)	(−2.74)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.32)	(−2.08)	(−2.03)	(−2.34)	(−2.10)	(−2.04)
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	No	Yes	No	No	Yes
Observations	54,331	54,331	54,331	54,331	54,331	54,331
Adjusted R^2	0.07	0.08	0.10	0.07	0.08	0.10

Table 6: The Fund–Centrality Premium when Funds Experience Severe Redemptions

This table examines whether the fund–centrality premium is more pronounced when funds' trading activities are primarily driven by liquidity reasons, such as to accommodate large investor redemptions. Specifically, we interact an indicator variable for contemporaneous large outflows with lagged brokerage network centrality in our baseline specification as follows:

$$\begin{aligned} \textit{Return Gap}_{i,t} = \delta \times \textit{Centrality}_{i,t-1} \times \mathbb{1}(\textit{Outflow}_{i,t} > 5\%) + \beta \times \textit{Centrality}_{i,t-1} \\ + \rho \times \mathbb{1}(\textit{Outflow}_{i,t} > 5\%) + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned}$$

where $\mathbb{1}(\textit{Outflow}_{i,t} > 5\%)$ is an indicator variable that is equal to 1 if fund i's outflow during half-year t exceeds five percent and the rest of the model is the same as in Table 5. Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Dependent variable:	Return Gap (%)			
	(1)	(2)	(3)	(4)
Degree Centrality × 𝟙(Outflow > 5%)	0.14***	0.17***		
	(2.94)	(3.19)		
Eigenvector Centrality × 𝟙(Outflow > 5%)			0.04***	0.04***
			(2.89)	(2.70)
Degree Centrality	0.10***	0.03		
	(2.89)	(0.67)		
Eigenvector Centrality			0.03***	0.01
			(2.75)	(0.79)
𝟙(Outflow > 5%)	−0.03***	−0.03***	−0.03***	−0.03***
	(−3.32)	(−3.25)	(−3.23)	(−2.78)
log(Fund TNA)	−0.01***	−0.03***	−0.01***	−0.03***
	(−6.87)	(−9.71)	(−6.87)	(−9.65)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***
	(6.20)	(2.59)	(6.13)	(2.60)
Expense Ratio (%)	−0.005	0.02	−0.005	0.02
	(−0.91)	(1.26)	(−0.86)	(1.30)
Commission Rate (%)	−0.04***	−0.07***	−0.04***	−0.07***
	(−2.94)	(−3.96)	(−2.98)	(−4.00)
Trading Volume, as % of TNA	0.0000	0.0001**	0.0000	0.0001**
	(0.90)	(1.97)	(0.93)	(2.00)
Size Percentile	−0.001***	0.001	−0.001***	0.001
	(−4.11)	(1.35)	(−4.13)	(1.36)
Value Percentile	−0.002***	−0.001***	−0.002***	−0.001***
	(−10.67)	(−2.71)	(−10.70)	(−2.73)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.38)	(−2.12)	(−2.40)	(−2.14)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	Yes	No	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.07	0.11	0.07	0.11

Table 7: The Fund–Centrality Premium For Valuable Clients

This table examines whether the fund–centrality premium is larger for more valuable clients, especially when the client funds are forced to trade to accommodate large investor redemptions. In unconditional tests presented in Panel A, we interact a measure of brokerage revenue generating potential with brokerage network centrality in our baseline specification as follows:

$$\begin{aligned} \textit{Return Gap}_{i,t} = \delta \times \textit{Centrality}_{i,t-1} \times \textit{Broker Incentive}_{i,t-1} + \beta \times \textit{Centrality}_{i,t-1} \\ + \rho \times \textit{Broker Incentive}_{i,t-1} + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned}$$

where $\textit{Broker Incentive}_{i,t-1}$ is our proxy for fund i's brokerage revenue generating potential as measured by an indicator variable that is equal to one if fund i's aggregate dollar commissions during half-year $t-1$ is greater than its top quartile value. As a robustness check, we replace an indicator variable with its continuous counterpart, log of aggregate dollar commissions in columns (3) and (4). The rest of the model is the same as in Table 5. The independent variable of interest is $\textit{Centrality}_{i,t-1} \times \textit{Broker Incentive}_{i,t-1}$ to tease out the effect of brokers' incentives on the fund–centrality premium. In conditional tests presented in Panel B, we add an indicator variable for contemporaneous large outflows as an additional interaction term. Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Panel A: Baseline

Dependent variable:	Return Gap (%)			
Broker Incentive:	$\mathbb{1}$(Dollar Commission $> Q_3$)		log(Dollar Commission)	
	(1)	(2)	(3)	(4)
Degree Centrality × Broker Incentive	0.26***		0.04*	
	(3.23)		(1.96)	
Eigenvector Centrality × Broker Incentive		0.05*		0.003
		(1.66)		(0.52)
Degree Centrality	0.03		0.16***	
	(0.53)		(2.74)	
Eigenvector Centrality		0.02		0.04*
		(1.15)		(1.94)
Broker Incentive	−0.06***	−0.04**	−0.01***	−0.01*
	(−3.70)	(−2.26)	(−2.74)	(−1.86)
log(Fund TNA)	−0.03***	−0.03***	−0.03***	−0.03***
	(−8.60)	(−8.58)	(−6.09)	(−5.98)
log(Family TNA)	0.01***	0.01***	0.01***	0.01**
	(2.66)	(2.65)	(2.62)	(2.56)
Expense Ratio (%)	0.02	0.02	0.01	0.02
	(1.17)	(1.21)	(1.14)	(1.19)
Commission Rate (%)	−0.05***	−0.06***	−0.04*	−0.04*
	(−3.00)	(−3.08)	(−1.96)	(−1.91)
Trading Volume, as % of TNA	0.0001**	0.0001**	0.0001**	0.0001**
	(2.28)	(2.29)	(2.30)	(2.37)
Size Percentile	0.001	0.001	0.001	0.001
	(1.32)	(1.34)	(1.30)	(1.32)
Value Percentile	−0.001***	−0.001***	−0.001***	−0.001***
	(−2.70)	(−2.71)	(−2.72)	(−2.74)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.00)	(−2.05)	(−1.97)	(−2.03)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	Yes	Yes	Yes	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.11	0.11	0.11	0.10

Table 7–*Continued*

Panel B: Triple Interaction

Dependent variable:	Return Gap (%)			
Broker Incentive:	$\mathbb{1}$(Dollar Commission $> Q_3$)		log(Dollar Commission)	
	(1)	(2)	(3)	(4)
Degree Centrality × Broker Incentive × $\mathbb{1}$(Outflow > 5%)	0.28**		0.05*	
	(2.08)		(1.86)	
Eigenvector Centrality × Broker Incentive × $\mathbb{1}$(Outflow > 5%)		0.12***		0.02**
		(2.74)		(2.00)
Degree Centrality × Broker Incentive	0.14		0.01	
	(1.52)		(0.62)	
Degree Centrality × $\mathbb{1}$(Outflow > 5%)	0.11*		0.30***	
	(1.90)		(3.54)	
Eigenvector Centrality × Broker Incentive		−0.004		−0.004
		(−0.14)		(−0.64)
Eigenvector Centrality × $\mathbb{1}$(Outflow > 5%)		0.02		0.09***
		(1.27)		(3.30)
Broker Incentive × $\mathbb{1}$(Outflow > 5%)	−0.06**	−0.08***	−0.01***	−0.01***
	(−2.24)	(−2.82)	(−2.63)	(−2.65)
Degree Centrality	−0.01		0.05	
	(−0.15)		(0.69)	
Eigenvector Centrality		0.01		−0.0001
		(0.61)		(−0.01)
Broker Incentive	−0.04*	−0.01	−0.01	−0.002
	(−1.95)	(−0.39)	(−1.29)	(−0.48)
$\mathbb{1}$(Outflow > 5%)	−0.02**	−0.01	−0.06***	−0.06***
	(−1.98)	(−1.40)	(−3.70)	(−3.46)
log(Fund TNA)	−0.03***	−0.03***	−0.03***	−0.03***
	(−8.63)	(−8.54)	(−6.19)	(−6.02)
log(Family TNA)	0.01***	0.01***	0.01***	0.01**
	(2.62)	(2.61)	(2.58)	(2.51)
Expense Ratio (%)	0.02	0.02	0.02	0.02
	(1.24)	(1.33)	(1.18)	(1.26)
Commission Rate (%)	−0.05***	−0.06***	−0.04**	−0.04*
	(−3.00)	(−3.12)	(−1.98)	(−1.96)
Trading Volume, as % of TNA	0.0001**	0.0001**	0.0001**	0.0001**
	(2.37)	(2.41)	(2.39)	(2.48)
Size Percentile	0.001	0.001	0.001	0.001
	(1.31)	(1.32)	(1.31)	(1.33)
Value Percentile	−0.001***	−0.001***	−0.001***	−0.001***
	(−2.69)	(−2.73)	(−2.71)	(−2.75)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.06)	(−2.12)	(−2.08)	(−2.14)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	Yes	Yes	Yes	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.11	0.11	0.11	0.11

Table 8: The Fund–Centrality Premium For Relationship Clients

This table examines whether the fund–centrality premium is larger for the clients that have established trading relationships with their brokers, especially when the client funds are forced to trade to accommodate large investor redemptions. In unconditional tests presented in Panel A, we interact a measure of existing trading relationships with brokerage network centrality in our baseline specification as follows:

$$\begin{aligned} \textit{Return Gap}_{i,t} = \delta \times \textit{Centrality}_{i,t-1} \times \textit{Trading Relationship}_{i,t-1} + \beta \times \textit{Centrality}_{i,t-1} \\ + \rho \times \textit{Trading Relationship}_{i,t-1} + \gamma \times \textit{Covariates}_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned}$$

where $\textit{Trading Relationship}_{i,t-1}$, or simply, $\textit{Relationship}_{i,t-1}$ is our proxy for fund i's strength of trading relationships with its current set of brokers, as measured by taking the minimum of a fraction of fund i' commissions paid to its broker k during half-year $t-1$ (current) and that during $t-3$ (a year before) and then summing it over all brokers currently employed by the fund. Intuitively, $\textit{Relationship}_{i,t-1}$ measures the extent to which fund i's current set of brokers overlap with the set of brokers the fund traded through a year before. The rest of the model is the same as in Table 5. The independent variable of interest is $\textit{Centrality}_{i,t-1} \times \textit{Relationship}_{i,t-1}$ to tease out the effect of prior trading relationships on the fund–centrality premium. In conditional tests presented in Panel B, we add an indicator variable for contemporaneous large outflows as an additional interaction term. Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Panel A: Baseline

Dependent variable:	Return Gap (%)			
	(1)	(2)	(3)	(4)
Degree Centrality × Relationship	0.13	0.12		
	(1.42)	(1.15)		
Eigenvector Centrality × Relationship			0.05*	0.05*
			(1.82)	(1.69)
Degree Centrality	0.07	0.03		
	(1.15)	(0.50)		
Eigenvector Centrality			0.01	0.001
			(0.66)	(0.06)
Relationship	−0.01	−0.02	−0.01	−0.03
	(−0.55)	(−1.07)	(−0.86)	(−1.51)
log(Fund TNA)	−0.01***	−0.03***	−0.01***	−0.03***
	(−6.98)	(−9.71)	(−6.99)	(−9.72)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***
	(6.18)	(2.59)	(6.07)	(2.61)
Expense Ratio (%)	−0.01	0.02	−0.01	0.02
	(−1.08)	(1.21)	(−1.04)	(1.25)
Commission Rate (%)	−0.04***	−0.07***	−0.04***	−0.07***
	(−2.99)	(−4.00)	(−3.03)	(−4.04)
Trading Volume, as % of TNA	0.0000	0.0001*	0.0000	0.0001*
	(0.73)	(1.92)	(0.74)	(1.93)
Size Percentile	−0.001***	0.001	−0.001***	0.001
	(−4.09)	(1.35)	(−4.09)	(1.38)
Value Percentile	−0.002***	−0.001***	−0.002***	−0.001***
	(−10.64)	(−2.72)	(−10.63)	(−2.73)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.29)	(−2.05)	(−2.30)	(−2.06)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	Yes	No	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.07	0.10	0.07	0.10

Table 8–*Continued*

Panel B: Triple Interaction

Dependent variable:	Return Gap (%)			
	(1)	(2)	(3)	(4)
Degree Centrality × Relationship × 𝟙(Outflow > 5%)	0.43**	0.44**		
	(2.12)	(2.06)		
Eigenvector Centrality × Relationship × 𝟙(Outflow > 5%)			0.12*	0.11*
			(1.87)	(1.66)
Degree Centrality × Relationship	−0.01	−0.04		
	(−0.08)	(−0.33)		
Degree Centrality × 𝟙(Outflow > 5%)	−0.09	−0.06		
	(−0.77)	(−0.45)		
Eigenvector Centrality × Relationship			0.01	0.01
			(0.40)	(0.35)
Eigenvector Centrality × 𝟙(Outflow > 5%)			−0.02	−0.01
			(−0.50)	(−0.24)
Relationship × 𝟙(Outflow > 5%)	−0.08**	−0.10***	−0.08**	−0.09**
	(−2.32)	(−2.68)	(−2.17)	(−2.37)
Degree Centrality	0.10	0.05		
	(1.40)	(0.68)		
Eigenvector Centrality			0.02	0.004
			(0.82)	(0.17)
Relationship	0.02	0.02	0.01	0.01
	(0.88)	(0.75)	(0.58)	(0.26)
𝟙(Outflow > 5%)	0.01	0.02	0.01	0.02
	(0.72)	(0.99)	(0.57)	(0.86)
log(Fund TNA)	−0.01***	−0.03***	−0.01***	−0.03***
	(−6.94)	(−9.77)	(−6.94)	(−9.71)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***
	(6.12)	(2.59)	(6.02)	(2.60)
Expense Ratio (%)	−0.005	0.02	−0.005	0.02
	(−0.92)	(1.27)	(−0.86)	(1.33)
Commission Rate (%)	−0.04***	−0.07***	−0.04***	−0.07***
	(−2.96)	(−3.98)	(−3.01)	(−4.03)
Trading Volume, as % of TNA	0.0000	0.0001*	0.0000	0.0001**
	(0.88)	(1.96)	(0.91)	(1.99)
Size Percentile	−0.001***	0.001	−0.001***	0.001
	(−4.09)	(1.37)	(−4.09)	(1.39)
Value Percentile	−0.002***	−0.001***	−0.002***	−0.001***
	(−10.65)	(−2.69)	(−10.68)	(−2.71)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.36)	(−2.12)	(−2.38)	(−2.14)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	Yes	No	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.07	0.11	0.07	0.11

Table 9: The Fund–Centrality Premium When Funds Submit Uninformed Large Orders

This table attempts to generalize our main results in Table 6 by examining whether the fund–centrality premium is larger when funds' trading activities are primarily driven by liquidity reasons, for instance, when funds submit large uninformed orders. First, we identify periods of heavy information-motivated buying and selling activities following Alexander, Cici, and Gibson (2007). We calculate BF and SF metrics as follows:

$$BF_{i,t} = \frac{BUYS_{i,t} - FLOW_{i,t}}{TNA_{i,t-1}} \quad \& \quad SF_{i,t} = \frac{SELLS_{i,t} + FLOW_{i,t}}{TNA_{i,t-1}}$$

where $BUYS_{i,t}$ is fund i's dollar volume of stock purchases during half-year t, $SELLS_{i,t}$ is fund i's dollar volume of stock sales during half-year t, $FLOW_{i,t}$ is fund i's net investor flow (inflow minus outflow) during half-year t, and $TNA_{i,t-1}$ is fund i's total net assets at the end of half-year $t-1$. Exploiting within-fund variation in BF and SF metrics, Alexander, Cici, and Gibson (2007) show that buy (sell) portfolios with high BF (SF) tend to outperform buy (sell) portfolios with low BF (SF). Since we cannot separately evaluate trading performance associated with buys and sells, we assign half-years where both BF and SF fall below its respective top quartile value as periods of uninformed trading. In Panel A, we interact an indicator variable for period of uninformed trading with brokerage network centrality as follows:

$$\begin{aligned} Return\ Gap_{i,t} = {} & \delta \times Centrality_{i,t-1} \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3) + \beta \times Centrality_{i,t-1} \\ & + \rho \times \mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3) + \gamma \times Covariates_{i,t-1} + \alpha_i + \theta_t + \varepsilon_{i,t} \end{aligned}$$

where $\mathbb{1}(BF_{i,t} < Q_3\ \&\ SF_{i,t} < Q_3)$ is an indicator variable that is equal to 1 if both $BF_{i,t}$ and $SF_{i,t}$ fall below its respective top quartile value during half-year t and the rest of the model is the same as in Table 5. Next, we proxy for average order sizes using average trade sizes inferred from consecutive portfolio disclosures, adjusting for trading volume in the market as follows:

$$\overline{Trade\ Size}_{i,t} = \frac{1}{N_{i,t}} \sum_{k} \frac{\mid Shares_{i,k,t} - Shares_{i,k,t-1} \mid}{\overline{VOL}_{k,t}^{CRSP}}$$

where $Shares_{i,k,t}$ is the split-adjusted number of shares held in stock k by fund i at the end of half-year (or quarter) t, $\overline{VOL}_{k,t}^{CRSP}$ is the average CRSP monthly volume between portfolio disclosures, and the averages are taken over stocks for which $Shares_{i,k,t} \neq Shares_{i,k,t-1}$. To arrive at the semi-annual figure, we take the average of quarterly numbers, if two quarterly observations are available. In Panel B, we add as an additional interaction term $\overline{Trade\ Size}_{i,t}$ as an indicator variable that is equal to 1 if $\overline{Trade\ Size}_{i,t}$ is above its quartile value or as a continuous variable to examine whether the fund–centrality premium is larger when funds submit uninformed large orders. Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Table 9–*Continued*

Dependent variable:	Return Gap (%)	
	(1)	(2)
Degree Centrality × $\mathbb{1}(\text{BF} < Q_3 \ \& \ \text{SF} < Q_3)$	0.17***	
	(2.85)	
Eigenvector Centrality × $\mathbb{1}(\text{BF} < Q_3 \ \& \ \text{SF} < Q_3)$		0.04**
		(2.04)
Degree Centrality	−0.01	
	(−0.12)	
Eigenvector Centrality		0.004
		(0.19)
$\mathbb{1}(\text{BF} < Q_3 \ \& \ \text{SF} < Q_3)$	−0.02	−0.01
	(−1.44)	(−0.79)
log(Fund TNA)	−0.03***	−0.03***
	(−9.90)	(−9.89)
log(Family TNA)	0.01***	0.01***
	(2.66)	(2.68)
Expense Ratio (%)	0.02	0.02
	(1.20)	(1.22)
Commission Rate (%)	−0.07***	−0.07***
	(−3.98)	(−4.00)
Trading Volume, as % of TNA	0.0001**	0.0001**
	(2.26)	(2.25)
Size Percentile	0.001	0.001
	(1.36)	(1.35)
Value Percentile	−0.001***	−0.001***
	(−2.63)	(−2.68)
Momentum Percentile	−0.001*	−0.001*
	(−1.89)	(−1.93)
Time Fixed Effects	Yes	Yes
Fund Fixed Effects	Yes	Yes
Observations	54,331	54,331
Adjusted R^2	0.11	0.11

Table 9–*Continued*

Dependent variable:	Return Gap (%)			
Brokerage Network Centrality:	Degree Centrality		Eigenvector Centrality	
	(1)	(2)	(3)	(4)
Centrality × $\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3) \times \mathbb{1}(\overline{\text{Trade Size}} > Q_3)$	0.31**		0.08*	
	(2.11)		(1.67)	
Centrality × $\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3) \times \overline{\text{Trade Size}}$		0.13*		0.04
		(1.68)		(1.44)
Centrality × $\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3)$	0.09	0.09	0.02	0.02
	(1.36)	(1.28)	(0.91)	(0.75)
$\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3) \times \mathbb{1}(\overline{\text{Trade Size}} > Q_3)$	−0.05*		−0.05	
	(−1.91)		(−1.56)	
$\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3) \times \overline{\text{Trade Size}}$		−0.02		−0.02
		(−1.16)		(−1.01)
Centrality × $\mathbb{1}(\overline{\text{Trade Size}} > Q_3)$	−0.13		−0.07	
	(−0.93)		(−1.51)	
Centrality × $\overline{\text{Trade Size}}$		−0.02		−0.02
		(−0.26)		(−0.93)
Centrality	0.02	0.002	0.02	0.02
	(0.36)	(0.03)	(0.97)	(0.77)
$\mathbb{1}(\text{BF} < Q_3$ & $\text{SF} < Q_3)$	−0.002	−0.01	0.002	0.0001
	(−0.18)	(−0.47)	(0.19)	(0.01)
$\mathbb{1}(\overline{\text{Trade Size}} > Q_3)$	0.02		0.04	
	(0.74)		(1.27)	
$\overline{\text{Trade Size}}$		−0.01		0.0002
		(−0.71)		(0.01)
log(Fund TNA)	−0.03***	−0.03***	−0.03***	−0.03***
	(−9.83)	(−8.68)	(−9.86)	(−8.73)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***
	(2.67)	(2.75)	(2.68)	(2.75)
Expense Ratio (%)	0.02	0.02	0.02	0.02
	(1.18)	(1.23)	(1.20)	(1.26)
Commission Rate (%)	−0.07***	−0.07***	−0.07***	−0.07***
	(−3.98)	(−3.92)	(−4.01)	(−3.93)
Trading Volume, as % of TNA	0.0001**	0.0001**	0.0001**	0.0001**
	(2.28)	(2.42)	(2.26)	(2.40)
Size Percentile	0.001	0.001	0.001	0.001
	(1.30)	(0.99)	(1.37)	(1.09)
Value Percentile	−0.001***	−0.001**	−0.001***	−0.001***
	(−2.64)	(−2.57)	(−2.67)	(−2.58)
Momentum Percentile	−0.001*	−0.001**	−0.001*	−0.001**
	(−1.92)	(−2.00)	(−1.94)	(−2.00)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	Yes	Yes	Yes	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.11	0.11	0.11	0.11

Table 10: A List of Brokerage Mergers (1995-2015)

This table reports a list of twenty six brokerage mergers, including the names of brokers involved in the merger, the merger effective date, the average brokerage shares pre- and post-merger, and changes in average broker shares around the merger. A broker share is defined as a fraction of the commission payments to the given broker by the fund. Broker shares are first averaged across funds each month on a rolling basis and then averaged over months $t-18$ to $t-7$ for the pre-merger and over months $t+7$ and $t+18$ for the post-merger. We highlight five largest mergers that will be used in our natural experiment.

	Acquiring Broker				Acquired Broker			
		Average Broker Shares (%)				Average Broker Shares (%)		
Effective Date	Broker Name	Before	After	Change	Broker Name	Before	After	Change
1997-05-31	MORGAN STANLEY	4.76	5.65	0.89	DEAN WITTER REYNOLDS	1.47	0.57	−0.90
1997-09-02	BT NEW YORK (SUCCESSOR: DEUTSCHE)	0.28	0.44	0.16	ALEX BROWN	1.04	1.16	0.12
1997-11-28	SMITH BARNEY (TRAVELERS)	4.83	5.69	0.86	SALOMON BROTHERS	3.94	0.78	−3.16
1998-06-30	SOCIETE GENERALE SECURITIES	0.18	0.18	−0.004	COWEN	0.54	0.66	0.12
2000-02-24	INSTINET	3.28	2.67	−0.61	LYNCH JONES RYAN	0.42	0.35	−0.07
2000-11-02	GOLDMAN SACHS GROUP	5.72	7.23	1.52	SPEAR LEEDS KELLOGG	0.22	0.35	0.12
2000-11-03	**CREDIT SUISSE FIRST BOSTON**	4.02	6.40	2.38	**DONALDSON LUFKIN JENRETTE**	4.40	0.75	−3.65
2000-11-03	**UBS WARBURG DILLON READ**	2.12	4.31	2.20	**PAINE WEBBER**	3.75	0.89	−2.85
2001-04-30	ABN-AMRO	1.15	0.77	−0.39	ING BARING-US	7.27	8.83	1.56
2001-09-04	WACHOVIA	0.41	0.50	0.09	FIRST UNION CAPITAL MARKETS	0.18	0.15	−0.03
2002-02-04	BANK OF NEW YORK	0.08	0.27	0.19	AUTRANET	1.02	0.44	−0.58
2003-07-01	WACHOVIA	0.47	0.86	0.40	PRUDENTIAL	1.30	1.05	−0.25
2003-10-31	LEHMAN BROTHERS	5.99	7.33	1.34	NEUBERGER BERMAN	0.14	0.02	−0.12
2003-12-08	UBS AG	5.69	5.11	−0.58	ABN-AMRO	0.90	1.14	0.24
2005-03-31	INSTINET	1.72	1.49	−0.24	BRIDGE TRADING	0.61	0.22	−0.39
2007-02-02	NOMURA HOLDINGS	0.23	0.20	−0.03	INSTINET	1.39	2.26	0.87
2007-10-01	WACHOVIA	0.26	0.13	−0.13	A.G. EDWARDS SONS	0.28	0.004	−0.28
2008-05-30	**JPMORGAN CHASE**	4.14	7.83	3.69	**BEAR STEARNS**	4.63	0.17	−4.46
2008-09-22	**BARCLAYS**	0.04	3.02	2.98	**LEHMAN BROTHERS**	7.53	0.12	−7.41
2009-01-01	**BANK OF AMERICA**	0.96	1.09	0.13	**MERRILL LYNCH**	8.69	6.23	−2.45
2009-10-02	MACQUARIE GROUP	0.42	0.69	0.27	FOX PITT KELTON	0.09	0.002	−0.09
2009-12-31	WELLS FARGO SECURITIES	0.04	0.16	0.12	WACHOVIA	0.12	0.11	−0.01
2010-07-01	STIFEL	0.52	0.60	0.08	THOMAS WEISEL PARTNERS	0.20	0.02	−0.18
2012-04-02	RAYMOND JAMES FINANCIAL	0.37	0.44	0.07	MORGAN KEEGAN	0.27	0.15	−0.12
2013-02-15	STIFEL	0.63	0.73	0.10	KEEFE BRUYETTE WOODS	0.22	0.15	−0.07
2014-09-03	KEYBANK	0.09	0.11	0.03	PACIFIC CREST SECURITIES	0.04	0.01	−0.03

Table 11: Testing for Matching Balance

This table reports the cross-sectional means and differences in means of the pre-treatment outcome variables and other pre-event fund-characteristics for the treated mutual funds and (matched) controls before and after the matching. We take top ten percent of funds with the largest expected changes in *Degree Centrality* as the treatment group. Among the remaining 90% of the sample, we construct the control group by matching on pre-treatment (pre-event) outcome variables and fund characteristics, using *Genetic Matching* algorithm proposed by Diamond and Sekhon (2013). The pre-treatment outcome variables include *Degree Centrality* and *Return Gap* and pre-event fund characteristics include *Log(Fund TNA)*, *Expense Ratio*, *Commission Rate*, *Trading Volume, as % of TNA*, *Index Fund (Yes=1)*, *Size Percentile*, *Value Percentile*, and *Momentum Percentile*. We choose one year just prior to the event window as the pre-event period. Return gaps are averaged over the twelve months in the pre-event period and mid-point values are taken for other variables. The event timelines are as depicted in Figure 4. The t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

		Before Matching			After Matching		
Variable	Treated	Control	Difference	(p-value)	Control	Difference	(p-value)
Panel A: 2000 Brokerage Mergers (Number of treated funds = 102)							
Pre-treatment outcomes							
Degree Centrality	0.21	0.16	0.05***	(< 0.001)	0.20	0.001	(0.49)
Return Gap (%)	−0.10	−0.08	−0.02	(0.84)	−0.11	0.01	(0.54)
Covariates							
log(Fund TNA)	5.67	5.47	0.19	(0.35)	5.99	−0.33	(0.18)
Expense Ratio (%)	0.01	0.01	−0.001***	(0.01)	0.01	−0.0001	(0.60)
Commission Rate (%)	0.12	0.14	−0.01	(0.42)	0.14	−0.01	(0.57)
Trading Volume, as % of TNA	160.99	174.71	−13.72	(0.24)	178.77	−17.78	(0.28)
Index Fund (Yes=1)	0.07	0.04	0.03	(0.33)	0.04	0.03	(0.32)
Size Percentile	90.20	88.03	2.17**	(0.04)	88.77	1.43	(0.22)
Value Percentile	26.92	29.75	−2.83**	(0.02)	26.86	0.06	(0.82)
Momentum Percentile	66.37	64.42	1.95	(0.13)	66.69	−0.32	(0.77)
Panel B: 2008 Brokerage Mergers (Number of treated funds = 160)							
Pre-treatment outcomes							
Degree Centrality	0.17	0.15	0.01***	(< 0.001)	0.17	−0.001	(0.63)
Return Gap(%)	0.15	0.14	0.01	(0.49)	0.15	0.003	(0.62)
Covariates							
log(Fund TNA)	6.35	5.82	0.53***	(< 0.001)	6.23	0.12	(0.18)
Expense Ratio(%)	0.01	0.01	−0.0005*	(0.09)	0.01	−0.0002	(0.44)
Commission Rate(%)	0.08	0.22	−0.14***	(< 0.001)	0.08	−0.003	(0.35)
Trade Volume, as % of TNA	130.52	106.52	23.99**	(0.01)	126.83	3.69	(0.21)
Index Fund (Yes=1)	0.11	0.10	0.003	(0.91)	0.14	−0.03	(0.20)
Size Percentile	85.26	83.84	1.42	(0.12)	85.92	−0.66	(0.19)
Value Percentile	40.23	41.35	−1.12	(0.17)	40.03	0.19	(0.40)
Momentum Percentile	60.89	60.69	0.20	(0.76)	61.15	−0.26	(0.48)

Table 12: Do brokerage networks improve trading performance? DiD Results

This table reports the difference-in-differences (DiD) results for *Degree Centrality* and *Return Gap* before and after brokerage mergers for the treated mutual funds and their matched controls. The selection of treatment and control groups, the matching procedure, and the construction of pre-event outcome variables are the same as in Table 11. We choose one year immediately following the event window as the post-event period. Return gaps are averaged over the twelve months in the post-event period and mid-point values are taken for *Degree Centrality*. The event timelines are as depicted in Figure 4. If we denote the average outcome variables in the treatment (T) and control (C) groups in the pre- and post-event periods by $O_{T,1}$, $O_{T,2}$, $O_{C,1}$, and $O_{C,2}$, respectively, the partial effect of change due to the mergers can be estimated as

$$DiD = (O_{T,2} - O_{T,1}) - (O_{C,2} - O_{C,1}).$$

The t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

	Treated		Matched Control		DiD	
Outcome Measures	Before	After	Before	After	Mean	(t-stat)
Panel A: 2000 Brokerage Mergers						
Degree Centrality	0.206	0.235	0.205	0.222	0.013**	(2.01)
Return Gap (%)	−0.097	0.063	−0.106	−0.038	0.093*	(1.67)
Panel B: 2008 Brokerage Mergers						
Degree Centrality	0.167	0.193	0.168	0.160	0.034***	(8.39)
Return Gap (%)	0.150	0.104	0.147	0.033	0.068*	(1.87)

Appendix

Table A1: Sample of CRSP-Thomson-NSAR Matched Funds

This table reports the total number and aggregate total net assets (TNA) of our CRSP-Thomson-NSAR matched funds each half-year.

Year	First Half		Second Half	
	Total Number of Funds	Aggregate TNA ($ billion)	Total Number of Funds	Aggregate TNA ($ billion)
1994	421	243.2	512	310.7
1995	617	426.6	759	551.2
1996	812	632.2	868	759.6
1997	941	917.2	991	1,142.8
1998	1,065	1,356.2	1,122	1,559.0
1999	1,260	1,916.8	1,306	1,999.6
2000	1,426	2,162.4	1,476	2,057.5
2001	1,572	2,047.9	1,608	1,847.8
2002	1,679	1,834.2	1,773	1,608.0
2003	1,774	1,702.2	1,803	2,059.3
2004	1,825	2,293.2	1,845	2,298.6
2005	1,845	2,426.4	1,897	2,610.3
2006	1,950	2,811.4	2,039	2,947.2
2007	2,106	3,166.0	2,137	3,227.0
2008	2,127	2,852.7	2,089	2,224.0
2009	2,017	1,998.4	2,012	2,415.2
2010	1,968	2,480.1	1,921	2,567.0
2011	1,891	2,931.1	1,852	2,622.4
2012	1,823	2,798.1	1,757	2,847.8
2013	1,730	3,115.2	1,684	3,685.1
2014	1,649	4,143.0	1,629	4,072.0
2015	1,614	4,137.0	1,568	4,132.1
2016	1,565	3,880.6		

Table A2: When Funds Experience Severe Redemptions: Robustness Checks

This table provides robustness checks for the results reported in Table 6. In Panel A, we use a different cutoff (10% instead of 5%) to identify large outflow events and the rest of the model is the same as in Table 6. In Panel B, we repeat our analysis in Table 6 using a sub-sample of funds with fund family TNA is below its top quartile value. Standard errors are clustered at the fund level and the resulting t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

Panel A: Using a larger cutoff to define large outflow events

Dependent variable:	Return Gap (%)			
	(1)	(2)	(3)	(4)
Degree Centrality × 𝟙(Outflow > 10%)	0.22***	0.24***		
	(3.60)	(3.53)		
Eigenvector Centrality × 𝟙(Outflow > 10%)			0.07***	0.07***
			(3.93)	(3.42)
Degree Centrality	0.11***	0.05		
	(3.31)	(1.07)		
Eigenvector Centrality			0.03***	0.01
			(3.09)	(1.04)
𝟙(Outflow > 10%)	−0.04***	−0.04***	−0.05***	−0.04***
	(−3.76)	(−3.30)	(−4.00)	(−3.17)
log(Fund TNA)	−0.01***	−0.03***	−0.01***	−0.03***
	(−6.86)	(−9.69)	(−6.85)	(−9.65)
log(Family TNA)	0.01***	0.01***	0.01***	0.01***
	(6.15)	(2.58)	(6.08)	(2.59)
Expense Ratio (%)	−0.01	0.02	−0.01	0.02
	(−1.01)	(1.24)	(−0.95)	(1.28)
Commission Rate (%)	−0.04***	−0.07***	−0.04***	−0.07***
	(−2.95)	(−3.99)	(−2.99)	(−4.01)
Trading Volume, as % of TNA	0.0000	0.0001**	0.0000	0.0001**
	(0.95)	(1.99)	(0.97)	(2.00)
Size Percentile	−0.001***	0.001	−0.001***	0.001
	(−4.19)	(1.34)	(−4.21)	(1.35)
Value Percentile	−0.002***	−0.001***	−0.002***	−0.001***
	(−10.67)	(−2.74)	(−10.71)	(−2.75)
Momentum Percentile	−0.001**	−0.001**	−0.001**	−0.001**
	(−2.38)	(−2.09)	(−2.39)	(−2.10)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	Yes	No	Yes
Observations	54,331	54,331	54,331	54,331
Adjusted R^2	0.07	0.11	0.07	0.11

Table A2–*Continued*

Panel B: Excluding funds that belong to large fund families

Dependent variable:	Return Gap (%)			
	(1)	(2)	(3)	(4)
Degree Centrality × 𝟙(Outflow > 5%)	0.13**	0.14**		
	(2.51)	(2.40)		
Eigenvector Centrality × 𝟙(Outflow > 5%)			0.05***	0.04**
			(2.77)	(2.23)
Degree Centrality	0.13***	0.03		
	(3.18)	(0.54)		
Eigenvector Centrality			0.03***	0.01
			(2.87)	(0.46)
𝟙(Outflow > 5%)	−0.03***	−0.03***	−0.03***	−0.03***
	(−3.05)	(−2.96)	(−3.26)	(−2.78)
log(Fund TNA)	−0.01***	−0.04***	−0.01***	−0.04***
	(−6.10)	(−8.85)	(−6.11)	(−8.81)
log(Family TNA)	0.01***	0.01**	0.01***	0.01**
	(3.89)	(2.07)	(3.90)	(2.09)
Expense Ratio (%)	−0.003	0.01	−0.003	0.01
	(−0.49)	(0.34)	(−0.43)	(0.36)
Commission Rate (%)	−0.04***	−0.07***	−0.04***	−0.07***
	(−2.84)	(−4.02)	(−2.89)	(−4.05)
Trading Volume, as % of TNA	−0.0000	0.0001	−0.0000	0.0001
	(−0.27)	(1.00)	(−0.25)	(1.01)
Size Percentile	−0.001***	0.001	−0.001***	0.001
	(−4.20)	(1.51)	(−4.22)	(1.51)
Value Percentile	−0.002***	−0.002***	−0.002***	−0.002***
	(−10.16)	(−3.13)	(−10.19)	(−3.15)
Momentum Percentile	−0.001***	−0.001*	−0.001***	−0.001*
	(−2.63)	(−1.89)	(−2.65)	(−1.91)
Time Fixed Effects	Yes	Yes	Yes	Yes
Fund Fixed Effects	No	Yes	No	Yes
Observations	40,743	40,743	40,743	40,743
Adjusted R^2	0.06	0.10	0.06	0.10

Table A3: Testing for Matching Balance

This table reports the cross-sectional means and differences in means of the pre-treatment outcome variables and other pre-event fund-characteristics for the treated mutual funds and (matched) controls before and after the matching. We take top ten percent of funds with the largest expected changes in *Eigenvector Centrality* as the treatment group. Among the remaining 90% of the sample, we construct the control group by matching on pre-treatment (pre-event) outcome variables and fund characteristics, using *Genetic Matching* algorithm proposed by Diamond and Sekhon (2013). The pre-treatment outcome variables include *Eigenvector Centrality* and *Return Gap* and pre-event fund characteristics include *Log(Fund TNA)*, *Expense Ratio*, *Commission Rate*, *Trading Volume, as % of TNA*, *Index Fund (Yes=1)*, *Size Percentile*, *Value Percentile*, and *Momentum Percentile*. We choose one year just prior to the event window as the pre-event period. Return gaps are averaged over the twelve months in the pre-event period and mid-point values are taken for other variables. The event timelines are as depicted in Figure 4. The t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

		Before Matching			After Matching		
Variable	Treated	Control	Difference	(p-value)	Control	Difference	(p-value)
Panel A: 2000 Brokerage Mergers (Number of treated funds = 102)							
Pre-treatment outcomes							
Eigenvector Centrality	0.68	0.50	0.18***	(< 0.001)	0.67	0.01	(0.37)
Return Gap (%)	−0.03	−0.09	0.05	(0.54)	−0.03	−0.01	(0.82)
Covariates							
log(Fund TNA)	5.27	5.52	−0.25	(0.18)	5.29	−0.02	(0.70)
Expense Ratio (%)	0.01	0.01	−0.001	(0.13)	0.01	−0.0001	(0.78)
Commission Rate (%)	0.18	0.13	0.06**	(0.04)	0.18	0.003	(0.31)
Trading Volume, as % of TNA	158.43	175.00	−16.57	(0.18)	152.24	6.19	(0.32)
Index Fund (Yes=1)	0.07	0.04	0.03	(0.33)	0.03	0.04	(0.16)
Size Percentile	89.72	88.09	1.63	(0.13)	89.03	0.69	(0.57)
Value Percentile	27.76	29.66	−1.90	(0.13)	28.53	−0.77	(0.47)
Momentum Percentile	65.69	64.49	1.19	(0.36)	64.89	0.80	(0.30)
Panel B: 2008 Brokerage Mergers (Number of treated funds = 161)							
Pre-treatment outcomes							
Eigenvector Centrality	0.60	0.49	0.10***	(< 0.001)	0.60	0.001	(0.71)
Return Gap(%)	0.17	0.13	0.03	(0.11)	0.16	0.01	(0.41)
Covariates							
log(Fund TNA)	6.50	5.80	0.70***	(< 0.001)	6.46	0.03	(0.45)
Expense Ratio(%)	0.01	0.01	−0.0005	(0.11)	0.01	0.0001	(0.85)
Commission Rate(%)	0.08	0.22	−0.15***	(< 0.001)	0.13	−0.05	(0.13)
Trade Volume, as % of TNA	132.59	106.27	26.32***	0.003	131.14	1.45	(0.53)
Index Fund (Yes=1)	0.11	0.10	0.01	(0.73)	0.12	−0.01	(0.82)
Size Percentile	85.10	83.85	1.25	(0.16)	85.64	−0.54	(0.66)
Value Percentile	40.08	41.37	−1.29	(0.12)	40.18	−0.10	(0.60)
Momentum Percentile	60.88	60.69	0.19	(0.76)	61.71	−0.84	(0.28)

Table A4: Do brokerage networks improve trading performance? DiD Results

This table reports the difference-in-differences (DiD) results for *Eigenvector Centrality* and *Return Gap* before and after brokerage mergers for the treated mutual funds and their matched controls. The selection of treatment and control groups, the matching procedure, and the construction of pre-event outcome variables are the same as in Table A3. We choose one year immediately following the event window as the post-event period. Return gaps are averaged over the twelve months in the post-event period and mid-point values are taken for *Eigenvector Centrality*. The event timelines are as depicted in Figure 4. If we denote the average outcome variables in the treatment (T) and control (C) groups in the pre- and post-event periods by $O_{T,1}$, $O_{T,2}$, $O_{C,1}$, and $O_{C,2}$, respectively, the partial effect of change due to the mergers can be estimated as

$$DiD = (O_{T,2} - O_{T,1}) - (O_{C,2} - O_{C,1}).$$

The t-statistics are reported in parentheses. Statistical significance at the 10%, 5%, and 1% level is indicated by *, **, and ***, respectively.

	Treated		Matched Control		DiD	
Outcome Measures	Before	After	Before	After	Mean	(t-stat)
Panel A: 2000 Brokerage Mergers						
Eigenvector Centrality	0.677	0.662	0.671	0.680	0.053**	(2.59)
Return Gap (%)	−0.034	0.026	−0.029	−0.076	0.108*	(1.68)
Panel B: 2008 Brokerage Mergers						
Eigenvector Centrality	0.596	0.679	0.595	0.592	0.090***	(5.68)
Return Gap (%)	0.167	0.107	0.159	0.024	0.075**	(1.98)

Dynamic Adverse Selection and Liquidity*

Ioanid Roşu†

February 1, 2019

Abstract

Does a larger fraction of informed trading generate more illiquidity, as measured by the bid-ask spread? We answer this question in the negative in the context of a dynamic dealer market where the fundamental value follows a random walk, provided we consider the long run (stationary) equilibrium. More informed traders tend to generate more adverse selection and hence larger spreads, but at the same time cause faster learning by the market makers and hence smaller spreads. These two effects offset each other in the long run.

KEYWORDS: Learning, adverse selection, dynamic model, stationary distribution.

*We thank Bruno Biais, Denis Gromb, Augustin Landier, Stefano Lovo, Talis Putnins, and Daniel Schmidt for their suggestions. We are also grateful to finance seminar participants at HEC Paris and the Bucharest Academy of Economic Studies, as well as conference participants at the AFFI/Eurofidai conference in Paris for valuable comments.

†HEC Paris, Email: rosu@hec.fr.

1 Introduction

A traditional view of market liquidity, going at least as far back as Bagehot (1971), posits that one of the causes of illiquidity is adverse selection: *"The essence of market making, viewed as a business, is that in order for the market maker to survive and prosper, his gains from liquidity-motivated transactors must exceed his losses to information motivated transactors. [...] The spread he sets between his bid and asked price affects both: the larger the spread, the less money he loses to information-motivated, transactors and the more he makes from liquidity-motivated transactors."*

This intuition has later been made precise by models such as Glosten and Milgrom (1985, henceforth GM85), in which a competitive risk-neutral dealer sequentially sets bid and ask prices in a risky asset, and makes zero expected profits in each trading round. Traders are selected at random from a population that contains a fraction ρ of informed traders, and must trade at most one unit of the asset. The asset liquidates at a value v that is constant and is either zero or one. In equilibrium, the bid-ask spread is wider when the informed share ρ is higher: there is more adverse selection, hence the dealer must set a larger bid-ask spread to break even.

This intuition, however, must be modified once we consider the dynamics of the bid-ask spread. A larger informed share also means that orders carry more information, which over time reduces the uncertainty about v and thus puts downward pressure on the bid-ask spread. We call this last effect *dynamic efficiency*. This effect is already present in GM85, who observe that a larger informed share causes initially a larger bid-ask spread, but also causes the bid-ask spread to decrease faster to its eventual value, which is zero (when v is fully learned).

A natural question is then: to what extent does dynamic efficiency reduce the traditional adverse selection? To answer this question, we extend the framework of GM85 to allow v to move over time according to a random walk v_t.[1] To obtain closed-form results, we assume that the increments of v_t are normally distributed with volatility σ_v, called the *fundamental*

[1] In GM85 both types of traders are willing to trade in each period (the informed because v is always outside the bid-ask spread, the uninformed for exogenous reasons), hence a trader is chosen at random among the informed and uninformed. When v_t is moving, there are times when the informed traders are not willing to trade (v_t is within the spread), hence a trader is chosen at random among the uninformed.

volatility. We chose a moving value for two reasons. First, this is a realistic assumption in modern financial markets, where relevant information arrives essentially at a continuous rate. Second, we want to study the long-term evolution of the bid-ask spreads, and this long-term analysis is trivial when v_t is constant, as the dealer eventually fully learns v. Note that we are interested in the long run because (as we show later) in the short run the equilibrium is similar to GM85, but in the long run it converges to the *stationary* equilibrium, which has novel properties.

The first property of the stationary equilibrium is that the dealer's uncertainty about v_t is constant. More precisely, we define the *public density* at t to be the dealer's posterior density of v_t just before trading at t. We also define the *public mean* and *public volatility* to be, respectively, the mean and standard deviation of the public density.[2] Thus, in a stationary equilibrium, the public volatility (which is a measure of the dealer's uncertainty) is constant. The second property of the stationary equilibrium is that the informed share is inversely related to the public volatility. The intuition is simple: when the informed share is low, the order flow carries little information, and thus the public volatility is large.[3]

A surprising property of the stationary equilibrium is that the informed share has no effect on the bid-ask spread. To understand this result, consider a small informed share, say 1%. Suppose a buy order arrives, and the dealer estimates how much to update the public mean (in equilibrium this update is half of the bid-ask spread). There are two opposite effects. First, it is very unlikely that the buy order comes from an informed trader (with only 1% chance). This is the *adverse selection effect*: a low informed share makes the dealer less concerned about adverse selection, which leads to a smaller update of the public mean, and hence decreases the bid-ask spread. But, second, if the buy order does come from an informed trader, a large public volatility translates into the dealer knowing that, on average, the informed trader must have observed a value far above the public mean. This is the *dynamic efficiency effect*: a low informed share leads to a larger update of the public mean, and hence increases the bid-ask

[2]To simplify the analysis, we assume here (as in Section 4) that the dealer always considers the public density to be normal: after observing the order flow at t, the dealer computes correctly the first two moments of the public density at $t+1$, but not the higher moments, and thus considers the new density to be normal as well. In Section 5, we show that the main results remain robust with exact learning.

[3]The existence itself of a stationary limit is not entirely obvious. Indeed, it might be possible for the public volatility to grow indefinitely, without any finite limit.

spread. At the other end, a large informed share means that the dealer learns well about the asset value (the public volatility is small), and therefore the bid-ask spread tends to be small.

It turns out that the two opposite effects exactly offset each other. As explained above, this translates into the fact that the magnitude of the updates in public mean caused by order flow is independent of the informed share. This result depends crucially on the equilibrium being stationary. To understand why, consider an equilibrium which is not necessarily stationary. If there was no order flow at t, then the dealer's uncertainty (the public volatility) would increase from t to $t+1$ as the asset value diffuses. But the order flow at t contains information and hence reduces the uncertainty at $t+1$. In a stationary equilibrium the uncertainty increase caused by diffusion must cancel the uncertainty decrease caused by order flow. Thus, as the value diffusion is independent of the informed share, the information content of the order flow must also be independent of the informed share. But this implies that the magnitude of public mean updates is independent of the informed share.

Our next result is that, for any initial public volatility, the equilibrium converges to the stationary equilibrium. In particular, consider a wide initial public volatility. Then, as the order flow starts providing information to the dealer, the public volatility starts decreasing toward its stationary value. The same is true for the bid-ask spread, which in a non-stationary equilibrium is always proportional to the public volatility. This phenomenon is similar to the GM85 equilibrium, except that there the stationary public volatility and bid-ask spread are both zero. This illustrates the statement made above, that the non-stationary equilibrium (the "short run") resembles GM85, while the stationary equilibrium (the "long run") is different and produces novel insights.

Studying the equilibrium behavior after various types of shocks provides a few testable implications. First, consider a positive shock to the informed share (e.g., the stock is now studied by more hedge funds). Then, the adverse selection effect suddenly becomes stronger, and as a result the bid-ask spread temporarily increases. In the long run, though, the bid-ask spread reverts to its stationary value, which does not change. At the same time, the public volatility gradually decreases to its new level, which is lower due to the increase in the informed share. Second, consider a negative shock to the current public volatility (e.g., public

news about the current asset value). Then, the bid-ask spread follows the public volatility and drops immediately, after which it increases gradually to its old stationary level. Third, consider a positive shock to the fundamental volatility (e.g., all future uncertainty about the asset increases). Then, the bid-ask spread follows the public volatility and increases gradually to its new stationary level.

Based on our results, the picture on dynamic adverse selection that emerges is that liquidity is more strongly affected not by the informed share (the intensive margin), but by the fundamental volatility (the extensive margin). By contrast, price discovery (measured by the public volatility) is strongly affected by both informed share and fundamental volatility. This suggests that the presence of privately informed traders can be more precisely identified by proxies of the current level of uncertainty, rather than by illiquidity measures such as the bid-ask spread (which is used by Collin-Dufresne and Fos, 2015).

A surprising outcome of our theory is that a lower level of uncertainty (lower public volatility) can occur if either the informed share becomes larger (more privately informed traders arrive), or more precise public news arrives.[4] We can disentangle the two scenarios, however, by examining the effect on the bid-ask spread: more precise public news should reduce it, while a larger informed share should have no effect.

Our paper contributes to the literature of dynamic models of adverse selection.[5] To our knowledge, this paper is the first to study the effect of stationarity in dealer models of the Glosten and Milgrom (1985) type.[6] By contrast, several stationary models of the Kyle (1985) type are analyzed for instance by Chau and Vayanos (2008) and Caldentey and Stacchetti (2010). The focus of these models, however, is not liquidity but price discovery: it turns out that in the limit the market in this models becomes strong-form efficient, as the insider trades very aggressively.

The paper speaks to the literature on the identification of informed trading and in particular on the identification of insider trading. Collin-Dufresne and Fos (2015, 2016) show

[4] In Section 6.2 we solve a simple extension of our model with public news, and show that more precise public news translates into a lower public volatility.

[5] See for instance the survey of Foucault, Pagano, and Röell (2013) and the references therein.

[6] Glosten and Putnins (2016) study the welfare effect of the informed share in the Glosten and Milgrom (1985) model, but they do not consider the effect of stationarity.

both empirically and theoretically that times when insiders trade coincide with times when liquidity is actually stronger (and in particular bid-ask spreads decline). They attribute this finding to the action of discretionary insiders who trade when they expect a larger presence of liquidity (noise) traders. During those times the usual positive effect of noise traders on liquidity dominates, and thus bid-ask spreads decline despite there being more informed trading. By contrast, our effect works even when the noise trader activity is constant over time, as long as there is enough time for the equilibrium to become stationary.

The paper is organized as follows. Section 2 describes the model (in which the value follows a random walk). Section 3 shows how to compute the equilibrium when the dealer is fully Bayesian. Section 4 studies in detail the model in which the dealer is approximately Bayesian, and describes the stationary and non-stationary equilibria. Section 5 verifies how well the approximate equilibrium approaches the exact equilibrium. Section 7 concludes. All proofs are in the Appendix. The Internet Appendix contains a discussion of general dealer models, and an application to a model in which the fundamental value switches randomly between zero and one.

2 Environment

The model is similar to GM85, except that the fundamental value moves according to a random walk:

$$v_{t+1} \;=\; v_t + \varepsilon_{t+1}, \quad \text{with} \quad \varepsilon_t \overset{IID}{\sim} \mathcal{N}\left(\cdot, 0, \sigma_v\right). \tag{1}$$

There is a single risky asset, and time is discrete and infinite. Trading in the risky asset is done on an exchange, where before each time $t = 0, 1, 2, \ldots$ a dealer posts two quotes: the *ask price* (or simply *ask*) A_t, and the *bid price* (or simply *bid*) B_t. Thus, a buy order at t executes at A_t, while a sell order at t executes at B_t. The dealer (referred to in the paper as "she") is risk neutral and competitive, and therefore makes zero expected profits from each trade.

The buy or sell orders are submitted by a trading population with a fraction $\rho \in (0, 1)$ of informed traders and a fraction $1 - \rho$ of uninformed traders. At each $t = 0, 1, \ldots$ a trader is selected at random from the population willing to trade, and can trade at most one unit of

the asset. An uninformed trader at t is always willing to trade, and is equally likely to buy or to sell.[7] An informed trader at t who observes the value v_t either (i) submits a buy order if $v_t > A_t$, (ii) submits a sell order if $v_t < B_t$, (iii) is not willing to trade if $v_t \in [B_t, A_t]$. If case (iii) occurs, an uninformed trader is selected, as no informed trader is willing to trade.

The dealer's uncertainty about the fundamental value is summarized by the *public density*, which is the density of v_t just before trading at t, conditional on all the order flow available at t, that is, the sequence of orders submitted at times $0, 1, \ldots, t-1$. Denote by ϕ_t the public density, by μ_t its mean (called the *public mean*) and by σ_t its standard deviation (called the *public volatility*). The initial density ϕ_0 is assumed to be rapidly decaying at infinity.[8] In the rest of the paper, by "density" we typically include the requirement that the density be rapidly decaying. To avoid cumbersome language, we make this requirement explicit only when we state the formal results.

3 Equilibrium

We prove the existence of an equilibrium of the model in two steps. First, for each $t = 0, 1, 2, \ldots$ we start with an public density ϕ_t, an ask A_t, a bid $B_t < A_t$, and compute the public density ϕ_{t+1} after a buy or sell order. Second, for any public density ϕ_t we show that there exists an *ask-bid pair* (A_t, B_t), meaning that the ask A_t and the bid B_t satisfy the dealer's pricing conditions which require that her expected profit from trading at t is zero. The ask-bid pair (A_t, B_t) is not necessarily unique, and we choose the pair with the ask closest to the public mean.

[7] A standard way to endogenize this assumption is to introduce relative private valuations for the uninformed traders. For instance, if a trader expects the value to be μ_t and has a relative private valuation larger than $A_t - \mu_t$ (which in equilibrium is half the bid-ask spread), the trader is always willing to buy at A_t.

[8] A function f is rapidly decaying (at infinity) if it is smooth and satisfies $\lim_{v\to\pm\infty} |v|^M f_0^{(N)}(v) = 0$, where $f^{(N)}$ is the N-th derivative of f. The space $\mathcal{S}$ of rapidly decaying functions is called the Schwartz space. Any normal density belongs to $\mathcal{S}$, and the convolution of two densities in $\mathcal{S}$ also belongs to $\mathcal{S}$.

3.1 Evolution of the Public Density

Let ϕ_t be the public density of v_t before trading at t, and let $A_t > B_t$ be, respectively, the ask and bid at t (not necessarily satisfying the dealer's pricing conditions). Suppose a buy or sell order $\mathcal{O}_t \in \{\mathrm{B}, \mathrm{S}\}$ arrives at t. Let $\mathbf{1}_P$ be the indicator function, which is one if P is true and zero if P is false. Conditional on $v_t = v$, the probability of observing the a buy order at t is

$$g_t(\mathrm{B}, v) \;=\; \rho \mathbf{1}_{v>A_t} + \tfrac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]} + \tfrac{1-\rho}{2}. \tag{2}$$

To see this, consider the following cases:

- If $v \in [B_t, A_t]$, the informed traders are not willing to trade, and an uninformed trader submits a buy order with probability $\frac{1}{2}$. Then, $g_t(\mathrm{B}, v) = \rho \times 0 + \frac{\rho}{2} \times 1 + \frac{1-\rho}{2} = \frac{1}{2}$.
- If $v \notin [B_t, A_t]$, an informed trader (chosen with probability ρ) submits a buy order with probability $\mathbf{1}_{v>A_t}$, while an uninformed trader (chosen with probability $1-\rho$) submits a buy order with probability $\frac{1}{2}$. Then, $g_t(\mathrm{B}, v) = \rho\mathbf{1}_{v>A_t} + \frac{\rho}{2} \times 0 + \frac{1-\rho}{2}$

Similarly, the probability of observing a sell order at t is

$$g_t(\mathrm{S}, v) \;=\; \rho \mathbf{1}_{v<B_t} + \tfrac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]} + \tfrac{1-\rho}{2}. \tag{3}$$

The next result describes the evolution of the public density.

Proposition 1. *Consider a rapidly decaying public density ϕ_t, and an ask-bid pair with $A_t > B_t$. After observing an order $\mathcal{O}_t \in \{\mathrm{B}, \mathrm{S}\}$, the density of v_t is $\psi_t(v|\mathcal{O}_t)$, where*

$$\begin{aligned} \psi_t(v|\mathrm{B}) \;&=\; \frac{\left(\rho\mathbf{1}_{v>A_t} + \frac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]} + \frac{1-\rho}{2}\right)\cdot\phi_t(v)}{\frac{\rho}{2}(1-\Phi_t(A_t)) + \frac{\rho}{2}(1-\Phi_t(B_t)) + \frac{1-\rho}{2}}, \\ \psi_t(v|\mathrm{S}) \;&=\; \frac{\left(\rho\mathbf{1}_{v<B_t} + \frac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]} + \frac{1-\rho}{2}\right)\cdot\phi_t(v)}{\frac{\rho}{2}\Phi_t(A_t) + \frac{\rho}{2}\Phi_t(B_t) + \frac{1-\rho}{2}}, \end{aligned} \tag{4}$$

where Φ_t is the cumulative density function corresponding to ϕ_t. The public density at $t+1$

is rapidly decaying, and satisfies

$$\phi_{t+1}(w|\mathcal{O}_t) \;=\; \int_{-\infty}^{+\infty} \psi_t(v|\mathcal{O}_t)\mathcal{N}(w-v,0,\sigma_v)\mathrm{d}v \;=\; \Big(\psi_t(\cdot|\mathcal{O}_t) \,*\, \mathcal{N}(\cdot,0,\sigma_v)\Big)(w), \qquad (5)$$

where "$*$" denotes the convolution of two densities.

Proposition 1 shows how the public density evolves once a particular order (buy or sell) is submitted at t. Note, however, that this result does not assume anything about the ask and bid other than $A_t > B_t$, so in principle these can be chosen arbitrarily. In equilibrium, however, these prices must satisfy the dealer's pricing conditions, namely that the dealer's expected profits at t must be zero.

In the next section (Section 3.2) we impose these conditions and we show how to determine the equilibrium ask and bid. Then, Proposition 1 allows us to describe the whole evolution of the public density, conditional on the initial density ϕ_0 and the sequence of orders $\mathcal{O}_0, \mathcal{O}_1, \ldots$ that have been submitted.

3.2 Ask and Bid Prices

Let ϕ_t be the public density of v_t before trading at t. We define an *ask-bid pair* (A_t, B_t) as a pair of ask and bid satisfying the pricing conditions of the dealer. As the dealer is risk neutral and competitive, the pricing conditions are: (i) the ask A_t is the expected value of v_t conditional on a buy order at t, and (ii) the bid B_t is the expected value of v_t conditional on a sell order at t. Using the previous notation, the dealer's pricing conditions are that A_t is the mean of $\psi_t(v|\mathrm{B})$, the posterior density of v_t after observing a buy order at t; and B_t is the mean of $\psi_t(v|\mathrm{S})$, the posterior density after observing a sell order at t. Thus, the dealer's pricing conditions are equivalent to

$$A_t \;=\; \int_{-\infty}^{+\infty} v\psi_t(v|\mathrm{B})\mathrm{d}v, \qquad B_t \;=\; \int_{-\infty}^{+\infty} v\psi_t(v|\mathrm{S})\mathrm{d}v. \qquad (6)$$

For future use, we record the following straightforward result.

Corollary 1. *The pair (A_t, B_t) is an ask-bid pair if and only if the following equations are satisfied:*

$$A_t \;=\; \mu_{t+1,\mathrm{B}}, \quad B_t \;=\; \mu_{t+1,\mathrm{S}}, \quad \textit{with} \quad \mu_{t+1,\mathcal{O}_t} \;=\; \int_{-\infty}^{+\infty} w\phi_{t+1}(w|\mathcal{O}_t)\mathrm{d}w, \quad \mathcal{O}_t = \{\mathrm{B},\mathrm{S}\}. \tag{7}$$

The next result shows that the existence of an ask-bid pair is equivalent to solving a 2×2 system of nonlinear equations. Suppose μ_t is the mean of ϕ_t. For $(A,B)\in(\mu_t,\infty)\times(-\infty,\mu_t)$, define the functions:

$$\begin{aligned} F(A,B) \;&=\; \frac{\Theta_t(A)+\Theta_t(B)}{A-\mu_t} - \frac{1+\rho}{\rho} + \Phi_t(A) + \Phi_t(B), \\ G(A,B) \;&=\; \frac{\Theta_t(A)+\Theta_t(B)}{\mu_t-B} - \frac{1-\rho}{\rho} - \Phi_t(A) - \Phi_t(B), \end{aligned} \tag{8}$$

where Φ_t is the cumulative density associated to ϕ_t, and Θ_t is defined by

$$\Theta_t(v) \;=\; \int_{-\infty}^{v} (\mu_t - w)\phi_t(w)\mathrm{d}w. \tag{9}$$

The function Θ_t is strictly positive everywhere and approaches zero at infinity on both sides.[9]

Proposition 2. *Consider a rapidly decaying public density ϕ_t, with mean μ_t. Then, the existence of an ask-bid pair is equivalent to finding a solution $(A,B)\in(\mu_t,\infty)\times(-\infty,\mu_t)$ of the system of equations:*

$$F(A,B) \;=\; 0, \qquad G(A,B) \;=\; 0. \tag{10}$$

A solution of (10) always exists. Among the set of ask-bid pairs (A,B) there is a unique one for which A is closest to μ_t.

The last statement in Proposition 2 shows that one can choose a unique ask-bid pair based on the criterion that the ask A be the closest to the public mean μ_t. Denote this pair by (A_t, B_t). In the rest of the paper, we assume that this is indeed the ask-bid pair chosen by the dealer.[10]

[9] As ϕ_t is rapidly decaying, $\Theta_t(-\infty)$ is equal to zero. The definition of μ_t implies that $\Theta_t(+\infty) = \int_{-\infty}^{+\infty}(\mu_t - w)\phi_t(w) = \mu_t - \int_{-\infty}^{+\infty} w\phi_t(w) = 0$. Also, $\Theta_t'(v) = (\mu_t - v)\phi_t(v)$, hence $\Theta_t(v)$ is increasing below μ_t and decreasing above μ_t. As $\Theta_t(\pm\infty) = 0$, the function Θ_t is strictly positive everywhere.

[10] In principle, the equations in (10) might have multiple solutions, meaning that one could manufacture an

4 Equilibrium with Approximate Bayesian Inference

In this section, we assume that at each step the dealer approximates the public density with a normal density such that the first two moments are correctly computed. Specifically, suppose that the dealer regards v_t to be distributed as

$$\phi_t^a(v) \;=\; \mathcal{N}(v, \mu_t, \sigma_t). \tag{11}$$

After the dealer observes an order $\mathcal{O}_t$ at t, denote by $\phi_{t+1}(w|\mathcal{O}_t)$ the exact density of v_{t+1} conditional on the past order flow including $\mathcal{O}_t$, and by $\mu_{t+1,\mathcal{O}_t}$ and $\sigma_{t+1,\mathcal{O}_t}$ its mean and standard deviation, respectively. Then, before trading at $t+1$ the dealer regards v_{t+1} to be distributed as

$$\phi_{t+1}^a(w|\mathcal{O}_t) \;=\; \mathcal{N}\big(w, \mu_{t+1,\mathcal{O}_t}, \sigma_{t+1,\mathcal{O}_t}\big). \tag{12}$$

Thus, we assume that the dealer continues to make the approximation at each step:

$$\phi_t \;=\; \phi_t^a. \tag{13}$$

Section 5 discusses the accuracy of this approximation. For simplicity, we continue to refer to $\phi_t(v)$ as the *public density*, μ_t as the *public mean*, and σ_t as the *public volatility*.

4.1 Evolution of the Public Density

Proposition 3. *Suppose the public density at $t = 0, 1, 2, \ldots$ is $\phi_t(v) = \mathcal{N}(v, \mu_t, \sigma_t)$. After observing $\mathcal{O}_t \in \{\mathrm{B}, \mathrm{S}\}$, the posterior mean and volatility at $t+1$ satisfy*

$$\mu_{t+1,\mathrm{B}} \;=\; \mu_t + \delta\sigma_t, \quad \mu_{t+1,\mathrm{S}} \;=\; \mu_t - \delta\sigma_t, \quad \sigma_{t+1,\mathrm{B}} \;=\; \sigma_{t+1,\mathrm{S}} \;=\; \sqrt{(1-\delta^2)\sigma_t^2 + \sigma_v^2}. \tag{14}$$

public density ϕ_t for which there is more that one corresponding ask-bid pair. Numerically, we have computed the sequence of public densities that starts with a normal density ϕ_0 and is associated by an arbitrary sequence of orders, but we have not yet been able to encounter a non-unique ask-bid pair. Nevertheless, we must account for the possibility that such non-uniqueness may in fact arise.

where δ is defined by:

$$\delta \;=\; g^{-1}(2\rho), \quad \textit{with} \quad g(x) \;=\; \frac{x}{\mathcal{N}(x,0,1)}. \tag{15}$$

There is a unique ask: $A_t = \mu_t + \delta\sigma_t$ and unique bid: $B_t = \mu_t - \delta\sigma_t$, and the bid-ask spread is

$$s_t \;=\; A_t - B_t \;=\; 2\delta\sigma_t. \tag{16}$$

We now investigate whether the public density reaches a steady state, in the sense that its shape converges to a particular density. As the mean μ_t evolves according to a random walk, we must demean the public density and focus on its standard deviation σ_t. The next result shows that the public volatility σ_t converges to a particular value, σ_*, regardless of the initial value σ_0.

Proposition 4. *For any $t = 0, 1, 2, \ldots$ the public volatility satisfies*

$$\sigma_t^2 \;=\; \sigma_*^2 + \left(\sigma_0^2 - \sigma_*^2\right)(1-\delta^2)^t, \tag{17}$$

where

$$\sigma_* \;=\; \frac{\sigma_v}{\delta} \;=\; \frac{\sigma_v}{g^{-1}(2\rho)}. \tag{18}$$

For any initial value σ_0 and any sequence of orders, the public volatility σ_t monotonically converges to σ_*, and the bid-ask spread monotonically converges to

$$s_* \;=\; 2\sigma_v. \tag{19}$$

Thus, Proposition 4 shows that in the long run the equilibrium approaches a particular stationary equilibrium, which we analyze next.

4.2 Stationary Equilibrium

We define a *stationary equilibrium* an equilibrium in which the public volatility σ_t is constant. According to Proposition 4, if the initial density is $\phi_0(v) = \mathcal{N}(v, \mu_0, \sigma_*)$, then all subsequent

public densities have the same volatility, namely the stationary volatility σ_*. We now analyze the properties of the stationary equilibrium.

Corollary 2. *In the stationary equilibrium, the public volatility σ_* is decreasing in the fraction of informed trading ρ, while the bid-ask spread s_* does not depend on ρ. Both σ_* and s_* are increasing in the fundamental volatility σ_v.*

Intuitively, an increase in the fundamental volatility σ_v raises the public volatility as the dealer's knowledge about the fundamental value becomes more imprecise. It also increases the adverse selection overall for the dealer, hence she increases the bid-ask spread. Moreover, a decrease in the fraction of informed trading ρ means that the order flow becomes less informative, and therefore the dealer's knowledge about the fundamental value is more imprecise (σ_* is large).

The surprising result is that the stationary bid-ask spread is independent of ρ. This is equivalent to the public mean update being independent of ρ. Indeed, the public mean evolves according to

$$\mu_{t+1,\mathrm{B}} \;=\; \mu_t + \sigma_v, \qquad \mu_{t+1,\mathrm{S}} \;=\; \mu_t - \sigma_v. \tag{20}$$

Thus, the bid-ask spread is $s_* = (\mu_t + \sigma_v) - (\mu_t - \sigma_v) = 2\sigma_v$. To understand the intuition behind this result, consider the case when ρ is low. Suppose the dealer observes a buy order at t. As ρ is low, there are two effects on the size of the public mean update. The first effect is negative: the trader at t is unlikely to be informed, which decreases the size of the update. This is the traditional *adverse selection effect* from models such as GM85. The second effect is positive: when the trader at t *is* informed, he must have observed a large fundamental value v_t, as the uncertainty in v_t (measured by the public volatility σ_*) is also large. This we call the *dynamic efficiency effect*: more informed traders create over time a more precise knowledge about the fundamental value, and thus reduce the effect of informational updates.

It turns out that the dynamic efficiency effect exactly cancels the adverse selection effect in a stationary setup, and as a result the magnitude of the public mean updates due to order flow is independent of ρ. To understand why, consider an equilibrium which is not necessarily stationary. If there was no order flow at t, then the dealer's uncertainty (the public volatility) would increase from t to $t+1$ as the fundamental value diffuses. But there

is order flow at t, which provides information to the dealer and hence reduces uncertainty at $t+1$. In a stationary equilibrium the public uncertainty stays constant. Thus, as the increase in uncertainty due to value diffusion is independent of the informed share ρ, the decrease in uncertainty due to order flow should also be independent of ρ. But an order flow information content that is independent of ρ translates into the magnitude of public mean updates also being independent of ρ.

Formally, the decrease in uncertainty due to the order $\mathcal{O}_t$ at t can be evaluated by comparing the prior public density $\phi_t(v)$ and the posterior density $\psi_t(v|\mathcal{O}_t)$. One measure of the decrease in uncertainty is how much the public mean is updated after a buy or sell order (which are equally likely). But (20) implies that this update is $\pm\sigma_v$, which from the point of view of the information at t is a binary distribution, with standard deviation σ_v which is indeed independent of ρ. Note that we have also essentially proved the following result.

Corollary 3. *In the stationary equilibrium, the volatility of the change in public mean is constant and equal to σ_v.*

This result is in fact true quite generally. Indeed, in Appendix B we prove that for any filtration problem in which the variance remains constant over time the volatility of the change in public mean must equal the fundamental volatility.

4.3 Liquidity Dynamics

In this section we analyze the evolution of the public volatility and the bid-ask spread after a shock to either the public volatility σ_t, the fundamental volatility σ_v, or the fraction of informed trading ρ. We are also interested in how quickly the equilibrium converges to the stationary equilibrium. In general, the speed of convergence of a sequence x_t that converges to a limit x_* is defined as the limit ratio

$$S \;=\; \lim_{t\to\infty} \frac{|x_t^2 - x_*^2|}{|x_{t+1}^2 - x_*^2|}, \tag{21}$$

provided that the limit exists. The next result computes the speed of convergence for several variables of interest.

Corollary 4. *The public volatility, public variance and bid-ask spread have the same speed of convergence:*

$$S = \frac{1}{1-\delta^2}. \tag{22}$$

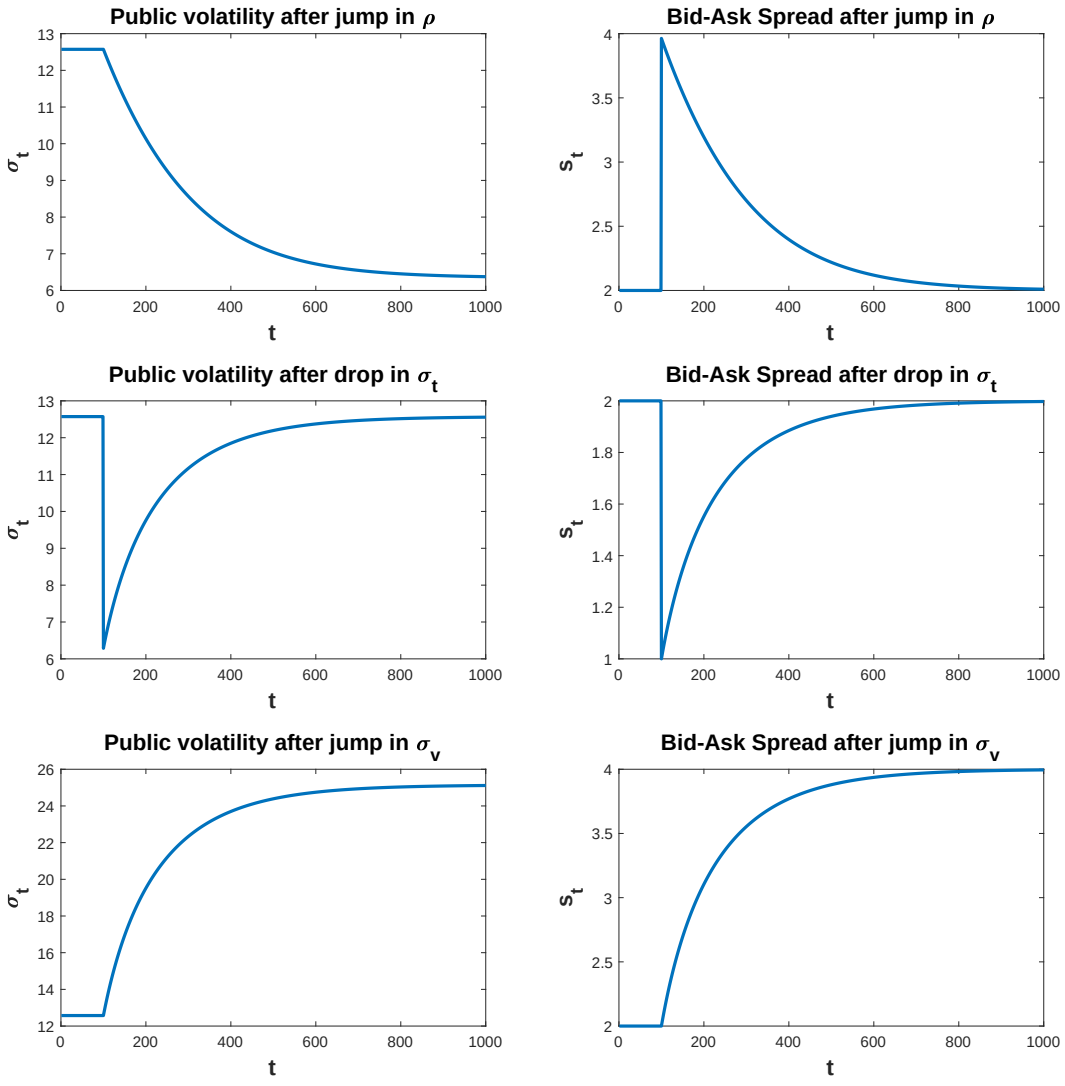
Moreover, S is increasing in the fraction of informed trading ρ.

Corollary 4 shows that the variables of interest have the same speed of convergence S, and we can thus call S simply as the *convergence speed* of the equilibrium. Another result of Corollary 4 is that a larger fraction of informed trading ρ implies a faster convergence speed of the equilibrium to its stationary value. This is intuitive, as more informed trading helps the dealer make quicker dynamic inferences. Note that when $\rho = 1$, equation (15) implies that $\delta = g^{-1}(2) \approx 0.647$, thus the maximum value of δ is less than one. Therefore, the maximum convergence speed is finite.

We now consider the effect of various types of shocks to our stationary equilibrium. In the first row of Figure 1 we plot the effects of a positive shock to the fraction of informed trading, meaning that ρ suddenly jumps to a higher value ρ'. This generates an increase in δ, which jumps to its new value $\delta' = g^{-1}(\rho')$, and it also generates a drop in the stationary public volatility, which is now $\sigma'_* = \sigma_v/\delta'$. Nevertheless, as there is no new information above the fundamental value, the current public volatility σ_t remains equal to its old stationary value, $\sigma_* = \sigma_v/\delta$. Proposition 4 shows that the public volatility starts decreasing monotonically toward its stationary value σ'_*. Note that according to Corollary 4 the speed of convergence to the new stationary equilibrium is $S' = 1/(1-\delta'^2)$, which is higher than the old convergence speed. We also describe the evolution of the bid-ask spread, which according to Proposition 3 satisfies $s_t = 2\delta'\sigma_t$. Initially, the bid-ask spread jumps to reflect the jump to δ'. But then, as σ_t converges to $\sigma'_* = \sigma_v/\delta'$, the bid-ask spread starts decreasing to $s_* = 2\sigma_v$, which does not depend on ρ.

To summarize, after a positive shock to ρ, the public volatility starts decreasing monotonically to its now lower stationary value, while the bid-ask spread initially jumps and then decreases monotonically to the same stationary value (that does not depend on ρ). Intuitively, a positive shock to the fraction of informed trading leads to a sudden increase in adverse selection for the dealer, reflected in an initially larger bid-ask spread, after which the bid-ask

Figure 1: Public Volatility and Bid-Ask Spread after Shocks.
This figure plots the effect of three types of shocks on the public volatility σ_t, and on the bid-ask spread s_t (each shock occurs at $t_0 = 100$). The initial parameters are: $\sigma_v = 1$, and $\rho = 0.1$ (hence $\delta = 0.0795$, $\sigma_* = 12.573$, $s_* = 2$). In the first row, the fraction of informed trading ρ jumps from 0.1 to 0.2 (hence σ_* drops from 12.573 to 6.345). In the second row, the public volatility drops from $\sigma_* = 12.573$ to half of its value (6.286). In the third row, the fundamental volatility jumps from 1 to 2.



spread reverts to its fundamental value, which is independent of informed trading. At the same time, more informed trading leads to more precision for the dealer in the long run, which is reflected in a smaller public volatility.

In the second row of Figure 1 we plot the effects of a negative shock to the public volatility, meaning that σ_t suddenly drops from the stationary value σ_* to a lower value. This drop can be caused for instance by public news about the value of the asset v_t. Then, according to Proposition 4, the public volatility increases monotonically back to the stationary value. The bid-ask spread is always proportional to the public density: $s_t = 2\delta\sigma_t$, hence s_t also drops initially and then increases monotonically toward the stationary value s_*. Intuitively, public news has the effect of helping the dealer initially to get a more precise understanding about the fundamental value. This brings down the bid-ask spread, as temporarily the dealer faces less adverse selection. But this decrease is only temporary, as the value diffuses and the same forces increase the public volatility and the bid-ask spread toward their corresponding stationary values, which are the same as before.

In the third row of Figure 1 we plot the effects of a positive shock to the fundamental volatility, meaning that σ_v suddenly jumps to a higher value σ'_v. This implies that every value increment $v_{t+1} - v_t$ now has higher volatility, but the uncertainty in v_t, which is measured by the public volatility σ_t, stays the same.[11] Proposition 4 shows that the stationary public volatility changes to $\sigma'_* = \sigma'_v/\delta$, and the stationary bid-ask spread changes to $s'_* = 2\sigma'_v$. Therefore, the public density increases monotonically from the initial stationary value to the new stationary value, and the same is true for the bid-ask spread. Intuitively, a larger fundamental volatility increases overall adverse selection for the dealer, and as a result both the public density and the bid-ask spread eventually increase.

5 Equilibrium with Exact Bayesian Inference

In this section, we analyze in more detail the evolution of the public density ϕ_t when the dealer is fully Bayesian. In particular, we are interested in computing the average shape of the public density over all possible future paths of the game.[12] Note that when computing the average

[11] One can mix this type of shock with a shock to the public volatility σ_t, which was already analyzed.

[12] As we see in Internet Appendix (Sections 1 and 2), one expects a well defined stationary density for the continuous time Markov chain associated to our game. The only problem is that the fundamental value v_t is no longer stationary in our case, but follows a random walk. One can show that it is still possible to define a stationary density as long as one does not require it to integrate to one over v. But we are interested in the simpler problem of computing the marginal stationary density of $v_t - \mu_t$, which we solve numerically.

shape of a density, we consider the average of the various densities after demeaning them. We then show numerically that this average exists and is not too far from the stationary public density described in Section 4.2, which is a normal density with mean zero and standard deviation equal to $\sigma_* = \sigma_v/\delta$.

We thus demean the variables and densities involved in the previous formulas, and in addition we normalize them by σ_*:

$$\begin{aligned} \tilde{A}_t &= \frac{A_t - \mu_t}{\sigma_*}, \qquad \tilde{B}_t = \frac{B_t - \mu_t}{\sigma_*}, \qquad \tilde{v}_t = \frac{v_t - \mu_t}{\sigma_*}, \qquad \tilde{\phi}_t(\tilde{v}) = \sigma_* \phi_t(\mu_t + \sigma_* \tilde{v}), \\ \tilde{\Phi}_t(\tilde{v}) &= \int_{-\infty}^{\tilde{v}} \tilde{\phi}_t(w) \mathrm{d}w, \qquad \tilde{\psi}_t(\tilde{v}|\mathcal{O}_t) = \sigma_* \psi_t(\mu_t + \sigma_* \tilde{v}|\mathcal{O}_t). \end{aligned} \tag{23}$$

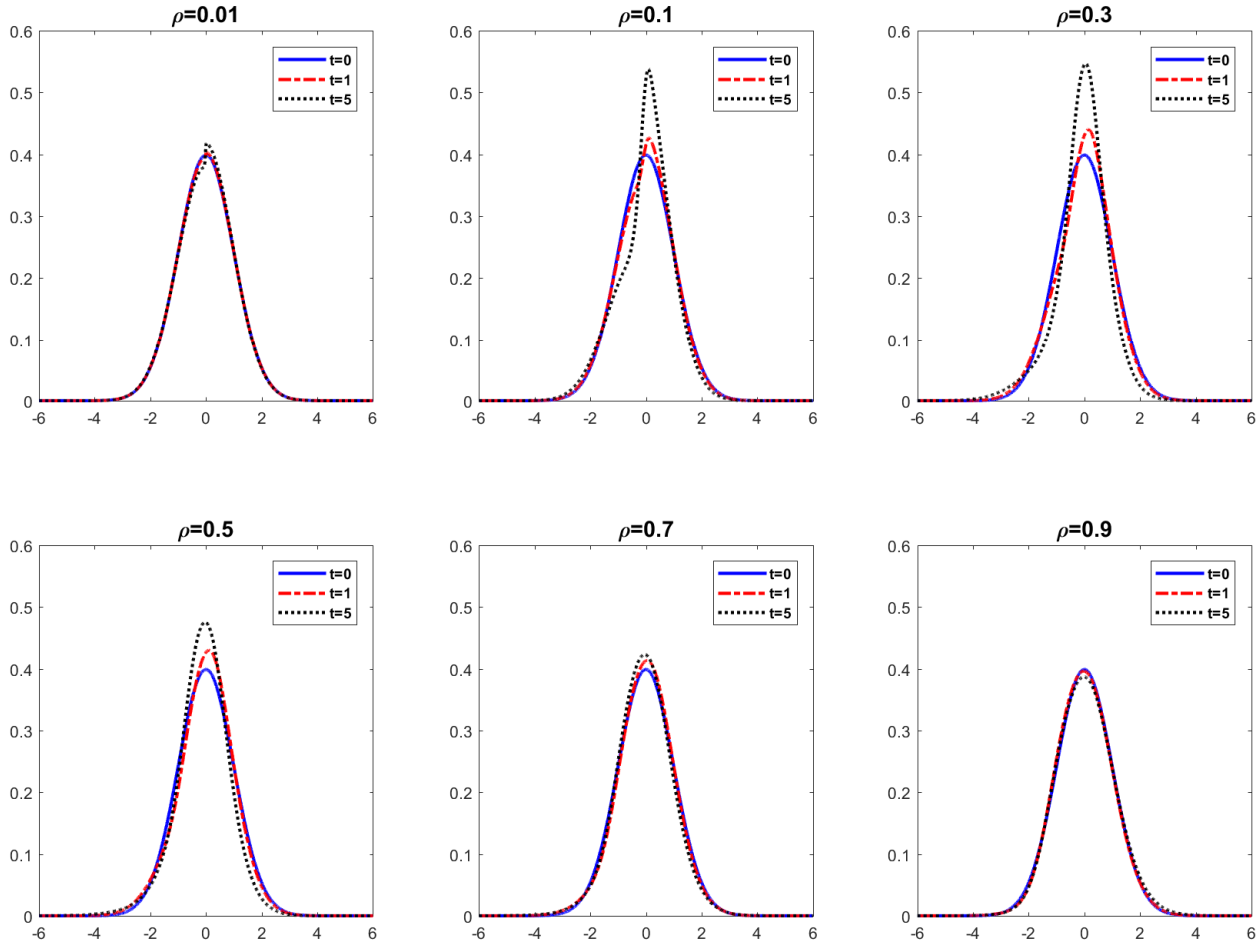
With this new notation, the equations (4) and (5) from Proposition 1 imply the following result.

Corollary 5. *Consider a rapidly decaying public density ϕ_t with normalization $\tilde{\phi}_t$, and an ask-bid pair (A_t, B_t) with normalization $(\tilde{A}_t, \tilde{B}_t)$. After observing an order $\mathcal{O}_t \in \{\mathrm{B}, \mathrm{S}\}$, the normalized density at $t+1$ is $\tilde{\phi}_{t+1}(\tilde{w}|\mathcal{O}_t)$, where*

$$\begin{aligned} \tilde{\phi}_{t+1}(\tilde{w}|\mathrm{B}) &= \int_{-\infty}^{+\infty} \mathcal{N}\Big(\frac{\tilde{w} - \tilde{v} + \tilde{A}_t}{\delta}\Big) \frac{\big(\rho \mathbf{1}_{\tilde{v} > \tilde{A}_t} + \frac{\rho}{2} \mathbf{1}_{\tilde{v} \in [\tilde{B}_t, \tilde{A}_t]} + \frac{1-\rho}{2}\big) \cdot \tilde{\phi}_t(\tilde{v})}{\frac{\rho}{2}(1 - \tilde{\Phi}_t(\tilde{A}_t)) + \frac{\rho}{2}(1 - \tilde{\Phi}_t(\tilde{B}_t)) + \frac{1-\rho}{2}} \, \mathrm{d}\tilde{v}, \\ \tilde{\phi}_{t+1}(\tilde{w}|S) &= \int_{-\infty}^{+\infty} \mathcal{N}\Big(\frac{\tilde{w} - \tilde{v} + \tilde{B}_t}{\delta}\Big) \frac{\big(\rho \mathbf{1}_{\tilde{v} < \tilde{B}_t} + \frac{\rho}{2} \mathbf{1}_{\tilde{v} \in [\tilde{B}_t, \tilde{A}_t]} + \frac{1-\rho}{2}\big) \cdot \tilde{\phi}_t(\tilde{v})}{\frac{\rho}{2}\tilde{\Phi}_t(\tilde{A}_t) + \frac{\rho}{2}\tilde{\Phi}_t(\tilde{B}_t) + \frac{1-\rho}{2}} \, \mathrm{d}\tilde{v}. \end{aligned} \tag{24}$$

Figure 2 displays the normalized public density after $t = 0$, $t = 1$, and $t = 5$ buy orders for various values of the informed share ρ. We notice by visual inspection that the normalized public density is close to the standard normal density even after a sequence of 5 buy orders (this sequence happens with probability 2^{-5}, which is approximately 3.13%). The deviation of the normalized public densities from the standard normal density is at its smallest level when the fraction of informed trading ρ is either small or large, and it peaks for an intermediate value ρ near 0.2. When ρ is small, the order flow is uninformative, hence the posterior is not far from the prior. When ρ is large, the order flow is very informative, hence the posterior depends strongly on the increment, which is normally distributed.

Figure 2: Exact Normalized Public Density after Series of Buy Orders.
Each of the 6 plots represents the evolution of the normalized public density $\tilde{\phi}_t$ after $t = 0$, $t = 1$ and $t = 5$ buy orders. The initial normalized public density in all cases (at $t = 0$) is the standard normal density with mean zero and volatility one. The 6 plots correspond to the fraction of informed trading $\rho \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9\}$.



It turns out, however, that the stationary shape of the public density is not precisely normal, but it has "fat tails," that is, its fourth centralized moment (kurtosis) is larger than 3. Table 1 displays, for each $\rho \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9\}$, several moments of the average normalized density computed after 200 different random paths. As the starting density (at $t = 0$) is standard normal for all the different ρ, we need to make sure that we choose a path length long enough for the average density to stabilize. Numerically, we see that it is enough

Table 1: Average Normalized Public Density after Series of Random Orders.
For each informed share $\rho \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9\}$, consider 200 random series of 20 orders chosen among buy or sell with equal probability, and denote by $\tilde{\phi}_S$ the normalized public density computed after observing the series $S = 1, 2, \ldots, 200$. The table displays four estimated moments of the average $\psi = \frac{\tilde{\phi}_1 + \tilde{\phi}_2 + \cdots + \tilde{\phi}_{200}}{200}$: the mean $\mu = \int_{-\infty}^{+\infty} x\psi(x)\mathrm{d}x$, the standard deviation $\sigma = \left(\int_{-\infty}^{+\infty}(x-\mu)^2\psi(x)\mathrm{d}x\right)^{1/2}$, the skewness $\int_{-\infty}^{+\infty}\left(\frac{x-\mu}{\sigma}\right)^3\psi(x)\mathrm{d}x$, and the kurtosis $\int_{-\infty}^{+\infty}\left(\frac{x-\mu}{\sigma}\right)^4\psi(x)\mathrm{d}x$. It also displays the average bid-ask spread normalized by $s_* = 2\sigma_v$ (N.Spread).

ρ	0.01	0.1	0.3	0.5	0.7	0.9
Mean	-0.000	0.000	-0.001	-0.001	-0.010	-0.002
St.Dev.	1.000	1.002	1.039	1.056	1.041	1.009
Skewness	-0.001	0.018	0.016	-0.003	-0.012	-0.009
Kurtosis	3.005	3.419	4.587	4.597	4.089	3.343
N.Spread	1.003	0.966	0.959	0.988	1.014	1.004

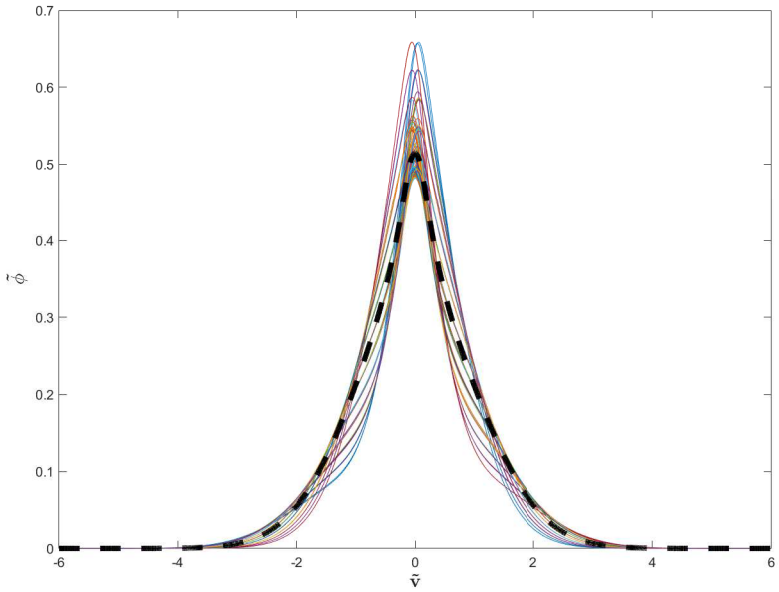
to choose $t = 20$.[13] Thus, in Table 1 we display the first four centralized moments for the average normalized public density at $t = 20$, computed over 200 random paths.

The first three moments of the average density at $t = 20$ are similar to the moments of the standard normal density: the mean and the skewness (centralized third moment) are close to zero, and the standard deviation is close to one. The kurtosis, however, is larger than 3, indicating that the stationary public density has indeed fat tails. Nevertheless, the deviation from the standard normal density is not large, especially when ρ is small or large. Moreover, the last row in Table 1 implies that the average bid-ask spread in each case is quite close to $s_* = 2\sigma_v$, which is the stationary value in the approximate Bayesian case: see equation (19). Thus, we argue that the normal approximation made in Section 4 is reasonable, especially when it comes to our main liquidity measure, the bid-ask spread.

The question remains how different the normalized public density can be from the *average* density. This question is already discussed tangentially in Figure 2, where we observe the normalized public density after five buy orders. But to understand this issue in more detail,

[13] We have checked that the average density at $t = 20$ is in absolute value less than 0.01 apart from the average density at $t = 25$ or $t = 30$.

Figure 3: Normalized Public Density after Series of Random Orders.
For an informed share $\rho = 0.1$, consider 200 random series of 20 orders chosen among buy or sell with equal probability, and denote by $\tilde{\phi}_S$ the normalized public density computed after observing the series $S = 1, 2, \ldots, 200$. The table displays the densities $\tilde{\phi}_S$, as well as their average $\psi = \frac{\tilde{\phi}_1 + \tilde{\phi}_2 + \cdots + \tilde{\phi}_{200}}{200}$. The average density is displayed with a thick dashed line.



we choose one particular value of the informed share, $\rho = 0.1$, for which the normalized public density after five buy orders appears more different than the normal density. Figure 3 displays the normalized public density after each of the 200 random series of 20 orders, along with the average density. Then, the results in Table 1 and Figure 3 can be summarized by observing that the normalized public density does not deviate too far from its average value, and in turn this average value does not deviate too far from the standard normal density.

6 Robustness and Extensions

6.1 Discussion

In this section we discuss whether our main result (that liquidity is not affected by the fraction ρ of informed trading) remains true if we modify the model assumptions. For that, we recall the intuition behind that result, as described in Section 4.2. Consider the case when ρ is low,

and suppose the dealer observes a buy order at t. First, we have the *adverse selection effect:* a low ρ means that the buyer is unlikely to be informed, which implies that the update of public mean (and hence the dealer's bid-ask spread) should be small. Second, there is an opposing *dynamic efficiency effect:* in the rare case when the buyer *is* actually informed, he must have observed a large fundamental value v_t, as the uncertainty in v_t (measured by the public volatility σ_*) is also large.

Note that for this offsetting argument to fully work, the public volatility σ_* must be very large when ρ is very small. This is possible only if the range of the fundamental value is not restricted to become very large. Such restrictions can occur in two ways: either (i) the fundamental value is directly assumed to be bounded, or (i) the signals received by the dealer essentially bound the dealer's uncertainty.

Situation (i) occurs if we require the fundamental value to lie in a bounded interval such as $[0, 1]$.[14] We analyze such a model in the Internet Appendix Section 2, where the fundamental value is either zero or one (as in Glosten and Milgrom, 1985), and it switches every period between these two values with probability $\nu < 1/2$. In that case, the dynamic efficiency effect no longer offsets the adverse selection effect. Nevertheless, even if ρ is small, as long as ρ is large relative to the switching parameter ν, the dynamic efficiency effect is relatively strong, and as a result the dependence of the average bid-ask spread on ρ is weaker, and the equilibrium approaches the one in the diffusing-value model where the average bid-ask spread is independent of ρ.

Situation (ii) occurs if the dealer receives at every t signals about the level v_t. Note that this is not by itself enough to bound dealer's uncertainty about v_t. Indeed, in Section 6.2 we consider an extension of our model in which, in addition to observing the order flow, the dealer receives at every t signals about the increment $v_t - v_{t-1}$. In that case, we see that the main result goes through. The reason is that the dealer's uncertainty about the level v_t becomes large when ρ is small: the dealer only learns about v_t once, after which she learns only about future value increments. If, by contrast, the dealer received at every t signals about the level v_t, then the uncertainty would remain bounded even when ρ is very small,

[14] This setup of course cannot occur if the fundamental value follows a random walk.

and the main result would no longer hold.

6.2 Public News

We now analyze an extension of the model in Section 2, in which the dealer receives news every period (this could be interepreted as the dealer receiving public news). Specifically, suppose that before each $t = 1, 2, \ldots$ the dealer receives a signal $\Delta s_t = s_t - s_{t-1}$ about the increment $\Delta v_t = v_t - v_{t-1}$:[15]

$$\Delta s_t \;=\; \Delta v_t + \Delta \eta_t, \quad \text{with} \quad \Delta \eta_t \;=\; \eta_t - \eta_{t-1} \overset{IID}{\sim} \mathcal{N}\big(\cdot, 0, \sigma_\eta\big). \tag{25}$$

Denote, respectively, by μ_t and σ_t the public mean and public volatility just before trading at $t = 0, 1, 2, \ldots$ (but after the signal Δs_t is observed). Note that this extension generalizes the model in Section 2: when σ_η approaches infinity, it is as if the dealer receives no signal at t. The next result generalizes Proposition 4.

Proposition 5. *For any $t = 0, 1, 2, \ldots$ the public mean and volatility satisfy*

$$\mu_{t+1} \;=\; \mu_t \pm \delta \sigma_t + \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \Delta s_{t+1}, \qquad \sigma_t^2 \;=\; \sigma_*^2 + \big(\sigma_0^2 - \sigma_*^2\big)\,(1 - \delta^2)^t, \tag{26}$$

where $\delta = g^{-1}(2\rho)$, as in equation (15), and

$$\sigma_* \;=\; \frac{\sigma_{v\eta}}{\delta} \quad \textit{with} \quad \sigma_{v\eta} \;=\; \frac{\sigma_v \sigma_\eta}{\sqrt{\sigma_v^2 + \sigma_\eta^2}}. \tag{27}$$

For any initial value σ_0 and any sequence of orders, the public volatility σ_t monotonically converges to σ_*, and the bid-ask spread monotonically converges to

$$s_* \;=\; 2\sigma_{v\eta}. \tag{28}$$

[15] Alternatively, but perhaps less realistically, the dealer receives in each period t a signal about the level v_t. In this case, there is an upper bound for the dealer's uncertainty even if the informed share is very small. Hence, the public volatility is no longer increasing indefinitely with the informed share, and therefore the dynamic efficiency effect is reduced. As a result, the adverse selection effect dominates the dynamic efficiency effect, and the stationary bid-ask spread is increasing in the informed share.

Proposition 5 is essentially the same result as Proposition 4, except that the fundamental volatility σ_v is replaced here by $\sigma_{v\eta}$. The parameter $\sigma_{v\eta}$ represents the increase in dealer uncertainty from t to $t+1$, conditional on her receiving the signal Δs_{t+1}.[16] When σ_η is zero, the dealer learns perfectly the increment Δv, hence even if the dealer does not know the initial value v_0, she ends up by learning v_t almost perfectly (she also learns about v_t from the order flow). When σ_η approaches infinity, the dealer receives uninformative signals, $\sigma_{v\eta}$ approaches σ_v, and the equilibrium behavior is described as in Proposition 4.

The stationary bid-ask spread s_* is twice the parameter $\sigma_{v\eta}$. Thus, the bid-ask spread is increasing in the news uncertainty parameter σ_η, and ranges from zero (when $\sigma_\eta = 0$) to $2\sigma_v$ (when $\sigma_\eta = \infty$). The relation between the bid-ask spread and σ_η is intuitive: with more imprecise news, the dealer is more uncertain about the asset value, and sets a larger stationary bid-ask spread.

Note that even in this more general context the stationary bid-ask spread s_* does not depend on the informed share ρ. The intuition is the same as for Proposition 4, and is discussed at the end of the proof of Proposition 5. This intuition is based on the general result (proved in Appendix B) that for any filtration problem in which the variance remains constant over time, the variance of the change in public mean must equal the fundamental variance. But the latter variance is independent of ρ, as is the variance of the signal Δs, hence the bid-ask spread is also independent of ρ.

7 Conclusion

In this paper we have presented a dealer model in which the asset value follows a random walk. The stationary equilibrium of the model has novel properties. Our main finding is that the stationary bid-ask spread no longer depends on the informed share (the fraction of traders that are informed). This result is driven by two offsetting effects: (i) the traditional adverse selection effect: the dealer sets higher bid-ask spreads to protect from a larger number of informed traders, and (ii) the dynamic efficiency effect: the dealer learns faster from the order flow when there are more informed traders, and this reduces the bid-ask spread.

[16] Indeed, its square $\sigma_{v\eta}^2$ is equal to the conditional variance $\mathsf{Var}\big(\Delta v_{t+1} | \Delta s_{t+1}\big)$.

The non-stationary equilibria converge to the stationary equilibrium, regardless of the initial state. The evolution of the non-stationary equilibrium after various types of shocks provides additional testable implications of our model. For instance, after a positive shock to the informed share (e.g., if more informed investors start trading in that stock) the bid-ask spread jumps but then it decreases again to its stationary level. This type of liquidity resilience occurs purely for informational reasons, without any additional market maker jumping in to provide liquidity.

Appendix A. Proofs of Results

Proof of Proposition 1. Using Bayes' rule, the posterior density of v_t after observing $\mathcal{O}$ is

$$\psi_t(v|\mathcal{O}) \;=\; \frac{\mathsf{P}(\mathcal{O}_t = \mathcal{O} \mid v_t = v)\cdot \mathsf{P}(v_t = v)}{\int_v \mathsf{P}(\mathcal{O}_t = \mathcal{O} \mid v_t = v)\cdot \mathsf{P}(v_t = v)} \;=\; \frac{g_t(\mathcal{O}, v)\cdot \phi_t(v)}{\int_v g_t(\mathcal{O}, v)\cdot \phi_t(v)}, \tag{A1}$$

where $\int_v F(v)$ is shorthand for $\int_{-\infty}^{+\infty} F(v)\mathrm{d}v$. Substituting $g_t(\mathcal{O}, v)$ from (2) and (3) in the above equation, we obtain (4).

Let $f(w, v) = \mathsf{P}(v_{t+1} = w|v_t = v) = \mathcal{N}(w - v, 0, \sigma_v)$ be the transition density of v_t. To compute the posterior density of v_t after observing $\mathcal{O}_t = \mathcal{O}$, note that

$$\begin{aligned}\phi_{t+1}(w|\mathcal{O}) \;&=\; \int_v \mathsf{P}(v_{t+1} = w \mid v_t = v, \mathcal{O}_t = \mathcal{O})\cdot \mathsf{P}(v_t = v \mid \mathcal{O}_t = \mathcal{O}) \\ &=\; \int_v \mathsf{P}(v_{t+1} = w \mid v_t = v)\cdot \mathsf{P}(v_t = v \mid \mathcal{O}_t = \mathcal{O}) \;=\; \int_v f(w, v)\cdot \psi_t(v|\mathcal{O}),\end{aligned} \tag{A2}$$

which proves (5).

To simplify notation, we omit conditioning on the order $\mathcal{O}_t$. From (4), it follows that the posterior density ψ_t is equal to ϕ_t multiplied by a piecewise constant function. The prior density ϕ_t is rapidly decaying, hence it is bounded. Therefore ψ_t is also bounded and continuous, although it is no longer smooth. Nevertheless, when we convolute $\psi_t(\cdot)$ with $\mathcal{N}(\cdot, 0, \sigma_v)$ the result ϕ_{t+1} becomes smooth. Indeed, the N'th derivative $\mathrm{d}^N\phi_{t+1}(w)/\mathrm{d}w^N$ involves differentiating the smooth function $\mathcal{N}(w - v, 0, \sigma_v)$ under the integral sign. As the

remaining term $\psi_t(v)$ is bounded, the integrals are well defined, and hence ϕ_{t+1} is a smooth function. The fact that ϕ_{t+1} is also rapidly decaying can be seen in the same way, using again the fact that ψ_t is bounded. □

Proof of Corollary 1. By definition of the ask-bid pair, A_t is the mean of the posterior density of v_t after observing a buy order at t. But the increment $v_{t+1} - v_t$ has zero mean and is independent of the previous variables until t. Therefore, A_t is also the mean of the posterior density of v_{t+1} after observing a buy order at t. Similarly, B_t is the mean of the posterior density of v_{t+1} after observing a sell order at t. This proves the equations in (7). □

Proof of Proposition 2. Define the following function:[17]

$$H_t(v) \;=\; \int_{-\infty}^{v} w\phi_t(w)\mathrm{d}w \;=\; v\Phi_t(v) - \int_{-\infty}^{v} \Phi_t(w)\mathrm{d}w. \tag{A3}$$

Note that $H_t(-\infty) = 0$ and $H_t(+\infty) = \int_{-\infty}^{\infty} w\phi_t(w)\mathrm{d}w = \mu_t$. Also, note that

$$\Theta_t(v) \;=\; \mu_t\Phi_t(v) - H_t(v). \tag{A4}$$

To prove the desired equivalence, start with an ask-bid pair (A_t, B_t). This pair must satisfy the dealer's pricing conditions: A_t is the mean of $\psi_t(\cdot|\mathrm{B})$, and B_t is the mean of $\psi_t(\cdot|\mathrm{S})$. Using the formulas in (4) for $\psi_t(v|\mathcal{O})$, we compute

$$\begin{aligned} A_t \;&=\; \frac{\rho\big(\mu_t - H_t(A_t)\big) + \frac{\rho}{2}\big(H_t(A_t) - H_t(B_t)\big) + \frac{1-\rho}{2}\mu_t}{\frac{\rho}{2}(1-\Phi_t(A_t)) + \frac{\rho}{2}(1-\Phi_t(B_t)) + \frac{1-\rho}{2}}, \\ B_t \;&=\; \frac{\rho H_t(B_t) + \frac{\rho}{2}\big(H_t(A_t) - H_t(B_t)\big) + \frac{1-\rho}{2}\mu_t}{\frac{\rho}{2}\Phi_t(A_t) + \frac{\rho}{2}\Phi_t(B_t) + \frac{1-\rho}{2}}. \end{aligned} \tag{A5}$$

[17] In the formula for H_t we use integration by parts, and also the fact that $\lim_{v\to-\infty} v\Phi_t(v) = 0$. To prove this last fact, suppose $v = -x$ with $x > 0$. Since ϕ_t is rapidly decaying, $\phi_t(-x) < Cx^{-3}$ for some constant C. Then $x\Phi_t(-x) = x\int_{-\infty}^{-x}\phi_t(w)\mathrm{d}w < x\frac{Cx^{-2}}{2}$, which implies $\lim_{x\to\infty} x\Phi_t(-x) = 0$.

Using (A4), we compute the following differences:

$$\begin{aligned} A_t - \mu_t &= \frac{\frac{\rho}{2}\Theta_t(A_t) + \frac{\rho}{2}\Theta_t(B_t)}{\rho(1-\Phi_t(A_t)) + \frac{\rho}{2}(\Phi_t(A_t) - \Phi_t(B_t)) + \frac{1-\rho}{2}}, \\ \mu_t - B_t &= \frac{\frac{\rho}{2}\Theta_t(A_t) + \frac{\rho}{2}\Theta_t(B_t)}{\rho\Phi_t(B_t) + \frac{\rho}{2}(\Phi_t(A_t) - \Phi_t(B_t)) + \frac{1-\rho}{2}}. \end{aligned} \tag{A6}$$

As Θ_t is strictly positive everywhere (see Footnote 9), we have the following inequalities: $A_t > \mu_t > B_t$, or equivalently $A_t \in (\mu_t, +\infty)$ and $B_t \in (-\infty, \mu_t)$. The equations (A6) can be written as

$$F(A_t, B_t) \;=\; 0, \qquad G(A_t, B_t) \;=\; 0, \tag{A7}$$

where the functions F and G are defined in (8). Conversely, suppose we have a solution (A_t, B_t) of (A7), with $A_t > \mu_t > B_t$. Then, this pair satisfies the equations in (A6), which are the dealer's pricing conditions. Thus, (A_t, B_t) is an ask-bid pair.

We now show that a solution of (A7) exists. The partial derivatives of F and G are

$$\begin{aligned} \frac{\partial F}{\partial A} &= -\frac{\Theta_t(A) + \Theta_t(B)}{(A-\mu_t)^2}, \qquad \frac{\partial F}{\partial B} = \frac{A-B}{A-\mu_t}\,\phi_t(B), \\ \frac{\partial G}{\partial A} &= -\frac{A-B}{\mu_t - B}\,\phi_t(A), \qquad \frac{\partial G}{\partial B} = \frac{\Theta_t(A) + \Theta_t(B)}{(\mu_t - B)^2}. \end{aligned} \tag{A8}$$

From (8) we see that $F(A, B)$ has well defined limits at $B = \pm\infty$, which follows from the formulas: $\Theta_t(\pm\infty) = 0$, $\Phi_t(-\infty) = 0$, and $\Phi_t(+\infty) = 1$. Thus we extend the definition of F for all $B \in \bar{\mathbb{R}} = [-\infty, +\infty]$. Now fix $B \in \bar{\mathbb{R}}$. We show that there is a unique solution $A = \alpha(B)$ of the equation $F(A, B) = 0$. From (A8) we see that $\frac{\partial F}{\partial A} < 0$ for all $A \in (\mu_t, \infty)$. From (8) we see that when $A \searrow \mu_t$, $F(A, B) \nearrow \infty$; while when $A \nearrow \infty$, $F(A, B) \searrow -\frac{1+\rho}{\rho} + 1 + \Phi_t(B) = -\frac{1}{\rho} + \Phi_t(B) < 0$ (recall that $\rho \in (0, 1)$). Thus, for any B there is a unique solution of $F(A, B) = 0$ for $A \in (\mu_t, \infty)$. Denote this unique solution by $\alpha(B)$. Differentiating the equation $F\big(\alpha(B), B\big) = 0$ implies that for all B the derivative of $\alpha(B)$ is $\alpha'(B) = -\frac{\partial F}{\partial B}(\alpha(B), B)/\frac{\partial F}{\partial A}(\alpha(B), B) > 0$. Define $\underline{A} = \alpha(-\infty)$ and $\overline{A} = \alpha(\mu_t)$. The results above imply that both $\underline{A}$ and $\overline{A}$ belong to (μ_t, ∞), and α is a bijective function between $[-\infty, \mu_t]$ and $[\underline{A}, \overline{A}]$.

A similar analysis shows that for all $A \in \bar{\mathbb{R}}$, there is a unique solution $B = \beta(A)$ of the

equation $G(A,B)=0$. Moreover, the function β is increasing, and if we define $\underline{B}=\beta(\mu_t)$ and $\overline{B}=\beta(\infty)$, it follows that both $\underline{B}$ and $\overline{B}$ belong to $(-\infty,\mu_t)$, and the function α is bijective between $[\mu_t,\infty]$ and $[\underline{B},\overline{B}]$.

Next, define the function $f:\bar{\mathbb{R}}\to\bar{\mathbb{R}}$ by

$$f(A) \;=\; \alpha\big(\beta(A)\big). \tag{A9}$$

Consider the set

$$S \;=\; \{(A,B)\mid A-f(A) \;=\; 0\;,\; B \;=\; \beta(A)\}. \tag{A10}$$

It is straightforward to show that S coincides with the set of all ask-bid pairs. Indeed, $(A,B)\in S$ is equivalent to $A=\alpha(B)$ and $B=\beta(A)$, which, from the discussion above, is equivalent to $F(A,B)=0$ and $G(A,B)=0$. Therefore, the existence of an ask-bid pair is equivalent to there being at least one solution of $A-f(A)=0$.

We now show that the equation $A-f(A)=0$ has at least one solution. The function $f(A)$ is increasing and bijective between $[\mu_t,\infty]$ and $[\alpha(\underline{B}),\alpha(\overline{B})]$. As $\underline{B},\overline{B}\in(-\infty,\mu_t)$, it follows that $[\alpha(\underline{B}),\alpha(\overline{B})]\subset(\underline{A},\overline{A})\subset(\mu_t,\infty)$. When $A\searrow\mu_t$, $A-f(A)\to\mu_t-\alpha(\underline{B})<0$, while when $A\nearrow\infty$, $A-f(A)\to\infty-\alpha(\overline{B})>0$. Thus, there exists a solution of $A-f(A)=0$ on (μ_t,∞).

Finally, define

$$A_t \;=\; \inf\Big\{A\in(\mu_t,\infty)\;\Big|\;A-f(A)=0\Big\},\qquad B_t \;=\; \beta(A_t). \tag{A11}$$

As f is continuous, A_t also satisfies $A_t-f(A_t)=0$, hence among all possible ask-bid pairs the ask closest to μ_t is attained at A_t. □

Proof of Proposition 3. Denote by $\phi(\cdot)=\mathcal{N}(\cdot,0,1)$ the standard normal density, and by $\Phi(\cdot)$ its cumulative density. Recall that $\phi_t(\cdot)$ is the density of v_t just before trading at t, and $\psi_t(\cdot|\mathcal{O}_t)$ is the density of v_t after trading at t. In this proposition, we assume that we start with a normal density

$$\phi_t(v) \;=\; \frac{1}{\sigma_t}\,\phi\left(\frac{v-\mu_t}{\sigma_t}\right), \tag{A12}$$

with mean μ_t and volatility σ_t. Define the normalized ask and bid, respectively, by

$$a_t \;=\; \frac{A_t-\mu_t}{\sigma_t}, \qquad b_t \;=\; \frac{B_t-\mu_t}{\sigma_t}. \tag{A13}$$

We now compute the mean and volatility of $\phi_{t+1}(\cdot|\mathcal{O}_t)$. As the increment $v_{t+1}-v_t \sim \mathcal{N}(0,\sigma_v^2)$ is independent of past variables, the mean and volatility of $\phi_{t+1}(\cdot|\mathcal{O}_t)$ satisfy

$$\mu_{t+1,\mathcal{O}_t} \;=\; \int_v v\,\psi_t(v|\mathcal{O}_t), \qquad \sigma^2_{t+1,\mathcal{O}_t} \;=\; \sigma_v^2 \;+\; \int_v \big(v-\mu_{t+1,\mathcal{O}_t}\big)^2 \psi_t(v|\mathcal{O}_t). \tag{A14}$$

From (4), we have $\psi_t(v|\mathrm{B}) = \frac{\rho \mathbf{1}_{v>A_t}+\frac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]}+\frac{1-\rho}{2}}{\frac{\rho}{2}(1-\Phi_t(A_t))+\frac{\rho}{2}(1-\Phi_t(B_t))+\frac{1-\rho}{2}}\,\phi_t(v)$. With the change of variables $z=\frac{v-\mu_t}{\sigma_t}$, we compute posterior mean conditional on a buy order:

$$\begin{aligned}
\mu_{t+1,\mathrm{B}} \;&=\; \mu_t + \int_{-\infty}^{+\infty} (v-\mu_t)\,\frac{\rho\mathbf{1}_{v>A_t}+\frac{\rho}{2}\mathbf{1}_{v\in[B_t,A_t]}+\frac{1-\rho}{2}}{\frac{\rho}{2}(1-\Phi_t(A_t))+\frac{\rho}{2}(1-\Phi_t(B_t))+\frac{1-\rho}{2}}\,\frac{1}{\sigma_t}\phi\big(\tfrac{v-\mu_t}{\sigma_t}\big)\,\mathrm{d}v \\
&=\; \mu_t + \sigma_t \int_{-\infty}^{+\infty} z\,\frac{\rho\mathbf{1}_{z>a_t}+\frac{\rho}{2}\mathbf{1}_{z\in[b_t,a_t]}+\frac{1-\rho}{2}}{\frac{\rho}{2}(1-\Phi(a_t))+\frac{\rho}{2}(1-\Phi(b_t))+\frac{1-\rho}{2}}\,\phi(z)\,\mathrm{d}z \\
&=\; \mu_t + \sigma_t\,\frac{\phi(-a_t)+\phi(-b_t)}{\Phi(-a_t)+\Phi(-b_t)+\frac{1-\rho}{\rho}}.
\end{aligned} \tag{A15}$$

Similarly, the posterior mean conditional on a sell order is

$$\mu_{t+1,\mathrm{S}} \;=\; \mu_t - \sigma_t\,\frac{\phi(a_t)+\phi(b_t)}{\Phi(a_t)+\Phi(b_t)+\frac{1-\rho}{\rho}}, \tag{A16}$$

To compute $\sigma^2_{t+1,\mathcal{O}_t}$, we notice that

$$\int_v \big(v-\mu_t\big)^2\psi_t(v|\mathcal{O}_t) \;=\; \int_v \big(v-\mu_{t+1,\mathcal{O}_t}\big)^2\psi_t(v|\mathcal{O}_t) \;+\; \big(\mu_{t+1,\mathcal{O}_t}-\mu_t\big)^2, \tag{A17}$$

where we use the fact that $\int_v \big(v-\mu_{t+1,\mathcal{O}_t}\big)\psi_t(v|\mathcal{O}_t)=0$. Using (A14) and (A17), a similar

calculation as in (A15) implies that the posterior variance conditional on a buy order satisfies

$$\begin{aligned}\sigma^2_{t+1,\mathrm{B}} \;-\; \sigma^2_v \;+\; \big(\mu_{t+1,\mathrm{B}} - \mu_t\big)^2 \;&=\; \int_v \big(v-\mu_t\big)^2 \psi_t(v|\mathrm{B}) \\ &=\; \sigma^2_t \int_{-\infty}^{+\infty} z^2 \, \frac{\rho \mathbf{1}_{z>a_t} + \frac{\rho}{2}\mathbf{1}_{z\in[b_t,a_t]} + \frac{1-\rho}{2}}{\frac{\rho}{2}(1-\Phi(a_t)) + \frac{\rho}{2}(1-\Phi(b_t)) + \frac{1-\rho}{2}} \, \phi(z)\,\mathrm{d}z \\ &=\; \sigma^2_t \left(1 \;+\; \frac{a_t\phi(a_t) + b_t\phi(b_t)}{\Phi(-a_t) + \Phi(-b_t) + \frac{1-\rho}{\rho}}\right).\end{aligned} \tag{A18}$$

Similarly, the posterior variance conditional on a sell order satisfies

$$\sigma^2_{t+1,\mathrm{S}} \;-\; \sigma^2_v \;+\; \big(\mu_{t+1,\mathrm{S}} - \mu_t\big)^2 \;=\; \sigma^2_t \left(1 \;-\; \frac{a_t\phi(a_t) + b_t\phi(b_t)}{\Phi(a_t) + \Phi(b_t) + \frac{1-\rho}{\rho}}\right). \tag{A19}$$

We now use the fact that a_t and b_t are the normalized ask and bid. Equation (7) implies that the ask is $A_t = \mu_{t+1,\mathrm{B}}$ and the bid is $B_t = \mu_{t+1,\mathrm{S}}$. If we normalize these equations, we have $a_t = \frac{\mu_{t+1,\mathrm{B}} - \mu_t}{\sigma_t}$ and $b_t = \frac{\mu_{t+1,\mathrm{S}} - \mu_t}{\sigma_t}$. Using (A15) and (A16), we obtain

$$a_t \;=\; \frac{\phi(-a_t) + \phi(-b_t)}{\Phi(-a_t) + \Phi(-b_t) + \frac{1-\rho}{\rho}}, \qquad b_t \;=\; -\frac{\phi(a_t) + \phi(b_t)}{\Phi(a_t) + \Phi(b_t) + \frac{1-\rho}{\rho}}, \tag{A20}$$

We show that this system has a unique solution. We use the notation from the proof of Proposition 2, adapted to this particular case. For $(a,b) \in (0,\infty)\times(-\infty,0)$, define $F(a,b) = \frac{\phi(a)+\phi(b)}{a} - \Phi(-a) - \Phi(-b) - \frac{1-\rho}{\rho}$ and $G(a,b) = \frac{\phi(a)+\phi(b)}{-b} - \Phi(a) - \Phi(b) - \frac{1-\rho}{\rho}$. As in the proof of Proposition 2, for $b \in [-\infty, 0]$ define $\alpha(b)$ as the unique solution of $F(\alpha(b), b) = 0$; and for $a \in [0,\infty]$ define $\beta(a)$ as the unique solution of $G(a, \beta(a)) = 0$. For $a \in (0,\infty)$, define $f(a) = \alpha(\beta(a))$. We show that any solution a of the equation $a - f(a) = 0$ must satisfy $a < 1$. Let $b = \beta(a)$. Since $a = \alpha(b)$, by definition $F(a,b) = 0$. As in the proof of Proposition 2, one shows that F is decreasing in a, and that $F(0,b) = +\infty > 0$. As $b < 0$, $F(1,b) = \phi(1) - \Phi(-1) + \phi(b) - \Phi(-b) - \frac{1-\rho}{\rho} < \phi(1) - \Phi(-1) + \phi(0) - \Phi(0) \approx -0.0177 < 0$. As $F(a,b) = 0$ and F is decreasing in a (and a is positive), we have just proved that $a \in (0,1)$. A similar argument (adapted to the function G) shows that $b = \beta(a) \in (-1,0)$.

As in equation (A9), define $f(a) = \alpha\big(\beta(a)\big)$. We need to show that the equation $a - f(a) = 0$ has a unique solution in $(0,\infty)$. By contradiction, suppose there are at least two solutions

$a_1 < a_2$, and suppose a_1 is the smallest such solution and a_2 the largest. As f is continuous and takes values in some compact interval $[\underline{b}, \overline{b}]$ (see the proof of Proposition 2), a_1 and a_2 are well defined. Also, since f is increasing, f is a bijection of $[a_1, a_2]$. The argument above then shows that both a_1 and a_2 are in $(0, 1)$. If we prove that $f' < 1$ on $[a_1, a_2]$, it follows that $a - f(a)$ is increasing on $[a_1, a_2]$ and cannot therefore be equal to zero at both ends. This contradiction therefore proves uniqueness, as long as we show that indeed $f' < 1$ on $(0, 1)$. Let $a \in (0, 1)$ and denote $b = \beta(a)$ and $a' = \alpha(b)$. Then by the chain rule $f'(a) = \alpha'(b)\beta'(a)$. Differentiating the equations $F(\alpha(b), b) = 0$ and $G(a, \beta(a) = 0$, we have $\alpha'(b) = \frac{\phi(b)(a'-b)a'}{\phi(a')+\phi(b)}$ and $\beta'(b) = \frac{\phi(a)(a-b)(-b)}{\phi(a)+\phi(b)}$. Both these derivatives are of the form $\frac{\phi(x_1)(x_1+x_2)x_1}{\phi(x_1)+\phi(x_2)}$ with $x_1, x_2 \in (0, 1)$. This function is increasing in x_2, hence it is smaller than $\frac{\phi(x_1)(x_1+1)x_1}{\phi(x_1)+\phi(1)}$, which is increasing in x_1, hence smaller than one, which is the value corresponding to $x_1 = 1$. Thus, $f' < 1$ on $(0, 1)$ and the uniqueness is proved.

To find the unique solution, note that by symmetry we expect $a_t = -b_t$. If we impose this condition, we have $\Phi(a_t) + \Phi(b_t) = \Phi(-a_t) + \Phi(-b_t) = 1$. Therefore, we need to solve the equation $a_t = 2\rho\phi(a_t)$ for $a_t > 0$, or equivalently $g(a_t) = 2\rho$, where $g(x) = \frac{x}{\phi(x)}$. As the derivative of ϕ is $\phi'(x) = -x\phi(x)$, the derivative of g is $g'(x) = \frac{1+x^2}{\phi(x)} > 0$ for all x. Moreover, $g(0) = 0$ and $\lim_{x\to\infty} g(x) = \infty$, hence g is increasing and a one-to-one and mapping of $(0, \infty)$. Thus, if we define $\delta = g^{-1}(2\rho)$, which is the same formula as in (15), we have $g(\delta) = 2\rho$. The solution of (A20) is then

$$a_t \;=\; -b_t \;=\; \delta, \qquad \text{or} \qquad A_t \;=\; \mu_t + \delta\sigma_t, \quad B_t \;=\; \mu_t - \delta\sigma_t. \tag{A21}$$

Thus, the posterior mean satisfies

$$\mu_{t+1,\mathrm{B}} \;=\; \mu_t + \delta\sigma_t, \qquad \mu_{t+1,\mathrm{S}} \;=\; \mu_t - \delta\sigma_t, \tag{A22}$$

which proves the first part of equation (14).

Equation (A22) also implies that $(\mu_{t+1,\mathcal{O}_t} - \mu_t)^2 = \delta^2\sigma_t^2$ for $\mathcal{O}_t \in \{\mathrm{B}, \mathrm{S}\}$. As $a_t = -b_t$, equations (A18) and (A19) imply that $\sigma^2_{t+1,\mathcal{O}_t} - \sigma_v^2 + \delta^2\sigma_t^2 = \sigma_t^2$. Thus, the posterior variance

satisfies

$$\sigma^2_{t+1,\mathrm{B}} \;=\; \sigma^2_{t+1,\mathrm{S}} \;=\; (1-\delta^2)\sigma_t^2 + \sigma_v^2, \tag{A23}$$

which proves the second part of equation (14). □

Proof of Proposition 4. Recall that the function $g : [0,\infty) \to [0,\infty)$ is increasing and $\delta = g^{-1}(2\rho)$, with $\rho \in (0,1)$. Hence, $\delta < g^{-1}(2) \approx 0.647$, and in particular $\delta < 1$. Equation (A23) implies that the public variance evolves according to $\sigma_t^2 = (1-\delta^2)\sigma_{t-1}^2 + \sigma_v^2$ for any $t \geq 0$ (by convention, $\sigma_{-1} = 0$). Iterating this equation, we obtain $\sigma_t^2 = (1-\delta^2)^t\sigma_0^2 + \frac{1-(1-\delta^2)^t}{\delta^2}\sigma_v^2$. Using $\sigma_* = \frac{\sigma_v}{\delta}$, we obtain $\sigma_t^2 = \sigma_*^2 + (1-\delta^2)^t(\sigma_0^2 - \sigma_*^2)$, which proves (17). As $\delta \in (0,1)$, it is clear that σ_t^2 converges monotonically to σ_*^2 for any initial value σ_0. The bid-ask spread satisfies $s_t = 2\sigma_t\delta$, hence it converges to $2\sigma_*\delta = 2\frac{\sigma_v}{\delta}\delta = 2\sigma_v = s_*$. □

Proof of Corollary 2. Following the proof of Proposition 4, recall that g is increasing on $(0,\infty)$. Its inverse g^{-1} is therefore also increasing, and $\sigma_* = \sigma_v/g^{-1}(2\rho)$ is decreasing in ρ. The dependence on σ_v is straightforward. □

Proof of Corollary 3. Conditional on the information at t, each order (buy or sell) is equally likely. Therefore, the change in the public mean $\mu_{t+1,\mathcal{O}_t} - \mu_t$ has a binary distribution with probability $1/2$, which has standard deviation equal to σ_v, which is the fundamental volatility. □

Proof of Corollary 4. Equation (A23) shows that the public variance σ_t^2 evolves according to $\sigma_{t+1}^2 = (1-\delta^2)\sigma_t^2 + \sigma_v^2$. Taking the limit on both sides, we get $\sigma_*^2 = (1-\delta^2)\sigma_*^2 + \sigma_v^2$. Subtracting the two equations above, we get $\sigma_{t+1}^2 - \sigma_*^2 = (1-\delta^2)(\sigma_t^2 - \sigma_*^2)$, which proves the speed of convergence formula (22) for the public variance. As $\sigma_t^2 - \sigma_*^2 = (\sigma_t - \sigma_*)(\sigma_t + \sigma_*)$, the formula (22) is true for the public volatility as well. Finally, the bid-ask spread is $s_t = 2\delta\sigma_t$, which proves (22) for the bid-ask spread. □

Proof of Corollary 5. This follows directly from equations (4) and (5) from Proposition 1, making the change of variables from equation (23). □

Proof of Proposition 5. The only difference from the setup of Section 2 is that after trading at t (but before trading at $t+1$) the dealer receives a signal $\Delta s_{t+1} = \Delta v_{t+1} + \Delta\eta_{t+1}$. By

notation, just before trading at t, v_t is distributed as $\mathcal{N}(\cdot, \mu_t, \sigma_t)$. We thus follow the proof of Propositions 3 and 4, and infer that after trading at t the dealer regards v_t to be distributed as $\mathcal{N}(\cdot, \mu'_t, \sigma'_t)$, where $\mu'_t = \mu_t \pm \delta\sigma_t$ and $\sigma'^2_t = (1-\delta^2)\sigma^2_t$.[18] After observing $\Delta s_{t+1} = \Delta v_{t+1} + \Delta\eta_{t+1}$, the dealer computes $\mathsf{E}\big(\Delta v_{t+1}|\Delta s_{t+1}\big) = \frac{\sigma_v^2}{\sigma_v^2+\sigma_\eta^2}\Delta s_{t+1}$ and $\mathsf{Var}\big(\Delta v_{t+1}|\Delta s_{t+1}\big) = \frac{\sigma_v^2\sigma_\eta^2}{\sigma_v^2+\sigma_\eta^2} = \sigma^2_{v\eta}$. Hence, after observing the signal, the dealer regards v_{t+1} to be distributed as $\mathcal{N}(\cdot, \mu_{t+1}, \sigma_{t+1})$, with

$$\mu_{t+1} \;=\; \mu'_t + \frac{\sigma_v^2}{\sigma_v^2+\sigma_\eta^2}\Delta s_{t+1}, \qquad \sigma^2_{t+1} \;=\; \sigma'^2_t + \sigma^2_{v\eta} \;=\; (1-\delta^2)\sigma_t + \sigma^2_{v\eta}. \tag{A24}$$

The recursive equation for σ_t is the same as (A23), except that instead of σ_v we now have $\sigma_{v\eta}$. Then, the same proof as in Propositions 3 and 4 can be used to derive all the desired results.

Note that equation (26) implies that the change in public mean is $\Delta\mu_{t+1} = \pm\delta\sigma_t + \frac{\sigma_v^2}{\sigma_v^2+\sigma_\eta^2}\Delta s_{t+1}$. Thus, in the stationary equilibrium, $\mathsf{Var}(\Delta\mu_{t+1}) = \delta^2\sigma_*^2 + \frac{\sigma_v^4}{(\sigma_v^2+\sigma_\eta^2)^2}(\sigma_v^2+\sigma_\eta^2) = \sigma^2_{v\eta} + \frac{\sigma_v^4}{\sigma_v^2+\sigma_\eta^2} = \sigma_v^2 = \mathsf{Var}(\Delta v_{t+1})$. This verifies the result in Appendix B that in any stationary filtration problem the variance of the change in public mean must equal the fundamental variance. Moreover, the half spread is equal to $\delta\sigma_* = \sigma_{v\eta}$, which does not depend on the informed share ρ. □

Appendix B. Stationary Filtering

We show that in a filtration problem that is stationary (in a sense to be defined below) the variance of value changes is the same as the variance of the public mean changes. Let v_t be a discrete time random walk process with constant volatility σ_v. Suppose each period the market gets (public) information about v_t. Let $\mathcal{I}_t$ be the public information set available at time t. Denote by $\mu_t = \mathsf{E}(v_t|\mathcal{I}_t) = \mathsf{E}_t(v_t)$ the public mean at time t, i.e., the expected asset value given all public information. This filtration problem is called *stationary* if the public

[18] The sign $\pm$ is plus if a buy order is submitted at t, and minus if a sell order is submitted at t.

variance is constant over time:

$$\mathsf{Var}_t(v_t) = \mathsf{Var}_{t+1}(v_{t+1}). \tag{B1}$$

The next result gives a necessary and sufficient for the filtration problem to be stationary.

Proposition 6. *The filtration problem is stationary if and only if*

$$\mathsf{Var}(v_{t+1} - v_t) \;=\; \mathsf{Var}(\mu_{t+1} - \mu_t).$$

Proof. Since $\mu_t = \mathsf{E}_t(v_t)$, we have the decomposition $v_t = \mu_t + \eta_t$, where η_t is orthogonal on the information set $\mathcal{I}_t$. Moreover, $\mathsf{Var}(\eta_t) = \mathsf{Var}_t(v_t)$. Similarly, $v_{t+1} = \mu_{t+1} + \eta_{t+1}$, and $\mathsf{Var}(\eta_{t+1}) = \mathsf{Var}_{t+1}(v_{t+1})$. Thus, the stationary condition reads $\mathsf{Var}(v_{t+1} - \mu_{t+1}) = \mathsf{Var}(v_t - \mu_t)$. We can decompose $v_{t+1} - \mu_t$ in two ways:

$$\begin{aligned} v_{t+1} - \mu_t \;&=\; (v_{t+1} - \mu_{t+1}) + (\mu_{t+1} - \mu_t) \\ &=\; (v_{t+1} - v_t) + (v_t - \mu_t). \end{aligned} \tag{B2}$$

We verify that these are orthogonal decompositions. The first condition is that $\mathsf{cov}(v_{t+1} - \mu_{t+1}, \mu_{t+1} - \mu_t) = 0$, i.e., that $\mathsf{cov}(\eta_{t+1}, \mu_{t+1} - \mu_t) = 0$. But η_{t+1} is orthogonal on $\mathcal{I}_{t+1}$, which contains μ_{t+1} and μ_t. The second condition is that $\mathsf{cov}(v_{t+1} - v_t, v_t - \mu_t) = 0$. But v_t has independent increments, so $v_{t+1} - v_t$ is independent of v_t and anything contained in the information set at time t. (This is true as long as the market does not get at t information about the asset value at a future time.)

The total variance of the two orthogonal decompositions in (B2) must be the same, hence $\mathsf{Var}(v_{t+1} - \mu_{t+1}) + \mathsf{Var}(\mu_{t+1} - \mu_t) = \mathsf{Var}(v_{t+1} - v_t) + \mathsf{Var}(v_t - \mu_t)$. But being stationary is equivalent to $\mathsf{Var}(v_{t+1} - \mu_{t+1}) = \mathsf{Var}(v_t - \mu_t)$, which is then equivalent to $\mathsf{Var}(v_{t+1} - v_t) = \mathsf{Var}(\mu_{t+1} - \mu_t)$.

□

REFERENCES

Bagehot, W. (1971). "The only game in town." *Financial Analysts Journal*, 27(2), 12–14,22.

Caldentey, R. and E. Stacchetti (2010). "Insider trading with a random deadline." *Econometrica*, 78(1), 245–283.

Chau, M. and D. Vayanos (2008). "Strong-form efficiency with monopolistic insiders." *Review of Financial Studies*, 21(5), 2275–2306.

Collin-Dufresne, P. and V. Fos (2015). "Do prices reveal the presence of informed trading?" *Journal of Finance*, 70(4), 1555–1582.

Collin-Dufresne, P. and V. Fos (2016). "Insider trading, stochastic liquidity and equilibrium prices." *Econometrica*, 84(4), 1441–1475.

Foucault, T., M. Pagano, and A. Röell (2013). *Market Liquidity: Theory, Evidence, and Policy.* Oxford University Press, New York.

Glosten, L. R. and P. R. Milgrom (1985). "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." *Journal of Financial Economics*, 14, 71–100.

Glosten, L. R. and T. J. Putnins (2016). "Welfare costs of informed trade."

Kyle, A. S. (1985). "Continuous auctions and insider trading." *Econometrica*, 53(6), 1315–1335.

Trading Volume, Illiquidity and Commonalities in FX Markets*

Angelo Ranaldo† Paolo Santucci de Magistris ‡

May 7, 2019

Abstract

We provide a unified model for foreign exchange (FX), trading volume, and volatility in a multi-currency environment. Tied by no-arbitrage conditions, FX rate movements are determined by common information and differences in traders' reservation prices, or disagreement, that induce trading. Using unique (intraday) data representative for the global FX spot market, the empirical analysis validates our theoretical predictions. We find that (i) volume and volatility are driven by disagreement, (ii) our volatility-volume ratio as in Amihud (2002) is an effective measure of FX illiquidity, and (iii) the commonalities of FX global volume, volatility, and illiquidity that vary across currencies and time can be explained by no-arbitrage.

Keywords: FX Trading Volume, Volatility, Illiquidity, Commonalities, Co-Jumps, Arbitrage

J.E.L. classification: C15, F31, G12, G15

*We are grateful to Tim Bollerslev, Massimiliano Caporin, Federico Carlini, Nina Karnaukh, Lukas Menkhoff, Federico Nucera, Paolo Pasquariello, Mark Podolskij, Roberto Renó, Fabricius Somogyi and Vladyslav Sushko for their relevant remarks on our work. We would also like to thank the participants at 7th Workshop on Financial Determinants of FX Rates Norges Bank, at the 2018 SoFiE conference, at the 2018 DEDA conference and at seminars at City Hong Kong University, at LUISS and at University of the Balearic Islands for useful comments.

†University of St. Gallen, Switzerland. angelo.ranaldo@unisg.ch

‡LUISS "Guido Carli" University, Department of Economics and Finance, Viale Romania 32, 00197 Roma, Italy; CREATES, Department of Economics and Business Economics, Aarhus University, Denmark. sdemagistris@luiss.it

1 Introduction

Since the demise of the post-war Bretton Woods system in the 1970s, the international financial system has witnessed a growing capital mobility and wider movements of foreign exchange (FX) rates. In such a regime of floating FX rates and open economies, anyone dealing with a currency other than that of the base currency is concerned with the (adverse) evolution of FX rates, their volatility, and market dynamics such as trading volume and illiquidity. It is thus a natural question how FX rates, volatility, and trading volume interrelate.

In this paper, we provide a simple theoretical framework to jointly explain FX rates, trading volume, and volatility in a multi-currency environment. Tied together by triangular no-arbitrage conditions, FX rate movements are determined by common information and differences in traders' reservation prices, or disagreement, that induce trading. In such a unified setting, our model outlines two main drivers within and across currencies: First, investors' disagreement is the common determinant of trading volume and volatility of each FX rate. Second, the no-arbitrage condition is the "glue" across currencies creating commonality in trading volume, volatility, and illiquidity. Our model also provides an intuitive closed-form solution for measuring illiquidity in terms of price impact (Amihud, 2002). Using new and unique intraday data representative for the global FX spot market, the empirical analysis validates our main theoretical predictions, that is, (i) more disagreement increases FX trading volume, volatility, and illiquidity, (ii) stronger commonalities pertain to more efficient (arbitrage-free) currencies, and (iii) our illiquidity proxy is effective in measuring FX illiquidity.

The joint analysis of FX volume and volatility is important for at least three reasons. First, the FX market is the world largest financial market with USD 5.1 trillion of daily traded volume (Bank of International Settlements, 2016). Despite its importance and apparent enormous liquidity, an in-depth understanding of FX volume is still missing. This can be explained by at least two reasons. On the one hand, FX rates are commonly traded over-the-counter, which is notoriously opaque and fragmented.[1] On the other hand, there has been a paucity of comprehensive volume data at a global scale. Second, FX rates are key for pricing many assets including international

[1]The microstructure of the FX market is explained in detail in e.g. Lyons (2001) and King et al. (2012). The recent developments of the FX markets are discussed in Rime and Schrimpf (2013) and Moore et al. (2016).

stocks, bonds, and derivatives, and for assessing their risk. They are also relevant for policy making such as conducting (unconventional) monetary policy and FX interventions. A better understanding of whether and how FX volume, volatility and illiquidity determine FX rates can improve all these tasks. Third, distressed markets such as currency crises are characterized by sudden FX rates movements, drops in liquidity, and raises in volatility. It could thus be supportive of financial stability to highlight the sources of volatility and illiquidity, how they reinforce each other, and across currencies.

Our analysis proceeds in two steps: theory and empirics. Our theory builds upon an equilibrium model in which the evolution of the FX rate is driven by the arrival of new information and by the trading activity. The trading volume is induced by the deviation of individual agent's reservation prices from the observed market price. The continuous-time feature of the model allows us to obtain consistent measurements of the underlying unobservable quantities, such as volatility and illiquidity, and to relate them to the trading volume. Furthermore, agents trade in a multi-currency environment in which direct FX rates are tied to cross rates by triangular no-arbitrage conditions. This implies that direct and arbitrage-related (or synthetic) rates must equate in equilibrium, while the trading volume reflects the dependence on the aggregated information flows across FX rates. Thus, trading volume is the driving force processing information and reservation prices in currency values and attracting FX rates to arbitrage-free prices.

Three basic propositions arise from our theoretical framework: First, trading volume and volatility are driven by traders' disagreement. Second, the combination of volatility and volume provides a closed-form intuitive expression for measuring illiquidity in terms of price impact such as the widespread proxy proposed in Amihud (2002). Third, trading volume, volatility and liquidity across FX rates are linked by no-arbitrage conditions, which lead to the commonalities across FX rates. Since arbitrage passes through the trading activity (volume), more liquid currencies should reveal stronger commonalities and price efficiency (in terms of smaller deviations from triangular arbitrage condition).

Set against this background, we test the main empirical predictions derived from our theory. To do this, we utilize two data sets. First, trading volume data come from CLS Bank International (CLS), which operates the largest payment-versus-payment (PVP) settlement service in the world. Hasbrouck and Levich (2017) provide a very comprehensive description of the CLS

institutional setting and Gargano et al. (2019) show that CLS data cover around 50% of the FX global turnover compared to the BIS triennial surveys. Trading volume is measured at the hourly level, across 29 currency pairs over a 5-year period from November 2011 to November 2016.[2] For the same FX panel, we obtain intraday spot rates from Olsen data. For each FX rate and each minute of our sample, we observe the following quotes: ask, bid, low, high, close, and midquote. By merging these two data sets, we can analyze the hourly time series of trading volume, realized volatility, and FX rate and bid-ask spread evolutions.

To test the empirical predictions, we carry out the following analysis. First, we perform a descriptive analysis that uncovers some (new) stylized facts. For instance, we find that FX trading volume and illiquidity follow intraday patterns and seasonalities indicating market fragmentation across geographical areas and FX rates consistent with the OTC nature of the FX global market. Then, we perform various regressions to test the three above-mentioned theoretical propositions. Three main results emerge: First, trading volume and volatility are linked by a very strong positive relationship both within and across FX rates. To provide more direct evidence that both are governed by disagreement between the agents, we show that volume and volatility increase with heterogeneous beliefs as measured in Beber et al. (2010). In contrast, large and directional FX moves associated with little disagreement identified by co-jumps (Caporin et al., 2017) do not generate abnormal trading volume, while being associated with above-average volatility. Consistent with our theory, this finding suggests that new common information such as macroeconomic announcements (see e.g. Bollerslev et al., 2016) on which everyone agrees might give rise to above-average volatility but not abnormal trading volume. Second, we provide evidence that our illiquidity measure is effective in capturing FX illiquidity episodes and correlate with well-accepted measures of FX illiquidity. Finally, using three methods, namely the Principal Component Analysis, regression analysis, and the connectedness index of Diebold and Yilmaz (2014), we perform a comprehensive analysis of commonalities in FX volume, volatility, and illiquidity. After documenting and measuring them, we provide evidence that more liquid currencies have stronger commonalities and obey more to (triangular) arbitrage conditions.

Our paper contributes to two strands of the literature: First, we contribute to prior research

[2]The entire set includes 33 currency pairs but the Hungarian forint (HUF) joined the CLS system later. Therefore, EURHUF and USDHUF are available only since 07 November 2015.

on trading and liquidity in financial markets. While most of the previous studies on volume has mainly focused on stocks,[3] there is a growing literature on trading and liquidity in FX markets (e.g. Mancini et al., 2013 and Karnaukh et al., 2015). Most previous studies focus on specific aspects of FX liquidity such as transaction costs[4] or order flow, which is as the net of buyer-initiated and seller-initiated orders. Following the seminal paper by Evans and Lyons (2002), order flow has drawn much attention as the main determinant of FX rate formation.[5] In contrast, the literature on trading volume is scant due to the paucity of comprehensive data on the FX global volume. Prior research has focused on the interdealer segment in which Electronic Broking Services (EBS) and Reuters are the two predominant platforms. For instance, Evans (2002) uses Reuters D2000-1 data, Payne (2003) analyze data from D2000-2 while Mancini et al. (2013) and Chaboud et al. (2007) utilize data from EBS.[6] Only with the recent access to CLS data, research on FX global volume at relatively high frequencies (e.g. daily) became possible.[7] Fischer and Ranaldo (2011) look at global FX trading around central bank decisions. Hasbrouck and Levich (2017) measure FX illiquidity using volume and volatility data. Gargano et al. (2019) analyze the profitability of FX trading strategies exploiting the predictive ability of FX volume. We add to the extant literature theoretically and empirically. On the one hand, we build a continuous-time model in a multiple-currency setting, which serves the purpose of defining a theoretical foundation for FX price determination in connection to FX volume, volatility, and illiquidity. Although abstracting from some market "imperfections" such as liquidity frictions, our model provides a closed-form and intuitive solution for illiquidity in terms of price impact proxies such as in Amihud (2002). Furthermore, we are the first providing a joint empirical analysis of intraday FX global volume, (realized) volatilities, and illiquidity that support two empirical predictions from our theory: First, disagreement drives trading volume and volatility; Second, our FX measure in

[3] For a recent literature survey, see Vayanos and Wang (2013).

[4] Transaction costs are typically measured in terms of bid-ask spreads that tend to increase with volatility. FX transaction costs in spot and future markets are studied in Bessembinder (1994), Bollerslev and Melvin (1994), Christiansen et al. (2011), Ding (1999), Hartmann (1999), Huang and Masulis (1999), Hsieh and Kleidon (1996), Mancini et al. (2013).

[5] Among others, order flow is studied in Bjønnes and Rime (2005), Berger et al. (2008), Frömmel et al. (2008), Breedon and Ranaldo (2013), Evans and Lyons (2002), Evans (2002), Mancini et al. (2013), Payne (2003), and Rime et al. (2010).

[6] Other sources of trading volume data are proprietary data sets from some specific banks (see e.g. Bjønnes and Rime (2005) and Menkhoff et al. (2016)), central banks, or FX futures or forward contracts (see e.g. Bjønnes et al. (2003), Galati et al. (2007), Grammatikos and Saunders (1986), Levich (2012), and Bech (2012)).

[7] Except from CLS, the only source of global FX trading volume is the triennial survey of central banks conducted by the BIS. It provides a snapshot of FX market volume on a given day once every three-years.

the spirit of Amihud proxy is effective in measuring FX illiquidity.

Second, we contribute to the literature on commonalities in liquidity, which has extensively studied liquidity co-movements of stocks (e.g. Chordia et al., 2000, Hasbrouck and Seppi, 2001, and Karolyi et al., 2012). In FX markets, this issue is empirically analyzed in Mancini et al. (2013) and Karnaukh et al. (2015). We contribute to this strand of literature by studying commonality in trading *volume* and the proposed FX illiquidity measure, as well as some pricing implications stemming from commonality in FX liquidity. Prior research has also provided some theoretical explanations for liquidity commonality. For instance, when dealers are active in two markets (or assets), they tend to reduce their liquidity supply in case of trading losses (Kyle and Xiong, 2001) or under funding constraints (Cespa and Foucault, 2014). From an asset pricing perspective, investors require higher expected returns and invest less in assets exposed to liquidity risk (e.g. Acharya and Pedersen, 2005); additionally, illiquidity and low asset prices might endogenously result from erosion of arbitrageurs' wealth (Kondor and Vayanos, 2018). Even if our theory abstracts from these frictions, commonality in trading volume naturally arises from agents' disagreement and arbitrage trading. Empirically, we find consistent results with the adage that "liquidity begets liquidity" (e.g. Foucault et al., 2013) and that liquidity begets price efficiency in the sense that more liquid currencies have stronger commonality and are less subject to arbitrage deviations.

This paper is organized as follows. Section 2 presents the simple theoretical setting for an unified analysis of volatility, volume and illiquidity on the FX rates, and their commonalities. Section 3 introduces the dataset and discusses summary statistics. Section 4 presents the empirical analysis. Section 5 concludes the paper.

2 A unified model for FX rates, volatility and volume

We depart from the Mixture-of-Distribution Hypothesis (MDH) of Clark (1973) and Tauchen and Pitts (1983), which provides a stylized representation of the supply/demand mechanism on the market at the intradaily level.[8] Let's first consider a world with two currencies, x (base) and y (quote). We assume that the market consists of a finite number $J \geq 2$ of active traders, who take

[8] See also the empirical analysis in Andersen (1996) and the survey in Karpoff (1987). According to Bauwens et al. (2006) only one out of the 19 studies of MDH is on exchange rates.

long or short positions on the FX rate $x|y$. Within a given trading period of unit length (e.g. an hour, a day, a week), the market for the currency pair $x|y$ passes through a sequence of $i = 1, \dots I$ equilibria. The evolution of the equilibrium price is motivated by the arrival of new information to the market. At intra-period i, the desired position of the j-th trader ($j = 1, \dots, J$) on the FX rate $x|y$ is given by

$$q_{i,j}^{x|y}(t) = \xi^{x|y}(p_{i,j}^{x|y,*} - p_i^{x|y}), \quad \xi^{x|y} > 0 \tag{1}$$

where $p_{i,j}^{x|y,*}$ is the reservation price of the j-th trader and $p_i^{x|y}$ is the current market price (both measured in logs). The reservation price of each trader might reflect individual preferences, liquidity issues, asymmetries in information sets and/or different expectations about the fundamental values of the FX rate. In general, the reservation price can deviate from the market price because of idiosyncratic reasons inducing the j-th trader to trade. The term $\xi^{x|y}$ is a positive constant capturing the market depth: The larger $\xi^{x|y}$, the larger quantities of x can be exchanged for y (and viceversa) for a given difference $p_{i,j}^{x|y,*} - p_i^{x|y}$. In other words, $\xi^{x|y}$ measures the capacity of the market to allow large quantities to be exchanged at the intersection between the demand and supply side, thus recalling the concept of resilience. Figure 1 illustrates the demand/supply mechanism of the j-th trader for the $x|y$ FX rates. If $(p_{i,j}^{x|y,*} - p_i^{x|y}) > 0$, this means that the j-th trader believes that the equilibrium trading price of $x|y$ is too low, i.e. currency x should be more expansive relatively to y, so he will buy x and sell y. On the contrary, if $(p_{i,j}^{x|y,*} - p_i^{x|y}) < 0$, the j-th trader will buy y and sell x. The amount associated with a unit change of $p_{i,j}^{x|y,*} - p_i^{x|y}$ is given by the slope $\xi^{x|y}$. The baseline assumptions of the MDH (linearity of the trading function and

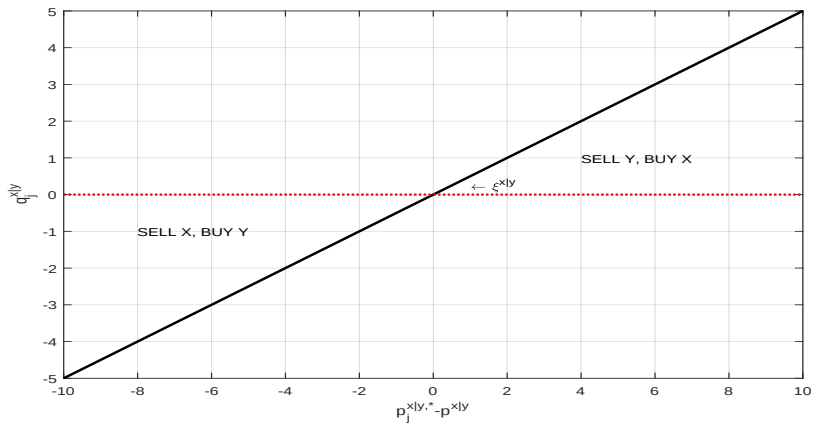


Figure 1: Trading function for the j-th trader for $x|y$ with $\xi^{x|y} = 0.5$.

constant number of active traders) are inevitably very stylized. As for the form of the equilibrium function in (1), note that the trades take place on short intradaily intervals of length $\Delta = 1/I$ and they are generally associated with small price variations. Therefore, it is not restrictive to assume the equilibrium function to be linear on small price changes. Furthermore, the assumption of J active traders observe one market price is more consistent with a centralized market or a fragmented one but with a reference price accessible to trading community, whereas FX rates can be dispersed and heterogeneous outside the interdealer segment, as emphasized by Evans and Rime (2016).

As new information arrives, the traders adjust their reservation prices, resulting in a change in the market price given by the average of the increments of the reservation prices. This means that the equilibrium condition is $\sum_j q_{i,j}^{x|y} = 0$. Hence, the average of the reservation prices clears the market, that is $p_i^{x|y} = \frac{1}{J}\sum_{j=1}^{J} p_{i,j}^{x|y,*}$, and the generated trading volume is

$$v_i^{x|y} = \frac{\xi^{x|y}}{2}\sum_{j=1}^{J} |\Delta p_{i,j}^{x|y,*} - \Delta p_i^{x|y}|,$$

where $\Delta p_{i,j}^{x|y,*} = p_{i,j}^{x|y,*} - p_{i-1,j}^{x|y,*}$ and $\Delta p_{i,j}^{x|y} = p_{i,j}^{x|y} - p_{i-1,j}^{x|y}$. The increments of the reservation log-prices are given by

$$\Delta p_{i,j}^{x|y,*} = \phi_i^{x|y} + \psi_{i,j}^{x|y}, \quad \text{with} \quad j = 1, \ldots, J,$$

where $\phi_i^{x|y}$ is the common information component about the FX rate $x|y$, stemming from public information events, such as those associated with central banks' announcements. The common term $\phi_i^{x|y}$could also be related to events that trigger common directional expectations among the practitioners about a specific currency. The term $\psi_{i,j}^{x|y}$ represents the investor's specific component about the FX rate between x and y. We assume the following continuous time version of the model to form the basis for volatility measurement, where the dynamics of the the investor-specific component about the FX rate is given by

$$d\psi_j^{x|y}(t) = \mu_j^{x|y}(t)dt + \sigma_j^{x|y}(t)dW_j^{x|y}(t), \quad j = 1, \ldots, J \tag{2}$$

where $W_j(t)$ is a Wiener process that is independent between each trader, i.e. $W_l(t) \perp\!\!\!\perp W_m(t)\ \forall l \neq m$ and the term $\sigma_j^{x|y}(t) \geq 0$ is the stochastic volatility process of the j-th trader which is assumed to

have locally square integrable sample paths. The term $\mu_j(t)$ is a predictable and finite variation drift process, which might represent the long-run expectation of the j-th trader about the FX rate and it could be function of fundamental quantities like interest rates differentials and long-term macroeconomic views. By also allowing $\sigma_j^{x|y}$ to be different across traders, we are implicitly introducing heterogeneity among them. This also reconciles with many realistic features including the evidence of long-memory in volatility that is obtained by the superposition of traders operating at different frequencies, see for instance the heterogeneous autoregressive model of Müller et al. (1997) and Corsi (2009). This setup is coherent with a representation of a frictionless market where each trader participates through its reservation price to the price discovery process by carrying new information. On the i-th discrete sub-interval of length $\Delta = \frac{1}{I}$,[9]

$$\psi_{i,j}^{x|y} = \int_{\Delta(i-1)}^{\Delta i} \mu_j(s)ds + \int_{\Delta(i-1)}^{\Delta i} \sigma_j^{x|y}(s)dW_j^{x|y}(s). \tag{3}$$

Proposition 1. *Over an interval of unit length (e.g. a day or a month), the trading volume, $v = \sum_{i=1}^{I} v_i^{x|y}$, and the aggregated volatility, as measured by the realized variance, $RV^{x|y} = \sum_{i=1}^{I} \left(\Delta p_i^{x|y}\right)^2$, or by the power variation of Barndorff-Nielsen and Shephard, 2003, $RPV^{x|y} = \sum_{i=1}^{I} |\Delta p_i^{x|y}|$, carry information about the investor disagreement on a given FX rate.*

Proof in Appendix A.1.

The extension to a continuous-time framework allows us to precisely measure the variability of the FX rates components in the limit for $I \to \infty$, and to relate it to the level of disagreement among investors leading to the observed trading volume. It should be stressed that the asymptotic results behind Proposition 1 are derived by abstracting from microstructural frictions (namely *microstructure noise*), like transaction costs in the form of bid-ask spread, clearing fees or price discreteness, which are intimately related and endogenous to the trading process; see the recent works of Darolles et al. (2015, 2017) for an extension of reduced-form version of the MHD with liquidity frictions. From a statistical point of view, as $I \to \infty$, the microstructure noise dominates over the volatility signal, thus leading to distorted measurements of the variance. However, over moderate sampling frequencies, e.g. 5-minute intervals over 24 hours ($I = 288$), the prices and quantities determined in equilibrium in each sub-interval can be considered (almost) free of

[9]For ease of exposition, we assume that trades happen on an equally spaced and uniform grid, $i = 1, 2, \ldots I$. This assumption can be relaxed allowing for random trading times.

microstructure noise contamination, and representative of new equilibria on the aggregated supply/demand functions. Rather than microstructural features, this setting outlines the aggregated disagreement on fundamentals leading to the price discovery process in which each trader participates to the equilibrium price variations in proportion to the information contained in her new reservation prices. As it is common in the literature on volatility measurement, see Bandi and Russell (2008) and Liu et al. (2015), in the following analysis we will work under the maintained assumption that sampling at 5-minute intervals is sufficient to guarantee that a new equilibrium price is determined. The latter is representative of the aggregated information contained on the demand and supply sides of the market. Furthermore, the assumption of independence between ϕ_i and $\psi_{i,j}$ and across traders does not allow for reversal or spill-over effects such as those studied in Grossman and Miller (1988) to investigate the mechanics of liquidity provision. The same type of sequential trading behavior has been recently proved to be responsible for crash episodes in Christensen et al. (2016) and associated with changes in the level of investors' disagreement around important news announcements, see Bollerslev et al. (2016). Despite the stylized set of assumptions, the next section shows how the theory outlined above can be successfully adopted as an encompassing framework to characterize the illiquidity and the commonalities in volatility and volume on the global FX markets.

2.1 Measuring FX Illiquidity

In light of Proposition 1 and analogously to the price impact illiquidity proxy in Amihud (2002), we can define a continuous-time version of the illiquidity index as

$$A^{x|y} := \frac{RPV^{x|y}}{v^{x|y}}, \tag{4}$$

which measures the price impact of a given trade, that is the amount of volatility of the FX rate associated with a unit of trading volume. The following proposition highlights the main determinants of market illiquidity.

Proposition 2. *Consider the illiquidity measure defined in* (4)*. In the limit for* $I \longrightarrow \infty$ *and under*

homogeneity of traders, i.e. $\sigma_j^{x|y} = \sigma^{x|y} \quad \forall j = 1, 2, \ldots, J,$

$$p \lim_{I \to \infty} A^{x|y} = \frac{2}{\xi^{x|y} J \sqrt{(J-1)}}. \tag{5}$$

Proof in Appendix A.2.

Proposition 2 shows that on a period of unit length, $A^{x|y}$ is inversely related to the slope, $\xi^{x|y}$, of the equilibrium function in (1). That is, for a given difference between the reservation price and the market price, $A^{x|y}$ decreases as this slope increases. In particular, for large values of $\xi^{x|y}$ large volume would be associated with small variations between the prevailing price and the reservation price for each trader, thus signaling market depth and liquidity. Instead, when $\xi^{x|y} \to 0^+$, i.e. in the limiting case of a flat equilibrium function in (1), the liquidity is minimal (and $A^{x|y}$ diverges), since no actual trade takes place. Under the assumption of homogeneity of the traders, i.e $\sigma_j^2(t) = \sigma^2(t)$ $\forall j = 1, \ldots, J$, Proposition 2 also highlights the inverse relationship between the number of active traders on the market and illiquidity.[10]

In the extreme case of only one observation per trading period $I = \Delta = 1$, the illiquidity measure in (4) reduces to the original Amihud index (up to the rescaling by $\sqrt{2/\pi}$),

$$A^{x|y,*} = \frac{|r|^{x|y}}{v^{x|y}}, \tag{6}$$

for which it is not trivial to obtain an expression as a function of the structural parameters analogous to the one in (5). For instance, the expected value of $|r|^{x|y}$ under Gaussianity is proportional to the daily (constant) volatility parameter, i.e. $E\left(|r|^{x|y}\right) = \sigma\sqrt{\frac{2}{\pi}}$, where $\sigma = \sqrt{Var(\phi) + Var(\psi)/J}$ in the original MDH theory. In the classic framework, inference on the structural parameters is performed through GMM by relying on the unconditional moments of the observable quantities which depend on the underlying (unobservable) information flow, see Richardson and Smith (1994) and Andersen (1996). The availability of high-frequency data coupled with the theory of quadratic variation makes the volatility and consequently the information flow *measurable* quantities. This means that inference on the structural parameters becomes more precise as we adopt moment conditions based on high-frequency data, see Li and

[10]Relaxing the assumption of homogeneity would result in the ratio of two aggregated volatility measures, each estimating the weighted average of the variance carried by each trader, see equation (30) in Appendix A.2.

Xiu (2016).

2.2 Commonalities in FX volume and volatility

In this section, we derive equilibrium relations between returns, trading volumes and volatilities *across* different FX rates. These relations are instrumental to the interpretation of commonalities in trading volumes and volatilities as well as information processing in global FX markets. Let's therefore consider a world with three currencies, x , y and z. The market for the currency pairs $x|y$, $x|z$ and $z|y$ also passes through a sequence of $i = 1, \dots I$ equilibria and the evolution of the equilibrium price of each currency pair is motivated by the arrival of new information to the market. By the triangular no-arbitrage parity it must hold that

$$p_i^{x|y} = p_i^{x|z} + p_i^{z|y}, \tag{7}$$

where $p_i^{x|z} = \sum_{j=1}^{J} p_i^{x|z,*}$ and $p_i^{z|y} = \sum_{j=1}^{J} p_i^{z|y,*}$. By imposing that $\Delta p_{i,j}^{x|z,*} = \phi_i^{x|z} + \psi_{i,j}^{x|z}$, and $\Delta p_{i,j}^{z|y,*} = \phi_i^{z|y} + \psi_{i,j}^{z|y}$, the *synthetic* return on $x|y$ results to be

$$\widetilde{r}_i^{x|y} = \phi_i^{x|z} + \phi_i^{z|y} + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{x|z} + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{z|y}. \tag{8}$$

Assuming that the common information component on the rate $x|y$, can be disentangled into two currency-specific terms ϕ_i^x and ϕ_i^y, with $\phi_i^{x|y} = \phi_i^x - \phi_i^y$,[11] it follows that

$$\widetilde{r}_i^{x|y} = \phi_i^x - \phi_i^y + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{x|z} + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{z|y},$$

where the common information part of $\widetilde{r}_i^{x|y}$ is the same as for $r_i^{x|y}$, that is $\phi_i^x - \phi_i^y$. It follows that the MDH coupled with the triangular no-arbitrage relation on the FX rates, i.e. $r_i^{x|y} = \widetilde{r}_i^{x|y}$, prescribes that

$$\frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{x|y} = \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{x|z} + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{z|y}, \tag{9}$$

[11] In Section 4.1 we discuss a strategy to separately identify ϕ_i^x and ϕ_i^y based on a cross section of FX rates and provide an empirical validation of such an assumption.

which means that the average of the traders' specific terms on $x|y$ must be equal to the sum of the average traders' specific terms of $z|y$ and $x|z$. This means that each trader can take a direct position on $x|y$ or operate on the synthetic rate by forming independent beliefs on $x|z$ and $z|y$, thus generating trading volume on each individual FX market.

Proposition 3. *Trading volume, volatility and liquidity across FX rates are linked by no-arbitrage constraints, which lead to the commonalities across FX rates. The synthetic volatility, as measured by* $\widetilde{RV}^{x|y} = \sum_{i=1}^{I} (\tilde{r}_i^{x|y})^2$*, and synthetic volume, denoted as*

$$\tilde{v}_i^{x|y} = \frac{\xi^{x|y}}{2} \sum_{j=1}^{J} |\Delta p_{i,j}^{x|z,*} - \Delta p_i^{x|z} + \Delta p_{i,j}^{z|y,*} - \Delta p_i^{z|y}|, \tag{10}$$

reveal the strength of the correlation across FX rates.

Proof in Appendix A.3.

Proposition 3 introduces the concept of *synthetic* volatility and volume, which are associated with the no-arbitrage equilibrium constraints and depend on the extent of the individual disagreement on the FX rates of $x|z$ and $z|y$. Furthermore, both synthetic volatility and volume are functions of the aggregated correlation in beliefs between $x|z$ and $z|y$, and hence expression of the commonalities in the global FX rates. For a given level of traders' disagreement on $x|y$ (leading to trading volume on $x|y$), we can measure the associated synthetic volume on $x|z$ and $z|y$, which is proportional to the correlation between the aggregated reservation prices on $x|z$ and $z|y$. The same holds true for the synthetic volatility, as measured by the realized variance of the synthetic return.

3 Data and Preliminary Analysis

3.1 Data Sets

Our empirical analysis relies on two data sets covering 29 currency pairs (15 currencies) over the period from November 2011 to November 2016.[12] First, trading volume data come from CLS,

[12]The full dataset contains data for 18 major currencies and 33 currency pairs. To maintain a balanced panel, we exclude the Hungarian forint (HUF), which enters the dataset only on 07 November 2015. Moreover, we discard USDILS and USDKRW due to very infrequent trades. We obtain very similar results by including them. The remaining

which is the largest payment system for the settlement of foreign exchange transactions launched in 2002. By means of a payment-versus-payment mechanism, this infrastructure supports FX trading by removing settlement risk and supporting market efficiency. For each hour of our sample period and each currency pair, we observe the settlement value and number of settlement instructions. Following the literature (e.g. Mancini et al., 2013), we exclude observations between Friday 10PM and Sunday 10PM since only minimal trading activity is observed during these nonstandard hours.[13] In 2017, the core of CLS was composed of 60 settlement members including the top ten FX global dealers, and thousands of third parties (other banks, non-bank financial institutions, multinational corporations and funds), which are customers of settlement members. The total average daily traded volume submitted to CLS was more than USD 1.5 trillion, which is around 30% of the total daily volume recorded in the last available BIS triennial survey (Bank of International Settlements 2016). However, after adjusting for the large fraction of BIS volume originated from interbank trading across desks and double-counted prime brokered "give-up" trades, the CLS data should cover about 50% of the FX market (Gargano et al., 2019 & Hasbrouck and Levich, 2017). In our study, we focus on FX spot transactions. Except for some exceptions such as the Renminbi, the CLS spot FX rates in our sample are highly representative of the entire FX market. For instance, the currency pairs involving the USD and EUR cover more than 85% (94%) of the total trading volume of the BIS triennial survey.

To the best of our knowledge, only few papers have analyzed CLS volume data so far. First, Fischer and Ranaldo (2011) study five aggregated currencies (e.g. all CLS-eligible currencies against the U.S. dollar, Euro, Yen, Sterling, and Swiss franc) rather than currency pairs. Hasbrouck and Levich (2017) analyze every CLS settlement instruction during April 2013. Gargano et al. (2019) use the same dataset to perform an asset pricing analysis. Ranaldo and Somogyi (2019) analyze the heterogeneous price impact of CLS order flows decomposed by market participants .

The second data set is obtained from Olsen Financial Technologies, which is the standard source for academic research on intraday FX rates. By compiling historical tick data from the

29 currency pairs are: AUDJPY, AUDNZD, AUDUSD, CADJPY, EURAUD, EURCAD, EURCHF, EURDKK, EURGBP, EURJPY, EURNOK, EURSEK, EURUSD, GBPAUD, GBPCAD, GBPCHF, GBPJPY, GBPUSD, NZDUSD, USDCAD, USDCHF, USDDKK, USDHKD, USDJPY, USDMXN, USDNOK, USDSEK, USDSGD, and USDZAR.

[13] In this paper, times are expressed in GMT

main consolidators such as Reuters, Knight Ridder, GTIS and Tenfore, Olsen data are representative of the entire FX spot market rather than specific segments such as the interdealer FX market dominated by two electronic limit order markets: EBS and Reuters. For each minute of our sample period and each currency pair, we observe the following quotes: bid, ask, high, low, and midquotes. With these data at hand, we can analyze at least four aspects of FX rates: (i) the FX rate movements at one minute or lower frequencies; (ii) the realized volatility or other measures of return dispersion; (iii) the quoted bid-ask spread as a measure of transaction cost; and (iv) violations of triangular arbitrage conditions.

3.2 Descriptive Analysis

In this subsection, we highlight some (new) stylized facts characterizing the times series of volume, volatilities and illiquidity measures associated with the 29 FX rates under investigation. First, we look into intraday patterns and then we study the daily time series of FX volume, volatility, and illiquidity.

To start with the intraday analysis, Figure 2 displays the total hourly volume series, denoted as $v_t^{tot} = \sum_{l=1}^{L} v_t^l$, where v_t^l is the hourly volume on the l-th FX rate. This plot highlights the size

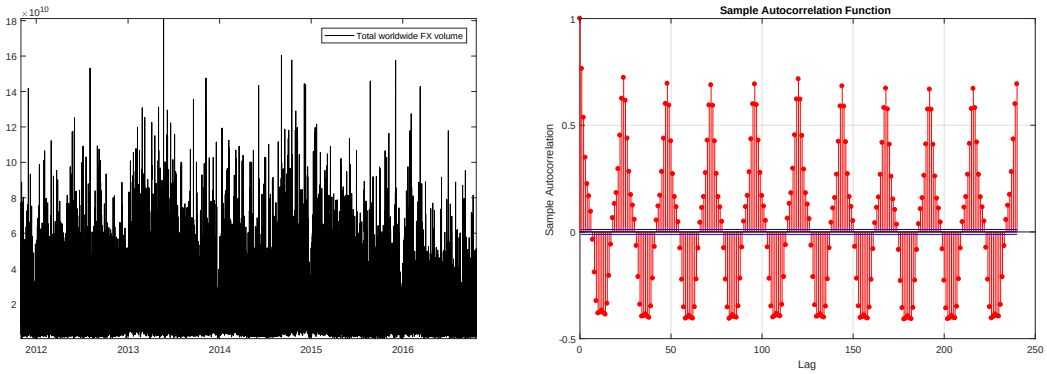


Figure 2: Time series and auto-correlation function of the total global volume.

and deepness of the FX market, with an average of around 20 billions USD dollar traded every hour. Moreover, the series of total volume is rather persistent and it clearly displays cyclical patterns, which can be associated with strong intradaily seasonality. We explicitly model the

intradaily patterns by estimating the following model with OLS

$$\log(v_t^l) = \delta_t \beta + \epsilon_t, \tag{11}$$

where δ_t contains hourly and day-of the-week dummies. We can also obtain the *filtered* volume as $v_t^l = \frac{v_t^l}{e^{\delta_t \hat{\beta}}}$. The hourly average of the total global volume is reported in Figure 3. The plot highlights that the average total volume is higher during the opening hours of the European and American stock markets, while it is very low between 10PM and 12AM as most of the largest stock markets are closed, while it has a relative peak associated with the opening of Tokyo (2 AM). Moreover, the total volume is the largest on average between 3PM and 4PM, i.e. before the WMR Fix, for which there is a well documented literature about the large traders submitting a rush of orders before the setting of the daily benchmarks for FX prices, see e.g. Marsh et al. (2017) and Evans (2018). Finally, Figure 3 shows that the filtering successfully removes the largest part of the seasonal pattern and that the filtered volume displays significant autocorrelation after many periods.

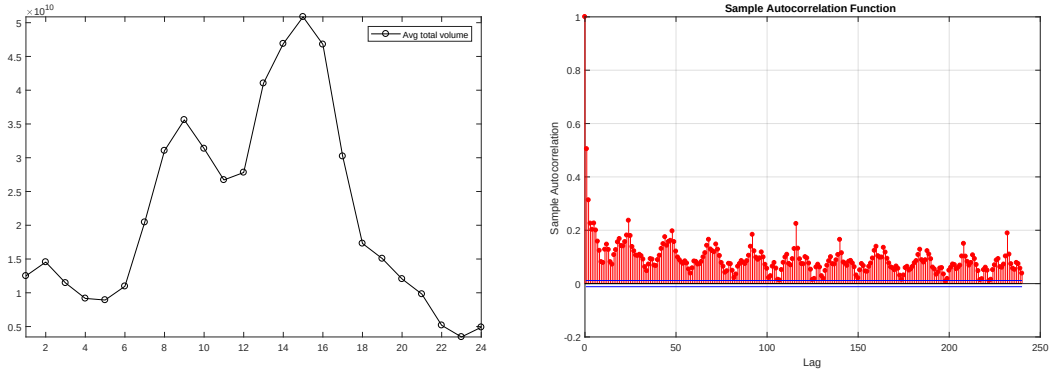


Figure 3: Hourly average of the total global volume and ACF of the *filtered* volume.

Turning the attention to individual FX rates, Figure 4 reports the hourly average share of the total volume of the five most liquid FX rates (by volume size). Firstly, as expected all the most liquid FX rates involve the USD as either base or quote currency. As for the total volume, the trading volume of the most liquid FX rates displays clear (intraday) seasonal patterns. For the individual FX rates, these patterns are suggestive of local effects in given geographical areas, coherent with the OTC segmented nature of FX markets. For instance, USDJPY covers around 30% of the total FX volume between 12PM and 4PM, that are the hours in which Far East markets

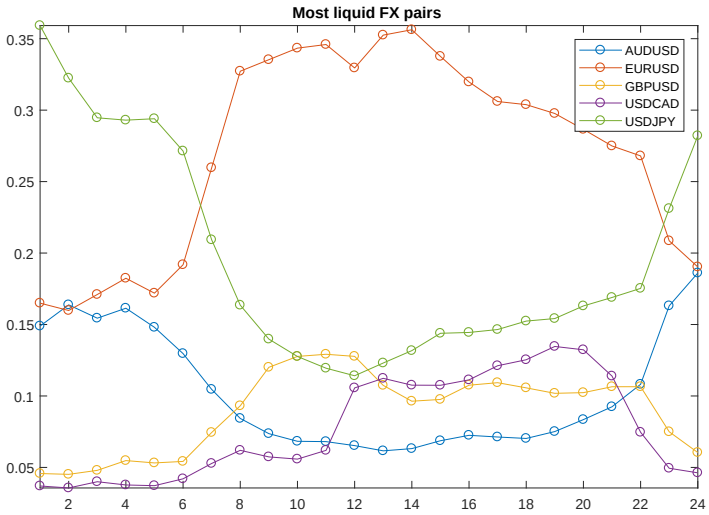


Figure 4: Hourly volume averages of the five most liquid FX rates, which are (in order) USDEUR, USDJPY, USDGBP, USDAUD, USDCAD.

are open. AUDUSD contributes with a 15% in the same hours, while its market share strongly declines to 7% during the central hours of the day. EURUSD is by far the most traded FX rate, with a share above 30% between 7AM to 6PM. A similar pattern characterizes also GBPUSD with an average share ranging between 5% and 10%. Finally, USDCAD is mostly traded at the opening of the business hours in North America, i.e. between 12PM and 10PM, with approximately 10% share on the total volume. These five FX rates amount for a share of more than 70% of the total global volume in every hour. Summarizing, the seasonal patterns are clearly discernible in two dimensions. First, on an intraday scale the trading volume follows the working time in each country or jurisdiction defining the currency pair. This means that round-the-clock, the trading volume of New Zeeland dollar is the first to increase, followed by Asian, European, and American currencies. Second, official banking holidays clearly reduce the trading activity. The seasonalities and calendar effects will be carefully considered in our empirical analysis.

Concerning the relationship between volatility and volume, Figure 5 shows that the hourly averages of realized volatility and volume for USDEUR and USDJPY follow the same patterns. At the intradaily level when the volatility on the FX rates is high, also the volume is high, which points to a wider variation of traders' reservation prices. Thus, Figure 5 provides *prima facie* evidence to Proposition 1 in our theoretical setting, that is, volatility and volume are mostly governed by a common latent factor, which seen through the lenses of the MDH represents the

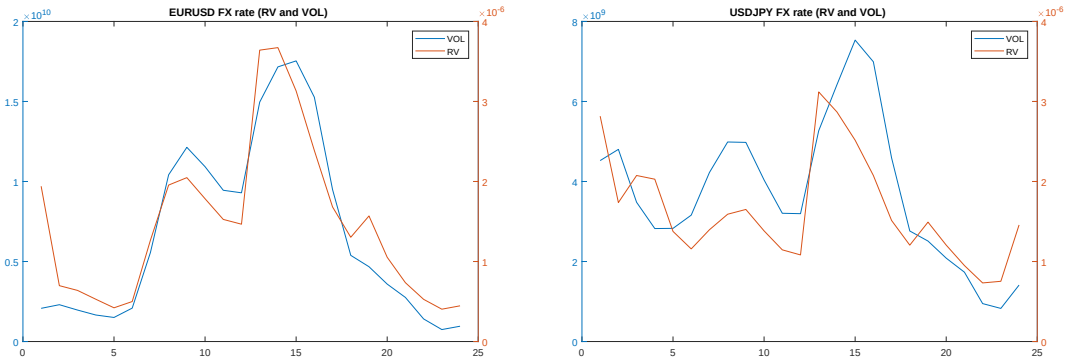


Figure 5: Hourly Averages of RV and VOL. In Panel a) USDEUR, in Panel b) USDJPY.

information flow proportional to the level of heterogeneous beliefs (disagreement) between the agents.

Before performing the empirical analysis and test our model predictions, we examine how daily changes in trading volume correlate with daily changes in realized volatility and other factors that proved to explain FX liquidity in the previous literature (e.g. Mancini et al., 2013 and Karnaukh et al., 2015) and trading activity in stock markets (e.g. Chordia et al., 2001).

Some of these variables are likely to determine each other endogenously. Rather than causation, the purpose of this analysis is to document some novel correlation patterns pertaining to FX trading volume. More specifically, we perform a panel regression of all currency pairs, in which the daily FX volume is explained by daily (realized) volatility, (average intraday) relative bid-ask spread (BAS), a dummy variable for the dollar appreciation, two common proxies of market stress such as the TED spread (the yield spread between the U.S. three-month Libor and T-bills) and FX VIX (i.e. the JP Morgan Global FX volatility index), and four weekday dummy variables equal to one if the trading day is on Monday, Tuesday, Thursday, and Friday, respectively. All variables except the dummy variables are taken in logs and changes and all regressions include the lagged dependent variable as additional regressor. For sake of comparison, we repeat similar regressions using dependent variables the realized volatility, the relative bid-ask spread, as well as the Amihud illiquidity measure, which will be studied in more details later.

Some novel patterns emerge from the analysis reported in the Table 1. On the one hand, FX trading volume increases with realized and implied FX volatility as well as TED spread, whereas it decreases with the relative bid-ask spread. Trading volume follows an inverted U-shape across

	(1) Δ Volume	(2) Δ RPV	(3) Δ Amihud	(4) Δ Relative BAS
Δ Volume	-	0.1374^a (98.778)	-	-0.04395^a (-43.356)
Δ RPV	1.301^a (92.12)	-	-	0.3995^a (148.69)
Δ Relative BAS	-0.9197^a (-40.87)	0.9049^a (154.33)	0.7961^a (42.501)	-
USD	0.00679 (1.6193)	-0.00852^a (-6.232)	-0.00866^a (-1.9637)	-0.00856^a (-9.4226)
Δ TED	0.1174^b (2.347)	0.0301^c (1.8444)	-0.1746^a (-3.3154)	-0.0754^a (-6.9454)
Δ VXY	0.1956^b (2.338)	0.6120^a (22.477)	-0.5974^a (-6.8248)	0.4271^a (23.521)
Monday	-0.3508^a (-50.68)	0.0324^a (14.572)	0.3285^a (45.258)	-0.0858^a (-60.692)
Tuesday	0.0206^a (3.01)	-0.0176^a (-8.1402)	-0.1089^a (-15.509)	-0.0362^a (-24.383)
Thursday	-0.1072^a (-16.65)	0.0005 (0.2354)	0.1300^a (19.102)	0.0054^a (3.9088)
Friday	-0.0766^a (-11.34)	-0.1072^a (-51.62)	0.1932^a (28.41)	0.0643^a (46.22)
Lagged Dep.	-0.3494^a (-81.065)	-0.0458^a (-35.95)	-0.4211^a (-91.694)	-0.2114^a (-55.475)
Constant	0.0986^a (19.167)	0.0231^a (14.227)	-0.1034^a (-19.324)	0.0146^a (13.516)
R^2	0.418	0.586	0.404	0.573
N	37055	37054	37054	37055

Table 1: Regressions of volume, volatility, illiquidity, and bid-ask spread. Volume and RPV are the daily trading volume and realized variance respectively, Amihud is the ratio between daily RPV and daily volume, and the bid-ask spread is the daily average of one-minute spreads. The *t*-statistics are in parentheses and the error variance are robust to heteroskedasticity and autocorrelation in the residuals. Except for dummy variables, all variables are taken in logs and changes. The superscripts *a*, *b* and *c* indicate significance at 1%, 5% and 10% significance level respectivelyy.

weekdays, that is, larger trading volumes tend to occur in the middle of the week. On the other hand, realized volatility increases with bid-ask spreads and tends to be lower when the U.S. dollar appreciates, possibly due to its status as international currency reserve and safe haven against several currencies (e.g. Ranaldo and Söderlind, 2010 and Maggiori, 2017). In addition to FX volume, (negative) autocorrelation and weekdays effects are discernible for FX volatility,

illiquidity, and relative bid-ask spread.

4 Empirical Analysis

Our theoretical setup in Section 2 offers three main propositions. For each of them we provide an in-depth empirical analysis in a separate subsection.

4.1 Determinants of FX trading volume and volatility

The first theoretical proposition postulates that volatility and trading volume are proportional to the level of heterogeneous beliefs between agents, that is traders' disagreement about the fundamental value of the FX rates. This proposition delivers two main empirical predictions: On the one hand, both volume and volatility should increase with disagreement. On the other hand, common news leading to a currency appreciation or depreciation with no or little disagreement can generates above-average volatility but no extraordinary trading volume.

To test the first empirical prediction, we follow Beber et al. (2010) and measure disagreement as heterogeneity in beliefs of market participants by using a detailed data set of currency forecasts made by a large cross-section of professional market participants. More specifically, we collect all Thomson Reuters surveys recorded at the beginning of every month during our sample period and compute measures of cross-sectional dispersion such as the (standardized) standard deviation of FX forecast and the high-low range from the distribution of FX forecasts of on average about 50 market participants.[14] This measure of heterogeneity in beliefs that we call *disagreement* is the main regressor in two panel regressions in which total trading volume and realized volatility are the dependent variables. In addition to our measure of disagreement, we include a constant, the lagged dependent variable, and FX illiquidity proposed in Karnaukh et al., 2015 as a control. All variables are taken in logs and changes.[15] As shown in column (1) and (2) of Table 2, both

[14]The total number of monthly observations included in the regression is 940, which includes the following 26 currency pairs: AUDJPY, AUDNZD, CADJPY, EURAUD, EURCAD, EURCHF, EURGBP, EURJPY, EURNOK, EURSEK, GBPCAD, GBPCHF, GBPJPY, USDAUD, USDCAD, USDCHF, USDEUR, USDGBP, USDHKD, USDJPY, USDMXP, USDNOK, USDNZD, USDSEK, USDSGD and USDZAR. Not for all currency pairs, forecasts are available from November 2011 onwards. The exact number of market participants depends on the currency pair. We report results using standard deviations of FX forecast. Using ranges, we obtain very similar results.

[15]We perform additional analyses including further regressors such as the TED and VXY and the results remain qualitatively the same.

	(1) Δ Volume	(2) Δ RV	(3) Δ Amihud	(4) Δ Relative BAS
Δ Disagreement	0.0503^b	0.1736^a	0.0397^b	0.0373^a
	(2.25)	(3.41)	(1.96)	(3.62)
Δ Illiquidity	0.0779^a	0.5431^a	0.1961^a	0.1405^a
	(5.29)	(8.12)	(5.66)	(10.65)
Lagged Dep.	-0.3499^a	-0.2625^a	-0.2389^a	0.0289
	(-10.03)	(-5.34)	(-4.38)	(0.97)
Constant	-0.0097	-0.0113	0.0041	-0.0067^c
	(-1.50)	(-0.81)	(0.61)	(-1.91)
R^2	0.146	0.367	0.254	0.345

Table 2: Monthly regression analysis - disagreement. The t-statistics are in parentheses and the error variance are robust to heteroskedasticity and autocorrelation in the residuals. Disagreement is the standardized standard deviations of Thomson Reuters forecasts, which are available on a monthly basis. Volume and RV are the daily trading volume and realized variance respectively, Amihud is the ratio between daily RPV and daily volume, bid-ask spread is the daily average bid-ask spread, and illiquidity is taken from Karnaukh, Ranaldo and Söderlind (2015). Except for illiquidity, all variables are taken in logs. The superscripts a, b and c indicate significance at 1%, 5% and 10% significance level respectively.

trading volume and volatility increase with disagreement providing evidence in support to our first empirical prediction. Moreover, both trading volume and volatility tend to increase with FX illiquidity, consistent with dealers' inventory imbalances and hot potato effects Lyons (1997).

The next empirical prediction is that common news or informational events sparking little disagreement across traders should not generate any extraordinary trading volume but it might result in above-average volatility. More specifically, the model prescribes that if new information is common as for macroeconomic announcements, then traders would promptly revise their reservation prices in the same manner and nearly no additional transaction volume should be generated. To avoid confounders and overlapping occurrences, the detection of such informational events needs an accurate identification econometric technique and granular (intraday) data. The recent advances in the literature on *jump* processes come to the aid of this analysis. Similarly to Bollerslev et al. (2016), we rely on a simple setup for the common news component, i.e. the "jumps", to separately identify it from the component of the variations in the FX rates due to the disagreement among traders.[16] For instance, $\phi_i^{x|y}$ can be modeled as compound Poisson

[16]Other studies associating large price jumps with news announcements are in Andersen et al. (2007) and Lee (2011).

processes as

$$\phi_i^{x|y} = \sum_{l=1}^{N_i^{x|y}} Z_l^{x|y}, \tag{12}$$

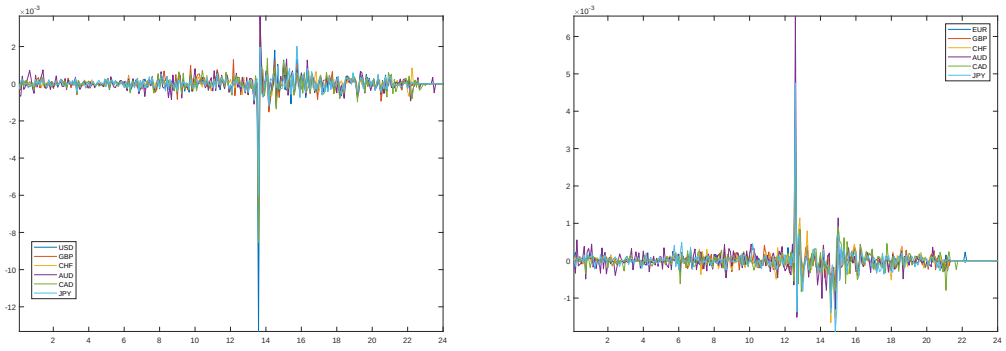
where $N_i^{x|y}$ is an independent Poisson random variable with intensity $\lambda_{x|y}\Delta$, where $\lambda_{x|y}$ is expressed with respect to the unit scale (e.g. daily). $Z_l^{x|y} \overset{iid}{\sim} D_{x|y}(\theta_{x|y}) \in \mathrm{R}$ where $\theta_{x|y}$ are the parameters associated with the distribution $D_{x|y}$. Furthermore, we assume that $\phi_i^{x|y}$ can be further decomposed into currency specific variations, that is $\phi_i^{x|y} = \phi_i^x - \phi_i^y$. For instance, we can assume that $\phi_i^x = \sum_{l=1}^{N_i^x} Z_l^x$ and $\phi_i^y = \sum_{l=1}^{N_i^y} Z_l^y$. The terms ϕ_i^x and ϕ_i^y cannot be uniquely identified by looking at a single FX rate since a large variation in the FX rate might be due to good (bad) news on x or bad (good) news on y. Therefore, we rely on the theory of *co-jumps*, as developed in Caporin et al. (2017), to identify ϕ_i^x given a cross section of FX rates with the same base currency x. In other words, the simultaneous occurrence of a jump in all the FX rates trading with a given base currency x allows us to identify episodes characterized by the ex-post realization of a currency-specific news common to all traders. In turns, this enables us to identify large and sudden directional appreciations or depreciations of one currency against the other currencies associated with no or little disagreement. The test for co-jumps proposed by Caporin et al. (2017) takes the form

$$C\mathcal{J} = \frac{1}{\zeta} \sum_{j=1}^{N} \frac{\left(SRV_j - \widetilde{SRV}_j\right)^2}{SQ_j}, \tag{13}$$

where N denotes the number of FX rates, ζ is a design parameter, SRV is the smoothed randomized realized variance of Podolskij and Ziggel (2010), $\widetilde{SRV}$ is the smoothed version of the truncated realized variance estimator of Mancini (2009) which is robust to jumps, while SQ is a smoothed estimator of the quarticity. Under the null hypothesis of absence of co-jumps, $C\mathcal{J}$ converges to a chi-square distribution with N degrees of freedom. Under the alternative hypothesis of at least one co-jump across all N series, $C\mathcal{J}$ diverges.

Figure 6 illustrates two representative episodes detected with the test for co-jumps developed in Caporin et al. (2017).[17] The left panel reports the log-returns of the FX rates of EUR against the six major currencies, USD, GBP, CHF, AUD, CAD and JPY on November 6, 2015. The sudden depreciation of the Euro occurred in reaction to a speech by the President of ECB, Mario Draghi

[17] We thank the authors for sharing with us their MATLAB code to detect co-jumps.



(a) EUR (b) USD

Figure 6: Co-jumps analysis. The figures reports the five-minute returns on six FX rates on days when the test of co-jumps of Caporin et al. (2017) has detected significant jumps at 0.01% significance level. The left plot reports the returns of the FX rates of USD, GBP, CHF, AUD, CAD and JPY against EUR on November 6, 2015. The right plot reports the returns of the FX rates of EUR, GBP, CHF, AUD, CAD and JPY against USD on May 1, 2014.

reinforcing traders' belief about the continuation of the Eurosystem's bond purchases (Quantitative Easing) as a stabilization tool to resolve the crisis situations in the financial market. The FX rate reacted with a sudden depreciation of EUR against all other currencies by approximately 1% on an interval of five minutes. The magnitude of such a variation is several times larger than the variation under *normal* market conditions, where the changes in the reservation prices of each individual trader is averaged over J traders. An analogous evidence arises for the appreciation of the USD against all major currencies on May 1, 2014, following the rumors on the beginning of a tapering policy by the Federal reserve.

To test our second empirical prediction, we examine whether trading volume significantly increases when the FX rates are hit by large and directional news. Using hourly time series, we perform the following panel regression with fixed effects

$$V_{i,t} = \alpha_i + \beta CJ_t + \delta BA_{i,t} + \gamma_h h_t + \gamma_w w_t + \rho V_{i,t-1} + \varepsilon_{i,t}, \tag{14}$$

where $V_{i,t}$ is the log-volume on the i-th FX rate trading against a given base currency, CJ is a dummy variable for a significant co-jump on the base currency. We control for illiquidity by including $BA_{i,t}$, i.e. the relative bid-ask spread on the i-th FX rate, and seasonal effects with h_t and w_t that are hourly and day-of-the-week dummies. The coefficient β captures the sudden in-

crease/reduction in the average trading volume associated with co-jumps. To guarantee enough counterparts to each currency, we analyze the four main currencies, i.e. co-jumps of USD, EUR, JPY, and GBP. Regression (14) can be considered the multiple-jumps analogous in the panel setting of the jump regression formalized in Li et al. (2017) and applied in Bollerslev et al. (2016) in the context of macroeconomic announcements. We replicate this analysis for realized volatility.

Table 3 reports the estimation results for four different base currencies, EUR, GBP, USD and JPY and 6 FX rates each (including also CHF, AUD and CAD). For the trading volume, the coefficient β is almost never significant at 5% level supporting the hypothesis that despite a sizable currency movement, common news with little disagreement does not induce abnormal trading volume. On the other hand, (realized) volatility is positively affected by the arrival of large directional news in almost all cases. To sum up, as prescribed by the theory common news that is similarly interpreted by all market participants induces price variation but no abnormal trading volume. On the other hand, both volume and volatility tend to increase with disagreement.

	EUR		**G**BP		**U**SD		**J**PY	
Volume	FE	PO	FE	PO	FE	PO	FE	PO
Baseline	0.0172	0.0170	-1.2007^b	0.7844	0.1082	0.1083	-0.2252^b	-0.2256^b
Controls	0.0412	0.0147	-0.0683	-0.1257	0.0609	0.0346	-0.0168	-0.0799
	EUR		**G**BP		**U**SD		**J**PY	
Volatility	FE	PO	FE	PO	FE	PO	FE	PO
Baseline	0.6586^a	0.6579^a	0.0511	0.0510	0.3966^a	0.3960^a	0.4233^a	0.4229^a
Controls	0.2757^a	0.3127^a	0.0643^b	-0.0545	0.0973^a	0.1217^a	0.2818^a	0.2798^a

Table 3: Common news, volume and volatility. Panel regression estimates with fixed effect (FE) and pooling (PO) of the parameter β in (14). The dependent variable are logarithm of the hourly trading volume and RV for six FX rates with different base currency EUR, GBP, USD and JPY. The regressors are the dummy variable for the co-jump (CJ) on the base currency (baseline specification), and a number of controls: the average relative bid-ask spread (BA), and hourly and day-of-the-week dummies and an AR(1) term. The superscripts a, b and c indicate significance at 1%, 5% and 10% significance level respectively.

4.2 FX Illiquidity

Proposition 2 in Section 2 provides a closed-form expression for illiquidity in the spirit of Amihud (2002), i.e. the ratio between volatility and trading volume. The empirical prediction is that illiquidity decreases with market depth and the number of active traders. It is difficult to accurately measure these quantities. However, the visual inspection of Figure 7 representing the intraday

development of our EURUSD Amihud measure suggests that illiquidity tends to decrease when international financial centers are open, that is, when the FX market is deep and populated by active traders. More precisely, it is discernible that FX illiquidity abruptly decreases at the opening of the European markets and it is minimal when both the European and the American markets are jointly open. After 8PM the illiquidity grows again and it is maximal during the night hours. A consistent pattern also holds for USDJPY (the right-hand side figure 7): market illiquidity reduces at the opening of the main financial markets Tokyo, London and New York and it sensibly increases again after 4PM. To shed further light on the measurement ability of our illiquidity

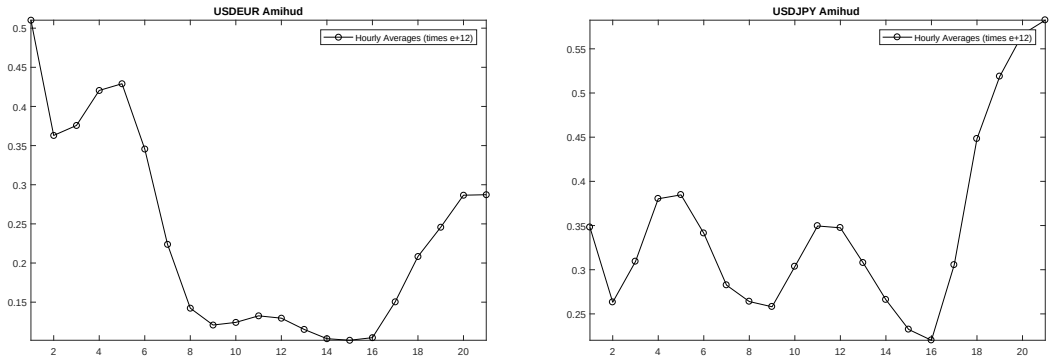


Figure 7: Hourly Averages of FX Amihud measures. In Panel a) USDEUR, in Panel b) USDJPY.

indicator, we perform various regressions similar to those shown in Table 1 and in Table 2. First, we regress changes in our daily illiquidity indicator on daily changes of bid-ask spreads. The results are exhibited in column (3) of in Table 1. Second, we regress monthly changes of our illiquidity indicator on a comprehensive measure of FX illiquidity proposed in Karnaukh et al., 2015 that proved to be highly correlated with precise high-frequency (intraday) data from Electronic Broking Services, which is the major interdealer trading platform for many currencies. The results are presented in column (3) of in Table 2. In both regressions, we include control variables.[18] Overall, we find that our illiquidity measure in the spirit of the Amihud indicator increases with other well-accepted measures of FX illiquidity.

So far, we have analyzed FX illiquidity on a global scale. Now, we ask the question whether our FX illiquidity measure is highly correlated with other illiquidity proxies in the FX interdealer segment. To do this, we obtain intraday data from Electronic Broking Services (EBS), the leading

[18] In addition to daily and monthly time intervals, we have performed the same regressions with weekly data and obtained consistent results.

platform for spot FX interdealer trading for various FX rates including the EURUSD. For the entire 2016, we access the depth of book at ten levels on both sides (bid and offer quotes) snapped every 100 milliseconds, the exact identification whether the deal is given or paid, transaction prices and amounts. We focus on EURUSD, which is primarily traded on this interdealer trading platform.[19] In the same spirit of Hasbrouck (2009), we analyze correlations between illiquidity measures. More specifically, we compute the following proxies: quoted spread (i.e. ask minus bid quotes), relative quoted spread (i.e. quoted spread divided by midquote), effective cost (i.e. the absolute value of the difference between transaction price and midquote), traditional Amihud measure (i.e. absolute return over trading volume), cost estimates implied by the Roll model (Roll, 1984), order flow price impacts (i.e. at five-minute intervals and trade-by-trade).

	A_t	BAS_t	Rel-BAS_t	EC_t	R_t	γ_1	γ_2	γ_3	Daily-A_t
Pearson correlation									
A_t	1.0000	0.5176	0.5628	0.8958	0.9000	0.6220	0.6945	0.7393	0.8527
BAS_t	0.5176	1.0000	0.9890	0.6092	0.5474	0.2520	0.2484	0.4524	0.4175
Rel-BAS_t	0.5628	0.9890	1.0000	0.6534	0.5933	0.2917	0.2965	0.4995	0.4685
EC_t	0.8958	0.6092	0.6534	1.0000	0.9329	0.5332	0.5426	0.6718	0.8132
R_t	0.9000	0.5474	0.5933	0.9329	1.0000	0.5741	0.6329	0.5883	0.8995
γ_1	0.6220	0.2520	0.2917	0.5332	0.5741	1.0000	0.7936	0.4199	0.6189
γ_2	0.6945	0.2484	0.2965	0.5426	0.6329	0.7936	1.0000	0.4363	0.6331
γ_3	0.7393	0.4524	0.4995	0.6718	0.5883	0.4199	0.4363	1.0000	0.5234
Daily-A_t	0.8527	0.4175	0.4685	0.8132	0.8995	0.6189	0.6331	0.5234	1.0000
Spearman rank correlation									
A_t	1.0000	0.8867	0.8617	0.8103	0.8266	0.3182	0.2208	0.2722	0.9403
BAS_t	0.8867	1.0000	0.9831	0.8938	0.8274	0.2225	0.1256	0.3557	0.8190
Rel-BAS_t	0.8617	0.9831	1.0000	0.8995	0.8343	0.2175	0.1066	0.3552	0.7988
EC_t	0.8103	0.8938	0.8995	1.0000	0.9253	0.2202	0.0690	0.4219	0.8174
R_t	0.8266	0.8274	0.8343	0.9253	1.0000	0.2727	0.1327	0.3633	0.8746
γ_1	0.3182	0.2225	0.2175	0.2202	0.2727	1.0000	0.7452	0.1628	0.3481
γ_2	0.2208	0.1256	0.1066	0.0690	0.1327	0.7452	1.0000	0.1032	0.2122
γ_3	0.2722	0.3557	0.3552	0.4219	0.3633	0.1628	0.1032	1.0000	0.2613
Daily-A_t	0.9403	0.8190	0.7988	0.8174	0.8746	0.3481	0.2122	0.2613	1.0000

Table 4: Correlation matrix for illiquidity measures on a daily basis. Sample: EBS data from 01-Jan-2016 to 17-Jul-2016. A_t: High-frequency Amihud measure, BAS_t: Bid-ask spread, Rel-BAS_t: Relative bid-ask spread, EC_t: effective cost, R_t: Roll measure, γ_1: 5min price impact coefficient,γ_2: order flow price impact coefficient,γ_3: trade-by-trade price impact coefficient, Daily-A_t; classic Amihud measure computed with the absolute value of daily log-return.

Table 4 delivers two main messages: First, it clearly shows that our FX illiquidity measure is

[19]The other main interdealer platform is Thomson Reuters. Some FX rates e.g. involving the British pound are mainly traded on it.

highly correlated with intraday illiquidity proxies based on EBS data, in particular the effective cost and order flow price impact. Second, it is also highly correlated with the traditional Amihud indicator suggesting that even approximating volatility with daily absolute returns (as in the traditional Amihud indicator) rather than gauging it with more accurate high-frequency measures realized power variation, as in our proxy), one can obtain a fairly accurate proxy of FX illiquidity. Spearman rank correlations confirm these results. Overall, we find that our illiquidity measure in the spirit of the Amihud indicator increases with high-frequency and well-accepted measures of FX illiquidity.

4.2.1 A natural experiment

Another method to assess the validity of our illiquidity measure is by means of a meaningful natural experiment. Through the lens of the theory developed in Section 2, the announcement of the cap removal of the Swiss franc by the Swiss National Bank (SNB) on January 15, 2015 represents an ideal natural experiment. Indeed, starting from September 6, 2011, the SNB set a minimum exchange rate of 1.20 francs to the euro (capping franc's appreciation) saying "the value of the franc is a threat to the economy", and that it was "prepared to buy foreign currency in unlimited quantities"'. This means that the SNB had a declared binding *cap* on the transaction price that was removed on January 15, 2015.[20]

In terms of our model, the SNB can be considered as the $(J+1)$-th trader. The SNB intervention strategy of selling CHF for EUR in potentially unlimited quantities is implemented if the average of the reservation prices of the J traders falls below the cap, that is if $\frac{1}{J}\sum_{j=1}^{J} p^*_{i,j} < \log(1.2)$. Indeed, despite the cap on the transaction price, the reservation prices of each individual trader might well be below the 1.20 threshold. For instance, a trader with a reservation price of 1.15, which observes a market price above 1.20, will sell EUR for CHF expecting the cap to be removed at some point in the future.[21] In other words, SNB buys (sells) foreign (domestic) currency to guarantee

[20] The SNB announcement was mostly unanticipated by market participants, see e.g. Jermann, 2017 and Mirkov et al., 2016

[21] The Thomson Reuters survey indicates dispersion of the beliefs of professional market participants around 1.20 along most of the capping period.

that the transaction price is above the threshold, that is

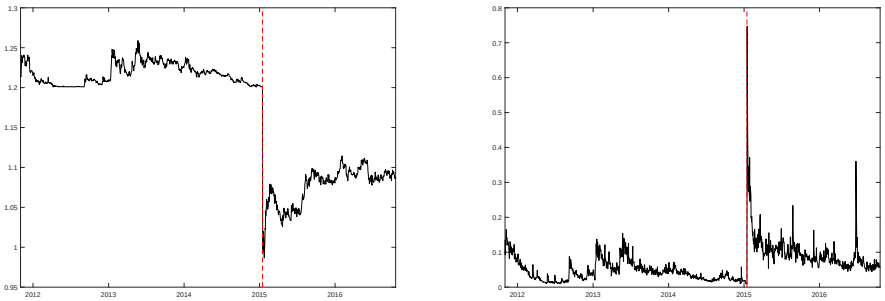
$$p_i = \frac{1}{J+1} \sum_{j=1}^{J+1} p_{i,j}^* \geq \log(1.2), \tag{15}$$

where $p_{i,J+1}^* = (\log(1.2) - \frac{1}{J} \sum_{j=1}^{J} p_{i,j}^*) \mathcal{I}(\sum_{j=1}^{J} p_{i,j}^* < 1.2)$, where $\mathcal{I}(\cdot)$ is the indicator function. The enforcement of the capping regime by SNB generates extra trading volume. In particular, the trading volume is

$$v_i = \frac{\xi^{x|y}}{2} \sum_{j=1}^{J} |\psi_{i,j} - \bar{\psi}_{i,j}| + v_i^{SNB}, \tag{16}$$

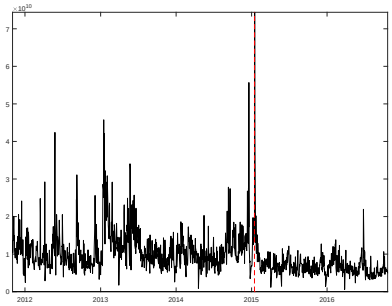
where v_i^{SNB} is the trading volume generated by the central bank to maintain the cap on the FX rate. Hence, the model prescribes a low volatility of the observed returns due to the implicit constraint given by the capping and a larger volume due to FX interventions. This implies that the Amihud illiquidity index is lower (higher) before (after) the removal of the FX capping regime.

Figure 8 provides graphical support for the prescriptions of the theoretical model. Indeed, volatility (realized power variation) is relatively low until January 15, 2015, it spikes on the day of the announcement of the un-capping and it remains high until the end of 2016. The trading volume has the opposite behavior, being relatively high during the capping period and reverting to a lower value after January 15, 2015. Finally, our FX Amihud measure displays a clear upward shift after the removal of the Swiss franc cap. To provide a statistical support, Table 5 reports the sample average of the main market variables before and after the cap removal. After the announcement, FX volatility significantly increases, trading volume decreases, and liquidity dries up (even discarding the announcement day). Furthermore, the average trading volume size significantly decreases, suggesting a reduction in market depth. The lack of statistical significance in the change of the dispersion (standard deviations and high-low ranges) in Thomson Reuters survey of forecasts before and after the announcement suggests that market participants do not disagree more (less) after (before) the currency cap removal. This also suggests that the liquidity dry-up after the cap removal cannot be explained by a stronger consensus regarding agents' reservation prices. All in all, the analysis of this natural experiment corroborates the empirical predictions of the theory, that is, the central bank's enforcement of its reservation price leads to lower volatility, larger trading volume, and higher liquidity. By abandoning this regime, opposite

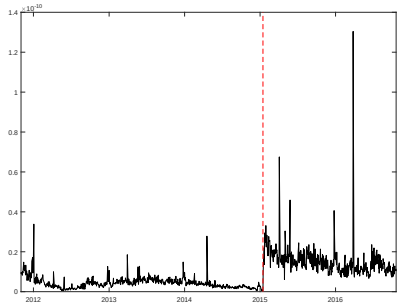


(a) Price

(b) RPV



(c) Volume



(d) Amihud

Figure 8: FX Rate (Figure a), Realized Power Variation (RPV, b), Trading Volume (c) and FX Amihud measure (d) of the EUR/CHF currency pair from 2012 to 2016 with the announcement of the cap removal of the Swiss franc by SNB on January 15, 2015 (red-dashed line).

patterns arise.

	Before	After	Test	p-value
RPV	4.548	9.7531	-12.71	0.000
RV	0.0825	0.3508	-5.842	0.000
VOL	1148.75	682.80	14.34	0.000
AMIHUD	0.4255	1.5002	-24.21	0.000
SIZE	264.84	207.65	27.55	0.000
DIS_1	0.0146	0.0166	-0.1704	0.865
DIS_2	0.0898	0.1015	-0.1252	0.908

Table 5: Sample averages of realized power variation (RPV), realized volatility (RV), trading volume (Volume), FX illiquidity measure (Amihud), and average trade size (Size) before (from Nov 1, 2011 to Jan 14, 2015) and after (from Jan 16, 2015 to Nov 30, 2016) announcement of un-capping (on Jan 15, 2015) - daily frequency. To proxy disagreement, we compute the average of the standard deviations (DIS_1) and high-low range (DIS_2) of monthly Thomson Reuters survey of forecasts on the EUR-CHF rate. The variables have been rescaled. Table also reports a test for the equality of the averages in the two sub-samples, $z = \frac{m_1 - m_2}{\sqrt{v1/n1 + v2/n2}}$, and associated p-values (one-tail) calculated accounting for the auto-correlation in the data.

4.3 Commonalities

Proposition 3 in Section 2 is about commonalities in FX trading volume, volatility and liquidity arising from the no-arbitrage condition. The purpose of this subsection is to empirically assess this idea. More precisely, we proceed in two steps: First, we analyze commonalities by means of three methods: (i) the factor analysis, (ii) the construction of a FX connectedness index, and (iii) the regression analysis. Second, we study the pricing implications stemming from arbitrage deviations and commonalities.

4.3.1 Factor Analysis

By means of the triangular no-arbitrage relation, Section 2.2 provides a theoretical underpinning that trading volume across FX rates are driven by common factors, which are function of the aggregated traders' specific components on different currency pairs. Notice that the FX-rate triangular condition can be extended to more than three FX rates. Actually, it is generalizable to any numbers of FX rates tied by triangular relationships. For instance, with four currencies, x, w, z, and y, the log-price is $p_i^{x|y} = p_i^{x|z} + p_i^{z|w} + p_i^{w|y}$, and the synthetic volume becomes $\widetilde{v}_i^{x|y} = \frac{\xi^{x|y}}{2} \sum_{j=1}^{J} |\psi_{i,j}^{x|z} - \bar{\psi}_i^{x|z} + \psi_{i,j}^{z|w} - \bar{\psi}_i^{z|w} + \psi_{i,j}^{w|y} - \bar{\psi}_i^{w|y}|$. This provides support for the existence of a factor structure in cross sections of FX rates of any order.

To the purpose of studying the commonality in volume, volatility and liquidity across multiple FX rates, we follow the common approach in the literature (e.g. Hasbrouck and Seppi, 2001) and apply the principal component analysis (PCA) to the panel of 29 FX rates introduced in Section 3.1. The goal is to identify a common factor structure across the volume, volatility, and illiquidity series of the FX rates and to study the exposure of each rate to it. Table 6 shows of these quantities for each individual FX rate load positively on the first principal component in all cases. Notably, the first component explains a large portion of the overall variation of volume, volatility and illiquidity measures of the panel of FX rates, being above 50% in many cases. Moreover, the weight associated with the volume and illiquidity measure of USDEUR is the highest signaling the leading role of the information on the USDEUR rate in determining the global FX volume. Instead, the loading on RPV for EURDKK is the smallest across all currencies, signaling that the volatility on EURDKK is strongly influenced by the pegging of DKK to EUR. These findings

	Hourly			Hourly Seasonally Adjusted			Daily		
	Volume	RPV	Amihud	Volume	RPV	Amihud	Volume	RPV	Amihud
AUDJPY	0.1555	0.1884	0.1526	0.2031	0.2143	0.1963	0.1830	0.1986	0.2230
AUDNZD	0.1288	0.1461	0.1418	0.1539	0.1774	0.1559	0.1781	0.1895	0.2195
CADJPY	0.1327	0.1966	0.1045	0.1528	0.2019	0.1133	0.1387	0.1884	0.1153
EURAUD	0.1854	0.1992	0.1852	0.1934	0.2144	0.1823	0.1968	0.2083	0.2277
EURCAD	0.1829	0.2138	0.1709	0.1711	0.2180	0.1627	0.1873	0.2111	0.1835
EURCHF	0.2173	0.1520	0.2065	0.1997	0.1542	0.1852	0.1677	0.1510	0.1871
EURDKK	0.1841	0.0863	0.1819	0.0910	0.0495	0.0879	0.1500	0.0623	0.0111
EURGBP	0.2285	0.2128	0.2419	0.2284	0.2139	0.2355	0.2155	0.2069	0.2508
EURJPY	0.1971	0.1954	0.2110	0.2112	0.1959	0.2367	0.1617	0.1677	0.2401
EURNOK	0.2142	0.1472	0.2217	0.1805	0.1203	0.1729	0.2118	0.1472	0.0612
EURSEK	0.2131	0.1393	0.2179	0.1764	0.1001	0.1571	0.2041	0.1314	0.0535
GBPAUD	0.1599	0.2034	0.1592	0.1672	0.2170	0.1821	0.1865	0.2163	0.2165
GBPCAD	0.1279	0.2045	0.1098	0.1284	0.2047	0.1398	0.1603	0.2067	0.1729
GBPCHF	0.1692	0.2148	0.1595	0.1488	0.2163	0.1648	0.1664	0.2094	0.1882
GBPJPY	0.1823	0.1984	0.1690	0.1754	0.1980	0.1724	0.1463	0.1774	0.1789
USDAUD	0.1839	0.1965	0.1880	0.2256	0.2129	0.2361	0.1951	0.2115	0.2417
USDCAD	0.2070	0.2066	0.2099	0.2144	0.2049	0.2158	0.2106	0.2079	0.2538
USDCHF	0.2236	0.2140	0.2227	0.2301	0.2131	0.2321	0.2184	0.2014	0.2516
USDDKK	0.1573	0.2127	0.1499	0.0979	0.2129	0.1089	0.1394	0.1979	0.1035
USDEUR	0.2320	0.2142	0.2482	0.2461	0.2155	0.2633	0.2011	0.1966	0.2690
USDGBP	0.2291	0.2131	0.2363	0.2384	0.2095	0.2575	0.2235	0.2050	0.2733
USDHKD	0.1555	0.0673	0.1650	0.1266	0.0917	0.1502	0.1515	0.1244	0.0891
USDJPY	0.1748	0.1676	0.1739	0.2263	0.1701	0.2116	0.1912	0.1487	0.2067
USDMXP	0.1459	0.1656	0.1442	0.1881	0.1538	0.1850	0.2211	0.1841	0.1138
USDNOK	0.1909	0.2026	0.1892	0.1518	0.1852	0.1484	0.1611	0.2014	0.0920
USDNZD	0.1669	0.1914	0.1606	0.1908	0.1966	0.1922	0.2003	0.2041	0.2365
USDSEK	0.1939	0.1974	0.1887	0.1585	0.1755	0.1457	0.1737	0.1861	0.0842
USDSGD	0.1528	0.1633	0.1403	0.1887	0.1677	0.1580	0.1671	0.1846	0.0927
USDZAR	0.2179	0.1681	0.2246	0.1989	0.1313	0.1915	0.2233	0.1705	0.0953
EXPL	0.5514	0.6135	0.4519	0.3440	0.5236	0.3066	0.4756	0.6494	0.3771

Table 6: PCA Analysis. The table reports the loadings for each currency pair for trading volume (Volume), volatility (realized power variation, RPV) and illiquidity (Amihud) to the first principal component. The bottom line reports the percentage of explained variance of the first principal component.

remain qualitatively the same for daily and hourly (seasonally un- or adjusted) time series.

Another way to analyze commonalities is by studying the dynamic interplay between the FX rates across currencies by means of the total connectedness index of Diebold and Yilmaz (2014).[22] The TCI is defined as

$$TCI = \frac{1}{N} \sum_{i,j=1 i \neq j}^{N} \tilde{d}_{i,j}, \tag{17}$$

[22] See also Greenwood-Nimmo et al. (2016) for an application of the connecteness measure in the context of returns and option-implied moments of FX rates.

where N denotes the number of variables in the system, and $\tilde{d}_{i,j}$ is the i,j entry of the standardized connectedness matrix $\tilde{D}$. The matrix $\tilde{D}$ is defined as

$$\tilde{d}_{i,j} = \frac{d_{i,j}}{\sum_{j=1}^{N} d_{i,j}}, \tag{18}$$

with

$$d_{i,j} = \frac{\sigma_{jj}^{-1} \sum_{h=0}^{H} (e_i A_h \Sigma e_j)^2}{\sum_{h=0}^{H} (e_i' A_h \Sigma A_h' e_i)}, \tag{19}$$

where A_h is the impulse-response matrix at horizon h associated with a VAR(p) model, Σ is the covariance matrix of the errors, and e_i, e_j are $N \times 1$ selection vectors. By construction, $\sum_{j=1}^{N} \tilde{d}_{i,j} = 1$ and $\sum_{i,j=1}^{N} \tilde{d}_{i,j} = N$. Equation (19) defines the generalized forecast error decomposition, as introduced by Pesaran and Shin (1998). In other words, the TCI measures the average portion over N variables of the forecast error variation of variable i coming from shocks arising from the other $j = 1, \ldots, N-1$ variables of the system. Although less standard in the literature on liquidity commonalities, the TCI approach provides an informative characterization of the connectedness of a system that is richer than the one obtained with a simple linear correlation coefficient. Indeed, the TCI combines information coming from both the contemporaneous and the dynamic dependence structure of the system trough Σ and A_h, respectively. Moreover, by estimating the VAR model over rolling windows, it is possible to characterize the evolution of the dependence structure between two or more variables by looking at the variations of the TCI over time.

As showed in Table 7, the connectedness analysis delivers two main findings: First, the overall level of connectedness of volume and volatility is very high and constant over time, being close to 90% for both volatility and volume at hourly and daily level. The connectedness remains very high also when volume and RPV are filtered from intradaily seasonality, being around 70%-80%. This picture corroborates the previous findings obtained from the Factor Analysis, that is, there is a strong commonality across FX volumes and volatilities. Second, the comparison between the most and least liquid FX rates indicates that a stronger connectedness of volume and volatility for the former set of currencies. Indeed, the connectedness on the most liquid FX rates is above 85% and it remains relatively high for hourly seasonally adjusted series. On the other hand, the connectedness level sensibly reduces when focusing on the least liquid FX rates. This result is

	Hourly				Hourly Seasonally Adjusted				Daily			
	Full	11/14	12/15	13/16	Full	11/14	12/15	13/16	Full	11/14	12/15	13/16
All FX rates												
Volume	0.884	0.880	0.885	0.890	0.726	0.719	0.730	0.731	0.891	0.889	0.891	0.898
RPV	0.910	0.907	0.910	0.916	0.846	0.844	0.850	0.856	0.921	0.920	0.928	0.930
10 Most Liquid												
Volume	0.875	0.873	0.883	0.880	0.709	0.702	0.714	0.722	0.862	0.864	0.863	0.870
RPV	0.893	0.890	0.893	0.904	0.814	0.815	0.818	0.838	0.919	0.920	0.922	0.935
10 Least Liquid												
Volume	0.621	0.607	0.634	0.649	0.275	0.270	0.284	0.289	0.623	0.608	0.617	0.659
RPV	0.808	0.810	0.821	0.819	0.718	0.710	0.728	0.732	0.846	0.829	0.861	0.873

Table 7: Connectedness. The table reports the value of the connectedness index of Diebold and Yilmaz (2014) of trading volume (Volume) and volatility (in terms of realized power variation, RPV) for different sampling periods (Full sample, 2011/2014, 2012/2015, and 2013/2016) and for different sets of FX rates. "10 Most Liquid" and "10 Least Liquid" refer to the ten most and least liquid FX rates in terms of total trading volume.

fully consistent the adage that "liquidity begets liquidity" (e.g. Foucault et al., 2013), in the sense that higher liquidity goes with stronger commonality. It is also consistent with the Proposition 3 in Section 2 in which FX rates are connected by arbitrage trading volume and this connection is stronger for liquid currencies (captured by the term $\xi^{x|y}$ in (10)). This result squares well with the idea that illiquid currency pairs are less (more) exposed to the common (specific) FX-factors as it emerges from the magnitude of the loadings of the first principal component in Table 6. In sum, liquid currencies appear to have stronger cross-currency commonalities than illiquid ones.

4.3.2 Measuring Commonalities

Common measures of liquidity commonalities are statistical measures such as R^2 or estimated slope coefficient when regressing liquidity of an asset on market liquidity (e.g. Chordia et al., 2000). Following the same reasoning but to be consistent with the arbitrage framework theorized in Section 2.2, we measure the strength of the pairwise commonality in volume between $x|y$, $x|z$ and $z|y$ through the following reduced-form model,

$$\log(v_t^{x|y}) = \beta_0 + \beta_1 \log\left(v_t^{x|z} + v_t^{z|y}\right) + \varepsilon_t, \quad t = 1, \ldots, T \tag{20}$$

where $v_t^{x|y}$, $v_t^{x|z}$ and $v_t^{z|y}$ are the log-volume on period t on the FX rates $x|y$, $x|z$ and $z|y$, respectively. In this regression, β_0 reflects the differential in the resiliency levels in the three markets,

while β_1 measures the magnitude of commonality in the volume of the three FX rates. The MDH theory outlined in Section 3 prescribes that $\beta_1 > 0$. The term ε_t can also be interpreted as the deviation from the long-run equilibrium between FX volume. Table 8 reports the estimates of regression (20) for the EUR/USD rate, where the aggregate volume combining $v_t^{x|z}$ and $v_t^{z|y}$ (*synthetic* volume) supports the triangular arbitrage with CHF, GBP, DKK, JPY, AUD, CAD, NOK, SEK.[23]

	Hourly			Daily			Daily (interacted)			
	β_0^H	β_1^H	R_H^2	β_0^D	β_1^D	R_D^2	β_0^i	β_1^i	β_2^i	R_i^2
Volume										
CHF	6.2791^a	0.7861^a	0.8068	9.0985^a	0.6973^a	0.5305	9.2331^a	0.6908^a	0.0246	0.5329
GBP	3.8371^a	0.8625^a	0.8195	7.8218^a	0.7209^a	0.4134	8.2372^a	0.6966^a	0.2057^a	0.4341
DKK	15.592^a	0.2962^a	0.5526	16.073^a	0.4452^a	0.3751	16.248^a	0.4293^a	0.2755^a	0.4054
JPY	2.8344^a	0.8799^a	0.4632	15.7280^a	0.3963^a	0.2099	17.185^a	0.3277^a	0.3040^a	0.2467
AUD	0.7486^a	1.0096^a	0.5051	7.1182^a	0.7599^a	0.5789	7.6327^a	0.7304^a	0.1926^a	0.6189
CAD	5.6124^a	0.7936^a	0.7248	10.634^a	0.6176^a	0.3875	11.372^a	0.5724^a	0.3874^a	0.4756
NOK	13.577^a	0.4635^a	0.6739	14.887^a	0.4782^a	0.2482	14.891^a	0.4781^a	-0.0014	0.2482
SEK	13.388^a	0.4700^a	0.6778	14.220^a	0.5051^a	0.2580	14.226^a	0.5050^a	-0.0029	0.2581
RPV										
CHF	-0.9108^a	0.9189^a	0.7752	-0.9071^a	0.7412^a	0.7091	-0.9387^a	0.7482^a	-0.8456^a	0.7130
GBP	-0.9777^a	0.9303^a	0.7552	-0.7973^a	0.8884^a	0.6257	-1.0890^a	0.9725^a	-8.0796^a	0.6892
DKK	0.2941^a	1.094^a	0.9320	0.0646^a	1.1223^a	0.9620	0.1065^a	1.1145^a	1.0515^b	0.9646
JPY	-1.2987^a	0.9156^a	0.5717	-1.2604^a	0.6790^a	0.3995	-1.6528^a	0.8527^a	-14.235^a	0.5317
AUD	-0.8686^a	1.0431^a	0.5542	-1.0562^a	0.9334^a	0.6033	-1.3224^a	1.0314^a	-8.0224^a	0.7045
CAD	-0.7850^a	0.9968^a	0.7238	-0.7546^a	0.9889^a	0.6680	-1.1531^a	1.0427^a	-9.3480^a	0.7634
NOK	-1.7326^a	0.8126^a	0.5366	-1.0991^a	0.8241^a	0.4756	-0.9015^a	0.7945^a	3.7473^a	0.4988
SEK	-1.3851^a	0.8718^a	0.5678	-0.7150^a	1.0555^a	0.5693	-0.5284^a	1.0459^a	3.2424^c	0.5845
Amihud										
CHF	-14.066^a	0.5552^a	0.6956	-14.4250^a	0.5417^a	0.7234	-14.0150^a	0.5570^a	0.0345^c	0.7277
GBP	-12.749^a	0.5985^a	0.7241	-8.7892^a	0.7538^a	0.7489	-9.9900^a	0.7117^a	-0.1495^a	0.7584
DKK	-24.583^a	0.1506^a	0.2566	-23.235^a	0.2086^a	0.2375	-23.244^a	0.2083^a	-0.0069	0.2375
JPY	-8.8037^a	0.7518^a	0.5280	-11.483^a	0.6518^a	0.5832	-12.621^a	0.6167^a	-0.2805^a	0.6191
AUD	-14.852^a	0.5419^a	0.3830	-13.826^a	0.5931^a	0.5401	-15.229^a	0.5418^a	-0.1812^a	0.5721
CAD	-18.857^a	0.3836^a	0.3733	-13.235^a	0.6319^a	0.5111	-14.574^a	0.5829^a	-0.2284^a	0.5382
NOK	-22.516^a	0.2321^a	0.4465	-17.387^a	0.4554^a	0.3539	-18.241^a	0.4210^a	-0.0485^b	0.3602
SEK	-22.151^a	0.2452^a	0.4510	-19.111^a	0.3772^a	0.2138	-20.320^a	0.3295^a	-0.0885^b	0.2304

Table 8: Commonalities in volume, volatility (realized power variation, RPV) and illiquidity (Amihud index, Amihud). For each currency, the table reports the intercept, slope and R^2 of the regression of the log volume/volatility/Amihud of EURUSD on the log of the sum of volume/volatility/Amihud index on the FX rate of the currency indicated in the first column against USD and EUR. The superscripts a, b and c indicate significance at 1%, 5% and 10% significance level, respectively.

[23] Besides these 8 FX rates providing triangular constructions with the EUR/USD rate, in our sample the following synthetic FX rates exist: (a) for the USDGBP, via AUD, CAD, CHF, EUR, and JPY; (b) for USDAUD, via EUR, GBP, JPY, and NZD; and (c) for EURCHF, via GBP and USD. We have analyzed all of them obtaining consistent results.

Overall, it emerges that regression (20) is able to explain a large portion of variability of $v^{EUR/USD}$, and this can be attributed to the portions of common information in $\psi_j^{USD/\cdot}$ and $\psi_j^{EUR/\cdot}$, which determine the synthetic volume in (35). At the hourly level, the estimated parameter β_0 reflects the average liquidity differential across currencies, with DKK, SEK and NOK being consistently less liquid than JPY, AUD and GBP. Notably, the parameter β_1 is positive in all cases and it is closer to 1 for the most liquid rates corroborating the idea that liquidity begets commonality. As expected, higher β_1 are associated with higher R^2. When removing the intradaily seasonality in volume or aggregating at the daily level, the R^2 slightly decreases but the result is qualitatively the same as for the raw hourly volume. The residuals display significant autocorrelation, suggesting that volume imbalances across FX markets are stationary but persistent. These long-lasting disequilibria in volume might be explained by the fragmented OTC structure of the FX market and prolonged time to incorporate agents' heterogeneous priors and (public and private) information into prices, as for conditional volatility (Engle et al., 1990).

When replacing volume with volatility (RPV) in (20), we note that also volatility displays a large degree of commonality across currencies. The R^2 is generally very well above 50% at both hourly and daily level. Interestingly, the R^2 and the slope coefficient of DKK are almost 1 consistent with the Danish Central Bank policy to keep EUR/DKK within a very narrow corridor (0.133-0.1346), thus the $Cov(p^{USD/DKK}, p^{USD/EUR}) \approx 1$. Consistent with the theory, the Danish central bank's intervention to fix the EUR/DKK rate reduces the commonality in volume and liquidity with the other currencies. Not surprisingly, the Amihud illiquidity measure, which combines information on both volatility and volume, also displays an analogous amount of commonality across currencies, being the highest for the most liquid ones.

The theory outlined in Section 2.2 suggests that the commonalities in trading volume across FX rates are driven by the level of correlation among the FX rates, where the synthetic volume is a function of the correlation of the aggregated traders' specific components on different currency pairs, see the the right-hand side of (35) in Appendix A. In other words, our theory predicts that the synthetic volume reveals the strength of the correlation across FX rates. To test this empirical prediction, we consider the following regression

$$\log(\tilde{v}_t^{x|y}) = \gamma_0 + \gamma_1 \log(\zeta_t) + \gamma_2 \tilde{v}_{t-1}^{x|y} + \varepsilon_t, \quad t = 1, \dots, T, \tag{21}$$

where (log) $\tilde{v}_t^{x|y}$ is the synthetic volume as measured by the fitted volume in regression (20), while $\zeta_t = \log(1 + |\rho_t|)$ and ρ_t is the realized correlation between $x|z$ and $z|y$. Hence, the term ζ_t measures the strength of the correlation in the FX rates $x|z$ and $z|y$, and the parameter γ_1 is expected to be positive. Table 9 contains the estimates of γ_1 based on regression (21) and on the extended version which controls for liquidity as measured by the bid-ask spreads on $x|z$ and $z|y$. At hourly frequency, the estimates of γ_1 are positive and highly significant in most cases, with the

	Hourly		Daily		Weekly	
	γ_0	γ_1	γ_0	γ_1	γ_0	γ_1
Baseline Regression						
CHF	4.2444	0.2141^a	13.9536	0.1697^a	8.4427	0.0993^c
GBP	3.7079	0.2310^a	18.2764	0.0897^c	12.1883	0.0485
DKK	6.8630	-0.1447^a	18.5609	-0.1159	11.2898	0.0615
JPY	6.1447	0.2810^a	10.9029	0.1511^a	7.0136	0.0331
AUD	6.8131	0.0642^a	12.5726	0.0609	8.4234	-0.0696
CAD	4.3665	-0.0764^a	22.5847	-0.0203	18.2935	0.0079
NOK	6.0960	0.7085^a	17.2974	0.4519^a	17.6356	0.1196^c
SEK	5.6549	0.5923^a	18.6241	0.1409^a	15.6949	-0.0160
Control for Liquidity						
CHF	5.3883	0.3172^a	16.6392	0.4082^a	12.1837	0.2538^b
GBP	4.8816	0.2439^a	18.7933	0.1463^a	12.9884	0.0658
DKK	7.0838	-0.1499^a	19.4777	-0.1202^a	13.8504	0.0317
JPY	7.1983	0.1077^a	14.6026	-0.0455	10.9355	-0.0353
AUD	7.6219	0.1614^a	17.4960	0.2028^a	16.9214	0.0394
CAD	4.5109	-0.0853^a	22.7906	0.0288	19.6981	0.0878
NOK	8.0024	0.7449^a	17.3807	0.4529^a	17.6188	0.1140^c
SEK	8.2093	0.6576^a	18.9913	0.1571^a	16.0763	-0.0129

Table 9: Synthetic volume and correlation. For each currency, the table reports the intercept and the slope of the regression of the log synthetic volume of EURUSD on the log of the correlation of the FX rates with USD and EUR.

notable exception of DKK. Again, the results suggest that the intervention of the central bank to peg DKK to EUR prevents the trading activity on EUR/DKK and DKK/USD from fully revealing the correlation structure of the investors' beliefs on EUR and USD. When aggregating over days and weeks, we still obtain generally positive estimates of γ_1 but they are often not significantly different from zero.

4.3.3 Commonality and Pricing Implications

One of our previous results is that liquidity begets liquidity across currencies. As the last step of our study, we address the question whether liquidity begets price efficiency as well. The rationale of this relationship is again our third theoretical proposition implying that arbitrage keeping FX rates tied to equilibrium relations passes through the trading activity (volume), which in turn it is sustained by liquidity. To do this, we build a simple measure of pricing errors to investigate whether high liquidity is associated with smaller *mispricing* errors. Specifically, $pe_{i,t}$ is the hourly cumulative no-arbitrage error at time time t for the i-th synthetic relation defined as

$$pe_{i,t} = \sum_{l=1}^{60} |r_{l,t}^{x|y} - \tilde{r}_{l,t}^{z_{n_i}}|,$$

where $r_{l,t}^{x|y}$ is the *direct* one-minute midquote log-return on the FX rate between the currency x and y, while $\tilde{r}_{l,t}^{z}$ is the *synthetic* one-minute log-return on the FX rate $x|y$ using the currency z_{n_i}. Empirically, we test the price-liquidity relation in two ways: First, by looking at the systematic relationship between arbitrage deviations and illiquidity; Second, by inspecting whether more liquidity facilitates the price adjustment process.

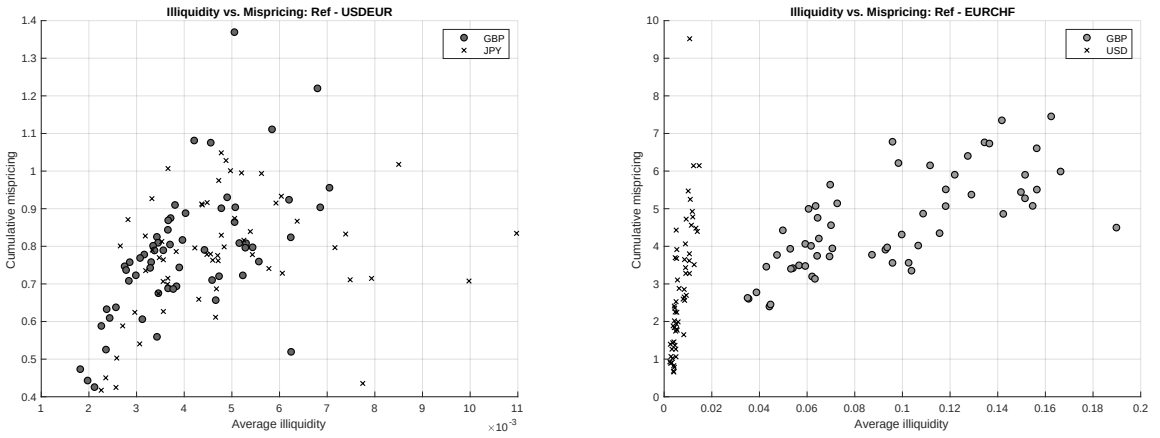
To analyze the systematic price-liquidity relationships, we apply two methods: First, we extend the previous commonality analysis in (20) by interacting synthetic volume and pricing error as follows:

$$\log(v_t^{x|y}) = \beta_0 + \beta_1 \log\left(v_t^{x|z} + v_t^{z|y}\right) + \beta_2 \log\left(v_t^{x|z} + v_t^{z|y}\right) pe_{i,t} + \varepsilon_t, \quad t = 1, \ldots, T \tag{22}$$

The results are showed on the right-hand side of Table 8. As predicted by our theory, we find a positive β_2 indicating that arbitrage deviations attract more trading volume to reestablish price equilibrium. We extend the analysis to volatility and illiquidity, for which we do not have clear empirical predictions. In both cases, we find a negative β_2 suggesting that the departure from arbitrage conditions goes with divergent liquidity and volatility patterns across currencies, consistent with the idea that illiquidity hinders the restoration of equilibrium prices.

The second method to study the systematic price-liquidity relationship is to compute the monthly average mispricing errors and *synthetic* illiquidity for each FX rate allowing for a tri-

angular FX construction. For instance, for EUR/USD, EUR/GBP, and GBP/USD we calculate the average deviations between direct (EUR/USD) and synthetic rate (via EUR/GBP, and GBP/USD) and the average Amihud measures of the two FX rates to operate triangular arbitrage (EUR/GBP and GBP/USD).



(a) Cumulative mispricing against illiquidity: USDEUR (b) Cumulative mispricing against illiquidity: EURCHF

Figure 9: Monthly cumulative mispricing ($pe_t^{x|y}$) against synthetic illiquidity, $\widetilde{A}_t^{x|y}$.

Figure 9 clearly shows a positive relationship between mispricing and illiquidity. Also, more liquid currencies have steeper curves suggesting that the same amount of additional liquidity is more effective in reducing arbitrage deviations in liquid currencies. We also carry out a statistical analysis to validate these findings and consider the following regression

$$pe_t^{x|y} = \alpha + \delta \widetilde{A}_t^{x|y} + \gamma \widetilde{BAS}_t + \varepsilon_t, \tag{23}$$

where $\tilde{Am}_t^{x|y}$ denotes the *synthetic* illiquidity on the FX rate $x|y$ computed with the same currency used to calculate $pe_t^{x|y}$. We expect the parameter δ to be positive and significant, signaling a positive relation between illiquidity and pricing errors. Analogously, the *synthetic* bid-ask spread, denoted as $\widetilde{BAS}$, is also computed in a similar way and it is added to the regression to control for deviations from the pricing equilibrium due another dimension of illiquidity that is the bid-ask spread. The results of regression (23) are reported Table 10, and the parameter estimates validate the findings observed in the scatter plots in Figure 9. In particular, by regressing monthly mispricing on (synthetic) FX illiquidity, we find compelling evidence that liquidity begets price

efficiency, i.e. limiting arbitrage deviations. This holds true also when controlling for bid-ask spread differentials, although the significance is reduced for EURUSD when combining with the least liquid currencies (e.g. NOK and SEK).

	EURUSD								EURCHF	
	CHF	**GBP**	**DKK**	**JPY**	**AUD**	**CAD**	**NOK**	**SEK**	**USD**	**GBP**
α	0.36^a	0.50^a	0.49^a	0.62^a	0.51^a	0.26^a	0.51^a	0.72^a	0.25	2.51^a
δ	54.77^a	65.12^a	1.01^a	34.06^b	21.54^a	17.37^a	16.15^a	9.07^a	357.9^a	-22.13^a
R^2	0.54	0.32	0.14	0.11	0.34	0.40	0.36	0.13	0.65	0.49
α	-0.30^a	0.04	0.59^a	0.14^b	0.13	-0.27^b	-0.10	-0.18^b	-2.48^a	-0.78^c
δ	23.21^a	11.15	1.27^a	30.77^a	10.18^b	2.95	2.92	-1.52	227.7^a	16.21^a
γ	30.05^a	31.25^a	-0.98	0.20^a	14.55^a	26.27^a	2.97^a	3.51^a	124.0^a	93.64^a
R^2	0.88	0.64	0.18	0.47	0.46	0.69	0.76	0.71	0.82	0.69

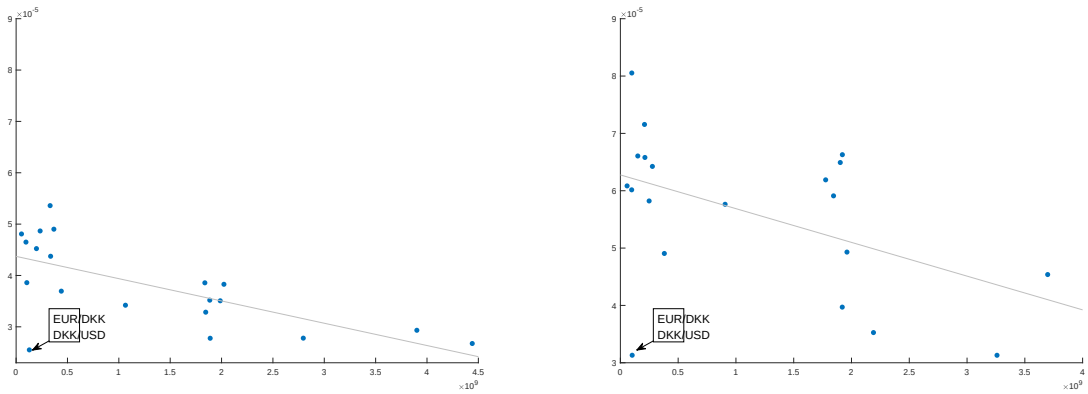
Table 10: Mispricing vs. Liquidity regression estimation. Table reports the estimates of the linear regression (23) for the FX rates EURUSD and EURCHF, when the triangular no-arbitrage condition is computed with of a third currency, that is CHF, GBP DKK, JPY, AUD, NOK and SEK for EURUSD; USD and GBP for EURCHF. The sample size is $N = 58$ months. The top (bottom) panel reports the estimates when *BAS* is excluded (included) among regressors in (23). The superscripts a, b and c indicate significance at 1%, 5% and 10% significance level respectively.

To study whether liquidity facilitates price adjustments, we benefit again from the identification of large price co-movements captured by the co-jumps. More specifically, we test whether the chances of mispricing are higher for less liquid currencies in reaction to directional FX movements measured by co-jumps. To carry out this test, we consider the following panel regression with fixed effects

$$pe_{i,t} = \alpha_i + \beta \bar{V}_{i,t}(1 + \zeta\ CJ_t) + \theta CJ_t + \delta BA_{i,t} + \gamma_h h_t + \gamma_w w_t + \varepsilon_{i,t}, \tag{24}$$

The term $\bar{V}_{i,t}$ is the aggregate or synthetic volume from the FX rates $x|z_{n_i}$ and $z_{n_i}|y$. Our sample consists of $n = 10$ currencies and allows us to consider $I = 20$ combinations of x, y and z_{n_i}.[24] The relation between the average volume $\bar{V}_{i,t}$ and the average no-arbitrage error $pe_{i,t}$ is depicted in Figure 10. The figure clearly displays a cross-sectional negative relation between the trading volume on the FX rates and the no-arbitrage pricing error. In other words, the pricing errors are higher for less liquid currencies, such as SEK and NOK. A notable exception is given by

[24]The combinations are: USDAUD/EURAUD, USDSEK/EURSEK, USDNOK/EURNOK, USDCHF/EURCHF, USDCAD/EURCAD, USDJPY/EURJPY, USDGBP/EURGBP, USDDKK/EURDKK, USDAUD/GBPAUD, USDCAD/GBPCAD, USDJPY/GBPJPY, USDCAD/JPYCAD, USDAUD/JPYAUD, EURCAD/GBPCAD, EURJPY/GBPJPY, EURCHF/GBPCHF, EURCAD/JPYCAD, USDAUD/JPYAUD, GBPAUD/JPYAUD, GBPCAD/JPYCAD.



(a) Unconditional (b) Co-Jump

Figure 10: Trading Volume and pricing errors. The figures show the scatter of the average volume (x-axis) versus the average triangular pricing error (y-axis) for 20 combination of currencies x, y and z. The left panel reports the unconditional relation, while the right panel is conditional to the event of a co-jump on the individual currencies, EUR, JPY, USD and GBP. The line represents the least squares fit.

DKK, which again it can be explained by the fixed exchange rate policy. When conditioning on the arrival of a large common news on the main individual currencies EUR, JPY, USD and GBP (right panel), the average mispricing error on the y-axis increases relatively to the left panel, suggesting that big news arrivals prompt price adjustment processes on individual currencies that can generate larger price dispersion and mispricing errors. However, the negative relation between magnitude of the mispricing and trading volume is maintained.

Table 11 reports the parameter estimates of (24) based on the sample of $I = 20$ combination of FX rates and for a sample of $T = 30720$ hours (24×1280 days). The results confirm our empirical prediction, that is, a negative relation between mispricing errors and volume, which is robust to the inclusion of the relative bid-ask spread as a control for transaction costs (where parameter δ is found significantly positive in all cases). As it also emerges from Figure 10, the co-jumps events are associated with a significant increase in the average level of mispricing ($\theta > 0$), and also with a significantly negative slope of volume ($\zeta < 0$). In sum, our results support the idea that liquidity begets price efficiency by reducing pricing errors, systematically and facilitating the information processing.

	FE	PO	FE	PO	FE	PO	FE	PO	FE	PO
Volume	-0.018^a	-0.028^a	-0.024^a	-0.028^a	-0.024^a	-0.028^a	0.004^a	-0.007^a	0.004^a	-0.007^a
Bid-Ask	–	–	0.142^a	0.210^a	0.141^a	0.209^a	-0.002	0.073^b	-0.002	0.073^b
CJ	–	–	–	–	0.001^a	0.001^a	0.001^a	0.001^a	0.001^a	0.001^a
CJ-Volume	–	–	–	–	–	–	–	–	-0.041^a	-0.046^a
Daily	no	no	no	no	no	no	yes	yes	yes	yes
Weekly	no	no	no	no	no	no	yes	yes	yes	yes
AR(1)	no	no	no	no	no	no	yes	yes	yes	yes

Table 11: No-arbitrage pricing errors and volume. Panel regression with fixed effect (FE) and pooling (PO). The dependent variable is the triangular pricing error accumulated at the hourly horizon for 20 combinations of FX rates. The regressors are the hourly aggregate (synthetic) trading volume of the two indirect FX rates (Volume) of the triangular arbitrage, the average relative bid-ask spread (Bid-Ask) of the direct FX rate, the dummy variable of the co-jump index on its own (CJ) and interacted with (synthetic) trading volume (CJ-Volume) as well as hourly and weekly dummies. The superscripts a, b and c indicate significance at 1%, 5% and 10% significance level respectively. The standard errors are computed with the White (1980) sandwich estimator for panel data models.

5 Conclusion

We provide a unified model for asset prices, trading volume, and volatility. The model is built in continuous-time and allows for multi-asset framework. We apply it to currency markets in which foreign exchange (FX) rates are tied by arbitrage conditions. Our model outlines new properties of the FX market including the relationships between trading volume and volatility of direct and arbitrage-related (or synthetic) FX rates. It also provides a theoretical foundation for common patterns (commonality) of trading volume, volatility, and illiquidity across currencies and time, and an intuitive closed-form solution for measuring illiquidity in the spirit of Amihud (2002).

We test the empirical predictions from our model using new and unique (intraday) data representative of the global FX spot market. A distinguishing characteristic of our data set is that it includes granular and intraday data on global FX trading volume. As predicted by our model, three main empirical findings arise: First, the difference in market participants' beliefs (disagreement) is the common source of trading volume and volatility. Second, our FX Amihud measure is effective in gauging FX illiquidity. Third, we find strong commonalities in FX volume, volatility, and illiquidity across time and FX rates. Consistent with the adage that "liquidity begets liquidity", we find that more liquid currencies reveal stronger commonality in liquidity. Furthermore, we find that liquidity begets price efficiency, in the sense that more liquid currencies obey more to the triangular arbitrage condition.

Several implications emerge from our study. First, by shedding light on the intricate interrelations between FX rates, volume, and volatility, our work should support an integrated analysis of FX rate evolution and risk. Our work also offers a straightforward method to measure FX illiquidity and commonality. For investors, these insights should increase the efficiency of trading and risk analysis. For policy makers, our work highlights the developments of FX global volume, volatility, and illiquidity across time and currencies, which can be important for the implementation of monetary policy and financial stability.

References

Acharya, V. V. and Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, 77:375–410.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5:31–56.

Andersen, T. G. (1996). Return volatility and trading volume: An information flow interpretation of stochastic volatility. *The Journal of Finance*, 51(1):169–204.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Vega, C. (2007). Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics*, 73(2):251–277.

Bandi, F. M. and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *The Review of Economic Studies*, 75(2):339–369.

Bank of International Settlements (2016). Foreign exchange and derivatives market activity in April 2016. Triennial Central Bank Survey.

Barndorff-Nielsen, O. E. and Shephard, N. (2002a). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B*, 64:253–280.

Barndorff-Nielsen, O. E. and Shephard, N. (2002b). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, 17(5):457–477.

Barndorff-Nielsen, O. E. and Shephard, N. (2003). Realized power variation and stochastic volatility models. *Bernoulli*, 9(2):243–265.

Bauwens, L., Rime, D., and Sucarrat, G. (2006). Exchange rate volatility and the mixture of distribution hypothesis. *Empirical Economics*, 30(4):889–911.

Beber, A., Breedon, F., and Buraschi, A. (2010). Differences in beliefs and currency risk premiums. *Journal of Financial Economics*, 98(3):415 – 438.

Bech, M. (2012). FX volume during the financial crisis and now. Staff report, BIS Quarterly Review.

Berger, D. W., Chaboud, A. P., Chernenko, S. V., Howorka, E., and Wright, J. H. (2008). Order flow and exchange rate dynamics in electronic brokerage system data. *Journal of International Economics*, 75:93–109.

Bessembinder, H. (1994). Bid-ask spreads in the interbank foreign exchange markets. *Journal of Financial Economics*, 35:317–348.

Bjønnes, G. H. and Rime, D. (2005). Dealer behavior and trading systems in foreign exchange markets. *Journal of Financial Economics*, 75:571–605.

Bjønnes, G. H., Rime, D., and Solheim, H. O. A. (2003). Volume and volatility in the FX-market: Does it matter who you are? CESifo Working Paper Series 786, CESifo Group Munich.

Bollerslev, T., Li, J., Xue, Y., et al. (2016). Volume, volatility and public news announcements. *The Review of Economic Studies*, 85(4):2005–2041.

Bollerslev, T. and Melvin, M. (1994). Bid-ask spreads and volatility in the foreign exchange market. *Journal of International Economics*, 36:355–372.

Brandt, M. W. and Diebold, F. X. (2006). A no-arbitrage approach to range-based estimation of return covariances and correlations. *The Journal of Business*, 79(1):61–74.

Breedon, F. and Ranaldo, A. (2013). Intraday patterns in FX returns and order flow. *Journal of Money, Credit and Banking*, 45(5):953–965.

Caporin, M., Kolokolov, A., and Renò, R. (2017). Systemic co-jumps. *Journal of Financial Economics*, 126(3):563 – 591.

Cespa, G. and Foucault, T. (2014). Illiquidity contagion and liquidity crashes. *Review of Financial Studies*, (6):1615–1660.

Chaboud, A. P., Chernenko, S. V., and Wright, J. H. (2007). Trading activity and exchange rates in high-frequency ebs data. International Finance Discussion Papers, Board of Governors of the Reserve System and Harvard University.

Chordia, T., Roll, R., and Subrahmanyam, A. (2000). Commonality in liquidity. *Journal of Finance*, 52:3–28.

Chordia, T., Roll, R., and Subrahmanyam, A. (2001). Market liquidity and trading activity. *Journal of Finance*, 56:501–530.

Christensen, K., Oomen, R. C., and Renò, R. (2016). The drift burst hypothesis. Technical report, CREATES WP Series.

Christiansen, C., Ranaldo, A., and Söderlind, P. (2011). The time-varying systematic risk of carry trade strategies. *Journal of Financial and Quantitative Analysis*, 46:1107–1125.

Clark, P. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–55.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7:174–196.

Darolles, S., Le Fol, G., and Mero, G. (2015). Measuring the liquidity part of volume. *Journal of Banking & Finance*, 50:92–105.

Darolles, S., Le Fol, G., and Mero, G. (2017). Mixture of distribution hypothesis: Analyzing daily liquidity frictions and information flows. *Journal of Econometrics*, 201(2):367–383.

Diebold, F. X. and Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Ding, D. K. (1999). The determinants of bid-ask spreads in the foreign exchange futures market: A microstructure analysis. *Journal of Futures Markets*, 19:307–324.

Engle, R. F., Ito, T., and Lin, W.-L. (1990). Meteor showers or heat waves? heteroskedastic intra-daily volatility in the foreign exchange market. *Econometrica*, 58(3):525–542.

Evans, M. D. (2002). FX trading and exchange rate dynamics. *Journal of Finance*, 57:2405–2447.

Evans, M. D. (2018). Forex trading and the WMR fix. *Journal of Banking and Finance*, 87:233 – 247.

Evans, M. D. and Lyons, R. K. (2002). Order flow and exchange rate dynamics. *Journal of Political Economy*, 110:170–180.

Evans, M. D. and Rime, D. (2016). Order flow information and spot rate dynamics. *Journal of International Money and Finance*, 69(1):45–68.

Fischer, A. and Ranaldo, A. (2011). Does FOMC news increase global FX trading? *Journal of Banking and Finance*, 35:2965–2973.

Foucault, T., Pagano, M., and Röell, A. (2013). *Market Liquidity: Theory, Evidence, and Policy*. Oxford University Press.

Frömmel, M., Mende, A., and Menkhoff, L. (2008). Order flows, news, and exchange rate volatility. *Journal of International Money and Finance*, 27(6):994–1012.

Galati, G., Heath, A., and McGuire, P. (2007). Evidence of carry trade activity. *BIS Quarterly Review*, September:27–41.

Gargano, A., Riddiough, S. J., and Sarno, L. (2019). Foreign exchange volume. Working paper.

Grammatikos, T. and Saunders, A. (1986). Futures price variability: A test of maturity and volume effects. *Journal of Business*, 59:319–330.

Greenwood-Nimmo, M., Nguyen, V. H., and Rafferty, B. (2016). Risk and return spillovers among the g10 currencies. *Journal of Financial Markets*, 31:43–62.

Grossman, S. J. and Miller, M. H. (1988). Liquidity and market structure. *Journal of Finance*, 43(3):617–633.

Hartmann, P. (1999). Trading volumes and transaction costs in the foreign exchange market: Evidence from daily dollar-yen spot data. *Journal of Banking & Finance*, 23:801–824.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *The Journal of Finance*, 64(3):1445–1477.

Hasbrouck, J. and Levich, R. M. (2017). FX market metrics: New findings based on CLS bank settlement data. Working paper.

Hasbrouck, J. and Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, 59:383–411.

Hsieh, D. A. and Kleidon, A. W. (1996). Bid-ask spreads in foreign exchange markets: Implications for models of asymmetric information. In Frankel, J. A., Galli, G., and A., G., editors, *The Microstructure of Foreign Exchange Markets*, pages 41–67. The University of Chicago Press, Chicago.

Huang, R. D. and Masulis, R. W. (1999). FX spreads and dealer competition across the 24-hour trading day. *Review of Financial Studies*, 12:61–93.

Jermann, U. J. (2017). Financial markets' views about the euro-swiss franc floor. *Journal of Money, Credit and Banking*, 49:553–565.

Karnaukh, N., Ranaldo, A., and Söderlind, P. (2015). Understanding FX liquidity. *Review of Financial Studies*, 28(11):3073–3108.

Karolyi, G. A., Lee, K.-H., and Dijk, M. A. V. (2012). Understanding commonality in liquidity around the world. *Journal of Financial Economics*, 105:82–112.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis*, 22(1):109–126.

King, M. R., Osler, C. L., and Rime, D. (2012). Foreign exchange market structure, players, and evolution. In James, J., Marsh, J. W., and Sarno, L., editors, *Handbook of Exchange Rates*. Wiley & Sons, Hoboken, New Jersey.

Kondor, P. and Vayanos, D. (2018). Liquidity risk and the dynamics of arbitrage capital. *Journal of Finance*, forthcoming.

Kyle, A. S. and Xiong, W. (2001). Contagion as a wealth effect. *Journal of Finance*, 56:1401–1440.

Lee, S. S. (2011). Jumps and information flow in financial markets. *The Review of Financial Studies*, 25(2):439–479.

Levich, R. M. (2012). FX counterparty risk and trading activity in currency forward and futures markets. *Review of Financial Economics*, 21:102–110.

Li, J., Todorov, V., and Tauchen, G. (2017). Jump regressions. *Econometrica*, 85(1):173–195.

Li, J. and Xiu, D. (2016). Generalized method of integrated moments for high-frequency data. *Econometrica*, 84(4):1613–1633.

Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5-minute RV? a comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1):293–311.

Lyons, R. K. (1997). A simultaneous trade model of the foreign exchange hot potato. *Journal of International Economics*, 42:275–298.

Lyons, R. K. (2001). *The microstructure approach to exchange rates*. MIT Press.

Maggiori, M. (2017). Financial intermediation, international risk sharing, and reserve currencies. *American Economic Review*, 107:3038–3071.

Mancini, C. (2009). Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scandinavian Journal of Statistics*, 36(2):270–296.

Mancini, L., Ranaldo, A., and Wrampelmeyer, J. (2013). Liquidity in the foreign exchange market: Measurement, commonality, and risk premiums. *Journal of Finance*, 68:1805–1841.

Marsh, I. W., Panagiotou, P., and Payne, R. (2017). The wmr fix and its impact on currency markets. Working paper series, Cass Business School.

Menkhoff, L., Sarno, L., Schmeling, M., and Schrimpf, A. (2016). Information flows in foreign exchange markets: Dissecting customer currency trades. *Journal of Finance*, 71:601–634.

Mirkov, N., Pozdeev, I., and Soderlind, P. (2016). Toward removal of the swiss franc cap: market expectations and verbal interventions. Working Paper Series 10/16, Swiss National Bank.

Moore, M., Schrimpf, A., and Sushko, V. (2016). Downsized FX markets: Causes and implications. Staff report, BIS Quarterly Review 35-52.

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V., and Von Weizsäcker, J. E. (1997). Volatilities of different time resolutions-analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2-3):213–239.

Payne, R. (2003). Informed trade in spot foreign exchange markets: An empirical investigation. *Journal of International Economics*, 61:307–329.

Pesaran, H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1):17–29.

Podolskij, M. and Ziggel, D. (2010). New tests for jumps in semimartingale models. *Statistical inference for stochastic processes*, 13(1):15–41.

Ranaldo, A. and Söderlind, P. (2010). Safe haven currencies. *Review of Finance*, 14(3):385–407.

Ranaldo, A. and Somogyi, F. (2019). Heterogeneous information content of global fx trading.

Richardson, M. and Smith, T. (1994). A direct test of the mixture of distributions hypothesis: Measuring the daily flow of information. *Journal of Financial and Quantitative Analysis*, 29(1):101–116.

Rime, D., Sarno, L., and Sojli, E. (2010). Exchange rate forecasting, order flow and macroeconomic information. *Journal of International Economics*, 80:72–88.

Rime, D. and Schrimpf, A. (2013). The anatomy of the global FX market through the lens of the 2013 triennial survey. Staff report, BIS Quarterly Review 27-43.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, 39:1127–1139.

Tauchen, G. E. and Pitts, M. (1983). The price variability-volume relationship on speculative markets. *Econometrica*, 51:485–505.

Vayanos, D. and Wang, J. (2013). Market liquidity-theory and empirical evidence. In *Handbook of the Economics of Finance*, volume 2, pages 1289–1361. Elsevier.

A Proofs

A.1 Proof of Proposition 1

The log-return and volume at trade i are given by

$$r_i^{x|y} = \Delta p_i^{x|y} = \phi_i^{x|y} + \frac{1}{J}\sum_{j=1}^{J} \psi_{i,j}^{x|y}, \tag{25}$$

and the volume at i-th trade is

$$v_i^{x|y} = \frac{\xi^{x|y}}{2}\sum_{j=1}^{J} |\psi_{i,j}^{x|y} - \bar{\psi}_i^{x|y}|, \tag{26}$$

where $\bar{\psi}_i^{x|y} = \frac{1}{J}\sum_{j=1}^{J}\psi_{i,j}^{x|y}$. We assume for the moment that the common news term is zero, i.e. $\phi_i^{x|y} = 0$. Based on the return on the i-th interval, we can consider the realized variance, defined as $RV^{x|y} = \sum_{i=1}^{I}(r_i^{x|y})^2$ with $\Delta = 1/I > 0$, as introduced by Andersen and Bollerslev (1998). Following Barndorff-Nielsen and Shephard (2002b,a), taking the limit for $\Delta \to 0$ (that is $I \to \infty$), we get

$$p\lim_{I\to\infty} RV^{x|y} = \frac{1}{J^2}\mathcal{V}_{\psi^{x|y}}, \tag{27}$$

where $\mathcal{V}_{\psi^{x|y}} = \sum_{j=1}^{J} V_{\psi^{x|y},j}$ is the variation of the FX rate on the unit interval generated by the aggregated individual components of $r^{x|y}$. The term $V_{\psi^{x|y},j} = \int_0^1 \left(\sigma_j^{x|y}(s)\right)^2 ds$ is the *integrated variance* associated with the j-th trader's specific component. The term $\mu_j(t)$ does not enter in the expression of $V_{\psi^{x|y},j}$ since the magnitude of the drift, when measured over infinitesimal intervals, is dominated by the diffusive component of $\psi_{i,j}$ that is driven by the Brownian motion. Following Barndorff-Nielsen and Shephard (2003), for a given $\Delta > 0$ we can also define the realized power variation of order one (or realized absolute variation) as $RPV^{x|y} = \sum_{i=1}^{I}|r_i|$. By the properties of the super-position of independent SV processes,[25] the limit for $\Delta \to 0$ of $RPV^{x|y}$ is

$$p\lim_{I\to\infty} \Delta^{1/2} RPV^{x|y} = \sqrt{\frac{2}{\pi}}\mathcal{S}_{\psi^{x|y}}, \tag{28}$$

[25] Similarly to Barndorff-Nielsen and Shephard (2002b), $\bar{\psi}_i^{x|y}(t) = \frac{1}{J}\sum_{j=1}^{J}\psi_{i,j}^{x|y}$ is equivalent in law to $\bar{\psi}_i^{x|y,*} = \int_{\Delta(i-1)}^{\Delta i}\bar{\sigma}^{x|y}(t)dW^{x|y,*}(t)$, where $\bar{\sigma}^{x|y}(t) = \frac{1}{J}\sqrt{\sum_{j=1}^{J}{\sigma_j}^{x|y^2}(t)}$.

where $\mathcal{S}_{\psi^{x|y},j} = \int_0^1 \bar{\sigma}^{x|y}(s)ds$ is the integrated average standard-deviation, where the latter is defined as $\bar{\sigma}^{x|y}(t) = \frac{1}{J}\sqrt{\sum_{j=1}^{J} {\sigma_j^{x|y}}^2(t)}$. Given equation (26), the aggregated volume of $x|y$ on a unit (daily) interval is $v^{x|y} = \sum_{i=1}^{I} v_i^{x|y}$, and letting $I \to \infty$, we get

$$p \lim_{I\to\infty} \Delta^{1/2} v^{x|y} = \frac{\xi^{x|y}}{2}\sqrt{\frac{2}{\pi}}\bar{\mathcal{S}}_{\psi_{x|y}}, \tag{29}$$

with $\bar{\mathcal{S}}_{\psi_{x|y}} = \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \tilde{\sigma}_j^{x|y}(s)ds$, where $\tilde{\sigma}_j^{x|y}(t) = \sqrt{(J-1)^2{\sigma_j^{x|y}}^2(t) + \sum_{s\neq j}{\sigma_s^{x|y}}^2(t)}$.

A.2 Proof of Proposition 2

Given Proposition 1, we get that

$$p \lim_{I\to\infty} A^{x|y} = \frac{2\mathcal{S}_{\psi^{x|y}}}{\xi^{x|y}\bar{\mathcal{S}}_{\psi_{x|y}}}, \tag{30}$$

which reflects the ratio of the total average standard deviation carried by each trader. Under homogeneity of the traders, we get that

$$\bar{\mathcal{S}}_{\psi_{x|y}} = J\sqrt{J-1}\mathcal{S}_{\psi^{x|y}}, \tag{31}$$

and Proposition 2 follows directly.

A.3 Proof of Proposition 3

By imposing the no-arbitrage restriction as in Brandt and Diebold (2006), it follows from (8) that the squares of the synthetic returns at the i-th trade can be written as

$$(\widetilde{r}_i^{x|y})^2 = (r_i^{x|z} + r_i^{z|y})^2 = (r_i^{x|z})^2 + (r_i^{z|y})^2 + 2r_i^{x|z}r_i^{z|y}.$$

Under the maintained assumption that $\phi_i^{x|y} = 0$, the synthetic return can be expressed as $\widetilde{r}_i^{x|y} = \bar{\psi}_i^{x|z} + \bar{\psi}_i^{z|y}$, so that we can define the *synthetic* realized variance as $\widetilde{RV}^{x|y} = \sum_{i=1}^{I}(\widetilde{r}_i^{x|y})^2$. By letting $I \to \infty$, we get

$$p \lim_{I\to\infty} \widetilde{RV}^{x|y} = \frac{\mathcal{V}_{\psi^{x|z}} + \mathcal{V}_{\psi^{z|y}} + 2\mathcal{CV}_{\psi^{x|z},\psi^{z|y}}}{J^2}, \tag{32}$$

where $\mathcal{V}_{\psi^{x|z}} = \sum_{j=1}^{J} \int_0^1 \left(\sigma_j^{x|z}(s)\right)^2 ds$ and $\mathcal{V}_{\psi^{z|y}} = \sum_{j=1}^{J} \int_0^1 \left(\sigma_j^{z|y}(s)\right)^2 ds$ are the components of the return variation generated by the cumulative individual variations of the reservation prices on $x|z$ and $z|y$. The term $\mathcal{CV}_{\psi^{x|z},\psi^{z|y}}$ is given by

$$\mathcal{CV}_{\psi^{x|z},\psi^{z|y}} = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} \int_{\Delta(i-1)}^{i\Delta} \sigma_j^{x|z}(s)\sigma_j^{z|y}(s)\rho_j^{x|z,z|y}(s)ds \right),$$

where $\rho_j^{x|z,z|y}(t) = Corr\left(dW_j^{x|z}(t), dW_j^{z|y}(t)\right)$ is the correlation between the individual components on $x|z$ and $z|y$. All the other covariance terms are zero due to independence. For what concerns the trading volume, for the i-th trade on $x|y$ and $x|y$ we have

$$v_i^{x|z} = \frac{\xi^{x|z}}{2} \sum_{j=1}^{J} |\Delta p_{i,j}^{x|z,*} - \Delta p_i^{x|z}|, \quad v_i^{z|y} = \frac{\xi^{z|y}}{2} \sum_{j=1}^{J} |\Delta p_{i,j}^{z|y,*} - \Delta p_i^{z|y}|.$$

Moreover, by the triangular no-arbitrage, $\Delta\tilde{p}_{i,j}^{x|z,*} = \phi_i^x - \phi_i^y + \psi_{i,j}^{x|z} + \psi_{i,j}^{z|y}$ and $\Delta\tilde{p}_i^{x|z} = \phi_i^x - \phi_i^y + \bar{\psi}_{i,j}^{x|z} + \bar{\psi}_{i,j}^{z|y}$, so that the *synthetic volume* of $x|y$ is given by

$$\tilde{v}_i^{x|y} = \frac{\xi^{x|y}}{2} \sum_{j=1}^{J} |\psi_{i,j}^{x|z} - \bar{\psi}_i^{x|z} + \psi_{i,j}^{z|y} - \bar{\psi}_i^{z|y}|, \tag{33}$$

which involves quantities that cannot be directly observed. However, by letting $I \longrightarrow \infty$, we get

$$p \lim_{I\to\infty} \Delta^{1/2}\tilde{v}^{x|y} = \frac{\xi^{x|y}}{2}\sqrt{\frac{2}{\pi}}\bar{\mathcal{S}}_{\psi_{x|z,z|x}}, \tag{34}$$

where $\bar{\mathcal{S}}_{\psi_{x|z,z|y}} = \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \tilde{\sigma}_j^{x|z,z|y}(s)ds$, and

$$\tilde{\sigma}_j^{x|z,z|y}(t) = \sqrt{\tilde{\sigma}_j^{x|z^2}(t) + \tilde{\sigma}_j^{z|y^2}(t) + 2\tilde{\sigma}_j^{x|z}(t)\tilde{\sigma}_j^{z|y}(t)\rho_j^{x|z,z|y}(t)}. \tag{35}$$

Equation (35) highlights that the synthetic volume reflects the aggregated trader-specific components on the individual FX rates, $x|z$ and $z|y$, as well as their aggregated correlation as measured by $\rho^{x|z,z|y}$, which reflects the correlation between $\psi_j^{x|z}$ and $\psi_j^{z|y}$.

The Real Effects of Secondary Market Trading Structure: Evidence from the Mortgage Market

Yesol Huh and You Suk Kim*

April 29, 2019

Abstract

A vast majority of mortgages in the U.S. are securitized into agency mortgage-backed securities (MBS), many of which are traded in the to-be-announced (TBA) forward market. By allowing different MBS to be traded based on a limited set of characteristics, TBA market generates liquidity, with the aggregate daily trading volume second only to the U.S. Treasury market. In this paper, we quantify the effect of the unique secondary market trading structure on individual borrowers' mortgage rates, demand for mortgages, and consumer spending. With a simple model, we show that the benefit of access to the TBA market is higher for loans with less desirable prepayment characteristics. Then, exploiting sharp discontinuities in the probability of a loan to be included in an MBS eligible for TBA delivery, we estimate that TBA eligibility reduces mortgage rates by 10–40 basis points, depending on the prepayment risk of the loan. Furthermore, we also provide evidence that TBA eligibility affects borrowers' refinancing decisions and subsequent durable consumption.

*Federal Reserve Board, yesol.huh@frb.gov, you.kim@frb.gov. We thank seminar participants at the Federal Reserve Board, Federal Reserve MBS Analytics Forum, Korea University, Yonsei University, and University of Washington for helpful comments. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the staff, by the Board of Governors, or by the Federal Reserve System.

1 Introduction

Do financial markets matter for the real economy? More specifically, does liquidity and trading structure of the secondary market matter? One may argue that trading structure in the secondary market only affects the investors in those markets and that it does not impact the economy more broadly.

From a theoretical perspective, better liquidity in the secondary market would result in better prices in the primary market, leading to lower costs of capital for those raising funding. If investors value liquidity, assets with better liquidity would have higher prices, and thus, investors would be willing to pay higher prices in the primary market as well. A few recent studies (Brugler et al., 2018a,b; Davis et al., 2018) use the introduction of post-trade transparency in the corporate bond market and the change in trading rule at NASDAQ to show that secondary market trading structure impacts the cost of capital for the firms issuing corporate bonds or seasoned equity.

In this paper, we focus on the market for agency mortgage-backed securities (MBS), which are secured by mortgages in pools guaranteed by government-sponsored enterprises (Fannie Mae and Freddie Mac) or the U.S. government (Ginnie Mae). Specifically, we study the impact of liquidity and trading structure of the agency MBS market on mortgage rates for individual borrowers, demand for mortgages, and consumer spending. The mortgage market is different from the markets studied by the aforementioned papers in that it impacts a large set of population directly. In fact, a vast majority of mortgages, particularly after the 2008 financial crisis, end up in agency MBS. Also, although it has not been studied as much in the academic literature, the agency MBS market is the second most actively traded fixed-income market.

The unique feature of the agency MBS market is the to-be-announced (TBA) market, through which 90% of the trading is done. A TBA trade is a forward contract for a future delivery of MBS, where parties do not specify the CUSIP but agree only on six parameters at the time of the trade: agency (Fannie, Freddie, or Ginnie), coupon, maturity, price, par amount, and settlement date. Thus, if an MBS meets the six parameters specified in the TBA trade and the eligibility criteria for TBA delivery set by the Securities Industry and Financial Markets Association, then the TBA seller can deliver any of such MBS. As a result, the TBA trading structure concentrates trading of MBS with heterogeneous prepayment risks into a handful of TBAs and makes the market liquid. Although the TBA trading structure is a vital part of the MBS market, no studies so far have quantified the impact of TBA trading on mortgage borrowers in the primary mortgage market.

The goal of this paper is to quantify the impact of this unique trading structure on mortgage rates for individual borrowers, demand for mortgages, and the real economy, exploiting cutoffs that determine the probability that a loan is included in an MBS eligible for TBA delivery (i.e., TBA-eligible MBS). Given the uncertain future of the TBA market due to a potential housing finance reform, some argue that the TBA market structure should be preserved, citing its benefit for mortgage borrowers.[1] This paper provides quantitative evidence on how much the TBA market

[1] For example, see Bright and DeMarco (2016).

matters for individual mortgage borrowers.

TBA eligibility provides a number of benefits for an MBS investor. First, TBA eligibility gives an MBS access to a more liquid market with a large investor base. A TBA-ineligible MBS must be traded in the much less liquid specified pool (SP) market, where the individual CUSIP is specified at the time of trade.[2] Moreover, TBA pricing only depends on six parameters, thus is relatively simple; this simplicity increases the investor base for TBA trades. Second, TBA eligibility decreases downside risks for the MBS holder. All agency MBS, including TBA-eligible ones, can be traded in SP in principle. In fact, despite the high SP trading cost, TBA-eligible MBS with better prepayment characteristics often trade as SP to receive prices higher than the cheapest-to-deliver TBA prices. However, having the option to trade in TBA shields such MBS from the risk of not being able to find an buyer in the SP market. As a result, even an SP trade for TBA-eligible MBS is found to be more liquid than that for TBA-ineligible MBS (Gao et al., 2017). These benefits may fully or partially be passed down to primary market mortgage borrowers as lower mortgage rates.[3]

We begin our analysis with a simple model that describes the decision problem of an MBS seller that can sell an MBS as TBA or SP. If traded as TBA, the seller receives the cheapest-to-deliver price that does not depend on the prepayment risk of the MBS. If traded as SP, the seller receives the price that reflects the prepayment risk at the expense of a stochastic trading cost, which can be potentially very high. Thus, an implication of the model is that the option to easily sell the MBS as TBA protects the seller from the downside risk of having a very large realized SP trading cost. However, the value of TBA eligibility will depend on the prepayment risk of the MBS. An MBS with lower prepayment risk is more likely to be traded as SP despite high trading costs, and thus the option value of TBA trading will be lower for MBS. Moreover, loans with better prepayment characteristics tend to be pooled together into the same MBS empirically. Thus, TBA eligibility will be more valuable for loans with higher prepayment risks.

With these implications of the model in mind, we then estimate the impact of TBA eligibility on the mortgage rate. Our empirical strategy exploits two cutoff-based rules that determine the probability that a loan is included in a TBA-eligible MBS. An important difference between the two cutoffs is that they affect TBA eligibility for loans located in the opposite ends of the prepayment risk distribution. The model predicts that the estimated impact of TBA eligibility will be higher for the cutoff that is more relevant for loans with higher prepayment risks. Moreover, estimating the value of TBA eligibility at the two cutoffs will give us the range of TBA-eligibility benefit for the mortgages in between.

The first cutoff is the national conforming loan limit (CLL), which determines the maximum loan size that the government-sponsored enterprises (GSEs) can purchase and securitize. The GSEs can securitize only "conforming" mortgages, whose sizes are not greater than the CLL. Starting in

[2]Bessembinder et al. (2013) find that the trading cost of TBA and SP trades are 1 basis points and 40 basis points, respectively.

[3]The TBA market also positively impacts TBA-ineligible MBS because investors price TBA-ineligible MBS based on TBA prices and may also hedge with TBAs. Hence, the effect we measure here is a lower bound of the total impact of the TBA market.

2008, the GSEs began purchasing "high-balance" loans, which are larger than the national CLL but still not greater than the high-cost CLL.[4] The high-cost CLL, which became effective in 2008, is an increased loan limit for counties with high home prices. If an MBS contains more than 10% of its pool value in high-balance loans, the MBS is ineligible for TBA delivery. Indeed, we find that the probability to be included in a TBA-eligible MBS drops discontinuously from almost 100% to around 65% for a loan securitized by the GSE with the size just above the national CLL. We also find that a GSE loan with the size around the national CLL tend to have higher prepayment risks than a majority of other loans securitized by the GSEs, most of which are smaller than the national CLL. This is because a borrower with a larger loan usually has higher incentive to refinance because the same decrease in interest rates would result in larger savings.

The second cutoff is the loan-to-value ratio (LTV) of 105. A TBA-eligible MBS is not allowed to include even a single loan with LTV greater than 105. Thus, all loans with LTVs greater than 105 are included in TBA-ineligible MBS. Loans with such high LTVs were originated and sold to the GSEs under the Home Affordable Refinance Program (HARP) in 2009. Because a borrower can take advantage of HARP only once, and because the high LTV makes it difficult for such a borrower to refinance without such a special government program, loans with LTVs around 105 empirically exhibit lower prepayment risks than a majority of other loans securitized by the GSEs.

Using an empirical strategy that exploits discontinuities at the two cutoffs, we find that TBA-eligibility reduces mortgage rates by 40 basis points for loans around the national CLL and 10 basis points for loans with LTVs around 105, respectively. The large difference in the estimated magnitudes for the two cutoffs is consistent with the prediction of the model. Loans around the national CLL tend to have higher prepayment risks than loans with LTVs around 105. Thus, the option value of TBA will be more valuable for the former than the latter, thereby resulting in a greater magnitude of the estimated benefit from TBA eligibility for the former.

The fact that we estimate the impact on the mortgage rate with the two cutoffs is important not only for testing the prediction of the model but also for estimating the upper and lower bounds of the value of TBA eligibility. A common criticism against research designs estimating local treatment effects based on discontinuities is that the resulting estimate can be very different from the true effect for the entire population. This concern would apply to our setup if we estimated the impact on the mortgage rate using only one of the two cutoffs. In fact, the two cutoffs result in very different magnitudes, and it would be difficult to apply any one of these estimates to loans with different prepayment risks. However, the two cutoffs affect loans near either end of the spectrum of prepayment risks. Thus, the estimated impact on the mortgage rates with the two cutoffs are likely to be close to the upper and lower bounds, and we expect that the benefit of TBA eligibility will fall between our two estimates for loans with prepayment risks toward the middle of the distribution of prepayment risks.

Next, we estimate the impact of TBA eligibility on demand for mortgages. Because TBA

[4]High-balance loans are often also referred to as jumbo-conforming or super-conforming loans. These loans are different from jumbo loans, which the GSEs are not allowed to securitize.

eligibility impacts mortgage rates, we would expect that it also impacts the demand for mortgages. Specifically, we investigate how much TBA eligibility affects refinancing decisions of borrowers with remaining loan balances around the national CLL.[5] Studying refinancing decisions is important because of their implications for monetary policy transmission and the real economy. In fact, there is a growing literature on the refinancing channel of monetary policy transmission, where lower interest rates induce mortgage borrowers to refinance and subsequently increase their consumption.[6] Using the data that link each mortgage to the borrower's credit record, we are able to identify whether a borrower refinances a mortgage and how much the mortgage balance increases after refinancing. We find that the monthly probability of plain refinancing, which does not involve a significant increase in the loan balance, discontinuously increases by 0.25 percentage points (50% of the unconditional mean) when remaining mortgage balance reaches the national CLL from above. This finding suggests that borrowers delay refinancing in order to refinance into a TBA-eligible mortgage. A borrower slowly pays off the remaining principal according to the amortization schedule, waiting until his balance reaches the national CLL. Once the borrower's balance reaches the national CLL, the borrower quickly refinances into a loan below the national CLL. Moreover, this waiting can be quite long. The average borrower in our sample would need to wait for 17 (32) months to pay down \$10,000 (\$25,000) to reach the national CLL.

Finally, we study whether the delay of refinancing stemming from TBA eligibility affects real economic outcomes outside the mortgage market. Specifically, we investigate how a borrower's durable consumption changes upon refinancing. Our data allow us to identify new auto loan originations, from which we can infer whether and when an individual purchases a new automobile. Among borrowers who refinance when their remaining balances are close to the national CLL, we find that the probability of a new auto sale sharply increases immediately after refinancing. Consistently, we also find that a borrower's auto new loan origination increases right after a borrower's remaining mortgage balance reaches the national CLL from above. At that point, a borrower is much more likely to refinance his mortgage and then purchase a new car with a new auto loan. Thus, when a borrower delays refinancing in order to refinance into a TBA-eligible loan, the borrower's durable consumption is also delayed. This finding implies that the unique trading structure of agency MBS also matters for monetary policy transmission and real economic outcomes by affecting borrowers' refinancing and subsequent durable consumption.

Literature Review This paper adds to the literature on the real effects of financial markets.[7] A few papers in this literature study the effect of secondary market trading structure and liquidity on firms' borrowing costs and investments. Both Brugler et al. (2018b) and Davis et al. (2018)

[5] We do not consider a refinancing decision of a borrower with updated LTV close to 105 because we do not observe the updated house value that would be used in underwriting a borrower refinance application.

[6] For examples, see Abel and Fuster (2018), Agarwal et al. (2017), Beraja et al. (2018), Di Maggio et al. (2016), Greenwald (2018), and Wong (2018).

[7] Bond et al. (2012) provides a survey of theoretical and empirical literature on the real effects of financial markets. A majority of papers in this area study the effect that financial markets have on firms' decisions because of the information or the incentives that financial markets provide.

show that the introduction of post-trade transparency in the secondary corporate bond market has decreased cost of capital in the primary market. Field et al. (2018) also use the same variation to show that that firms with greater bond liquidity engage in more merger and acquisition activities. Brugler et al. (2018a) study a specific rule change in NASDAQ that moved the market from a dealer-oriented market towards a more centralized one and argue that this rule change decreased the underpricing of seasoned equity offerings.

However, only a few papers in this literature study consumer financial markets. Fuster and Vickery (2014) show that there are fewer fixed-rate mortgages when securitization is difficult. Benmelech et al. (2016) find that the collapse of the asset-backed commercial paper market reduced automobile purchases by decreasing the auto loan supply from nonbank auto lenders that depended on the funding market. We contribute to this literature by showing how the trading structure of the secondary market affects the primary mortgage market and consumer spending.

This paper is also related to a small number of papers that study the trading structure and liquidity of the secondary market for agency MBS, with a particular focus on TBA and SP trading. Vickery and Wright (2013) provides a comprehensive overview of the institutional details of TBA market and discusses how TBA market generates liquidity and how TBA trades are used. They also argue that TBA market liquidity would likely impact the pricing in the primary market for mortgages. However, given that their paper mostly focuses on the secondary market, they only provide preliminary evidence that TBA eligibility affects mortgage rates and caution the readers that differences in prepayment risks are not controlled for. In this paper, we look at narrow bands around TBA-eligibility cutoffs and use discontinuity tests to tease out the impact of TBA eligibility.

In addition, Bessembinder et al. (2013) studies trading costs in structured credit products and finds that trading costs in TBA trades are very small (1 bp) while that of SP trades are much higher (40 bps). Gao et al. (2017) argues that TBA eligibility affects trading costs for SP trades because dealers can more easily hedge SP inventory for TBA-eligible MBS with TBA trades. Schultz and Song (2018) studies the impact of post-trade transparency in the TBA market.

This paper also contributes to the literature that studies monetary policy transmission through the mortgage market. A number of papers study the refinancing channel of monetary policy transmission; for example, see Abel and Fuster (2018), Agarwal et al. (2017), Beraja et al. (2018), Di Maggio et al. (2016), Greenwald (2018), and Wong (2018). Moreover, Di Maggio et al. (2017) studies consumption and deleveraging of borrowers with adjustable-rate mortgages, whose mortgage rates would be automatically decreased by an accommodative monetary policy. We add to this literature by showing that the secondary mortgage market trading structure is an important factor that affects refinancing, which is an important part of monetary policy transmission.

This paper is also related to studies that estimate the spread in mortgage rates between conforming and jumbo loans such as Passmore et al. (2005), Sherlund (2008), Kaufman (2014), and DeFusco and Paciorek (2017). These papers measure how much the GSEs subsidize the mortgage market by comparing mortgage rates of conforming loans just under the CLL and jumbo (not high-balance) loans just above the CLL. Because jumbo loans cannot be sold to the GSEs, the spread

reflects not only the value of TBA eligibility but also the value of credit guarantees from the GSEs. In our empirical strategy, in contrast, we compare GSE loans around the national CLL to estimate the value of TBA eligibility.

2 Institutional Details

2.1 Basic Facts about the TBA Market

TBA Eligibility A vast majority of mortgages in the U.S. are securitized and packaged into agency MBS. Most of agency MBS are backed by mortgages in pools guaranteed by Fannie Mae, Freddie Mac, or Ginnie Mae. Thus, these mortgages carry either implicit or explicit credit guarantees from the U.S. government.

TBA trade is in essence a forward contract where two parties agree on a price today for a future delivery of agency MBS. Moreover, instead of agreeing upon a specific CUSIP at the onset of the trade, parties only agree on six general parameters: agency (Freddie Mac, Fannie Mae, and Ginnie Mae), coupon, maturity, price, par amount, and settlement date. Only 48 hours before the delivery date, the seller is required to notify the buyer of the specific CUSIP(s) that he will deliver. Because the seller chooses what to deliver, there is a cheapest-to-deliver pricing for TBA trades. Given the large number of individuals CUSIPs in the agency MBS market and the relative homogeneity, this structure concentrates the trading into a handful of TBAs and generates liquidity. According to Vickery and Wright (2013), TBA trades account for 90 percent of trading volume in the agency MBS market.

However, not all MBS are allowed to be delivered for TBA settlement. There are largely three reasons why an MBS is not eligible for TBA settlement. First, MBS that include any loans with the original LTV greater than 105 are not TBA-eligible. Mortgages with such high LTVs are usually very difficult to be sold to the GSEs if not impossible. The GSEs began to buy and securitize these loans under the Home Affordable Refinancing Program (HARP). This program was set up in March 2009 to help refinancing for existing mortgage borrowers with depreciated home prices due to the housing market crisis at that time. With a very large decrease in home prices, many borrowers found themselves having remaining mortgage balances more than their the market values of their homes. In other words, their updated LTVs were greater than 100, which would have made it impossible for these borrowers to refinance into new loans to take advantage of historically low interest rates at that time. However, HARP made it possible for borrowers meeting its eligibility criteria with very high LTVs to refinance into a GSE loan.[8] Initially, HARP excluded loans with updated LTVs greater than 125, but the LTV limit was removed in December 2011. As for TBA eligibility of HARP loans, only HARP loans with LTVs up to 105 were allowed to be included in TBA-eligible MBS. Thus, any HARP loans with LTVs greater than 105 must be included in TBA-ineligible MBS.

Second, MBS that have more than 10 percent of the pool value in high-balance loans are not

[8] A borrower is eligible for HARP if he originated a mortgage sold to a GSE before May 31, 2009 and if he had not missed a mortgage payment for past 12 months.

eligible to be delivered for TBA settlement. High-balance loans refer to mortgages with loan size greater than the national conforming loan limit (CLL) but not greater than the county-specific high-cost CLL.[9] The GSEs are only allowed to purchase "conforming" loans that are not greater than the CLL. Until February 2008, the CLL was national except for Alaska, Guam, Hawaii, and Virgin Islands. For example, with the national CLL equal to $417,000 in 2007, the GSEs were able to buy only loans with size up to $417,000. In March 2008, however, Congress passed the Economic Stimulus Act (ESA) in response to the ongoing financial crisis, which raised the CLL in counties with high home prices. The new CLLs for the high-cost counties under the ESA were set equal to the greater of $417,000 and 125 percent of the county-level median home price with the ceiling of $725,750.[10] As a result, the ESA made it possible for the GSEs to buy and securitize high-balance loans. Initially, there was uncertainty about whether MBS including high-balance loans will be eligible for TBA settlement. Eventually, the SIFMA set the rule in August 2008 such that MBS with more than 10 percent of the pool value in high-balance loans are TBA-ineligible.

Lastly, MBS with greater than 15 percent of pool value in loans with other non-standard features such as co-op share loans, relocation loans, and loans with significant interest rate buydowns are not eligible for TBA delivery. As will be discussed in Section 2.2, only very few agency MBS are TBA-ineligible based on this criterion. Thus, we do not study the loans with these non-standard features in this paper.

Specified-Pool Market In a specified pool (SP) trade, parties agree and trade on the specific CUSIP, and each CUSIP is thinly traded. As a result, a SP trade usually has a higher trading cost than a TBA trade. In fact, Bessembinder et al. (2013) find that the trading cost of TBA and SP trades are 1 basis points and 40 basis points, respectively. Agency MBS that are not eligible for TBA delivery must be traded in the specified-pool (SP) market. TBA-eligible CUSIPs may also trade in the SP market; they may do so especially when the value of the CUSIP is high, that is, when the prepayment risk is low compared to other TBA-eligible CUSIPs.

2.2 TBA-Ineligible Pools

Figure 1 shows the evolution of dollar-weighted shares of loans (among 30-year fixed-rate mortgages sold to the GSEs) included in new agency MBS that are not eligible for TBA settlement. We categorize TBA-ineligible MBS into three broad groups: high-balance MBS, high-LTV MBS, and other TBA-ineligible MBS. First, high-balance MBS consist of high-balance loans only.[11] Thus, high-balance MBS are not eligible for TBA settlement. Second, high-LTV MBS consist of HARP loans with LTVs greater than 105. Because a TBA-eligible MBS cannot include any loan with the LTV greater than 105, such loans are packaged together into a high-LTV MBS. Third, other

[9] These loans are sometimes referred to as super-conforming or jumbo-conforming loans.

[10] The national CLL was $417,000 until the end of 2016. It was increased to $424,100 in 2017 and then to $453,100 in 2018.

[11] Note that not all high-balance loans are included in high-balance MBS. Because a TBA-eligible MBS is allowed to have up to 10% of its pool value in high-balance loans, many high-balance loans are still packaged into TBA-eligible MBS. We will discuss this in more details in Section 3.2.

TBA-ineligible MBS include various MBS that are not eligible for TBA settlement because loans in the MBS have other non-standard features such as co-op share loans, relocation loans, and loans with significant interest rate buydowns.

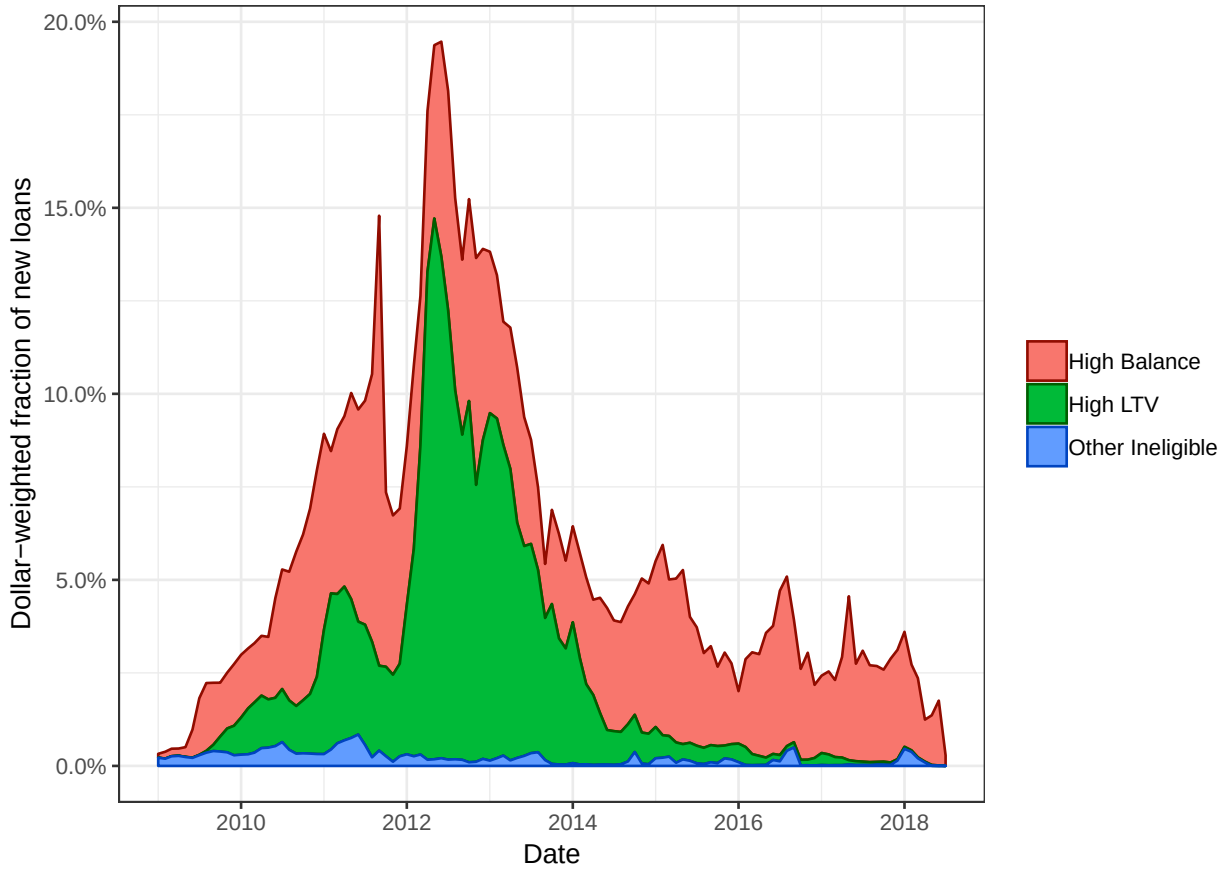
There are two main takeaways from Figure 1. First, the main reason why a loan is included in a TBA-ineligible MBS during the sample period is either because the loan has the original balance greater than the national CLL (a high-balance MBS) or because the LTV of the loan is greater than 105 (a high-LTV MBS). Most TBA-ineligible MBS are either high-balance or high-LTV MBS except in early 2009, although shares of the two types of MBS vary over time.

Second, the total share of TBA-ineligible MBS is not negligible during our sample period. In early 2009, the total share of TBA-ineligible MBS was close to zero, which means that all loans in the sample were included in TBA-eligible MBS. However, the share of loans in TBA-ineligible MBS grew substantially in the next few years, reaching close to 20% in mid-2012. The increase was mainly due to increasing originations of loans included in high-balance and high-LTV MBS.

The large shares of high-balance and high-LTV MBS in mid-2012 were in part because of a large refinance volume driven by historically low mortgage rates at that time. Many high-balance loan originations were due to refinancing by borrowers with jumbo mortgages that were originated pre-crisis (Bond et al., 2017). As mortgage rates continued to decrease in years after the 2008 financial crisis, many such borrowers refinanced into high-balance mortgages. The low interest rate environment, together with the slump in house prices, also resulted in a large number of mortgages being refinanced into HARP loans. In addition, a new version of HARP was implemented in December 2011 (called HARP 2.0) to increase the take-up of the program. Among many changes brought by HARP 2.0 to encourage borrowers to refinance into HARP loans, even a borrower with LTV greater than 125% were allowed to refinance into HARP mortgages without private mortgage insurance.

As mortgage rates increased in recent years, the refinancing volume decreased, and the shares of the high-balance and high-LTV MBS also decreased. In particular, there are barely any new issuances of the high-LTV MBS in 2018 as many borrowers eligible for HARP already took advantage of the program. Because the program is only available for borrowers who took out loans sold to the GSEs by early 2009, the number of borrowers eligible for the program will only decrease over time. Moreover, the house price appreciation in recent years has left very few borrowers with very high updated LTVs.

Figure 1: **Share of Mortgages in TBA-Ineligible MBS**: This figure plots the dollar-weighted share of loans in TBA-ineligible MBS, among 30-year fixed-rate mortgages in MBS securitized by the GSEs, that were originated in the period from 2008 to August 2018. Each month refers to the month of loan origination. The red area represents the share for loans in high-LTV MBS, which contain only loans with LTVs greater than 105. The green area represents the share for loans in high-balance MBS, which contain only high-balance loans. However, there are also high-balance loans included in TBA-eligible MBS. The blue area represents the share for loans in other TBA-ineligible MBS. The source of the data for this figure is eMBS.



2.3 Two Cutoff Rules

In our empirical analyses, we focus on two main TBA-eligibility cutoffs: loan size of national CLL and LTV of 105.[12] For both loan size and LTV dimensions, the probability that a loan is included in a TBA-eligible MBS changes discontinuously around the cutoffs, which we will show in Section 3.2. Our empirical strategies hinges on the discontinuities at the two cutoffs. For instance, to control for other characteristics that affect mortgage rates, we compare loans with sizes just under and above the national CLL. Similarly, we also compare loans with LTVs just under and above the threshold of 105.

[12] As discussed earlier, high-balance mortgages are greater than the national CLL but not greater than high-cost CLL.

3 Data and Summary Statistics

3.1 Data Description

We use multiple data sources to estimate the effect of TBA eligibility on the primary market. First, we use the eMBS data, which provides various information on agency MBS and mortgages underlying each of agency MBS. From this data, we obtain information on MBS-level characteristics such as coupon rate, issuer, pool issue amount, pool issue date, product type, TBA eligibility, prepayment history, the distribution of loan-level characteristics within an MBS, etc. The loan-level eMBS data provides information about loan-level characteristics and prepayment history. Moreover, the loan-level data provide a link between a loan and the CUSIP of the MBS that includes the loan. This information is crucial in correctly estimating the benefit of TBA eligibility on mortgage rates because some high-balance loans can be included in TBA-eligible pools, as discussed in Section 2. The eMBS loan-level data to which we have access covers loans in Fannie Mae pools that are issued in or before October 2013 and loans in Freddie Mac pools that are issued in or before August 2018. Thus, we are missing mortgages sold to Fannie Mae for for the period after late 2013.

The second data is the loan-level data from Fannie Mae and Freddie Mac. They provide a publicly available single family loan-level performance data for fixed-rate mortgages originated between between January 1, 1999 and September 30, 2017. Importantly for this paper, they also provide loan-level data for HARP mortgages and a link between a HARP mortgage and the original loan. With this link, we can track performance of an original loan and a HARP loan, which is crucial for our empirical test using the LTV cutoff.

We use the first two data sets for our analysis of the impact on mortgage rates. In our sample, we only keep 30-year fixed-year-mortgages originated in or after 2009 that are sold to the GSEs. In addition, we only keep loans originated for single-family houses to keep the sample relatively homogeneous. We also use different subsamples for different cutoffs to only compare loans near each cutoff. The subsample selection will be explained in more details in Section 5.

The third data set we use is Equifax Credit Risk Insight Servicing and Black Knight McDash Data (CRISM), which links loan-level mortgage data to each borrower's credit records from Equifax. We use this data to analyze the impact on refinancing and subsequent durable consumption through new auto loan originations.

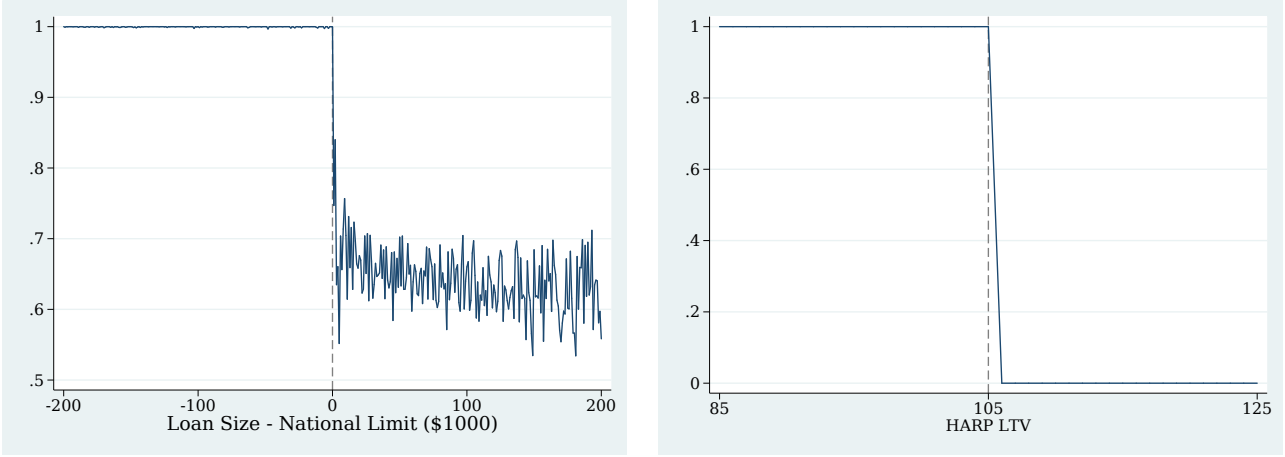
3.2 Summary Statistics

Figure 2 shows that the fraction of loans (among 30-year fixed-rate mortgages sold to the GSEs) that are included in TBA-eligible MBS changes substantially and discontinuously at the two cutoffs. In panel (a), the fraction is one for loans with size below the national CLL. However, the fraction decreases to around 0.6 for loans right above the CLL. This fraction does not decrease all the way to zero because high-balance loans can still be included in a TBA-eligible MBS as long as their share does not exceed 10%. In panel (b), the fraction decreases sharply to zero once the LTV exceeds 105 because any of such loans cannot be included in TBA-eligible MBS.

Note that these figures are created using only GSE loans. Consequently, jumbo loans, which are greater than the high-cost CLLs and thus cannot be securitized by the GSEs, are excluded from the data sample, and loans greater than the national CLL in the figure are high-balance loans that are securitized by the GSEs. Therefore, the fraction of loans included in TBA-eligible MBS decreases at the national CLL not because loans above the national CLL cannot be sold to the GSEs but because there is a limit on how much high-balance loans can be part of TBA-eligible MBS.

This is the main difference from papers that estimate spreads between jumbo and conforming loans in the period before the ESA introduced the high-cost CLLs in 2008 (e.g. Passmore et al. (2005); Sherlund (2008); Kaufman (2014); DeFusco and Paciorek (2017)). These papers aim to estimate how much the GSEs reduce mortgage rates by comparing loans that are eligible and ineligible for the GSE securitization. The effect of GSE eligibility will capture not only the value of having access to the TBA market, which is only available for agency MBS, but also the value of mortgage credit guarantees for GSE loans. In contrast, our data sample consists only of loans securitized by the GSEs, so we can estimate the effect of TBA-eligibility controlling for the effect of mortgage credit guarantees from the GSEs.

Figure 2: **Probability to Be Included in TBA-eligible Pools around the Cutoffs**: These figures plot the probability for a loan to be included in TBA-eligible MBS. Panel (a) plots the probability against the loan size. In the x-axis of this panel, the loan size is measured relative to the national CLL in thousand dollars. The source of the data for Panel (a) is eMBS. Panel (b) plots the probability against the LTV of a HARP loan. The source of the data for Panel (b) is loan-level data for HARP.



(a) Loan Size - CLL ($1000)

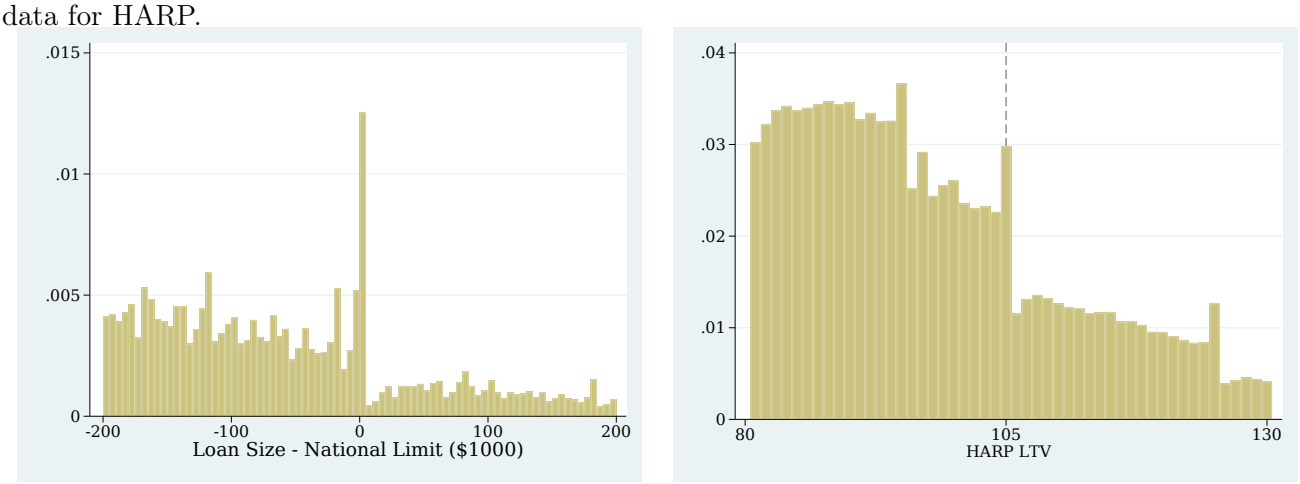
(b) Loan-To-Value

Figure 3 presents loan-level density around the two cutoffs. Bunching at the cutoffs shown in both panels is indicative of pricing differentials between loans below and above the cutoffs, possibly because of TBA eligibility. If mortgage rates are lower for loans that are more likely to be included in TBA-eligible MBS, borrowers that are slightly above the cutoff could to adjust the mortgage (e.g., by putting higher downpayments) to be at or below the cutoff. Previous papers that estimate the rate spread between jumbo and conforming loans also report a pattern similar to panel (a)

at the national CLL before 2008; the pattern there is driven by pricing differential between loans securitized by the GSEs and loans that are not.

At the same time, the bunching at the cutoffs also poses challenges to estimation of the rate spreads at the cutoffs because those who bunch might have different unobserved characteristics from those who originate loans just above the cutoffs. If that is the case, then at least some of the rate spreads may be accounted for by the potential difference in unobserved characteristics of borrowers. We discuss how we address this challenge in Section 5.

Figure 3: **Bunching at the Cutoffs**: These figures plot loan-level density. Panel (a) plots the density against the loan size. In the x-axis of this panel, the loan size is measured relative to the national CLL in thousand dollars. The source of the data for Panel (a) is eMBS. Panel (b) plots the density against the LTV of a HARP loan. The source of the data for Panel (b) is loan-level data for HARP.



(a) Loan Size - CLL ($1000)

(b) Loan-To-Value

4 Simple Model

We write down a simple model that describe the value of TBA eligibility for an MBS. Consider a risk-neutral originator (or Fannie Mae or Freddie Mac) that is selling an MBS with fundamental value m. The fundamental value m would mostly be driven by prepayment risk. If the MBS is TBA-eligible, the originator has two options. First is to sell in the TBA market at price P_{tba}. Because of the cheapest-to-deliver pricing in TBA trades, this price does not depend on m. Second is to sell in the SP market at price $P_{sp}(m) + \epsilon$, where $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$. The expected SP price, $P_{sp}(m)$, is an increasing function in m. The noise term ϵ can be thought of as coming from a random liquidity shock to the SP market or the difference in private valuation (or preferences) of the buyers. We assume that the originator observes ϵ before choosing which market to sell the MBS at.

The originator sells in the TBA market if $P_{sp}(m) + \epsilon < P_{tba}$. The expected value of this MBS is:

$$V(m) = \rho(m)P_{tba} + (1 - \rho(m))\mathbb{E}\left[P_{sp}(m) + \epsilon | \epsilon > P_{tba} - P_{sp}(m)\right] \tag{1}$$

where $\rho(m)$ is the probability that this MBS trades in the TBA market. The above equation illustrates that TBA eligibility decreases downside risk. When ϵ is low, that is, when the price that one can receive by selling in the SP market is low, one can sell the MBS in the TBA market and get a better price. This optionality in effect allows one to always sell the MBS at reasonable prices and makes the MBS more liquid (Gao et al., 2017).

We can rearrange Equation (1) to more clearly show the value of TBA eligibility. Given that the expected value of an MBS that is not TBA eligible is simply $P_{sp}(m)$, the value of TBA eligibility is:

$$V(m) - P_{sp}(m) = \rho(m)\mathbb{E}\left[P_{tba} - P_{sp}(m) - \epsilon | \epsilon < P_{tba} - P_{sp}(m)\right]. \tag{2}$$

From this expression, it can be easily seen that this value is always positive.

Given the simple structure of the model, we can solve for $\rho(m)$ and $V(m) - P_{sp}(m)$.

Lemma 1.

$$\rho(m) = 1 - \Phi\left(\frac{P_{sp}(m) - P_{tba}}{\sigma}\right)$$
$$V(m) - P_{sp}(m) = -\left\{1 - \Phi\left(\frac{P_{sp}(m) - P_{tba}}{\sigma}\right)\right\}(P_{sp}(m) - P_{tba}) + \sigma\phi\left(\frac{P_{sp}(m) - P_{tba}}{\sigma}\right)$$

where Φ is the standard normal CDF, and ϕ is the standard normal PDF.

We can easily show the following properties using Equation (1) and Lemma (1).

Proposition 1. *Probability of trading in the TBA market, $\rho(m)$, and the value of TBA eligibility, $V(m) - P_{sp}(m)$, have the following properties:*

(i) $\rho(m)$ is a decreasing function in m: Probability of trading in the TBA market is higher for MBS with higher prepayment risks.

(ii) $V(m) - P_{sp}(m) > 0$: The value of TBA eligibility is positive.

(iii) $-1 \leq \frac{\partial(V(m) - P_{sp}(m))}{\partial P_{sp}(m)} \leq 0$: The value of TBA eligibility is higher for MBS with higher prepayment risks.

These results and the interpretations are fairly intuitive. MBS with higher prepayment risks have lower value ($P_{sp}(m)$ and m are lower), and thus are more likely to be traded through the TBA market. Hence, the value added from TBA eligiblity is also higher for those MBS. Lastly, the value of TBA eligiblity is positive because TBA eligiblity gives an option to trade in the TBA market. Although we currently take $P_{sp}(m)$ to be exogenous, we can easily extend the model to make it endogenous. Proposition 1 still hold in the extended model. In rest of this paper, Proposition 1(iii), namely that the value of TBA eligibility is higher for MBS with higher prepayment risks, will be important. Lastly, while we do not model how the value of TBA eligibility, $V(m) - P_{sp}(m)$, gets passed to individual loans in the pool, we expect that it would fully or partially get passed down to the mortgage borrowers in the primary market.

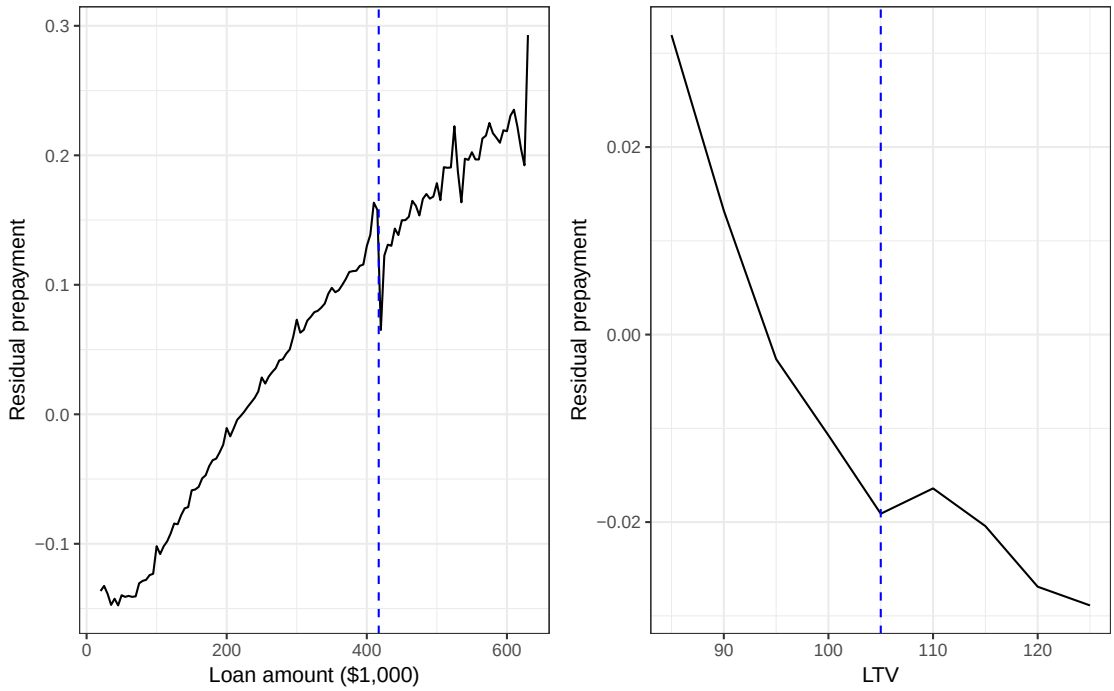
Value of TBA-eligibility and Prepayment Risks So far, we have shown that the value of TBA eligibility is greater for MBS with higher prepayment risks. What does this imply for the value of TBA-eligibility and loan-level prepayment risks? By definition, MBS with better prepayment characteristics (or lower prepayment risks) should have a larger share of loans with better prepayment characteristics than others. Moreover, we find that mortgage lenders tend to pool loans with better prepayment characteristics into the same MBS and trade the MBS in the SP market. Thus, loans with better prepayment characteristics will end up in MBS with better overall prepayment characteristics, and the value of TBA-eligibility will be also lower for such loans.

How is the relationship between prepayment risks and the value of TBA eligibility applied to our empirical setting? Figure 4 plots the relationship between ex-post prepayments and loan amounts (left panel) and between ex-post prepayments and LTVs (right panel). Ex-post prepayments in the figure are measured in terms of whether a loan was paid off completely by 36 months after the loan origination. Of course, this is a very specific measure of prepayments, but the pattern remains qualitatively unchanged when we use other loan ages. To control for the potential interactive effect of mortgage rates and interest rate path on prepayments, we consider residual prepayments, which are calculated by removing variation accounted for by the origination month $\times$ mortgage rate fixed effects.

The left-hand-side figure shows that prepayment risks and loan amounts are positively correlated. The vertical line is drawn at \$417,000, which is the national CLL until the end of 2016. Combined with the fact that a vast majority of loans are smaller than the national CLL as shown in Figure 3, this prepayment pattern suggests that loans around the national CLL have higher prepayment risks than a vast majority of loans securitized by the GSEs. Because the value of TBA eligibility will be higher for loans with higher prepayment risks, it is likely that the value of TBA eligibility for loans around the national CLL is close to the upper bound of the value of TBA eligibility.

On the other hand, the right-hand-side figure shows that prepayment risks and LTVs are negatively correlated. The vertical line is drawn at LTV 105, which is another cutoff used in the empirical analysis. Although the figure shows prepayments only for LTVs greater than 85, the prepayments for LTVs below 85 are higher than prepayments for LTVs greater than 85. This implies that loans with LTVs around 105 have lower prepayment risks than a vast majority of loans securitized by the GSEs. Thus, it is likely that the value of TBA eligibility for loans with LTVs around 105 is close to the lower bound of the value of TBA eligibility.

Figure 4: **Ex-post Prepayments by Loan Age 36 Months**: This figure displays the relationship between ex-post prepayment risks and loan amounts (on the left) and LTV (on the right). Vertical lines refer to the two cutoffs used in our empirical analysis: the national CLL (on the left) and LTV of 105 (on the right). Ex-post prepayments in the figure are measured in terms of whether a loan was paid off completely by loan age 36 months since origination. To control for potentially different prepayment behaviors depending on when a loan is originated and other loan characteristics, we consider residual prepayments, which are calculated by removing variation accounted for by the origination month × mortgage rate fixed effects.



5 Effects on Mortgage Rates

To quantify the benefit of TBA eligibility at the loan level, we would ideally compare interest rates between two identical loans, one of which is included in a TBA-eligible MBS while the other is included in a TBA-ineligible MBS. We can get close to the ideal situation by exploiting the rules that determine whether an MBS is eligible for TBA, which result in the discontinuities in the probability that a loan is included in TBA-eligible MBS around the cutoff values. We use the two cutoffs discussed earlier: the national CLL and LTV of 105.

5.1 High-Balance Loans

As shown by panel (a) in Figure 3, a high number of loans bunch at the national CLL although loans larger than the national CLL can be sold to the GSEs. This bunching poses a challenge to an identification strategy that utilizes the discontinuity in the probabilities for a loan to be included in TBA-eligible pools. Borrowers who bunch might have different unobserved characteristics from

those who take out loans just above the cutoff. In that case, the rate spreads could be due to the potential difference in unobserved characteristics of borrowers.

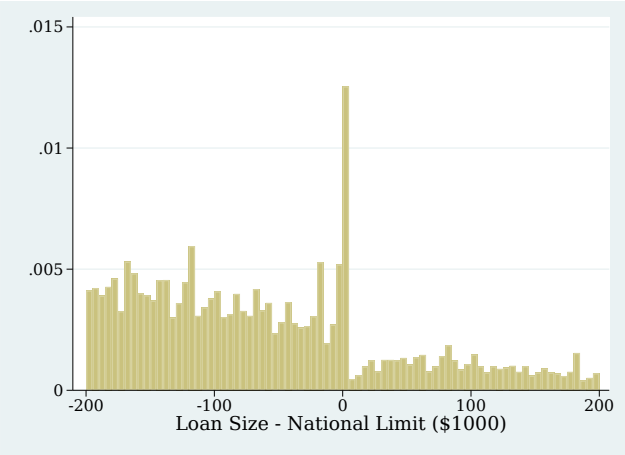
We overcome this challenge using an instrument variable strategy used by previous papers which estimate the impact of GSE purchase eligibility on mortgage rates in the period before the high-cost CLL was introduced.[13] The main idea of this empirical strategy is to utilize an alternative cutoff based on the home appraisal value instead of the loan size. The GSEs usually requires a borrower with less than a 20% down payment to have a mortgage insurance. In fact, a significant fraction of borrowers (36%) make exactly 20% down payments in our data. With this ubiquity of the 20% down payment, a borrower purchasing a home with the appraisal value not greater than 125% of the national CLL would most likely take out a conventional conforming mortgage. In contrast, a borrower purchasing a home with the appraisal value greater than 125% of the national CLL would take out a high-balance mortgage with a greater probability. Therefore, the probability of a loan to be included in a TBA-eligible MBS will change discontinuously depending on whether the home value is greater than 125% of the national CLL.

For this analysis, we impose the following additional sample selection criteria to keep the sample relatively homogeneous. First, we only keep purchase loans because our identification strategy is the most relevant for such loans. In fact, a majority of new originations with original LTV of 80 are purchase loans. Second, we exclude any loans with second mortgages (by comparing combined LTV and original LTV) or any loans with mortgage insurance (original LTV greater than 80).

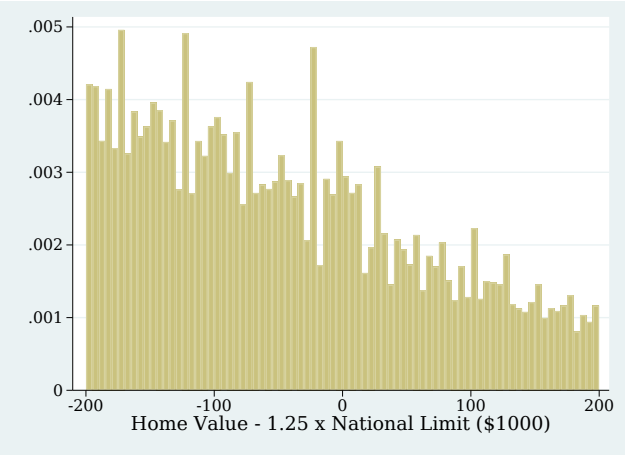
The alternative cutoff based on the home value leads to a smooth density around the cutoff. Figure 5 shows differences in sorting patterns around the two different cutoffs. It is very clear that whereas panel (a) exhibits bunching at the national CLL, panel (b) shows a relatively smooth density around the cutoff based on the home appraisal value.

[13] Examples of such papers are Adelino et al. (2012); Kaufman (2014); DeFusco and Paciorek (2017). Another related paper that used the identification strategy is Vickery and Wright (2013), which studies how securitization affects availability of fixed-rate mortgages.

Figure 5: **Sorting around the Cutoffs**: These figures plot loan-level density. Panel (a) plots the density against the loan size. In the x-axis of this panel, the loan size is measured relative to the national CLL in thousand dollars. Panel (b) plots the density against home value associated with each loan. In the x-axis of this panel, the home value is measured relative to the cutoff based on the home value in thousand dollars. The source of both figures is eMBS.



(a) Loan Size - CLL ($1000)



(b) Appraisal Value - CLLx1.25 ($1000)

Regression Specification To estimate the effect of TBA eligibility on mortgage rates using the IV strategy, we start with the following first-stage regression:

$$NoTBA_i = \alpha 1[h_i > h^*_{t(i)}] + g^-(h_i;\theta_0) + g^+(h_i;\theta_1) + Z_i\gamma + \xi_{s(i)\times l(i)\times t(i)} + \epsilon_i. \tag{3}$$

The dependent variable, $NoTBA_i$, is a dummy variable that equals one if loan i is included in a TBA-ineligible MBS. On the right hand side, h_i represents the house appraisal value associated with loan i, and $h^*_{t(i)}$ is the 125 percent of the national CLL that is effective for the year corresponding to the origination year-month $t(i)$ for loan i.[14] Thus, $1[h_i > h^*_{t(i)}]$, which is our instrument, is a dummy variable that is equal to one if the house appraisal value associated with loan i is greater than 125 percent of the national CLL. Based on Figure 5, we expect that the coefficient estimate for α is negative.

Next, $g^-(h_i;\theta_0)$ and $g^+(h_i;\theta_1)$ represent polynomials of the running variable h_i for values not greater than $h^*_{t(i)}$ and values greater than $h^*_{t(i)}$, respectively. Both θ_0 and θ_1 are the coefficients of the polynomials and will be estimated. We experiment with different degrees of polynomials to see how sensitive the estimate for the main coefficient, α, is.. Vector Z_i contains other loan and borrower characteristics relevant for loan pricing: credit score, loan-to-income ratio, whether a loan is originated by a broker, and whether a loan is originated by a correspondent lender. The next term $\xi_{s(i)\times l(i)\times t(i)}$ refers to the fixed effects for a combination of state $s(i)$, the lender $l(i)$, and origination year-month $t(i)$. With these fixed effects, we can flexibly control for any differences across states, mortgage lenders, and origination months.

[14]During our sample period, the national CLL was set at \$417,000 from 2009 to 2016, \$424,100 in 2017, and \$453,100 in 2018. Thus, the cutoff is \$521,250, \$530,125, and \$566,375, respectively.

Once we estimate the first stage regression, we estimate the following second stage regression:

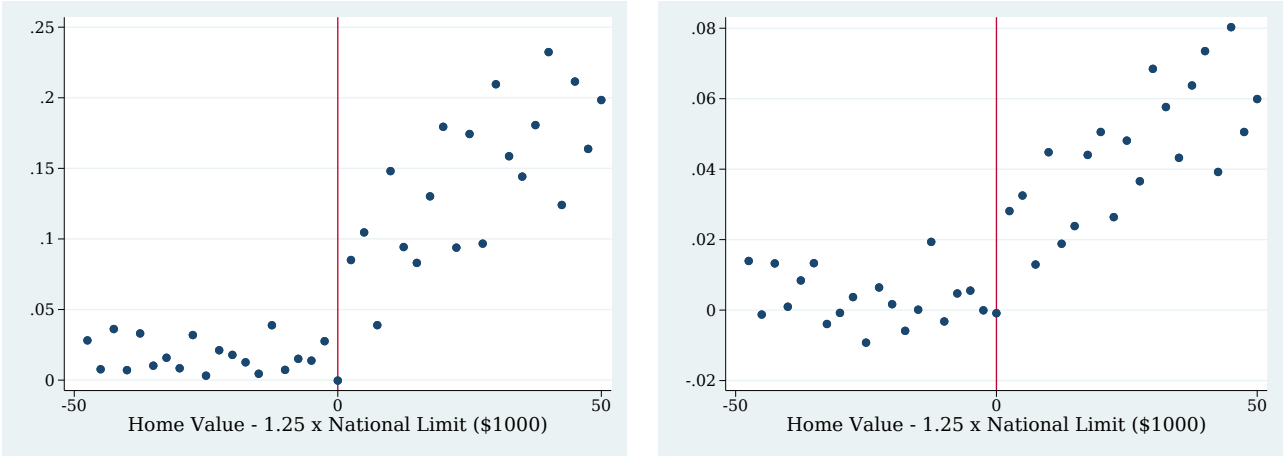
$$Rate_i = \beta \widehat{NoTBA}_i + g^-(h_i; \phi_0) + g^+(h_i; \phi_1) + Z_i\delta + \chi_{s(i)\times l(i)\times t(i)} + \omega_i \tag{4}$$

Based on the estimates of the first-stage regression, we calculate predicted value of the probability that loan i is not included in a TBA-eligible MBS ($\widehat{NoTBA}_i$). Then we use this variable in place of $1[h_i > h^*_{t(i)}]$ from Equation (3). The coefficient on $\widehat{NoTBA}_i$, β, estimates the impact of TBA eligibility on loan rates.

Graphical Examination of Discontinuities Before estimating the regressions, we first investigate whether there are visible discontinuities at the home appraisal value cutoff with respect to the probability of being included in TBA-ineligible MBS and mortgage rates. To precisely examine the relationship of the two variables and the home appraisal value, we remove variation in the two variables accounted for by the control variables Z_i and the fixed effects $\xi_{s(i)\times l(i)\times t(i)}$. For this purpose, we estimate a regression described by Equation (3) with the two variables ($NoTBA_i$ and $Rate_i$) as dependent variables, using the sample of loans with corresponding home values within the window of \$50,000 around the cutoff. Then by subtracting the estimate of $Z_i\gamma + \xi_{s(i)\times l(i)\times t(i)}$ from the dependent variable, we calculate the residual value of the dependent variable. We plot the residual dependent variable against the difference between the home appraisal value and 125 percent of the national CLL in Figure 6.

Discontinuities at the cutoff are clearly visible for both $NoTBA_i$ and $Rate_i$. Moreover, both panels have very similar patterns in terms of not only the jump at the cutoff but also the change in the slope. In both panels, the residual dependent variables do not change very much as the home appraisal value approaches to the cutoff from below. At the cutoff, both residual dependent variables increase discretely, and they increase as the home appraisal value moves upward from the cutoff. This similarity in the patterns shown in both panels indicates that TBA eligibility reduces mortgage rates.

Figure 6: **Probability to Be in TBA-ineligible MBS and Mortgage Rates (Cutoff 1)**: The figures plot the residual probability to be included in TBA-ineligible MBS (panel (a)) and the residual mortgage rate (panel (b)) against home values. The residual values are obtained by removing variation accounted by observable loan characteristics (Z_i) and the fixed effects ($\xi_{s(i)\times l(i)\times t(i)}$) after running regressions given by Equation (3) with $NoTBA_i$ and $Rate_i$ as dependent variables. Each dot in the plot represents the average value of each residual variable for each bin of the size of $2,500.



(a) NoTBA

(b) Mortgage Rate

IV Regression Results First, we estimate the first-stage regression described by Equation (3). Table 1 presents estimates of α, which measure the difference in the probability of being included in a TBA-ineligible MBS between loans just above the cutoff and loans just below the cutoff. Columns (1)–(3) display estimates with a subsample with loans for home values within the window of $50,000 around the cutoff. For instance, until 2016, this sample covers home values ranging from $471,250 to $571,250 with the national CLL equal to $417,000. Columns (4)–(6) display estimates with an even smaller subsample with the window of $25,000 around the cutoff. Until 2016, this sample covers home values ranging from $496,250 to $546,250. For each subsample, we experiment with different maximum numbers of polynomials for functions g^- and g^+ in Equation (3).

The table shows that loans just above the cutoff are more likely to be included in TBA-ineligible MBS than loans right below the cutoff. Although magnitudes of estimates are slightly different across specifications, the estimates show that borrowers purchasing homes just above the cutoff are more likely to originate high-balance loans, some of which will be included in TBA-ineligible MBS. With our preferred specification (column (3)), the probability to be included in TBA-ineligible MBS increases by 6 percentage points for loans just above the cutoff. This result is consistent with panel (a) of Figure 6, which shows a discrete jump in the probability by a similar magnitude at the cutoff.

Table 1: **First-Stage Results for Loans near Cutoff 1**: This table display estimates of coefficient α in Equation (3). Columns (1)–(3) are for the subsample of loans with home values within the window of \$50,000 around the cutoff. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively. Columns (4)–(6) are for the subsample of loans with home values within the window of \$25,000 around the cutoff. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All specifications include State x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of State x Lender x Month.

	Home Value: 1.25xCLL±\$50K			Home Value: 1.25xCLL±\$25K		
	(1)	(2)	(3)	(4)	(5)	(6)
	Polynomial=1	Polynomial=2	Polynomial=3	Polynomial=1	Polynomial=2	Polynomial=3
$1[h_i > h^*_{t(i)}]$	0.059***	0.040***	0.060***	0.026***	0.061***	0.088***
	(17.52)	(9.08)	(10.02)	(5.98)	(8.89)	(9.64)
STATExMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	78,535	78,535	78,535	37,891	37,891	37,891
Adj. R^2	0.22	0.22	0.22	0.16	0.16	0.16

Next, Table 2 displays the result from the second-stage regression. The table shows that the estimated effects of TBA eligibility on the mortgage rate are mostly similar across specifications. Our preferred estimate (Columns (3)) shows that TBA eligibility reduces the mortgage rate by around 40 basis points for loans near the national CLL. With the first-stage regression, we found that the probability to be included in TBA-eligible MBS increases by 6 percentage points for loans just above the cutoff. Then the second-stage estimate of 40 basis points implies that mortgage rates are higher for loans just above the cutoff by 2.4 basis points than loans just below the cutoff, which is consistent with the magnitude of the discrete jump shown in panel (b) in Figure 6.

Note that the estimate of 40 basis points does not measure the difference in mortgage rates between conventional conforming loans (not larger than the national CLL) and high-balance loans. Figure 2 shows that the about 65 percent of high-balance loans are still included in TBA-eligible MBS. Thus, the average rate spread between conventional conforming loans and high-balance loans should be about 14 basis points (= 0.35 × 40 basis points).[15]

[15]The actual spread between the two types of loans is time-varying. In fact, Vickery and Wright (2013) report that the spread was around 30 basis points in the beginning of 2009 and decreased to around 10 basis points by 2011. Our estimate of 14 basis points for the spread is the average across our sample period, which ranges from 2009 to mid-2018.

Table 2: **Second-Stage Results for Loans near Cutoff 1**: This table display estimates of coefficient β in Equation (4). Columns (1)–(3) are for the subsample of loans with home values within the window of \$50,000 around the cutoff. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively. Columns (4)–(6) are for the subsample of loans with home values within the window of \$25,000 around the cutoff. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All specifications include State x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of State x Lender x Month.

	Home Value: 1.25xCLL±\$50K			Home Value: 1.25xCLL±\$25K		
	(1)	(2)	(3)	(4)	(5)	(6)
	Polynomial=1	Polynomial=2	Polynomial=3	Polynomial=1	Polynomial=2	Polynomial=3
$\widehat{NoTBA}$	0.310***	0.275**	0.398***	0.521***	0.321**	0.321***
	(6.02)	(2.51)	(3.74)	(2.83)	(2.57)	(2.69)
STATExMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	77,898	77,898	77,898	37,565	37,565	37,565
Adj. R^2	0.87	0.87	0.86	0.86	0.87	0.87

Differences in Prepayments An important identifying assumption in our IV strategy is that unobserved characteristics correlated with mortgage rates are smooth at the cutoff. There is no direct way to test whether this assumption is true, but our data allow us to test it indirectly with ex-post prepayments. Because mortgages in our sample are guaranteed by the GSEs, default risks are viewed just as a source of prepayment risks from an MBS investor's perspective, and our ex-post prepayment measures include prepayments due to defaults.

We measure the ex-post prepayment by whether a loan was paid off, from an MBS investor's perspective, by the n-th month since the origination for $n \in \{24, 36, 48, 60, 72, 84\}$. Using these dummy variables as dependent variables, we estimate regressions similar to Equation (3) but with realized prepayment as the dependent variable. Because the loans in our sample are originated in 2009 or later, we only consider prepayment outcomes up to the 84th month since origination. Moreover, when considering whether a loan was paid off by the n-th month, we estimate the regression only with loans that could reach the loan age of n months without being paid off as of September 2018, when the latest prepayment data are available. For example, for $n = 48$, we exclude loans originated after September 2014 because the maximum loan age for such a loan would be 47 in September 2018, when the most recent performance data are available.

After estimating the regression for each n, we calculate the residual rate of prepayment by loan age n by removing variation accounted for by $Z_i\gamma + \xi_{s(i)\times l(i)\times t(i)}$. We then plot the residual rate of prepayment against the difference between the home value and the national CLL in Figure 7. The figures show no systematic changes in prepayment at the cutoff across all prepayment measures. Compared with the two panels in Figure 6, whose patterns lined up with each other, the patterns shown in Figure 7 do not seem to have any systematic relationships with the change in the probability of being included in a TBA-ineligible MBS shown in panel (a) of Figure 6. Therefore, this finding indicates that the discontinuity in mortgage rates at the cutoff is unlikely to be driven by changes in unobserved characteristics at the cutoff.

Figure 7: **Prepayment Probabilities around Cutoff 1**: These figures plot the residual probability that a loan is completely paid off by different loan ages in terms of months since origination. The x-axis represents the home value associated with each loan relative to the cutoff in thousand dollars. The residual variables are obtained by removing variation in corresponding original variables accounted by observable loan characteristics after running regressions given by Equation (3) with the original variables as dependent variables ($Z_i\gamma + \xi_{s(i)\times l(i)\times t(i)}$). Each dot in the plot represents the average value of each residual variable for each bin of the size of $2,500.

In addition, regression estimates reported in Table 3 show that there are no discontinuities in ex-post prepayments at the cutoffs. In the Appendix, Figure 16 and Table 10 show similar patterns with an alternative measure of ex-post prepayments, which is the ratio of original balance paid off by loan age n. This measure captures partial payoffs, whereas the original measure only captures complete payoffs. This set of evidence suggests that the estimated impact on the mortgage rate is unlikely to reflect differences in unobservables around the cutoff.

Table 3: **Regression Results for Prepayment Probabilities (Cutoff 1)**: This table displays the estimates of the regression similar to Equation (3), but where dependent variables are the dummy variable that is equal to one if a loan is completely paid off by loan age n for $n \in \{24, 36, 48, 60, 72, 84\}$. The maximum degree of polynomials included in the regressions are two for each regression. For all columns, we used the subsample of loans with corresponding home values within the window of $50,000 around the cutoff. For each column, we further restricted the subsample to loans that were originated at least n months before the most recent month available in the data (2018m9). All specifications include State x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of State x Lender x Month.

	(1) By Age 24	(2) By Age 36	(3) By Age 48	(4) By Age 60	(5) By Age 72	(6) By Age 84
$1[h_i > h^*_{t(i)}]$	-0.005	-0.020	-0.011	-0.034	-0.050	-0.021
	(-0.27)	(-0.86)	(-0.41)	(-1.20)	(-1.59)	(-0.69)
h_i	-0.000	0.004**	0.003	0.000	0.000	0.000
	(-0.16)	(2.11)	(1.24)	(0.16)	(0.19)	(0.21)
$1[h_i > h^*_{t(i)}] \times h_i$	0.001	-0.004	-0.002	0.003	0.007	0.001
	(0.29)	(-0.89)	(-0.54)	(0.53)	(1.31)	(0.22)
STATExMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	35,443	29,761	25,066	20,564	15,614	12,434
Adj. R^2	0.14	0.20	0.24	0.25	0.20	0.09

Exogenous Variables around the Cutoff An identifying assumption with the regression discontinuity design is that the sample is selected randomly around the cutoff. A way to test for a random selection is to test whether exogenous variables exhibits any discrete jumps at the cutoffs. Figure 8 plot the residual values of exogenous loan characteristics against the difference between the home appraisal value and 125 percent of the national CLL. We consider exogenous loan characteristics included in Z_i in Equations (3) and (4). The figure shows that there is no noticeable jump in any of the four variables. In appendix, Table 11 also confirms that there is no statistically significant jump at the cutoff for any of the four variables.

Figure 8: **Exogenous Variables around Cutoff 1**: The figures plot the residual values of exogenous loan characteristics against home values. The residual values are obtained by removing variation accounted by the fixed effects ($\xi_{s(i)\times l(i)\times t(i)}$) after running regressions given by Equation (3) with the exogenous loan characteristics as dependent variables. Each dot in the plot represents the average value of each residual variable for each bin of the size of $2,500.

5.2 Loan-To-Value 105

In the previous section, we showed that TBA eligibility reduces mortgage rates for loans near the national CLL. In this subsection, we similarly estimate the effect of TBA eligibility on mortgage rates for loans with LTVs near 105.

As shown by panel (b) in Figure 3, there is bunching at LTV of 105, and origination shares for LTVs above 105 seems to be discretely lower than origination shares for LTV right below 105. This discontinuity in the distribution of LTV at origination poses a challenge to an identification strategy that utilizes the cutoff of LTV 105. Those who originate loans at or right below the LTV 105 might have different unobserved characteristics from those who originate loans just above the cutoffs. Thus difference in the mortgage rates around the cutoff may be accounted for by the potential difference in unobserved characteristics.

We also address this problem with an instrument variable strategy, which utilizes an alternative cutoff based on the ending balance of the original loan that preceded the refinanced HARP loan. When refinancing into a HARP mortgage, the borrower needs to pay a closing cost. This cost varies

across lenders and can be thousands of dollars. A borrower can roll the closing cost into the new balance, which can make the new loan balance higher than the ending balance of the preceding loan. Freddie Mac imposes a limit on how much of the closing cost can be included in the balance of the new HARP loan: the lesser of 4% of the balance of the preceding loan and $5,000.[16] This rule restricts the size of the HARP loan to a limit that depends on the ending balance of the preceding loan. As a result, the maximum LTV of the new HARP loan is a function of the ending balance of the preceding loan:

$$PredLTV_i \equiv \frac{\min\{PrevBalance_i \times 1.04, PrevBalance_i + 5000\}}{UpdatedHomeValue_i} \times 100 \geq LTV_i.$$

We argue that $PredLTV_i$ predicts the probability that a HARP loan would be included in a TBA-ineligible MBS. If $PredLTV_i \leq 105$, then we expect that the LTV of a new HARP loan securitized by Freddie Mac is likely to be not greater than 105. Otherwise, the probability that the LTV of a HARP loan is greater than 105 will increase as $PredLTV_i$ increases beyond 105. Because any HARP loan with the LTV greater than 105 must be included in a TBA-ineligible MBS, we expect that the relationship between $PredLTV_i$ and the probability of being included in a TBA-ineligible MBS changes discontinuously at $PredLTV_i$ of 105.

Loan-level data from Freddie Mac allow us to calculate $PredLTV_i$ for each HARP origination. $UpdatedHomeValue_i$ can be obtained by multiplying the new LTV and the new loan amount for each HARP origination. Since the data provide a link between each HARP loan and its preceding loan, we obtain $PrevBalance_i$ by looking at the outstanding balance of the preceding loan in the month right before HARP refinancing.

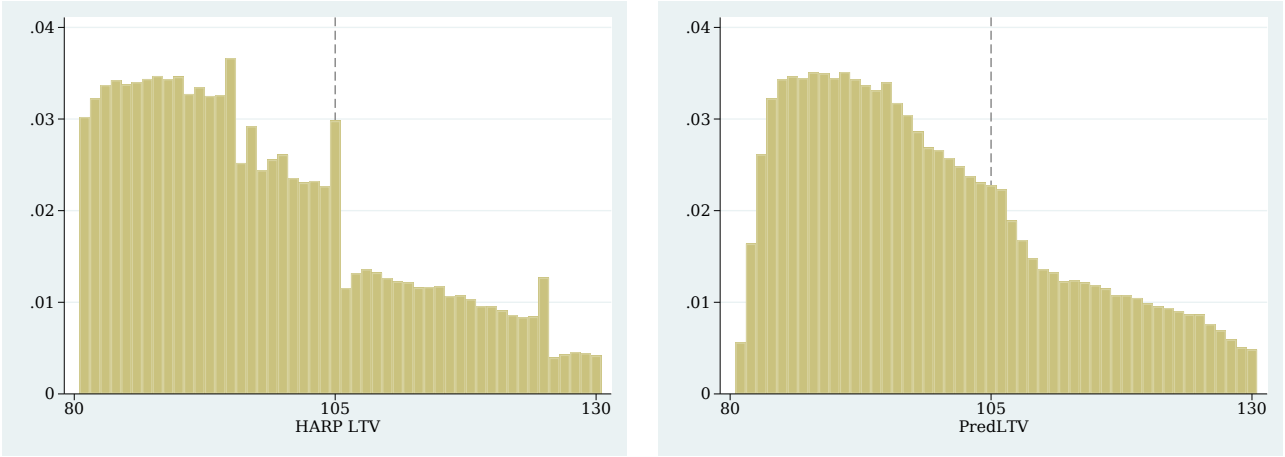
For this analysis, we impose the following additional sample selection criteria to keep the sample relatively homogeneous. First, we only keep HARP loans securitized by Freddie Mac because Fannie Mae did not have similar restrictions on closing costs. Second, we keep only the loans for owner-occupied single-family houses as we did for our analysis for loans near the national CLL in Section 5.1. Moreover, we exclude any loans with second mortgages (by comparing combined LTV and original LTV) or any loans with mortgage insurance.[17]

Figure 9 suggests that the density of $PredLTV_i$ is smooth at the cutoff of 105 (panel (b)) unlike the density of LTVs of HARP loans (panel (a)). The smooth density of $PredLTV_i$ suggests that there is no systematic manipulation of $PredLTV_i$, which satisfies a key condition for identification.

[16] This information on the limit on how much the closing cost can be included in the new balance is provided on page 15 of the evaluation report by the Office of Inspector General of the Federal Housing Finance Agency on the HARP program. The link to the report is: https://www.fhfaoig.gov/Content/Files/EVL-2013-006.pdf.

[17] HARP allowed borrowers to refinance without mortgage insurance although their updated LTVs are greater than 80. However, there are a small group of HARP borrowers who still had mortgage insurance.

Figure 9: **Sorting around the Cutoffs**: These figures plot loan-level density for HARP originations securitized by Freddie Mac. Panel (a) plots the density of the LTV at origination for a HARP loan. Panel (b) plots the density of $PredLTV_i$. The data source of both figures is the loan-level data from Freddie Mac.



(a)HARP LTV (b) PredLTV

Graphical Examination Panel (a) of Figure 10 displays the relationship between $PredLTV_i$ and the probability of being included in a TBA-ineligible MBS. As before, we calculate the residual probability of being included in a TBA-ineligible MBS with the following regression:

$$NoTBA_i = \alpha 1[PredLTV_i > 105] + g^-(PredLTV_i;\theta_0) + g^+(PredLTV_i;\theta_1) + Z_i\gamma + \xi_{zip3(i)\times l(i)\times t(i)} + \epsilon_i \quad (5)$$

The dependent variable, $NoTBA_i$ is a dummy variable that is equal to one if loan i is included in TBA-eligible MBS. Note that in this setting, a HARP loan with the initial LTV above 105 is not allowed to be included in TBA-eligible MBS. The dummy variable, $1[PredLTV_i > 105]$ is equal to one if a HARP loan i's $PredLTV$ is above 105. Similarly to the analysis for loans near the national CLL, we include up to the third-degree polynomials of $PredLTV_i$ interacted with $1[PredLTV_i > 105]$,which are captured by the two functions g^- and g^+. Next, Z_i include other loan characteristics: credit score, whether a loan is originated by a broker, whether a loan is originated by a correspondent lender, and the mortgage rate for the previous loan. Lastly, [18] $\xi_{zip3(i)\times l(i)\times t(i)}$ refers to the fixed effects for a combination of first three digits of zipcodes, mortgage lenders, and loan origination months.

After estimating Equation 5, we calculate the residual value of $NoTBA_i$ by removing variation in the probability accounted for by loan characteristics Z_i and fixed effects. The figure shows that

[18]Loan characteristics in Z_i in Equation 5 are not exactly the same as loan characteristics included for the analysis for loans near the national CLL. This is in part because different data sets are used for the analyses for loans near the national CLL and HARP loans. For example, a borrower's income is missing for the data set for HARP loans, whereas we can infer a borrower's income in the data set for loans near the national CLL. Thus, we do not include the loan-to-income ratio in Equation 5.

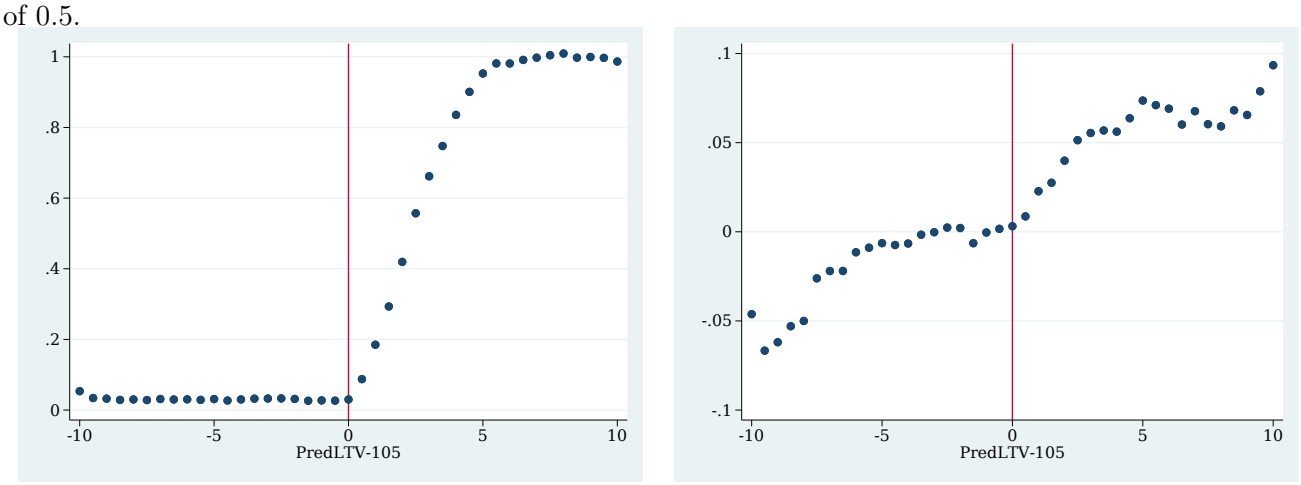
although there is no jump at the cutoff, the slope changes discontinuously, which results in a kink at the cutoff. As $PredLTV_i$ approaches to 105 from below, the slope of the graph is almost flat. In contrast, as $PredLTV_i$ moves away from the cutoff value, the slope of the graph becomes suddenly much steeper, indicating that the same amount of an increase in $PredLTV_i$ makes the LTV of HARP loan much more likely to be greater than 105. Based on this pattern, it is evident that there is a kink at the cutoff in the relationship between $PredLTV_i$ and the probability of being included in a TBA-ineligible MBS.

Panel (b) of Figure 10 displays the relationship between $PredLTV_i$ and residual mortgage rates of HARP loans. As in panel (a), we calculate the residual rate by removing variation in mortgage rates accounted for by loan characteristics Z_i and the fixed effects $\xi_{zip3(i)\times l(i)\times t(i)}$. As in panel (a), there is also a kink in the relationship between $PredLTV_i$ and the residual mortgage rate at the cutoff, which suggests that the change in the relationship with the residual mortgage rate at the cutoff is correlated with TBA eligibility.

An important pattern shown by Figure 10 is that there is a kink instead of a jump at the cutoff, whereas there is a jump at the cutoff for loans near the national CLL (displayed in Figure 6). The main reason for the pattern is that $PredLTV_i$ is calculated based not on actual closing costs but on the largest possible closing costs that can be rolled into a HARP balance. In contrast, the appraised home value, which is the IV used for the analysis for loans near CLLs, is the actual home value, not the maximum possible home value. If closing costs for HARP ends up being smaller than the upper limit, then LTVs for new HARP loans will be below $PredLTV_i$. Thus, if closings cost are usually below the upper limit, then actual LTVs for HARP loans will not be very different regardless of whether $PredLTV_i$ is below or above the cutoff of 105. In that case, there would not be a jump at the cutoff as shown in Figure 10. However, as $PredLTV_i$ increases away from the cutoff, even closing costs smaller than the upper limit will be more likely to result in actual LTVs for HARP loans higher loan 105, which will lead to the upward slopes for $PredLTV_i$ above 105 as shown in Figure 10.

Another notable pattern is that the relationships between $PredLTV_i$ and the two variables are non-linear for values of $PredLTV_i$ away from the cutoff. For example, the slope for the mortgage rate in panel (b) also seems to changes for $PredLTV_i$ near -5. Thus, it is important to control for these non-linear relationship by including higher-degree polynomials especially when using a subsample with nonlinearity. That is because the linear terms by themselves with the larger subsample cannot estimate the change in the slopes in the relationship between $PredLTV_i$ and the mortgage rate at the cutoff.

Figure 10: **TBA-eligible Probabilities and Mortgage Rates (Cutoff 2)**: The figures plot the residual probability to be included in TBA-ineligible MBS (panel (a)) and the residual mortgage rate (panel (b)) against $PredLTV_i$. The residual variables are obtained by removing variation in corresponding original variables accounted by observable loan characteristics after running regressions given by Equation (5) with the original variables as dependent variables ($Z_i\gamma + \xi_{zip3(i)\times l(i)\times t(i)}$). Each dot in the plot represents the average value of each residual variable for each bin of the size of 0.5.



(a) NoTBA

(b) Mortgage Rate

Regression Specifications Figures 10 suggests that the standard regression discontinuity design will not work well in this setup because there is no jump in the probability of being included in a TBA-ineligible pool at the cutoff. Instead, we use the regression kink design, which estimates the treatment effect by estimating changes in the slopes at the cutoff in the relationship between a running variable and dependent variable. In this empirical design, we run a two-stage-least-square regression with the first- and second-stage regressions as follows:

$$\begin{aligned} NoTBA_i = {} & \alpha_0 PredLTV_i + \alpha_1 PredLTV_i \times 1[PredLTV_i > 105] \\ & + g^-(PredLTV_i;\theta_0) + g^+(PredLTV_i;\theta_1) + Z_i\gamma + \xi_{zip3(i)\times l(i)\times t(i)} + \epsilon_i \end{aligned} \tag{6}$$

$$\begin{aligned} Rate_i = {} & \beta_0 PredLTV_i + \beta_1 \widehat{NoTBA}_i \\ & + g^-(PredLTV_i;\phi_0) + g^+(PredLTV_i;\phi_1) + Z_i\delta + \chi_{zip3(i)\times l(i)\times t(i)} + \omega_i \end{aligned} \tag{7}$$

In this empirical design, Equations (6) and (7) are the first- and second-stage regressions, respectively. In Equation (6), α_1 measures the change in slopes at the cutoff. The variable $PredLTV_i \times 1[PredLTV_i > 105]$, which only shows up in Equation (6), serves as an IV. The coefficient β_1 measures the treatment effect of being included in a TBA-ineligible MBS.

IV Regression Results First, we estimate the first-stage regression described by Equation (6). Table 4 presents coefficient estimates for the first stage with six different specifications, which measure how much slopes between $PredLTV_i$ and the probability to be included in TBA-ineligible

MBS changes at the cutoff. Columns (1)–(3) present estimates with the subsample with loans with $PredLTV_i \in [95, 115]$, which has the window size of 10 around the cutoff of 105. Columns (4)–(6) present estimates with the subsample with loans with $PredLTV_i \in [100, 110]$, which has the window size of 5 around the cutoff of 105. For each subsample, we experiment with different maximum numbers of polynomials for functions g^- and g^+ in Equation (6).

The table shows that the slope is more positive for $PredLTV_i$ above 105, which is consistent with Figure 10. These estimates shows that a marginal increase in $PredLTV_i$ is much more likely to result in a HARP loan with the LTV above 105 if $PredLTV_i$ is above 105. This result is consistent with panel (a) of Figure 10, which shows a kink in the probability.

Table 4: **First-Stage Results with Cutoff 2**: This table display estimates of coefficients in Equation (6). Columns (1)–(3) are for the subsample of loans with $PredLTV_i \in [95, 115]$. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively Columns (4)–(6) are for the subsample of loans with $PredLTV_i \in [100, 110]$. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All specifications include Zip3 x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of Zip3 x Lender x Month.

	$PredLTV$: [95,115]			$PredLTV$: [100,110]		
	(1) Polynomial=1	(2) Polynomial=2	(3) Polynomial=3	(4) Polynomial=1	(5) Polynomial=2	(6) Polynomial=3
$PredLTV$	0.014***	-0.005***	-0.023***	0.002	-0.014***	0.032***
	(37.75)	(-4.01)	(-8.22)	(1.36)	(-3.11)	(2.95)
$1[PredLTV > 105] \times PredLTV$	0.107***	0.284***	0.329***	0.205***	0.249***	0.089***
	(138.22)	(114.31)	(45.86)	(79.25)	(23.12)	(3.52)
ZIP3xMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	67,466	67,466	67,466	26,473	26,473	26,473
Adj. R^2	0.76	0.82	0.82	0.63	0.63	0.63

The second-stage regression results are reported in Table 5. Note that because the relationship between $PredLTV_i$ and the mortgage rate are highly non-linear, the estimate can be misleading withjust up to the first-degree polynomials. In Column (1), our point estimate is negative (but statistically insignificant) although Figure 10 is indicative of at least a positive coefficient. Our estimates become consistent with the figure as we include higher-degree polynomials (Columns (2) and (3)) or as we use a narrower sample window (Columns (4)–(6)). When we use the smaller sample window and include up to the third-degree polynomials (Column (6)), the point estimate of the impact of TBA eligibility on mortgage rates is the largest but statistically insignificant because the standard error of the estimate becomes very large with the third-degree polynomials with a relatively small number of observations with the second subsample. Our preferred estimate (Columns (3)) shows that TBA eligibility reduces the mortgage rate by 10 bps for HARP loans with LTVs around 105.

Table 5: **Second-Stage Regression Results with Cutoff 2**: This table display estimates of coefficients in Equation (7). Columns (1)–(3) are for the subsample of loans with $PredLTV_i \in [95, 115]$. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively Columns (4)–(6) are for the subsample of loans with $PredLTV_i \in [100, 110]$. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All specifications include Zip3 x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of Zip3 x Lender x Month.

	$PredLTV$: [95,115]			$PredLTV$: [100,110]		
	(1) Polynomial=1	(2) Polynomial=2	(3) Polynomial=3	(4) Polynomial=1	(5) Polynomial=2	(6) Polynomial=3
$\widehat{NoTBA}$	-0.005	0.079***	0.075***	0.061***	0.067**	0.148
	(-0.74)	(7.73)	(4.31)	(5.14)	(2.20)	(0.73)
ZIP3xMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	66,632	66,632	98,754	26,107	39,900	39,900
Adj. R^2	0.85	0.85	0.84	0.86	0.85	0.85

Differences in Prepayments To check whether there is a kink at the cutoff for the relationship between $PredLTV_i$ and unobserved characteristics, we investigate whether the relationship between $PredLTV_i$ and the ex-post prepayment changes at the cutoff. Similar to Figure 7, we also measure the ex-post prepayment by whether a borrower paid off the loan by the n-th month since the origination for $n \in \{24, 36, 48, 60, 72, 84\}$. Using these dummy variables as dependent variables, we estimate regressions similar to Equation (5). Similarly to the earlier analysis for high-balance loans, moreover, we also estimate the regression only with loans that could reach the loan age of n months without being paid off as of September 2018, when the latest prepayment data are available. After estimating the regression for each n, we calculate the residual rate of prepayment by loan age n by removing variation accounted for by loan characteristics Z_i and the fixed effects.

Table 6: **Regression Results for Prepayment Probabilities (Cutoff 2)**: This table displays the estimates of the regression similar to Equation (3), where dependent variables are the dummy variable that is equal to one if a loan is completely paid off by loan age n for $n \in \{24, 36, 48, 60, 72, 84\}$. The maximum number of polynomials included in the regressions are three for each column. For all columns, we used the subsample of loans with $PredLTV_i$ between 100 and 110. For each column, we further restricted the subsample to loans that were originated at least n months before the most recent month available in the data (2018m9). All specifications include Zip3 x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of Zip3 x Lender x Month.

	(1) By Age 24	(2) By Age 36	(3) By Age 48	(4) By Age 60	(5) By Age 72	(6) By Age 84
$1[PredLTV > 105]$	0.004	-0.005	0.003	0.002	-0.009	-0.024
	(1.51)	(-1.54)	(0.74)	(0.27)	(-0.80)	(-1.12)
$PredLTV$	-0.002	-0.002	-0.002	-0.003	-0.002	0.022**
	(-1.16)	(-1.09)	(-0.64)	(-0.72)	(-0.25)	(2.17)
$1[PredLTV > 105] \times PredLTV$	0.001	0.004	0.004	0.003	0.012	-0.019
	(0.41)	(1.41)	(1.20)	(0.56)	(1.08)	(-0.93)
ZIP3xMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	61,744	54,914	44,277	23,311	11,282	3,481
Adj. R^2	-0.02	0.01	-0.00	0.00	-0.03	-0.06

The regression estimates are reported in Table 6, and we also plot the residual rate of prepayment against $PredLTV_i$ in Figure 11. The table shows that any changes in the slopes are not statistically significant. The figure is consistent with the results displayed by the table. It is apparent that no systemic changes in the relationship between the ex-post prepayment and $PredLTV_i$ at the cutoff for all measures of prepayment we considered. Therefore, this finding indicates that the discontinuity in mortgage rates at the cutoff is unlikely to be driven by changes in unobserved characteristics at the cutoff.

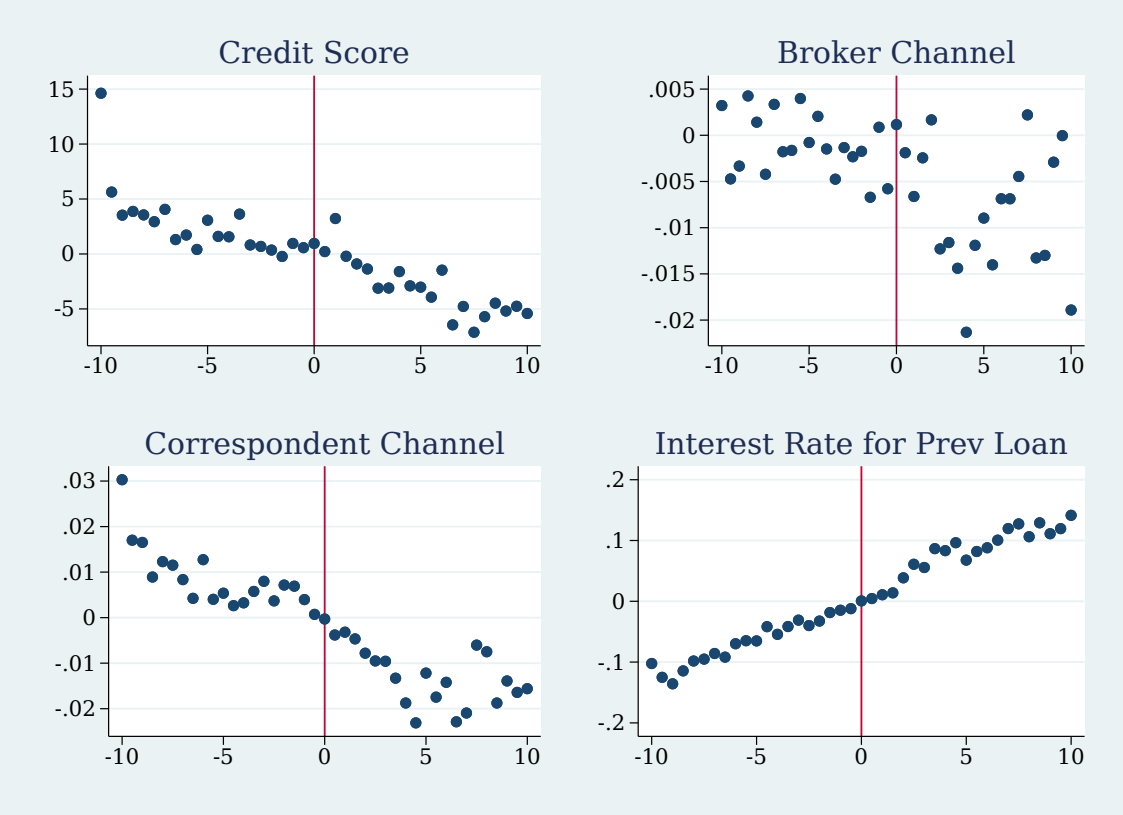
Figure 11: **Prepayment Probabilities around Cutoff 2**: These figures plot the residual probability that a loan is completely paid off by different loan ages in terms of months since origination. The residual variables are obtained by removing variation in corresponding original variables accounted by observable loan characteristics after running regressions given by Equation (5) with the original variables as dependent variables ($Z_i\gamma + \xi_{zip3(i)\times l(i)\times t(i)}$). Each dot in the plot represents the average value of each residual variable for each bin of the size of 0.5.

In the Appendix, Figure 17 and Table 12 show similar patterns with an alternative measure of ex-post prepayments, which is the ratio of original balance paid off by loan age n. This measure captures partial payoffs, whereas the original measure only captures a complete payoff. This set of evidence suggests that the estimated impact on the mortgage rate is unlikely to reflect difference in unobservables around the cutoff.

Exogenous Variables around the Cutoff We also test for a random selection with respect to exogenous variables. Figure 12 plot the residual values of exogenous loan characteristics against $PredLTV_i$. We consider exogenous loan characteristics were included in Z_i in Equations (5). The figure shows that there is no noticeable pattern in any of the four variables. In appendix, Table 13 report estimated coefficients for each dependent variable. Although the coefficient for the interaction between $1[PredLTV_i > 105]$ and $PredLTV_i$ is significant, the pattern is not robust to different

specifications. In fact, the corresponding figure suggests that there is no noticeable pattern for loans very close to the cutoff. In an alternative specification with only up to first- or second-degree polynomials of $PredLTV_i$ or the smaller subsample, the coefficient becomes statistically insignificant even at the 90% level.

Figure 12: **Exogenous Variables around Cutoff** 2: The figures plot the residual values of exogenous loan characteristics against $PredLTV_i$. The residual values are obtained by removing variation accounted by the fixed effects ($\xi_{zip3(i)\times l(i)\times t(i)}$) after running regressions given by Equation (5) with the exogenous loan characteristics as dependent variables. Each dot in the plot represents the average value of each residual variable for each bin of the size of 0.5.



5.3 Discussion

We have estimated the effect of TBA-eligibility on the mortgage rate, exploiting the two cutoffs that determine TBA-eligibility. We found that TBA eligibility reduces the mortgage rate by 40 basis points for loans near the national CLL and by 10 basis points for loans with LTVs near 105. The difference in magnitudes of the estimates is consistent with the prediction from our model that the option to trade in the TBA market is more valuable for loans with higher prepayment risks because they are less likely to trade in SP.

A common criticism against empirical designs based on discontinuities estimating local treatment effects is that the resulting estimate might be difficult to be extrapolated to the rest of the population. This concern would apply to our setup if we estimated the impact on the mortgage

rate using only one of the two cutoffs. However, the two cutoffs used in our empirical analysis are at opposite ends of the spectrum of prepayment risks. Thus, the estimated impact on the mortgage rates with the two cutoffs are likely to be close to the upper and lower bounds. Moreover, given the model prediction that the benefit of TBA eligibility is higher for loans with higher prepayment risks, we expect that the benefit of TBA-eligibility will fall between our two estimates for a majority of loans, which are likely to have prepayment risks toward the middle of the distribution of prepayment risks.

6 Effects on Refinancing Behavior and Consumer Spending

In the previous section, we have established that loans included in TBA-eligible pools have lower interest rates. Because TBA eligibility impacts the price, we expect that it would impact the quantity, or in other words, borrowers' demand for mortgages. In this section, we specifically investigate whether TBA eligibility affects borrowers' refinancing behavior. Previous research such as Agarwal et al. (2017) and Abel and Fuster (2018) find that refinancing is important for monetary policy transmission because consumer spending increases subsequent to refinancing. Thus, we also investigate how TBA eligibility affects consumer spending as well.

For this analysis, we focus only on refinancing behavior of borrowers with remaining loan balances near the national CLL. We do not investigate refinancing behavior with LTVs near 105 because of data limitations. Our data provides exact information about evolution of a borrower's loan balance over time. In contrast, we do not have good information about the evolution of updated LTV. Because our analysis hinges on differences in borrowers' behavior at the cutoffs, it is important to observe any differences in a borrower's decision to refinance depending on whether his updated LTV is above or below the cutoff. However, our data only allows us to observe updated LTVs only when a borrower refinances into a HARP loan, and we do not observe updated LTVs for borrowers who do not refinance at a given time.

6.1 Refinancing and National CLL

The period after when the GSEs were allowed to purchase and securitize high-balance loans (since March 2008) has experienced historically low interest rates, which resulted in a refinancing boom. We investigate whether TBA eligibility affects the refinancing decision of a borrower with a remaining loan balance near the national CLL after 2008 when the GSEs purchased high-balance loans. As in the earlier analysis on mortgage rates, we do not consider 2008 because there was significant uncertainty regarding TBA eligibility of high-balance loans.

A borrower seeking to refinance with a remaining balance above the national CLL faces the following trade-off: refinancing now into a high-balance loan with a higher rate versus refinancing later into a conventional conforming loan with a lower rate.[19] Having the trade-off in mind, we

[19] A borrower could always refinance into a conventional conforming loan if he makes a sufficient lump-sum mortgage payment. However, this possibility is not very likely.

investigate whether the refinancing probability increases for borrowers with remaining loan balances right below the national CLL relative to those with remaining balances right above the cutoff. We will interpret the difference in the refinancing probabilities as demand response to the spread in mortgage rates around the cutoff.

Another possibility is that a borrower cannot refinance into a high-balance loan because a higher rate associated with a high-balance loan makes debt-to-income (DTI) ratio binding. A DTI ratio is calculated as the monthly mortgage payment divided by a borrower's income. Thus, the larger mortgage rate is, the higher DTI is. Thus, even if a borrower would like to refinance into a high-balance loan, a lender might not be willing to extend a loan to the borrower because of the binding DTI. In this case, we would also expect an increase in the refinancing probability when a borrower's remaining balance decreases to right below the national CLL because of higher rates associated with high-balance loans. For our purpose, it is not very important to distinguish different reasons why refinancing volumes increase abruptly when the remaining mortgage balance reaches the national CLL, as long as the increase is due to the rate differential between conventional conforming and high-balance loans.

Sample Selection and Data For this analysis, we restrict the sample to pairs of a loan and a month with 30-year FRMs that were originated in 2007 or later with remaining balances greater than the national CLL at any point in March 2009 or later. Many of these loans in the sample were not securitized by the GSEs, including jumbo loans and loans kept on lenders' portfolios, unlike our analysis on mortgage rates in Section 5. Moreover, we only consider borrowers in high-cost counties where the Economic Stimulus Act of 2008 increased the CLL at least by $50,000. This geographical sample restriction is important because we would like to study a borrower's trade-off between a conventional conforming loan and high-balance loan, the latter of which is available only in high-cost counties.

We further restrict the sample to loan-month pairs with remaining balances within $50,000 around the national CLL at some point in our sample period (January 2009 or later). We exclude borrowers with adjustable-rate mortgages (ARMs) because their incentives to refinance are different from those with FRMs. ARM borrowers often refinance to avoid higher rates after the end of the initial period with fixed teaser rates, whereas FRM borrowers refinance to take advantage of lower current rates. Moreover, we only consider loans originated in 2007 or later because loans that were originated earlier and remain in the sample in our sample period may cause a selection bias.

For this analysis, we use the CRISM data, which provide loan-level mortgage performance information matched to borrower-level credit records. The main advantage of using CRISM over a typical loan-level performance data is that CRISM allows us to tell apart different reasons for a voluntary payoff of a mortgage such as plain refinancing, cash-out refinancing, moving to a different home, etc. Moreover, the data also provide information about a borrower's other credit activities such as auto financing, etc. In investigating the effects on consumption, we will focus on auto financing, which is used to purchase a car.

Another desirable feature of the CRISM data is that the data provide information for loans that are not securitized by the GSEs. Because GSEs were allowed to purchase high-balance loans starting in 2008, many loans with remaining balance above the national CLL in 2009 or later are jumbo loans originated before the high-cost loan limits were introduced and were not able to be purchased by GSEs. Hence, our sample includes many cases where the original loans were not securitized by the GSEs and were either packaged into private-label mortgage securities or kept on lenders' balance sheets.

Empirical Design We estimate the following regression:

$$y_{it} = \alpha 1[bal_{it} \leq CLL_t] + g^+(bal_{it}; \theta_0) + g^-(bal_{it}; \theta_1) + Z_{it}\gamma + \xi_{zip(i)\times t} + \epsilon_i. \tag{8}$$

Equation (8) looks quite similar to Equation (3), which was used for earlier analyses. Whereas the unit of analysis is a loan at the time of its origination in the previous analyses, the unit of analysis here is a loan-month pair. Thus, the dependent variable y_{it} has two subscripts: i for each loan and t for each month. On the right hand side, bal_{it} refers to the remaining balance of loan i as of time t, and $1[bal_{it} \leq CLL]$ is a dummy variable that is equal to one if the remaining balance of loan i in time t is not greater than the national CLL at that time. Functions g^+ and g^- give polynomials of bal_{it} depending on whether or not bal_{it} is greater than CLL_t, respectively. Vector Z_{it} includes the following loan characteristics: loan age, the purpose of the loan (refinance or purchase), whether a loan is kept on a lender's balance sheet, whether a loan is securitized by a GSE, updated estimated LTV, the fraction of the initial balance paid off as of time t, original loan balance, updated credit score, mortgage interest rate, LTV at origination, whether the loan is a first mortgage, whether the borrower is an owner-occupant, whether there is a prepayment penalty, whether the prepayment penalty period expired by time t, whether there is a delinquency in last twelve months, whether the loan is an interest-only loan, and whether the interest-only period expired by time t. Lastly, $\xi_{zip(i)\times t}$ refers to the fixed effects at the level of loan i's zipcode and time t. Because the CRISM data do not provide identities of lenders or servicers, we are not able to include lender or servicer characteristics unlike in the previous analyses on mortgage rates.

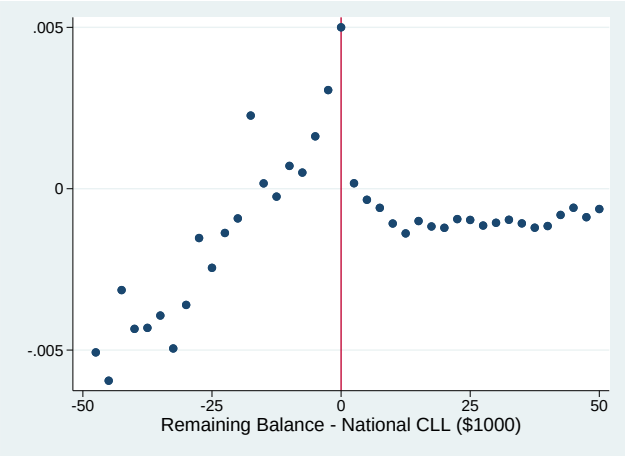
The dependent variable, y_{it}, is an indicator variable that equals one if loan i is refinanced at time t. We consider the following two types of refinancing separately: plain refinancing and cash-out refinancing. Plain refinancing refers to refinancing without a significant increase in the loan balance. Since the loan balance could increase because of the closing cost of a new loan, we view refinancing as plain refinancing if the loan balance does not increase more than 5% of the ending balance of the previous loan. Cash-out refinancing is refinancing in which a borrower increases the loan balance by more than 5% of the ending balance of the previous loan. We expect that the probability of plain refinancing increases discontinuously when the remaining balance of a loan is right below the national CLL because a borrower can refinance into a conventional conforming loan without making additional mortgage payments. In contrast, we do not expect to see a similar pattern for cash-out refinancing because cash-out refinancing for a borrower with the remaining balance right below the

national CLL will make the new loan balance greater than national CLL.

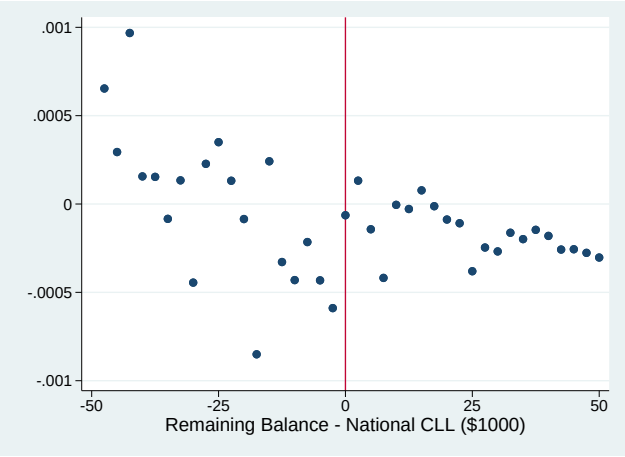
Graphical Examination Before presenting estimated coefficients, we first visually examine the relationship between the remaining loan balance and the probability of plain and non-plain refinancing. Similar to earlier analyses on mortgage rates, we also run the regression described in Equation (8) and calculate the residual probability of refinancing a loan by removing the variation in the dependent variable accounted for by loan characteristics Z_i and the fixed effects. Figure 13 displays the relationship between the remaining loan balance and the probability of plain and non-plain refinancing in panels (a) and (b), respectively. Panel (a) clearly shows that there is a jump in the probability of plain refinancing at the CLL. A borrower with a remaining balance just below the national CLL is about 0.50 pp more likely to plain-refinance than a borrower with a remaining balance slightly above the cutoff. Given that the average monthly probability to pay off the loan is 1.1 percent, the difference of 0.50 pp in the probability of plain refinancing at the cutoff is economically significant. As the remaining balance increases away from the national CLL, the probability does not change very much. In contrast, the probability decreases fast as the remaining loan balance decreases away from the national CLL. This pattern suggests that many borrowers wait for their loan balances to reach below the CLL to refinance into mortgages not greater than the national CLL. It also suggests that very few borrowers make extra mortgage payments to refinance into loans smaller than the national CLL.

In contrast, panel (b) shows that the probability of cash-out refinancing does not exhibit a similar pattern around the national CLL. Because cash-out refinancing of a loan with a remaining balance right below the cutoff will still make a borrower refinance into a high-balance loan, there is no reason for a discrete jump at the national CLL.

Figure 13: **Monthly Probability of Plain and Cash-out Refinancing around the National CLL**: This figure plots the relationship between the residual monthly probabilities of plain and cash-out refinancing and the remaining loan balance. The residual probability is obtained by removing variation in the original variable accounted for by observable characteristics Z_{it} and fixed effects $\xi_{zip(i)\times t}$ after running the regression given by Equation (8). Each dot represents the average value for each bin with the size of $2,500.



(a) Plain Refinancing



(b) Cash-out Refinancing

Regression Estimates Table 7 shows the results of regression (8) with plain refinancing as the dependent variable. Columns (1)–(3) show estimates with the sample consisting of pairs of borrower and months with remaining loan balances within the window of $50,000 around the national CLL. Columns (4)–(6) show estimates with the window of $25,000 around the national CLL. Regardless of the number of polynomials included and the size of the sample window, we find a statistically significant jump in the probability of plain refinancing at the cutoff. The coefficient estimates are around 0.50 pp, which is consistent with Figure 13.

Table 7: **Monthly Probability of Plain Refinancing around the National CLL**: The table displays the estimated coefficients of the regression given by Equation (8) with plain refinancing as the dependent variable. Columns (1)–(3) are for the subsample of loan-month pairs with remaining mortgage balances within the window of $50,000 around the cutoff. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively. Columns (4)–(6) are for the subsample of loan-month pairs with remaining mortgage balances within the window of $25,000 around the cutoff. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All columns include the Zipcode × Month fixed effects and the control variables described in the main text. The standard errors are clustered at the level of Zipcode × Month.

	Larger Window			Small Window		
	(1) Polynomial=1	(2) Polynomial=2	(3) Polynomial=3	(4) Polynomial=1	(5) Polynomial=2	(6) Polynomial=3
$1[bal_i \leq CLL]$	0.0048***	0.0048***	0.0051***	0.0040***	0.0048***	0.0058***
	(13.90)	(9.75)	(7.86)	(8.83)	(7.15)	(6.42)
bal_i	0.0000	-0.0001***	-0.0001**	-0.0001***	-0.0002***	-0.0003*
	(0.48)	(-3.28)	(-2.42)	(-4.78)	(-2.62)	(-1.67)
$1[bal_i \leq CLL] \times bal_i$	0.0002***	0.0004***	0.0007***	0.0003***	0.0008***	0.0016***
	(12.65)	(7.20)	(5.11)	(8.19)	(5.71)	(4.87)
ZIPxMONTH FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	3,684,465	3,684,465	3,684,465	1,422,031	1,422,031	1,422,031
Adj. R^2	0.013	0.013	0.013	0.025	0.025	0.025

Table 8 shows estimates for regressions with cash-out refinancing as the dependent variable. Regardless of the size of the sample window, the main coefficient becomes statistically insignificant with polynomials of degree two or higher.

Table 8: **Monthly Probability of Cash-out Refinancing around the National CLL**: The table displays the estimated coefficients of the regression given by Equation (8) with cash-out refinancing as the dependent variable. Columns (1)–(3) are for the subsample of loan-month pairs with remaining mortgage balances within the window of $50,000 around the cutoff. Columns (1)–(3) are for specifications with up to first-, second-, and third-degree polynomials, respectively. Columns (4)–(6) are for the subsample of loan-month pairs with remaining mortgage balances within the window of $25,000 around the cutoff. Columns (4), (5), and (6) are for specifications with up to first-, second-, and third-degree polynomials, respectively. All columns include the Zipcode × Month fixed effects and the control variables described in the main text. The standard errors are clustered at the level of Zipcode × Month.

	Larger Window			Small Window		
	(1) Polynomial=1	(2) Polynomial=2	(3) Polynomial=3	(4) Polynomial=1	(5) Polynomial=2	(6) Polynomial=3
$1[bal_i \leq CLL]$	-0.0004***	-0.0003	-0.0002	-0.0003*	0.0001	-0.0002
	(-2.77)	(-1.41)	(-0.77)	(-1.79)	(0.49)	(-0.65)
bal_i	-0.0000**	-0.0000	-0.0000	-0.0000	0.0001*	-0.0001*
	(-2.49)	(-0.83)	(-0.48)	(-0.42)	(1.90)	(-1.78)
$1[bal_i \leq CLL] \times bal_i$	-0.0000	0.0000	0.0000	0.0000	0.0000	0.0002*
	(-1.54)	(0.62)	(0.95)	(0.04)	(0.06)	(1.70)
ZIPxMONTH FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	3,684,465	3,684,465	3,684,465	1,422,031	1,422,031	1,422,031
Adj. R^2	-0.008	-0.008	-0.008	-0.014	-0.014	-0.014

The relationship between the remaining loan balance and the probability of plain refinancing shows that a borrower typically wait until the remaining loan balance falls below the national CLL. As discussed earlier, that is either because a borrower would like to take advantage of a lower rate with a conventional conforming loan or because a higher mortgage rate with a high-balance loan makes the DTI binding.

How long would a borrower have to wait for the remaining balance to reach the national CLL? In our sample, a borrower with a remaining balance of the national CLL plus $25,000 has to wait 32 months to reach the national CLL. For a borrower with a remaining balance of the national CLL plus $10,000, it takes about 17 months to reach the national CLL. Even for a borrower with a remaining balance of the national CLL plus $5,000, it takes about 11 months. This finding suggests that the rate spread between high-balance and conforming loans due to TBA-eligibility results in a significant delay in a borrower's refinancing.

6.2 Consumer Spending

Previous research such as Agarwal et al. (2017) and Abel and Fuster (2018) find that consumer spending increases subsequent to refinancing. In the previous subsection, we found that TBA eligibility delays a borrower's refinancing decision. Then a natural question is whether and how much a borrower's consumption behavior is affected by TBA eligibility. Although our data do not provide direct information about a mortgage borrower's spending, CRISM sheds light on part of consumption that typically involves financing such as an automobile purchase. For this reason, a

number of papers have relied on consumer credit data and used new auto loan originations to study durable consumption.[20]

In this subsection, we first investigate whether borrowers in our sample increase their auto loan originations subsequent to refinancing. Then we investigate whether TBA eligibility affects auto originations by estimating how much auto originations increase once borrowers' remaining mortgage balances fall below the national CLL.

6.2.1 Consumer Spending Subsequent to Refinancing

To examine whether auto loan originations increase after refinancing, we construct our estimation sample in the following way. First, we include borrowers who ever plain-refinanced within the estimation sample used to study a borrower's refinancing decision in Section 6.1. We then follow the borrowers for twelve months: six months before and after plain refinancing.

With this estimation sample, we estimate the following regression:

$$NewAutoAmt_{it} = \Sigma_{t'=-6}^{5} \beta_{t'} 1[t = t_i^* + t'] + Z_{it}\gamma + \xi_{zip(i)\times t} + \epsilon_i. \tag{9}$$

The dependent variable $NewAutoAmt_{it}$ denotes a new auto loan amount. If borrower i does not originate a new auto loan in time t, then the variable is equal to zero. If he does, then the variable is equal to the origination amount. This variable measures changes not only in the extensive margin (whether a borrower originates a new auto loan) but also in the intensive margin (whether a borrower takes out a larger auto loan possibly to buy a more expensive car). CRISM does not tell directly whether or how much a borrower takes out a new auto loan in a given month. Instead, we observe each borrower's outstanding auto loan balances in each month. We assume that a borrower takes out a new auto loan if his outstanding auto loan balance increases by more than \$3,000 and if the number of outstanding auto loan accounts increases at the same time.[21]

On the right-hand side, $\beta_{t'}$ are the main coefficients of interest. The dummy variable $1[t = t_i^* + t']$ is equal to one if calendar month t is t' months after t_i^*, the month in which borrower i plain-refinanced. We normalize β_{-1} (the month right before refinancing) to zero. Next, Z_{it} refers to the same set of observed characteristics included in Equation (8). Moreover, we also include Zipcode $\times$ Month fixed effects ($\xi_{zip(i)\times t}$).

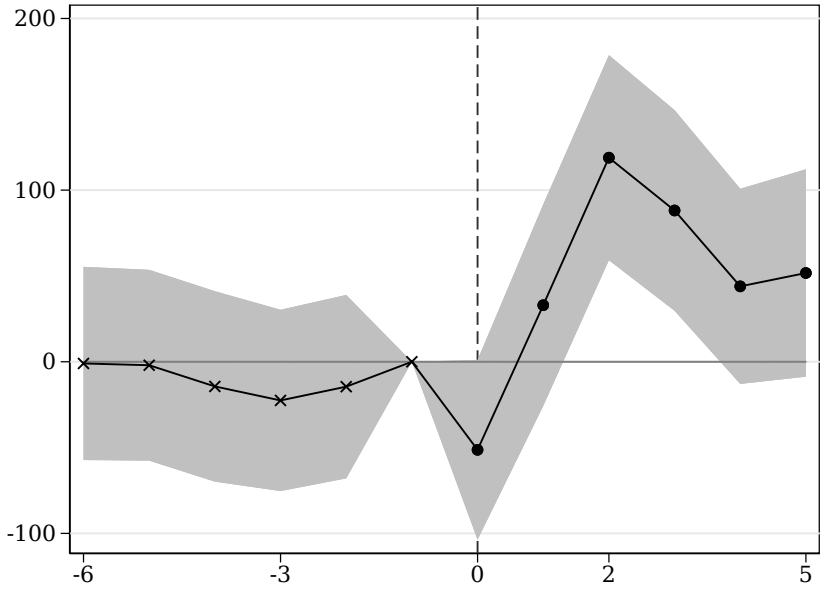
Figure 14 displays point estimates of $\beta_{t'}$ from Equation (9) and their 95% confidence intervals. We find that auto loan origination amounts increase after Month 0. The increases are especially high and statistically significant for for Months 2 and 3. Moreover, the increases in the two months are also substantial given that the average auto loan origination amount is \$300 in a month.[22] The figure also shows that the auto loan originations increases only after refinancing. We do not find any trends in new auto loan origination amounts before before Month 0.

[20] Examples of such papers are Agarwal et al. (2017), Di Maggio et al. (2017), and Abel and Fuster (2018).

[21] We also tried with a different threshold (\$5,000) to define an auto loan origination. The results are very robust to this alternative definition of auto loan originations.

[22] This number appears low because there are lots of borrowers who do not originate a new auto loan in a given month.

Figure 14: **Auto Loan Origination Amounts after Refinancing**: This figure plots point estimates of $\beta_{t'}$ from Equation (9) and their 95% confidence intervals. We normalize $\beta_{-1}$to zero. The dependent variable is a new auto loan amount. The regression includes the Zipcode × Month fixed effects and the control variables described in the main text. Standard errors for all specifications are clustered at the level of Zipcode × Month.



6.2.2 National CLL and Consumer Spending

Our findings so far suggest that TBA eligibility affects a borrower's refinancing decision, and those who refinanced increase their spending on automobiles after refinancing. Thus, we expect that TBA eligibility also affects a borrower's spending on automobiles.

In this subsection, we investigate whether a new auto origination amount increases for a borrowers with a remaining mortgage balance below the national CLL. We estimate the following regression, which is similar to Equation (8) but with $NewAutoAmt_{it}$ as the dependent variable. Our regression specification, including the set of variables included as controls and fixed effects, is exactly identical to the specification used in Section 6.1. The estimation sample is also identical.

$$NewAutoAmt_{it} = \alpha 1[bal_{it} \leq CLL_t] + g^{+}(bal_{it}; \theta_0) + g^{-}(bal_{it}; \theta_1) + Z_{it}\gamma + \xi_{zip(i)\times t} + \epsilon_i. \quad (10)$$

We first graphically examine patterns of residual value of $NewAutoAmt_{it}$ around the cutoff in Figure 15. The auto loan origination amounts generally decreases as remaining mortgage balances decreases. Patterns for borrowers with remaining mortgage balances below the cutoff are noisier than those for borrowers with balances above the cutoff. Comparing values right below and above the cutoff, however, we find that auto loan origination amounts increase at the cutoff by \$50, which is about 13% of the unconditional average (\$300).

Figure 15: **Auto Loan Origination Amounts around the National CLL**: This figure plots the relationship between the residual value of new auto origination amounts and the remaining loan balance. The residual value is obtained by removing variation in the original variable accounted for by observable characteristics Z_{it} and fixed effects $\xi_{zip(i)\times t}$ after running the regression given by Equation (10). Each dot represents the average value for each bin with the size of $2,500.

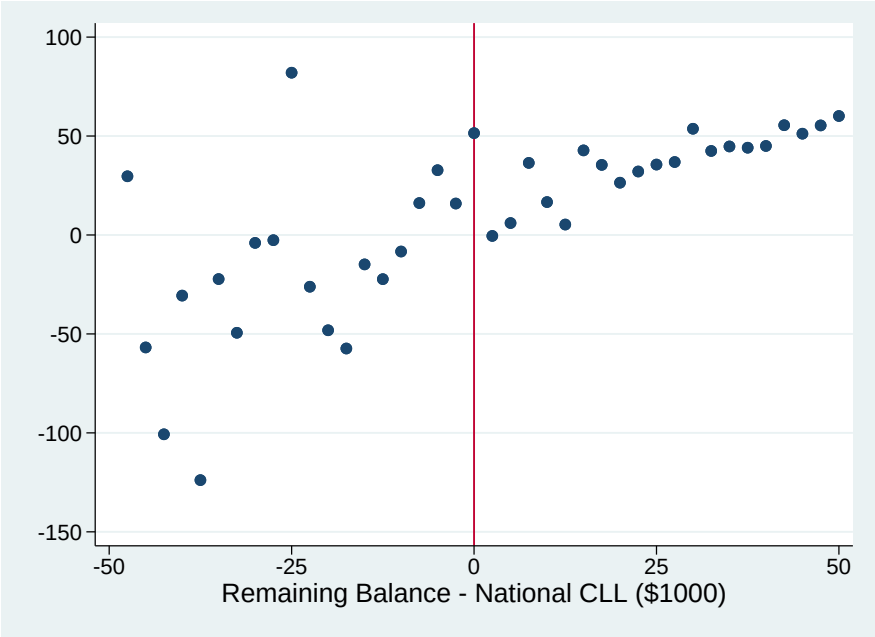


Table 9 provide regression estimates that are consistent with Figure 15. Although exact magnitudes are different, all estimates indicate that auto loan origination amounts increase at the cutoff. The estimate in Column (6) is not statistically significant, but its magnitude is quite consistent with other estimates. Additional polynomials seem to increase the standard error of the estimate very much.

Table 9: **Auto Loan Origination Amounts around the National CLL**: The table displays the estimated coefficients of the regression given by Equation (10) with new auto origination amounts as the dependent variable. Each column is different only with respect to the maximum number of polynomials. All columns include the Zipcode × Month fixed effects and the control variables described in the main text. The standard errors are clustered at the level of Zipcode × Month.

	Larger Window			Small Window		
	(1) Polynomial=1	(2) Polynomial=2	(3) Polynomial=3	(4) Polynomial=1	(5) Polynomial=2	(6) Polynomial=3
$1[bal_i \leq CLL]$	19.8052*	37.6581**	51.7643**	37.0721**	55.6822***	32.3898
	(1.82)	(2.42)	(2.57)	(2.54)	(2.59)	(1.15)
bal_i	0.9492***	1.4851**	2.8757*	1.0552*	3.2657	3.8933
	(5.63)	(2.24)	(1.72)	(1.96)	(1.57)	(0.73)
$1[bal_i \leq CLL] \times bal_i$	1.0813**	2.8484	3.8282	2.9645**	3.1970	-11.9429
	(1.99)	(1.59)	(0.94)	(2.52)	(0.74)	(-1.15)
ZIPxMONTH FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	3,684,465	3,684,465	3,684,465	1,422,031	1,422,031	1,422,031
Adj. R^2	0.004	0.004	0.004	0.002	0.002	0.002

The findings in this section show that TBA eligibility eventually affects a borrower's spending on automobiles through his refinancing decision. Facing a trade-off between refinancing now into a high-balance loan and refinancing later into a conventional conforming loan, many borrowers wait until their mortgage balances decrease to levels below the national CLL and then refinance into conventional conforming loans. As mentioned earlier, this waiting can be quite long. At the same time, a borrower's consumption spending is also tied to his refinancing decision. Durable spending, approximated by the new auto loan amount, typically increases in two or three months after refinancing. As a result, as a borrower waits for refinancing into a conventional conforming loan, his durable spending is also delayed.

This finding has an important implication for monetary policy transmission. One of the main channels for a lower interest rate to be translated into the real economy is through mortgage borrowers' refinancing (Agarwal et al., 2017; Abel and Fuster, 2018). Our finding suggests that liquidity of the secondary mortgage market, which is captured by eligibility for TBA delivery in our setting, is an important factor that affects how a lower interest rate is transmitted to the real economy. It also highlights that preserving the secondary market structure that improves liquidity of the market is important not only for welfare of borrowers but also for monetary policy transmission.

7 Conclusion

In this paper, we quantify the value of TBA eligibility for the mortgage borrowers. Being included in TBA-eligible pools reduces primary mortgage rates by 10–40 basis points, depending on the prepayment risk of the loan. Hence, the liquidity and trading structure of the secondary market can have direct impact on the primary market and in the real economy. Borrowers also delay refinancing in order to refinance into TBA-eligible loans. Given that refinancing is an important

channel in which monetary policy affects the real economy, the discontinuity in TBA-eligibility and the associated delay in refinancing may potentially slow the transmission of monetary policy.

References

Abel, J. and A. Fuster (2018): "How do mortgage refinances affect debt, default, and spending? Evidence from HARP," Tech. rep., Federal Reserve Bank of New York.

Adelino, M., A. Schoar, and F. Severino (2012): "Credit supply and house prices: evidence from mortgage market segmentation," Tech. rep., National Bureau of Economic Research.

Agarwal, S., G. Amromin, S. Chomsisengphet, T. Landvoigt, T. Piskorski, A. Seru, and V. W. Yao (2017): "Mortgage Refinancing, Consumer Spending, and Competition: Evidence from the Home Affordable Refinancing Program," Tech. rep., Columbia Business School.

Benmelech, E., R. R. Meisenzahl, and R. Ramcharan (2016): "The Real Effects of Liquidity During the Financial Crisis: Evidence from Automobiles*," *The Quarterly Journal of Economics*, 132, 317–365.

Beraja, M., A. Fuster, E. Hurst, and J. Vavra (2018): "Regional heterogeneity and the refinancing channel of monetary policy," *The Quarterly Journal of Economics*, 134, 109–183.

Bessembinder, H., W. F. Maxwell, and K. Venkataraman (2013): "Trading activity and transaction costs in structured credit products," *Financial Analysts Journal*, 69, 55–67.

Bond, P., A. Edmans, and I. Goldstein (2012): "The Real Effects of Financial Markets," *Annual Review of Financial Economics*, 4, 339–360.

Bond, P., R. Elul, S. Garyn-Tal, and D. K. Musto (2017): "Does Junior Inherit? Refinancing and the Blocking Power of Second Mortgages," *The Review of Financial Studies*, 30, 211–244.

Bright, M. and E. DeMarco (2016): "Toward a New Secondary Mortgage Market," .

Brugler, J., C. Comerton-Forde, and T. Hendershott (2018a): "Does Financial Market Structure Impact the Cost of Raising Capital?" .

Brugler, J., C. Comerton-Forde, and J. S. Martin (2018b): "Do you see what I see? Transparency and bond issuing costs," .

Davis, R., D. A. Maslar, and B. Roseman (2018): "Secondary Market Trading and the Cost of New Debt Issuance," .

DeFusco, A. A. and A. Paciorek (2017): "The interest rate elasticity of mortgage demand: Evidence from bunching at the conforming loan limit," *American Economic Journal: Economic Policy*, 9, 210–240.

Di Maggio, M., A. Kermani, B. J. Keys, T. Piskorski, R. Ramcharan, A. Seru, and V. Yao (2017): "Interest Rate Pass-Through: Mortgage Rates, Household Consumption, and Voluntary Deleveraging," *American Economic Review*, 107, 3550–88.

Di Maggio, M., A. Kermani, and C. Palmer (2016): "How quantitative easing works: Evidence on the refinancing channel," Tech. rep., National Bureau of Economic Research.

Field, L. C., A. Mkrtchyan, and Y. Wang (2018): "Bond Liquidity and Investment," *Northeastern University D'Amore-McKim School of Business Research Paper*.

Fuster, A. and J. Vickery (2014): "Securitization and the fixed-rate mortgage," *The Review of Financial Studies*, 28, 176–211.

Gao, P., P. Schultz, and Z. Song (2017): "Liquidity in a Market for Unique Assets: Specified Pool and To-Be-Announced Trading in the Mortgage-Backed Securities Market," *The Journal of Finance*, 72, 1119–1170.

Greenwald, D. (2018): "The mortgage credit channel of macroeconomic transmission," .

Kaufman, A. (2014): "The influence of Fannie and Freddie on mortgage loan terms," *Real Estate Economics*, 42, 472–496.

Passmore, S. W., S. M. Sherlund, and G. Burgess (2005): "The effect of housing government-sponsored enterprises on mortgage rates," .

Schultz, P. and Z. Song (2018): "Transparency and Dealer Networks: Evidence from the Initiation of Post-Trade Reporting in the Mortgage Backed Security Market," *Forthcoming, Journal of Financial Economics*.

Sherlund, S. M. (2008): "The Jumbo-Conforming Spread: A Semiparametric Approach," .

Vickery, J. I. and J. Wright (2013): "TBA Trading and Liquidity in the Agency MBS Market," *FRBNY Economic Policy Review*, May.

Wong, A. (2018): "Transmission of Monetary Policy to Consumption and Population Aging," Tech. rep., mimeo, Princeton University, Princeton.

A Appendix: Extra Figures

Figure 16: **Ratio of the Remaining Balance to the Original Balance around Cutoff 1**: These figures plot the residual ratio of the remaining balance to the original balance as of different loan ages in terms of months since origination. The residual variables are obtained by removing variation in corresponding original variables accounted by observable loan characteristics (Z_i) and fixed effects ($\xi_{s(i)\times l(i)\times t(i)}$) after running regressions given by Equation (3) with the original variables as dependent variables. Each dot in the plot represents the average value of each residual variable for each bin of the size of $2,500.

Figure 17: **Ratio of the Remaining Balance to the Original Balance around Cutoff 2** These figures plot the residual ratio of the remaining balance to the original balance as of different loan ages in terms of months since origination. The residual variables are obtained by removing variation in corresponding original variables accounted by observable loan characteristics (Z_i) and fixed effects ($\xi_{zip3(i)\times l(i)\times t(i)}$) after running regressions given by Equation (3) with the original variables as dependent variables. Each dot in the plot represents the average value of each residual variable for each bin of the size of 0.5.

B Appendix: Extra Tables

Table 10: **Regression Results for the Ratio of the Remaining Balance to the Original Balance (Cutoff 1)**: This table displays the estimates of the regression similar to Equation (3), where dependent variables are the ratio of remaining balance to the original balance as of loan age n for $n \in \{24, 36, 48, 60, 72, 84\}$. The maximum degree of polynomials included in the regressions are three for each column. For all columns, we used the subsample of loans with corresponding home values within the window of $50,000 around the cutoff. For each column, we further restricted the subsample to loans that were originated at least n months before the most recent month available in the data (2018m9). All specifications include State x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of State x Lender x Month.

	(1) By Age 24	(2) By Age 36	(3) By Age 48	(4) By Age 60	(5) By Age 72	(6) By Age 84
$1[h_i > h^*_{t(i)}]$	0.006	0.023	0.013	0.030	0.049*	0.023
	(0.36)	(1.06)	(0.55)	(1.17)	(1.81)	(0.90)
h_i	0.000	-0.004**	-0.003	-0.001	-0.001	-0.001
	(0.10)	(-2.17)	(-1.43)	(-0.33)	(-0.50)	(-0.36)
$1[h_i > h^*_{t(i)}] \times h_i$	-0.001	0.003	0.002	-0.002	-0.006	-0.001
	(-0.40)	(0.73)	(0.59)	(-0.37)	(-1.28)	(-0.35)
STATExMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	35,443	29,761	25,066	20,564	15,614	12,434
Adj. R^2	0.13	0.20	0.23	0.24	0.19	0.09

Table 11: **Regression Results for Exogenous Variables (Cutoff 1)**: This table displays the estimates of the regression similar to Equation (3) but with dependent variables are exogenous loan characteristics in Z_i. The maximum degree of polynomials included in the regressions are three for each column. For all columns, we used the subsample of loans with corresponding home values within the window of $50,000 around the cutoff. All specifications include State x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of State x Lender x Month.

	(1) Credit Score	(2) Loan-to-Income Ratio	(3) Broker Channel	(4) Correspondent Channel
$1[h_i > h^*_{t(i)}]$	0.592	0.063	0.001	-0.008
	(0.43)	(1.20)	(0.09)	(-0.62)
h_i	-0.176	-0.005	0.001	-0.000
	(-1.57)	(-1.10)	(0.92)	(-0.01)
$1[h_i > h^*_{t(i)}]$=1 $\times$ h_i	0.104	-0.007	-0.002	0.003
	(0.46)	(-0.75)	(-1.33)	(1.34)
STATExMONTHxSELLER FE	Y	Y	Y	Y
Other Controls	N	N	N	N
N. Obs.	78,683	78,589	78,737	78,737
Adj. R^2	0.06	0.11	0.48	0.43

Table 12: **Regression Results for the Ratio of the Remaining Balance to the Original Balance (Cutoff 2)**: This table displays the estimates of the regression similar to Equation (5), where dependent variables are the ratio of remaining balance to the original balance as of loan age n for $n \in \{24, 36, 48, 60, 72, 84\}$. The maximum degree of polynomials included in the regressions are three for each column. For all columns, we used the subsample of loans with $PredLTV_i$ between 100 and 110. For each column, we further restricted the subsample to loans that were originated at least n months before the most recent month available in the data (2018m9). All specifications include Zip3 x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of Zip3 x Lender x Month.

	(1) By Age 24	(2) By Age 36	(3) By Age 48	(4) By Age 60	(5) By Age 72	(6) By Age 84
$1[PredLTV > 105]$	-0.005	0.004	-0.003	0.002	0.009	0.021
	(-1.61)	(1.28)	(-0.80)	(0.38)	(0.82)	(1.01)
$PredLTV$	0.002	0.003	0.002	0.003	0.003	-0.018*
	(1.35)	(1.32)	(0.94)	(0.79)	(0.41)	(-1.80)
$1[PredLTV > 105] \times PredLTV$	-0.001	-0.004	-0.004	-0.007	-0.017	0.013
	(-0.31)	(-1.30)	(-1.15)	(-1.26)	(-1.61)	(0.65)
ZIP3xMONTHxSELLER FE	Y	Y	Y	Y	Y	Y
Other Controls	Y	Y	Y	Y	Y	Y
N. Obs.	61,744	54,914	44,277	23,311	11,282	3,481
Adj. R^2	-0.02	0.01	0.01	0.01	-0.01	-0.03

Table 13: **Regression Results for Exogenous Variables (Cutoff 2)**: This table displays the estimates of the regression similar to Equation (5) but with dependent variables are exogenous loan characteristics in Z_i. The maximum degree of polynomials included in the regressions are three for each column. For all columns, we used the subsample of loans with $PredLTV_i$ between 100 and 110. All specifications include Zip3 x Lender x Month Fixed effects and control variables described in the main text. Standard errors are clustered at the level of Zip3 x Lender x Month.

	(1) Credit Score	(2) Broker Channel	(3) Correspondent Channel	(4) Interest Rate for Prev Loan
$1[PredLTV > 105]$	1.511	0.003	0.000	-0.009
	(0.79)	(0.36)	(0.04)	(-0.52)
$PredLTV$	-0.777	0.003	-0.004	0.012
	(-0.69)	(0.79)	(-1.03)	(1.29)
$1[PredLTV > 105]{=}1 \times PredLTV$	-0.252	-0.013**	-0.003	0.021
	(-0.14)	(-1.99)	(-0.54)	(1.40)
ZIP3xMONTHxSELLER FE	Y	Y	Y	Y
Other Controls	N	N	N	N
N. Obs.	70,193	70,195	70,195	70,195
Adj. R^2	0.10	0.23	0.21	0.22

Statistical Arbitrage with Uncertain Fat Tails

Bo Hu *

University of Maryland, College Park

February 10, 2019

Abstract

I develop a model of statistical arbitrage trading in an environment with "fat-tailed" information. If risk-neutral arbitrageurs are uncertain about the variance of fat-tail shocks and if they implement max-min robust optimization, they will choose to ignore a wide range of pricing errors. Although model risk hinders their willingness to trade, arbitrageurs can capture the most profitable opportunities because they follow a linear momentum strategy beyond the inaction zone. This is equivalent to a machine-learning algorithm called LASSO. Arbitrageurs can also amass market power due to conservative trading under this strategy. Their uncoordinated exercise of robust control facilitates tacit collusion, protecting their profits from being competed away even if their number goes to infinity. In an extended model where an insider strategically interacts with those arbitrageurs, the insider can induce them to trade too aggressively, giving herself a reversal trading opportunity. Doing so distorts price informativeness and threatens market stability.

Keywords: fat tails, robust control, cartel effect, machine learning, price manipulations.

I am indebted to my dissertation committee: Albert Pete Kyle, Mark Loewenstein, Steve Heston, Yajun Wang, Shrihari Santosh, and John Chao for their help and support. I am grateful to Gurdip Bakshi, Utpal Bhattacharya, Cecilia Bustamante, Wen Chen, Julien Cujean, Nengjiu Ju, Dan Li, Jiasun Li, Tao Li, Danmo Lin, Xuewen Liu, Richmond Mathews, Anna Obizhaeva, Geoffrey Tate, Jun Pan, Nagpurnanand Prabhala, Alberto Rossi, Hyun-Song Shin, Pablo Slutzky, Andrea Vedolin, Tan Wang, Pengfei Wang, Wei Xiong, Liu Yang, Jialin Yu, Chu Zhang, and many seminar participants at the U.S. *Securities Exchange and Commission* (SEC), Bank for International Settlements, Warwick, HKUST, CityUHK, SAIF, CUHK(SZ), and George Mason University for valuable comments. All errors are mine.

*Department of Finance, Robert H. Smith School of Business, University of Maryland, College Park. Email: bohu25@rhsmith.umd.edu Website: https://sites.google.com/rhsmith.umd.edu/bo-hu

1 Introduction

In finance, extreme movements of asset prices occur much more frequently than predicted by the tail probabilities of a Gaussian distribution. Such *fat-tail* events have caused many problems, as exemplified by the failure of *Long Term Capital Management*. It is error-prone to predict fat-tail events or to deal with their higher-order statistics. These difficulties give rise to model risk[1] and drive traders to implement robust control. Model risk is a prominent concern for arbitrageurs whose activities are essential for market efficiency. Little is known about how model risk affects arbitrage trading in a fat-tail environment. This topic is both practically relevant and theoretically challenging. Answers to this question can provide new insights into many topics in asset pricing, risk management, and market regulation.

The existence of various anomalies such as momentum suggests that financial markets are not completely efficient[2]. Statistical arbitrage opportunities are also indicative of price inefficiency, because arbitrageurs can make profits given only public information[3]. To study statistical arbitrage trading, I introduce random fat-tail shocks to disrupt the efficient market of a two-period Kyle (1985) economy. In the standard Kyle model setup, an informed trader privately observes the stock liquidation value and trades sequentially to maximize her profits, under the camouflage of noise traders and against competitive market makers. A Gaussian information structure permits a unique linear equilibrium with an efficient linear pricing rule.

This paper models the stock value as a random realization drawn from the mixture of Gaussian and Laplacian distributions, which have the same mean and variance. It is only observed by an informed trader. The choice of a Laplacian distribution is empirically well-grounded[4]. It has fat tails on both sides since its probability density decays exponentially. This mixture setup allows the stock value to be *fat-tailed* with some probability. Market makers believe that they live in the Gaussian world and also regard it as the common belief among all agents. Market makers have the correct prior about the mean, variance, and skewness, but incorrect beliefs about higher moments of the stock value distribution. With Gaussian beliefs, they keep using a linear pricing rule[5], which can result in estimation bias if fat-tail shocks occur. This invites arbitrageurs to correct pricing errors. By assumption, arbitrageurs are sophisticated enough to distinguish the distribution types (i.e., mispricing

[1] Model risk is the risk of loss when traders use the wrong model or deal with uncertain model parameters.

[2] As documented by Jegadeesh and Titman (1993), the momentum strategy could earn abnormal returns.

[3] See Lehmann (1990), Campbell, Lo, and MacKinlay (1997), Bondarenko (2003), Hogan, Jarrow, Teo, and Warachka (2004), and Gatev, Goetzmann, and Rouwenhorst (2006) for discussions.

[4] The Laplace distribution can well characterize the distributions of stock returns sampled at different time horizons. This is documented, for example, by Silva, Prange, and Yakovenko (2004).

[5] The empirical price impact function, which measures the average price change in response to the size of an incoming order, is roughly linear with slight concavity. See Loeb (1983), Grinold and Kahn (2000) [p. 453], Gabaix, Gopikrishnan, Plerou, and Stanley (2006), and Kyle and Obizhaeva (2016).

cases), but they face uncertainty about the dispersion of Laplace priors. For robust control, arbitrageurs make trading decisions under the criterion of *max-min expected utility*[6].

My main finding is that model risk can motivate risk-neutral arbitrageurs to implement a machine-learning algorithm which mitigates their competition and ignores many mispricings. This result contains three points that are discussed in greater details below.

First, arbitrageurs' maximin robust strategy has a wide inaction zone: they start trading only when the observed order flow exceeds three standard deviations of noise trading. Yet this strategy is effective in catching the most profitable trades: arbitrageurs trade less than 2% of the time but can capture over 60% of the maximum profits they could earn in the absence of model risk. Under this strategy, arbitrageurs choose to ignore small mispricings. They focus on large events that involve little uncertainty about the trading direction. *Ex post*, an econometrician may find a lot of mispricings that persist in this economy and question arbitrageurs' rationality or capacity. In fact, arbitrageurs are rational and risk-neutral in my setting. They leave money on the table because of their aversion to *uncertainty*.

Second, this paper rationalizes a famous machine-learning method widely used in finance. The above-mentioned robust strategy is operationally equivalent to a simple algorithm called the *Least Absolute Shrinkage and Selection Operator* (LASSO)[7]. This is a powerful tool that can select a few key factors from a large set of regression coefficients. The standard statistical interpretation of LASSO involves a different mechanism, namely, the *Maximum a Posteriori* estimate. This learning rule lacks Bayesian rationality because it uses the posterior mode as point estimate, without summarizing all relevant information. In my setup, arbitrageurs are Bayesian-rational when they decide to use LASSO: they evaluate all possible states using Bayes' rule and dynamically maximize a well-defined utility with sequential rationality.

Third, the maximin robust strategy supports tacit collusion and impairs market efficiency. Arbitrageurs trade conservatively beyond the inaction zone. This enables them to accumulate market power, which is most prominent near the kinks of their robust strategy. Therefore, uncoordinated exercise of individual robust control facilitates tacit collusion among traders, without any communication device or explicit agreement. Remarkably, even as the number of arbitrageurs goes to infinity, their total profit does not vanish but converges to a finite level. This non-competitive payoff is due to the "cartel" effect which hinders price efficiency.

[6]The theory of max-min expected utility is a standard treatment for ambiguity-averse preferences. It is axiomatized by Gilboa and Schmeidler (1989), as a framework for robust decision making under uncertainty. Related discussions can be found in Dow and Werlang (1992) and Hansen and Sargent (2001) for example.

[7]LASSO is a machine-learning technique developed by Tibshirani (1996) to improve prediction accuracy and model interpretability. It is popular among algorithmic traders. This technique has recently been employed in many financial studies, such as Huang and Shi (2011), Kozak, Nagel, and Santosh (2017), Chinco, Clark-Joseph, and Ye (2017), and Freyberger, Neuhierl, and Weber (2017).

Finally, I extend the model to allow for strategic interaction between the informed trader and the group of arbitrageurs. The informed trader entices arbitrageurs to mimic past order flows; arbitrageurs' trend-following responses also tempt the informed trader to trick them: she may first trade a large quantity to trigger those arbitrageurs and then unwind her position against them. This strategy resembles several controversial schemes in reality. One example is *momentum ignition*, a trading algorithm that attempts to trigger many other algorithmic traders to run in the same direction so that the instigator can profit from trading against the momentum she ignited. Another scheme is *stop-loss hunting* which attempts to force some traders out of their positions by pushing the asset price to certain levels where they have set stop-loss orders. In my setup, this sort of strategies can impair pricing accuracy, exaggerate price volatility, and raise the average trading costs for common investors. Numerical results also generate empirically testable patterns regarding price overreactions and volatility spikes.

Contributions to the literature. This paper investigates strategic arbitrage trading in an uncertain fat-tail environment. This topic requires new methods and inspires fresh thinking. Results discussed in this paper can contribute in multiple ways to the vast literature of asset pricing, market microstructure, and behavioral finance.

First, this paper develops a new modeling framework for statistical arbitrage. The semi-strong-form market efficiency holds in the standard Kyle (1985) model where traders have common Gaussian beliefs about the economy. This simple assumption has been followed by most subsequent studies[8]. The present paper deviates from the literature by introducing fat-tail shocks to disrupt the Kyle equilibrium when market makers stick to Gaussian beliefs. Unexpected changes in the underlying distribution cause mispricings in the market. This gives room for arbitrageurs if they can foresee fat-tail shocks. Due to model risk, arbitrageurs are uncertain about the extent of mispricings. If they simply follow the maximin criterion, they may overemphasize the least favorable prior and become overly pessimistic in decision making. This paper implements a rational procedure that prevents such biases. Similar to the spirit of *rational expectations*, an internally consistent assumption is that arbitrageurs inside this model have the correct belief *on average* about the model structure, despite their uncertainty about some prior parameter. Recognizing this consistency, a rational arbitrageur only considers those strategies that converge to the optimal strategy (as averaged across all possible priors) and that preserve the convexity of their optimal strategy. Such constraints make their admissible strategies comparable to the ideal rational-expectations strategy[9].

[8]The literature includes Back (1992), Holden and Subrahmanyam (1992), Foster and Viswanathan (1994), Foster and Viswanathan (1996), Vayanos (1999), Back, Cao, and Willard (2000), Vayanos (2001), Huddart, Hughes, and Levine (2001), and Collin-Dufresne and Fos (2016), among many others.

[9]The rational-expectations strategy is the one that traders would use if they knew the true Laplace prior.

Second, this paper is the first to study how market efficiency gets hindered by model risk when arbitrageurs have fat-tail beliefs. This angle distinguishes the present paper from the existing literature on *limits to arbitrage*[10]. Previous studies have suggested various important frictions, including short-selling costs, leverage constraints, and wealth effects, which limit arbitrageurs' *ability* to eliminate mispricings. Excluding those frictions, the present paper identifies another mechanism that can strongly affect the *willingness* of arbitrageurs to trade. Specifically, model uncertainty of fat-tail priors make arbitrageurs hesitate to eradicate small mispricings, because of ambiguity about the trading direction.

Third, this work sheds light on interesting topics at the interface of behavioral finance and machine learning. This paper uses the max-min decision rule to rationalize the LASSO (*"soft-thresholding"*) strategy, which was taken by Gabaix (2014) as a behavioral assumption of the *anchoring-and-adjustment* mechanism. The LASSO algorithm has an inaction zone where agents choose to ignore whatever happened, similar to the *status quo bias*[11]. The strategy of arbitrageurs also resembles the behavior of feedback traders discussed in behavioral finance[12]. In the eyes of an observer who has a Gaussian prior, arbitrageurs are "irrational" because they show up randomly and all perform feedback trading based on historical prices. The observer's view is incorrect, given his misspecified prior in this economy.

This study can also help us to interpret empirical results about high-frequency traders (HFTs)[13]. My primary model of statistical arbitrage can describe the situation where an informed institutional investor executed large orders over time without anticipating that HFTs detected her footprints to catch the momentum train; see Lewis (2014) for a historical account. As an extension, I consider strategic interaction between an informed trader and a group of arbitrageurs. This extended model can describe the situation where institutional investors anticipate those HFTs and optimize their execution algorithms with strategic considerations. My model is consistent with the empirical implications reported in van Kervel and Menkveld (2017) on HFTs around institutional trading: (1)"*HFTs appear to lean against the wind when an order starts executing but if it executes more than seven hours, they seem to reverse course and trade with wind.*" (2)"*Institutional orders appear mostly information-motivated, in particular the ones with long-lasting executions that HFTs eventually trade along with.*" (3)"*Investors are privately informed and optimally trade on their signal in full awareness of HFTs preying on the footprint they leave in the market.*"

[10]Gromb and Vayanos (2010) is an excellent survey on this subject. See also Shleifer and Vishny (1997), Xiong (2001), Gabaix, Krishnamurthy, and Vigneron (2007), Kondor (2009), among others.

[11]See Kahneman, Knetsch, and Thaler (1991) and Samuelson and Zeckhauser (1988).

[12]For behavioral interpretations of feedback traders, see DeLong, Shleifer, Summers, and Waldmann (1990), Barberis, Greenwood, Jin, and Shleifer (2015), and Barberis, Greenwood, Jin, and Shleifer (2018).

[13]For recent research on high-frequency trading, see Hendershott, Jones, and Menkveld (2011), van Kervel and Menkveld (2017), Kirilenko, Kyle, Samadi, and Tuzun (2017), and Korajczyk and Murphy (2018).

The extended model also contributes to the body of literature on market manipulations[14]. In Allen and Gale (1992), a trade-based price manipulation is played by an *uninformed* trader who attempts to trick other traders into believing the existence of informed trading. In my model, the manipulative strategy is performed by an *informed* trader who trades in an unexpected way to distort the learning of other traders. The informed trader may hide her signal when it is strong and bluff when it is weak. In the linear equilibrium of Foster and Viswanathan (1994), the better informed trader may also hide her information in early periods and even trade against the direction of her superior signal. My analysis focuses on a nonlinear equilibrium where the informed trader hides her information to reduce competitive pressure from arbitrageurs. Several articles by Chakraborty and Yılmaz[15] show that if market makers face uncertainty about the existence of informed trades, then the informed trader will bluff in every equilibrium by directly adding noise to other traders' inference problem. The disruptive strategy in my model is different because (1) it occurs under a set of specific conditions, not state-by-state in every equilibrium; (2) it is a pure strategy that distorts the learning of other traders, not a mixed strategy that adds some noise[16]; (3) it produces bimodal distributions of prices, thereby magnifying both price volatility and trading costs.

Finally, the disruptive strategy in this paper shows that asset price "bubbles and crashes" can take place in a strategic environment where speculators have fat-tail beliefs. Under good enough liquidity conditions, a better-informed savvy trader may trade very aggressively to trigger those speculators whose subsequent momentum responses can give this savvy trader a reversal trading opportunity. This finding is related to the literature on market instability[17]. The mechanism here shares some similarity with the model of Scheinkman and Xiong (2003) where asset price bubbles reflect resale options due to traders' overconfidence. In my setup, speculators' over-aggressive trading implicitly grants the informed trader a "resale option" which could be exercised if condition permits. It is however worth remarking that traders in my (extended) model share a common fat-tail prior, without any overconfidence bias.

The rest of this paper is organized as follows. Section 2 focuses on the primary model where arbitrageurs exploit uncertain pricing errors in a robust manner. Section 3 studies the extended model where a savvy informed trader anticipates and exploits those arbitrageurs. Concluding remarks are made in Section 4. Major proofs are provided in Appendix A.

[14]See Allen and Gale (1992), Kumar and Seppi (1992), Jarrow (1992), van Bommel (2003), Huberman and Stanzl (2004), Huddart et al. (2001), Khwaja and Mian (2005), Jiang, Mahoney, and Mei (2005); Brunnermeier (2005), Brunnermeier and Pedersen (2005), Kyle and Viswanathan (2008), Goldstein and Guembel (2008), Jarrow (2015), and Fox, Glosten, and Rauterberg (2018).

[15]See Chakraborty and Yılmaz (2004a), Chakraborty and Yılmaz (2004b), Chakraborty and Yılmaz (2008).

[16]Mixed strategies are studied in modified Kyle models by Huddart et al. (2001) and Yang and Zhu (2017).

[17]See Kyle and Xiong (2001), Abreu and Brunnermeier (2003), Hong and Stein (2003), and Scheinkman and Xiong (2003), among others.

2 Model of Robust Arbitrageurs

In this section, an equilibrium model is developed to study how arbitrageurs' prior uncertainty about mispricing shocks affects arbitrage strategy and market efficiency. This model adds random fat-tail shocks to disturb the efficient market of a two-period Kyle (1985) model.

Table 1. The timeline and market participants in an economy of two auctions.

	$t=0$	$t=1$	$t=2$	$t=3$
Informed Trader	observe v	submit x_1	submit x_2	receive π_x
Noise Traders	...	submit u_1	submit u_2	...
Arbitrageurs	observe s	submit $z_{1,n}$	submit $z_{2,n}$	receive $\pi_{z,n}$
Market Makers	prior $\mathcal{N}(0,\sigma_v^2)$	set p_1	set p_2	observe v

Structure and Notation. Consider the market in Table 1 with two rounds of trading, indexed by $t=1,2$. The liquidation value of a risky asset, denoted $\tilde{v}$, is either Gaussian or Laplacian:

$$\tilde{v} = (1-\tilde{s})\cdot\tilde{v}_G + \tilde{s}\cdot\tilde{v}_L, \quad \text{where} \quad \tilde{v}_G \sim \mathcal{N}(0,\sigma_v^2), \quad \tilde{v}_L \sim \mathcal{L}(0,\xi_v), \quad \xi_v \equiv \frac{\sigma_v}{\sqrt{2}}. \tag{1}$$

Here, $\tilde{s}$ takes the integer value 1 with probability α and takes the value 0 with probability $1-\alpha$. The true Laplace scale parameter is set to be $\xi_v = \frac{\sigma_v}{\sqrt{2}}$ so that the variance of $\tilde{v}$ is always σ_v^2. The initial asset price is set as $p_0 = 0$ without loss of generality. The quantities traded by noise traders are Gaussian, denoted $\tilde{u}_1 \sim \mathcal{N}(0,\sigma_u^2)$ and $\tilde{u}_2 \sim \mathcal{N}(0,\gamma\sigma_u^2)$. The noise variances can be different, as tuned by the parameter $\gamma > 0$. All the random variables $\tilde{v}$, $\tilde{s}$, $\tilde{u}_1$, and $\tilde{u}_2$ are mutually independent. The parameters $\{\sigma_v, \sigma_u, \gamma\}$ are common knowledge.

A risk-neutral informed trader privately observes $\tilde{v}$ at $t=0$, submits market orders, $\tilde{x}_1$ and $\tilde{x}_2$, to buy or sell this asset before her private signal becomes public at $t=3$. The strategy is denoted by a vector of real-valued functions, $\mathbf{X} = \langle X_1, X_2\rangle$. Prices and volumes become public information right after the auctions take place. The information sets of informed trader before trading at $t=1,2$ are $\mathcal{I}_{1,x} = \{\tilde{v}\}$ and $\mathcal{I}_{2,x} = \{\tilde{v}, \tilde{p}_1\}$ where $\tilde{p}_1$ is the asset price at $t=1$. It is justified to write $\tilde{x}_1 = X_1(\tilde{v})$ and $\tilde{x}_2 = X_2(\tilde{v}, \tilde{p}_1)$. The informed trader's total profit from both periods can be written as $\tilde{\pi}_x = \sum_{t=1}^{2}(\tilde{v} - \tilde{p}_t)\tilde{x}_t$.

A number of risk-neutral arbitrageurs (indexed by $n = 1, ..., N$) observe $\tilde{s}$, which encodes the distribution type of $\tilde{v}$. Each arbitrageur can place market orders, $\tilde{z}_{1,n}$ and $\tilde{z}_{2,n}$, to exploit potential market inefficiency. Their strategy profile is represented by a matrix of real-valued functions, $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_N]$ where $\mathbf{Z}_n = \langle Z_{1,n}, Z_{2,n} \rangle$ is the n-th arbitrageur's strategy for $n = 1, ..., N$. The information sets of arbitrageurs are $\mathcal{I}_{1,z} = \{\tilde{s}\}$ and $\mathcal{I}_{2,z} = \{\tilde{s}, \tilde{p}_1\}$ before their trading at $t = 1, 2$. The quantities traded by the n-th arbitrageur are $\tilde{z}_{1,n} = Z_{1,n}(\tilde{s})$ and $\tilde{z}_{2,n} = Z_{2,n}(\tilde{s}, \tilde{p}_1)$. The total profit for the n-th trader is denoted $\tilde{\pi}_{z,n} = \sum_{t=1}^{2} (\tilde{v} - \tilde{p}_t)\tilde{z}_{t,n}$.

Uninformed competitive market makers clear the market by setting prices at which they strive to break even. Their pricing strategy is denoted by the vector of real-valued functions, $\mathbf{P} = \langle P_1, P_2 \rangle$. The total order flow $\tilde{y}_t \equiv \tilde{x}_t + \sum_{n=1}^{N} \tilde{z}_{t,n} + \tilde{u}_t$ is observed by market makers before they set the price $\tilde{p}_t$ at period $t \in \{1, 2\}$. We can write $\tilde{p}_1 = P_1(\tilde{y}_1)$ and $\tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2)$.

Belief System. Several assumptions are needed to clarify traders' beliefs in this model:

Assumption 2.1. *The informed trader and market makers think that it was common belief among all traders that the asset liquidation value was normally distributed,* $\tilde{v} \sim \mathcal{N}(0, \sigma_v^2)$.

Assumption 2.2. *Arbitrageurs have the correct Gaussian prior when* $\tilde{s} = 0$*, but they face uncertainty about the variance of fat-tail shocks when* $\tilde{s} = 1$*. Their Laplace prior is modeled as* $\mathcal{L}(0, \tilde{\xi})$ *where* $\tilde{\xi} \in \Omega$ *is a positive random variable. Arbitrageurs are ambiguity-averse and maximize the minimum expected payoff over all possible priors. On average, arbitrageurs are correct about the information structure, despite their prior uncertainty.*

Assumption 2.3. *Arbitrageurs know that market makers and the informed trader obey Assumption 2.1. Moreover, Assumption 2.2 is held as common knowledge among arbitrageurs.*

Since fat-tail shocks occur with probability α in this market, the higher-order moments of $\tilde{v}$ can differ from those of the Gaussian counterpart $\tilde{v}_G$. When $\alpha = 0$, the asset value $\tilde{v}$ is exactly Gaussian and the model reduces to the standard two-period Kyle (1985) model. The Laplace probability density, $f_L(v) = \frac{1}{2\xi_v} \exp\left(-\frac{|v|}{\xi_v}\right)$, has fat tails as it decays to zero at an exponential rate. Thus, the likelihood of observing extreme events under the Laplace distribution is much higher than under the Gaussian distribution with identical variance.

Knowledge of $\tilde{s}$ is valuable since it tells traders the distribution type of stock value. If market makers have fat-tail beliefs and observe $\tilde{s} = 1$, they should use a convex pricing rule (which is rarely seen in real data). The Gaussian prior in Assumption 2.1 permits linear pricing schedules compatible with empirical observations. Despite its simplicity, the linear pricing function can underestimate the fat-tail information in large order flows. This opens the door to arbitrageurs because market makers have mistakes with probability α.

Arbitrageurs are sophisticated traders who may use advanced technology to detect mispricings. Their privilege of observing $\tilde{s}$ represents their superior ability to identify statistical arbitrage opportunities. Nonetheless, arbitrageurs often face uncertainty about their trading models. The failure of *Long-Term Capital Management* (LTCM) demonstrates the critical role of model risk and the disastrous impact when the worst-case scenario hit. This motivates Assumption 2.2 that arbitrageurs care about the worst-case expected profits for robustness. As proved by Gilboa and Schmeidler (1989), the *max-min expected utility* theory rationalizes ambiguity-averse preferences. However, decisions derived from maximin optimization tend to follow the least favorable prior regardless of its likelihood. This appears too pessimistic. A more realistic assumption is that arbitrageurs' admissible strategies converge, in a rational manner[18], to the average of optimal strategies evaluated across all possible priors. Similar to the concept of *rational expectations*, I assume that arbitrageurs inside this model are correct on average about the model structure. Without systematic bias, the average of optimal strategies across all possible priors should converge to the *rational-expectations equilibrium* (REE) strategy which corresponds to the ideal case that they know the true prior ξ_v.

Assumptions 2.1, 2.2, and 2.3 capture salient features of real-life arbitrage. In a nearly efficient market, arbitrage opportunities should be rare and thus overlooked by most market participants. Such opportunities may be identified and exploited by a small number of traders (i.e., arbitrageurs who observe $\tilde{s}$). What may limit their trading is the model risk and their imperfect competition. Arbitrageurs are likely to have similar priors and preferences, given that they have similar forecasting technology and face similar pressures of robust control.

The belief system described in Assumption 2.1 can be denoted as $\mathcal{B} = \{\tilde{s} = 0\}$, which is shared by the informed trader and market makers. They think that it is common knowledge among all traders that $\tilde{v} \sim \mathcal{N}(0, \sigma_v^2)$. Arbitrageurs are aware of their Gaussian belief $\mathcal{B}$. By Assumptions 2.2 and 2.3, the belief system shared by arbitrageurs can be expressed as $\mathcal{A} = \{\tilde{s}, \tilde{\xi}\}$, where $\tilde{\xi}$ denotes the uncertain Laplace prior. Arbitrageurs' belief depends on the observed $\tilde{s}$ which tells them the type of prior to use:

$$\tilde{v} \sim \mathcal{N}(0, \sigma_v^2) \text{ if } \mathcal{A} = \{\tilde{s} = 0, \tilde{\xi}\} \quad \text{and} \quad \tilde{v} \sim \mathcal{L}(0, \tilde{\xi}) \text{ if } \mathcal{A} = \{\tilde{s} = 1, \tilde{\xi}\}. \tag{2}$$

Obviously, $\mathcal{A}$ and $\mathcal{B}$ are consistent when $\tilde{s} = 0$ but they are at odds when $\tilde{s} = 1$. Market makers believe that any uninformed trader holds the same Gaussian prior as they do. In fact, arbitrageurs can infer that market makers use the wrong prior when $\tilde{s} = 1$.[19].

[18] To avoid overfitting, their admissible strategy should preserve the convexity of their optimal strategies.

[19] This is not "agreement to disagree" because traders have inconsistent belief structures here. Han and Kyle (2017) discussed the situation where traders have inconsistent beliefs about the mean. In my model, traders agree on the mean but hold inconsistent beliefs about higher moments of $\tilde{v}$.

2.1 Equilibrium Definition and Conjecture

The trading of arbitrageurs affects the realized profit of informed trader $\tilde{\pi}_x$. To emphasize its dependence on all traders' strategies, we write $\tilde{\pi}_x = \tilde{\pi}_x(\mathbf{X}, \mathbf{P}, \mathbf{Z})$. Similarly, each arbitrageur takes into account the strategies played by other traders. To stress such dependence, we write $\tilde{z}_{t,n} = \tilde{z}_{t,n}(\mathbf{X}, \mathbf{P}, \mathbf{Z})$ and $\tilde{\pi}_{z,n} = \tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{P}, \mathbf{Z})$ for $n = 1, ..., N$. By Assumption 2.2, each arbitrageur seeks to maximize the minimum expected profit over all possible priors:

$$\max_{\mathbf{Z}_n \in \mathcal{Z}^2} \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}} \left[\tilde{\pi}_{z,n} \middle| \tilde{s}, \tilde{\xi} = \xi \right] = \max_{\mathbf{Z}_n \in \mathcal{Z}^2} \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}} \left[\sum_{t=1}^{2} (\tilde{v} - \tilde{p}_t) z_{t,n} \middle| \tilde{s}, \tilde{\xi} = \xi \right], \tag{3}$$

where $\mathbf{Z}_n = \langle z_{1,n}, z_{2,n} \rangle$. Both $z_{1,n}$ and $z_{2,n}$ are in the admissible set $\mathcal{Z}$ which requires asymptotic convergence to the REE without losing the convexity/concavity of the REE strategy.

Definition of Equilibrium. A sequential trading equilibrium in this model is defined as a tuple of strategies $(\mathbf{X}, \mathbf{P}, \mathbf{Z})$ such that the following conditions hold:

1. For any alternative strategy $\mathbf{X}' = \langle X_1', X_2' \rangle$ differing from $\mathbf{X} = \langle X_1, X_2 \rangle$, the strategy $\mathbf{X}$ yields an expected total profit no less than $\mathbf{X}'$, and also X_2 yields an expected profit in the second period no less than the single deviation X_2':

$$\begin{aligned} \mathrm{E}^{\mathcal{B}}[\tilde{\pi}_x(\mathbf{X}, \mathbf{P}, \mathbf{Z}) | \tilde{v}] &\geq \mathrm{E}^{\mathcal{B}}[\tilde{\pi}_x(\mathbf{X}', \mathbf{P}, \mathbf{Z}) | \tilde{v}], & (4) \\ \mathrm{E}^{\mathcal{B}}[(\tilde{v} - \tilde{p}_2(\langle X_1, X_2 \rangle, \mathbf{P}, \mathbf{Z})) X_2 | \tilde{v}, \tilde{p}_1] &\geq \mathrm{E}^{\mathcal{B}}[(\tilde{v} - \tilde{p}_2(\langle X_1, X_2' \rangle, \mathbf{P}, \mathbf{Z})) X_2' | \tilde{v}, \tilde{p}_1]. & (5) \end{aligned}$$

2. For all $n = 1, ..., N$ and any alternative strategy profile $\mathbf{Z}'$ differing from $\mathbf{Z}$ only in the n-th component $\mathbf{Z}_n' = \langle Z_{1,n}', Z_{2,n}' \rangle$, the strategy profile $\mathbf{Z}$ yields a utility level (i.e., the minimum expected profit over all possible priors) no less than $\mathbf{Z}'$, and also $Z_{2,n}$ yields a utility level in the second period no less than the single deviation $Z_{2,n}'$:

$$\begin{aligned} \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}}[\tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{P}, \mathbf{Z}) | \tilde{s}, \tilde{\xi} = \xi] &\geq \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}}[\tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{P}, \mathbf{Z}') | \tilde{s}, \tilde{\xi} = \xi]; & (6) \\ \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \tilde{p}_2(\cdot, Z_{2,n})) Z_{2,n} | \tilde{s}, \tilde{p}_1, \tilde{\xi} = \xi] &\geq \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \tilde{p}_2(\cdot, Z_{2,n}')) Z_{2,n}' | \tilde{s}, \tilde{p}_1, \tilde{\xi} = \xi] & (7) \end{aligned}$$

   where the strategy profile on the right hand side of Eq. (7) only differs from $(\mathbf{X}, \mathbf{P}, \mathbf{Z})$ at $Z_{2,n}$. Any strategy considered by arbitrageurs has to be in the admissible set $\mathcal{Z}$.

3. The prices, $\mathbf{P} = \langle P_1, P_2 \rangle$, are set by the market makers' posterior expectation of $\tilde{v}$:

$$\tilde{p}_1 = P_1(\tilde{y}_1) = \mathrm{E}^{\mathcal{B}}[\tilde{v} | \tilde{y}_1], \quad \text{and} \quad \tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2) = \mathrm{E}^{\mathcal{B}}[\tilde{v} | \tilde{y}_1, \tilde{y}_2]. \tag{8}$$

Equilibrium Conjecture. The full equilibrium $(\mathbf{X}, \mathbf{P}, \mathbf{Z})$ can be characterized separately. The informed trader and market makers believe that they were living in a two-period Kyle model (Assumption 2.1). They think that arbitrageurs held the same Gaussian belief and would not trade in a conjectured equilibrium with (semi-strong-form) market efficiency. This inspires them to conjecture a subgame perfect linear equilibrium $(\mathbf{X}, \mathbf{P})$.

Proposition 2.1. *Under Assumptions 2.1, there exists a unique subgame perfect linear equilibrium* $(\mathbf{X}, \mathbf{P})$ *identical to the linear equilibrium of a two-period Kyle (1985) model with normally distributed random variables. Market makers set the linear pricing rule:*

$$\tilde{p}_1 = P_1(\tilde{y}_1) = \lambda_1 \tilde{y}_1, \quad \tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2) = \tilde{p}_1 + \lambda_2 \tilde{y}_2, \quad \lambda_1 = \frac{\sqrt{2\delta(2\delta-1)}}{4\delta - 1}\frac{\sigma_v}{\sigma_u}, \quad \lambda_2 = \delta\lambda_1. \tag{9}$$

The equilibrium ratio $\delta = \frac{\lambda_2}{\lambda_1}$ *is determined by the largest root to the cubic equation:*

$$8\gamma\delta^3 - 4\gamma\delta^2 - 4\delta + 1 = 0. \tag{10}$$

The informed trader follows the linear trading strategy:

$$\tilde{x}_1 = X_1(\tilde{v}) = \frac{\tilde{v}}{\rho\lambda_1} = \frac{2\delta - 1}{4\delta - 1} \cdot \frac{\tilde{v}}{\lambda_1}, \qquad \tilde{x}_2 = X_2(\tilde{v}, \tilde{y}_1) = \frac{\tilde{v} - \lambda_1\tilde{y}_1}{2\delta\lambda_1}, \tag{11}$$

where $\rho \equiv \frac{4\delta-1}{2\delta-1}$ *is a liquidity-dependent parameter that reflects the trading intensity at* $t = 1$*. Informed trader and market makers believe that no arbitrageurs would trade under* $(\mathbf{X}, \mathbf{P})$*.*

Proof. This is an extension of Proposition 1 in Huddart et al. (2001). See Appendix A.1. □

To break even under different liquidity conditions, market makers can adjust the slopes of linear pricing schedules. For example, when noise trading volatility is constant (i.e., $\gamma = 1$), they can solve from Eq. (10) that $\delta \approx 0.901$; when $\gamma = \frac{3}{4}$, they can find that $\delta = 1$ and $\lambda_1 = \lambda_2 = \frac{\sqrt{2}}{3}\frac{\sigma_v}{\sigma_u}$; when liquidity evaporates ($\gamma \to 0$), the solution explodes: $\delta \to \infty$ so that $\lambda_1 = \frac{\sigma_v}{2\sigma_u}$ and $\lambda_2 \to \infty$. It is convenient to introduce a dimensionless parameter to denote the liquidity condition. Market depth is usually measured by the inverse of price impact parameter. To quantify the change of market depth in the second period, I define

$$\mu \equiv \frac{\lambda_1^{-1} - \lambda_2^{-1}}{\lambda_1^{-1}} = 1 - \frac{1}{\delta}. \tag{12}$$

In general, $\mu \in [-1, 1]$. For example, $\mu = 0.5$ indicates a 50% drop of market depth, while $\mu = 0$ reflects constant depth. Market depth becomes higher (i.e., $\mu < 0$) if $\gamma > \frac{3}{4}$.

If market makers know that $\tilde{v}$ is drawn from the mixture distribution, the linear pricing rule in Eq. (9) can still help them to break even, regardless of the mixture parameter α. Linear pricing preserves the symmetry of probability distributions so that market makers' unconditional expected profits are zero : $\mathrm{E}[(\tilde{p}_2 - \tilde{v})\tilde{y}_2] = 0$ and $\mathrm{E}[(\tilde{p}_1 - \tilde{v})\tilde{y}_1 + (\tilde{p}_2 - \tilde{v})\tilde{y}_2] = 0$. This shows the robustness of linear pricing strategy and may explain its popularity.

By Proposition 2.1, the informed trader and market makers believe that no arbitrageurs would trade in this market. Thus, any strategy profile $\mathbf{Z}$ does not affect the linear equilibrium strategies $\mathbf{X}$ and $\mathbf{P}$. Arbitrageurs can take Proposition 2.1 as given when solving their own dynamic optimization problems Eq. (6) and Eq. (7). Arbitrageurs know that the informed trader and market makers do not anticipate their trading. Arbitrageurs take into account the price impacts of all traders in the market. When $s = 0$, the belief structure of all traders is consistent and correct. In this case, arbitrageurs have no advantage over market makers.

Corollary 2.1. *When $s = 0$, arbitrageurs do not trade because the market is indeed efficient.*

Arbitrageurs are better "informed" than market makers in the presence of fat-tail shocks. Will they trade immediately? Let us conjecture now and verify later that arbitrageurs would not trade in the first period. This is intuitive given the symmetry of their priors and the linearity of pricing rule. It simplifies the procedure to solve this equilibrium. First, Eq. (7) can be used to derive the optimal strategy profile $\langle Z_{2,1}, ..., Z_{2,N}\rangle$ in the next period under the conjecture that $Z_{1,n} = 0$ for all $n = 1, ..., N$. Second, Eq. (6) can be used to verify that it is not a profitable deviation for any arbitrageur to trade in the first period. If no one would deviate, $\mathbf{Z} = [\langle 0, Z_{2,1}\rangle, ..., \langle 0, Z_{2,N}\rangle]$ will indeed be the equilibrium strategy for arbitrageurs.

2.2 Optimal Strategy without Model Risk

The linearity of informed trader's strategy $X_1(v) = \frac{v}{\rho\lambda_1}$ simplifies arbitrageurs' inference. Intuitively, the quantities traded by them in the presence of fat-tail shocks are proportional to their conditional expectation of the stock value mispriced by the market. Of course, the posterior estimate of $\tilde{v}$ depends on their fat-tail priors. It is helpful to study the ideal case that model risk vanishes. If there is no ambiguity in their prior, arbitrageurs become *subjective expected utility* optimizers, under their Laplace prior $\mathcal{L}(0, \xi)$ when $s = 1$.

Proposition 2.2. *In the absence of model risk, arbitrageurs maximize their expected profits. Over the liquidity regime $\mu > \mu_\epsilon \approx -0.2319$ where μ_ϵ is the largest root to the cubic equation $\mu^3 + 21\mu^2 + 35\mu + 7 = 0$, arbitrageurs do not trade at $t = 1$ and their optimal strategy at*

$t = 2$ is proportional to their posterior expectation of $\tilde{\theta} = \tilde{v} - p_1$ under the prior $\mathcal{L}(0, \xi)$:

$$Z^o_{2,n}(s, y_1; \xi) = s\frac{1-\mu}{N+1} \cdot \frac{\hat{v}(y_1;\xi) - \lambda_1 y_1}{2\lambda_1} = s\frac{1-\mu}{N+1} \cdot \frac{\hat{\theta}(y_1;\xi)}{2\lambda_1}, \qquad n = 1, ..., N. \tag{13}$$

The estimator $\hat{v}(y_1;\xi)$ is the posterior mean of $\tilde{v}$ under the prior that $\tilde{v}$ is drawn from $\mathcal{L}(0, \xi)$:

$$\hat{v} = \mathrm{E}^{\mathcal{A}}[\tilde{v}|y_1 = y'\sigma_u, \xi] = \frac{\kappa\xi(y' - \kappa)\mathrm{erfc}\left(\frac{\kappa - y'}{\sqrt{2}}\right)}{\mathrm{erfc}\left(\frac{\kappa - y'}{\sqrt{2}}\right) + e^{2\kappa y'}\mathrm{erfc}\left(\frac{\kappa + y'}{\sqrt{2}}\right)} + \frac{\kappa\xi(y' + \kappa)\mathrm{erfc}\left(\frac{\kappa + y'}{\sqrt{2}}\right)}{\mathrm{erfc}\left(\frac{\kappa + y'}{\sqrt{2}}\right) + e^{-2\kappa y'}\mathrm{erfc}\left(\frac{\kappa - y'}{\sqrt{2}}\right)}. \tag{14}$$

The rescaled estimator $\hat{v}/\xi$ is an increasing function of the rescaled quantity $y' = y_1/\sigma_u$, with one dimensionless shape parameter, $\kappa \equiv \frac{\rho\lambda_1\sigma_u}{\xi}$. The rational-expectations equilibrium (REE) corresponds to the case that their prior is correct, i.e., $\xi = \xi_v$. Under REE, $\kappa = \frac{2}{\sqrt{1+\mu}}$.

Proof. See Appendix A.2. □

Arbitrageurs only trade when fat-tail shocks occur. In the eyes of some econometrician who holds the Gaussian belief and trusts in market efficiency, those arbitrageurs seem to be "irrational" because they show up randomly and behave like feedback traders. This may raise various behavioral arguments, without recognizing the misspecification of priors.

Arbitrageurs' prior is symmetric (non-directional) at the beginning. They postpone arbitrage trading until they could tell the trading direction from past price movements, or equivalently, until their posterior beliefs become skewed. Proposition 2.2 confirms this no-trade conjecture in the first period. It also explains why this paper starts from a two-period setup. Even though arbitrageurs are better informed (with the knowledge of $\tilde{s}$) than market makers, their prior expectation of the stock value is identical to market makers'. Arbitrageurs have to watch the market first to see in which direction market makers incur pricing errors. This "wait-and-see" strategy suggests that arbitrage trading can be delayed for learning purposes so that mispricings may sustain for a longer period of time. The mechanism here is different from the delayed arbitrage discussed in Abreu and Brunnermeier (2002) where arbitrageurs face uncertainty about when their peers will exploit a common arbitrage opportunity.

The optimal strategy is symmetric with the past order flow: $Z^o_{2,n}(s, -y_1) = -Z^o_{2,n}(s, y_1)$. The rescaled strategy, $Z^o_{2,n}/\sigma_u$, is a function of the rescaled order flow $y' = y_1/\sigma_u$ in the fat-tail case. The optimal strategy becomes almost linear at large order flows. Its asymptotic slope is equal to the slope of linear strategy for traders who have a uniform prior ($\xi \to \infty$). Examination of the first and second derivatives leads to the following statement.

Corollary 2.2. *When $s = 1$, the optimal strategy $Z^o_{2,n}(s, y_1)$ is convex in the positive domain of y_1 and concave otherwise. It is asymptotically linear with a limit slope of $\frac{1-\mu}{(1+\mu)(N+1)}$.*

2.3 Robust Strategy under Model Risk

As indicated by Eq. (14), the estimator $\hat{v}$ depends on the dispersion of Laplace prior, ξ. How would arbitrageurs trade when they have uncertain priors? Model risk is a critical issue in statistical arbitrage, because using a wrong prior could yield a business disaster like the failure of LTCM. In the real world, traders often face the pressure to test the performance of their strategies in the worst-case scenario. This pressure can drive them to adopt alternative strategies that sacrifice some optimality for robustness.

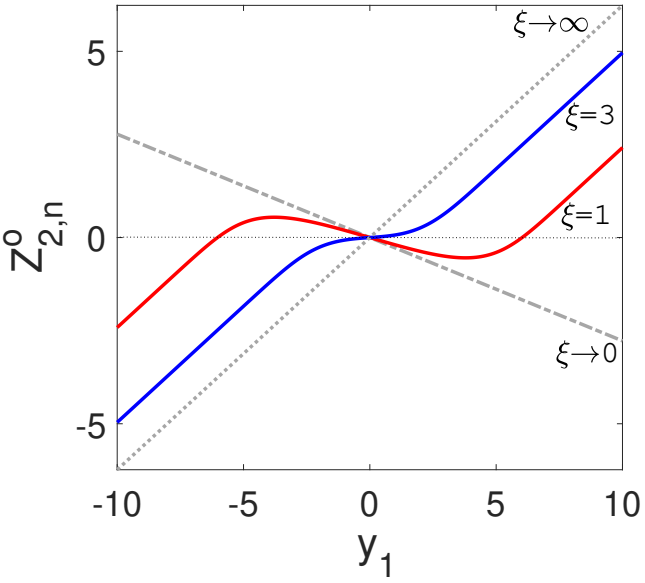


Figure 1. The optimal strategy $Z^o_{2,n}(s=1, y_1; \xi)$ in Eq. (13) under different values of ξ.

Fig. 1 shows the optimal strategy under different values of the Laplace parameter ξ. An arbitrageur with the prior $\xi \to 0$ believes that the stock value is unchanged (i.e., $\tilde{v} = 0$). This trader will attribute all the order flow y_1 to noise trading and trade against any price change. In contrast, an arbitrageur with the extreme prior $\xi \to \infty$ believes that the past order flow is dominated by informed trading and thus will chase the price trend straightly. For small ξ, arbitrageurs will engage in contrarian trading on small order flows which are dominated by noise trading under their belief. For large ξ, arbitrageurs always use a momentum strategy.

Suppose that arbitrageurs' uncertain prior $\tilde{\xi}$ is in the interval $[\xi_L, \xi_H]$, where both the highest and lowest priors, ξ_H and ξ_L, have non-zero chances. If the divergence between ξ_H and ξ_L is large enough, arbitrageurs can face ambiguity about the trading direction conditional on small order flows[20]: they may want to buy the asset under a high prior (for example,

[20]If the extreme priors satisfy $y_1 Z^o_{2,n}(s, y_1; \xi_H) > 0$ for any $y_1 \neq 0$ and $y_1 Z^o_{2,n}(s, y_1; \xi_L) \leq 0$ for a nonzero measure of y_1, then different fat-tail priors can give opposite trading directions at small order flows.

$\xi = 3$ in Fig. 1) but sell it under a low prior (for example, $\xi = 1$ in Fig. 1). If they use the wrong prior, they may trade in the wrong direction and undergo adverse fat-tail shocks.

By Assumption 2.2, arbitrageurs rank strategies based on the maximin decision criterion, i.e., each arbitrageur maximizes the minimum expected profit over a set of multiple priors. Pure maximin optimization can give very pessimistic decisions which stick to the least favorable prior even if it has a tiny chance to occur. To avoid over-pessimistic responses, I assume that arbitrageurs' admissible strategies converge to the averaged optimal strategy (across all priors) in a rational manner that preserves its convexity and/or concavity. Let's also enforce internal consistency: arbitrageurs inside this model "know" its structure in a statistical sense. On average, they are correct about the economy without systematic bias.

First, it is reasonable and important to invoke the convergence condition. If arbitrageurs observe an extremely large order flow y_1, they will be pretty sure that y_1 was dominated by informed trading in the fat-tail scenario. This resolves their ambiguity about trading directions and boosts their confidence to follow the averaged optimal strategy, $E^{\mathcal{A}}[Z^o_{2,n}(\tilde{s}, y_1; \tilde{\xi})|\tilde{s} = 1]$. Let Z^∞ denote the asymptotes of this averaged strategy. Simple derivation yields

$$Z^\infty(y_1, K_\xi) = \frac{1-\mu}{1+\mu} \cdot \frac{y_1 - \text{sign}(y_1)K_\xi}{N+1}, \quad \text{where} \quad K_\xi = \frac{\lambda_1 \rho^2 \sigma_u^2}{\rho - 1} \mathrm{E}^{\mathcal{A}}[\tilde{\xi}^{-1}]. \tag{15}$$

To ensure internal consistency, Eq. (15) should coincide with the asymptotes of the *rational-expectations equilibrium* (REE) strategy given the true prior ξ_v. This requires $\mathrm{E}^{\mathcal{A}}[\tilde{\xi}^{-1}] = \xi_v^{-1}$ under which the asymptotes becomes $Z^\infty(y_1, K^*)$ where

$$K^* = \frac{\lambda_1 \rho^2 \sigma_u^2}{(\rho - 1)} \xi_v^{-1} = \frac{3+\mu}{\sqrt{1+\mu}} \sigma_u = \frac{\sqrt{2}\sigma_v}{\lambda_1}. \tag{16}$$

The condition $\mathrm{E}^{\mathcal{A}}[\tilde{\xi}^{-1}] = \xi_v^{-1}$ means that arbitrageurs' average belief is correct regarding the *precision* of Laplace prior. Similar to the concept of *rational expectations*, arbitrageurs inside this model make unbiased predictions on average, despite their uncertainty about the model structure. Any candidate strategy should converge to $Z^\infty(y_1, K^*)$. This condition ensures that the strategy space of arbitrageurs is anchored to their REE strategy (benchmark).

Second, the admissible strategies should rationally preserve the convexity and/or concavity of the optimal strategy. By Corollary 2.2, any optimal strategy (without model risk) is convex in the positive domain and concave otherwise (Fig. 1). Thus, any candidate strategy must be convex in the regime of $y_1 > 0$ and concave in the regime of $y_1 < 0$. Without this convexity-preserving condition, traders would consider strategies with arbitrarily complex curvatures. This may cause over-fitting problems and make model interpretation difficult.

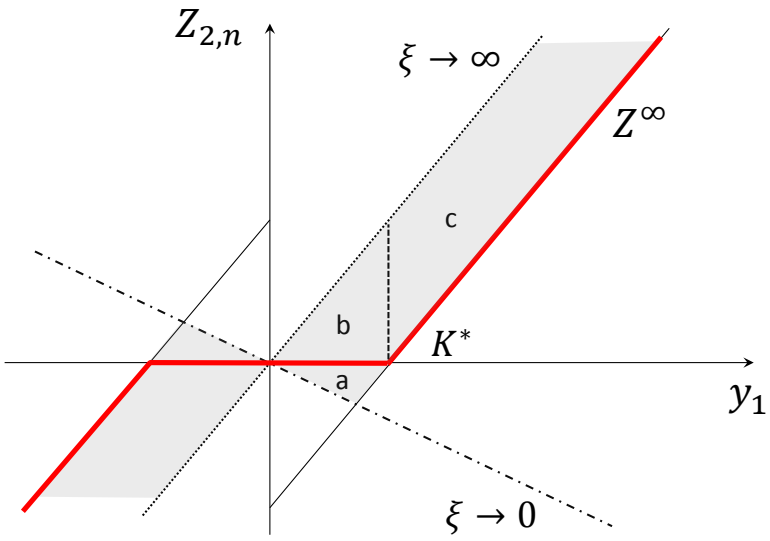


Figure 2. The robust strategy $Z_{2,n}$ in the presence of model risk.

Any strategy that converges to the REE strategy without losing its convex property must lie in the shaded areas of Fig. 2. Any strategy running outside this area violates either the convergence condition or the convexity-preserving rule. We can focus on the positive domain and divide the shaded area into three regions. For any $y_1 \in [0, K^*]$, arbitrageurs will not sell against y_1, because they may lose money if the highest prior ξ_H is true. This rules out any decision point inside the triangle "a". Similarly, arbitrageurs will not buy the stock since they may also lose money if the lowest prior ξ_L is true. This rules out any decision point inside the triangle "b". So the max-min choice criteria indicate a no-trade zone over $y_1 \in [0, K^*]$. Next, for any $y_1 > K^*$, ambiguity-averse traders should not trade a quantity more than the one prescribed by the REE asymptotes $Z^{\infty}(y_1, K^*)$; otherwise they may lose in the worst-case scenario. This argument rules out any decision point inside the region "c". By symmetry, the robust strategy turns out to be a piecewise linear function of y_1, with the trading threshold K^*. This simple strategy is labeled by the red line in Fig. 2.

Proposition 2.3. *If arbitrageurs face sufficient model uncertainty about the fat-tail priors and if they follow the max-min choice criteria to rank the admissible strategies defined before, then their robust strategy at $t = 2$ is a piece-wise linear function of the order flow at $t = 1$:*

$$Z_{2,n}(s, y_1; K^*) = sZ^{\infty}(y_1, K^*)\mathbf{1}_{|y_1|>K^*} = s\frac{1-\mu}{1+\mu} \cdot \frac{y_1 - \text{sign}(y_1)K^*}{N+1} \cdot \mathbf{1}_{|y_1|>K^*}, \tag{17}$$

which is along the REE asymptotes with the trading threshold K^* given by Eq. (16).

Proof. See Appendix A.3. □

The endogenous decision boundary K^* is independent of the number of arbitrageurs (N) or the variance of asset value (σ_v^2). For constant noise trading volatility ($\gamma = 1$), one can find $K^* \approx 3.063\sigma_u$ which is roughly three standard deviations of noise order flows. This indicates a very large inaction zone for the robust strategy. To see how inactive it is, let us examine the unconditional variance of the first-period total order flow, $\sigma_y^2 = \frac{\sigma_v^2}{(\rho\lambda_1)^2} + \sigma_u^2 = \frac{3+\mu}{2}\sigma_u^2$, which implies $K^* \approx 2.5483\sigma_y$. When the asset value $\tilde{v}$ is Laplacian, the probability that arbitrageurs get triggered to trade is very small, $P(|y_1| > K^*) \approx 1.33\%$. One might think that such a strategy is too inert to be profitable. This is not true. Numerically, the robust strategy can capture about 60% of the maximum profit recouped by the ideal REE strategy. This performance is surprisingly good given the idleness of the robust strategy. Fat-tail shocks create a disproportionate distribution of mispricings. The robust strategy is effective in picking up most profitable opportunities which correspond to those large fat-tail events.

So far, I have discussed various belief-related reasons for arbitrageurs' inaction. Their no-trade conditions are summarized as follows:

Corollary 2.3. *Arbitrageurs do not trade if any of the following conditions holds:*

(1) the market is efficient in the semi-strong form under their belief;
(2) their prior expectation of $\tilde{v}$ is identical to market makers' expectation;
(3) the past price change cannot drive them out of their inaction (ambiguity) zone.

Proof. Condition (1) holds at $\tilde{s} = 0$, Condition (2) holds for their decision making at $t = 1$, and Condition (3) is implied by Proposition 2.3. □

Given their idleness, it may well be the case that arbitrageurs are overlooked by the rest of the market. This is self-consistent with the implication of Assumptions 2.1, 2.2, and 2.3.

More importantly, given their no-trade strategy in the first period and inaction region in the second period, a lot of pricing errors can persist in this market. *Ex post*, an econometrician can run regressions on historical data to discover many mispricings in this economy. The econometrician may question the rationality or capability of arbitrageurs as they apparently leave money on the table. *Ex ante*, arbitrageurs assess all possible states using Bayes' rule. They are risk-neutral but ambiguity averse. For maximin robustness, they rationally ignore small profit opportunities which involve ambiguity about the trading direction. Neither financial constraints nor trading frictions exist here. There is no limit to arbitrageurs' trading ability. It is model risk that reduces their willingness to eliminate mispricings. This intrinsic friction is especially important in the fat-tail world where it leads to a large no-trade zone.

2.4 Equivalent Learning Rule and Alternative Interpretations

The optimal strategy without model risk uses the posterior *mean* estimate in Bayesian learning (Proposition 2.2). What is the learning mechanism behind the robust strategy? Arbitrageurs are Bayesian rational when they solve their maximin objectives, Eq. (6) and Eq. (7). It is noteworthy that the derived (robust) strategy is observationally equivalent to the *Least Absolute Shrinkage and Selection Operator* (LASSO), a famous machine-learning technique developed by Tibshirani (1996). The LASSO estimate can be interpreted as the posterior *mode* under independent Laplace prior. In statistics, the posterior *mode* is formally known as the *Maximum a Posteriori* (MAP) estimate. This learning rule itself lacks Bayesian rationality because it does not use all relevant information in forming expectations of unknown variables[21]. Nonetheless, the MAP estimate can "produce" the robust strategy.

Proposition 2.4. *If arbitrageurs know the true Laplace prior ξ_v but directly use the MAP learning rule to estimate the mispricing signal $\tilde{\theta} = \tilde{v} - p_1$, then their strategy in the second period will be operationally equivalent to the robust strategy in Proposition 2.3:*

$$Z_{2,n}(s=1, y_1; K^*) = \frac{\hat{\theta}_{map}}{2(N+1)\lambda_2} = \frac{(\hat{v}_{map} - \lambda_1 y_1)\mathbf{1}_{|y_1|>K^*}}{2(N+1)\lambda_2}. \tag{18}$$

Here, $\hat{\theta}_{map}$ is the MAP estimate of $\tilde{\theta}$. It contains $\hat{v}_{map}$ which is the MAP estimate of $\tilde{v}$ under the prior $\mathcal{L}(0, \xi_v)$. This is a soft-thresholding function with a threshold $\kappa\sigma_u = \frac{\rho\lambda_1\sigma_u^2}{\xi_v} = \frac{2\sigma_u}{\sqrt{1+\mu}}$:

$$\hat{v}_{map}(y_1; \xi_v) = \rho\lambda_1[y_1 - \text{sign}(y_1)\kappa\sigma_u]\mathbf{1}_{|y_1|>\kappa\sigma_u}. \tag{19}$$

Proof. See Appendix A.4. □

Fig. 3 compares the learning rules and their associated strategies. Both the posterior mean estimate $\hat{v}$ and the REE strategy $Z^o_{2,n}(s=1, y_1; \xi_v)$ are smooth and nonlinear. In contrast, the posterior mode estimate $\hat{v}_{map}$ is zero for $y_1 \in [-\kappa\sigma_u, \kappa\sigma_u]$ and linear beyond that zone. The robust strategy $Z_{2,n}$ has a similar pattern as it performs linear momentum trading beyond the inaction zone $[-K^*, K^*]$. Traders who follow this strategy only respond to large events and deliberately ignore small ones. This rational response is similar to various behavioral patterns, including limited attention, status quo bias, anchoring and adjustment, among others[22]. Again, it is worth stressing that arbitrageurs are Bayesian-rational here: they evaluate all possible states using Bayes rule and maximize their well-defined utility with

[21] The MAP estimate of a variable equals the mode of the posterior distribution. As a point estimate, it does not summarize all relevant information in the posterior distribution.

[22] Barberis and Thaler (2003) provide an excellent survey on those topics in behavioral finance.

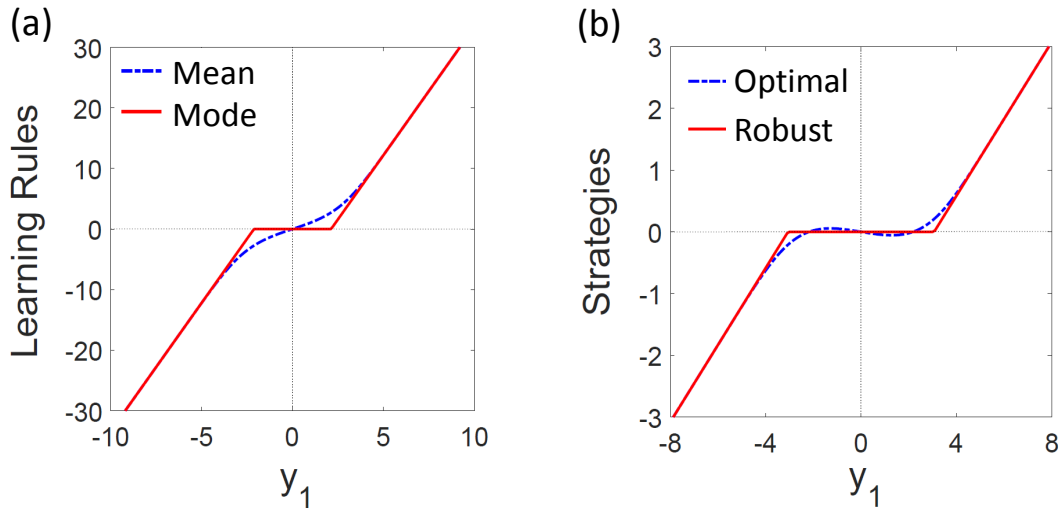


Figure 3. (a) The posterior mean versus the posterior mode of $\tilde{v}$ under the Laplace prior $\mathcal{L}(0, \xi_v)$. (b) the optimal (REE) strategy versus the robust strategy at $t = 2$ when $s = 1$.

sequential rationality. One can apply Propositions 2.3 and 2.4 to rationalize the behavioral assumption of Gabaix (2014). In his model, the *soft-thresholding* function like Eq. (19) is used to describe the anchoring bias. Such behavior also permits a rational interpretation.

In a multi-asset economy subject to uncertain fat-tail shocks, Proposition 2.4 implies that arbitrageurs can directly incorporate the LASSO algorithm into their trading system:

Corollary 2.4. *Suppose that arbitrageurs identify $M \geq 1$ assets with independent and identically distributed liquidation values, $\tilde{v}_i \sim \mathcal{L}(0, \xi_v)$ for $i = 1, ..., M$, and each of these assets is traded by a single informed trader in the two-period Kyle model with constant noise trading. For robust learning, arbitrageurs solve the LASSO objective in the Lagrangian form:*

$$\min_{\{v_1,...,v_M\}} \sum_{i=1}^{M} \left| p_{1,i} - \frac{v_i}{\rho} \right|^2 + \frac{2(\lambda_1 \sigma_u)^2}{\xi_v} |v_i|, \tag{20}$$

where $p_{1,i} = \lambda_1 y_{1,i}$ is the price change of the i-th asset and ρ^{-1} is the percentage of private signal that has been incorporated into the asset price at $t = 1$. This leads to a simple strategy

$$Z_{2,n}(p_{1,i}, \xi_v) = \frac{\rho - 1}{N + 1} \cdot \frac{p_{1,i} \pm 2\xi_v}{2\lambda_2} \cdot \mathbf{1}_{|p_{1,i}| \geq 2\xi_v}, \qquad \text{for} \quad i = 1, ..., M, \tag{21}$$

which is automatically triggered to trade the i-th asset if its price change $p_{1,i}$ exceeds $\pm 2\xi_v$.

Proof. See Appendix A.4. □

The objective of maximizing the posterior (under MAP) is equivalent to the minimization problem Eq. (20). It involves an l^1 penalty term that comes from the Laplace prior $\mathcal{L}(0, \xi_v)$. LASSO shrinks certain estimation coefficients to zero and effectively selects a simpler model that exclude those coefficients. This is a popular tool among quantitative traders because it picks up a small number of key features (factors) from a large set of candidate features. For traders who use LASSO, their trading models shall involve fat-tail (typically Laplace) priors. If traders use the Gaussian prior instead, they will incur an l^2 penalty in their objective. The resulted algorithm is *ridge regression* which uniformly shrinks the size of all coefficients but does not send any coefficients to zero. Even with parameter uncertainty about the Gaussian prior, traders will not get an inaction zone. This is because signal inference is linear when the posterior is Gaussian. For symmetric unimodal distributions, the mean coincides with the mode; the two learning rules will give identical predictions. Since different Gaussian priors only change the slopes of linear responses, the maximin robust strategy in a pure Gaussian-mixture model will be linear; see Appendix A.4 for more details.

Corollary 2.4 can help explain the momentum strategy and anomaly in asset pricing[23]. Short-term momentum traders can be viewed as statistical arbitrageurs who have uncertain fat-tail priors about mispriced stocks. Their robust trading is exactly the momentum strategy of buying winners and selling losers. Those traders usually focus on top market gainers and losers, instead of the entire universe of equities. Corollary 2.4 can also be used to interpret rule-based algorithmic trading which gets triggered at some predefined price levels. At first glance, such trading behavior seems to be mechanical and at odds with Bayesian rationality. It is possible that algorithmic traders are Bayesian-rational. They may use machine-learning techniques (such as LASSO) to manage unknown risks or improve prediction accuracy.

The robust LASSO strategy can also be used by market makers for error self-correction. Market makers can split their pricing logic into two programs. The first one is the linear pricing strategy which allows them to almost break even, despite their occasional mistakes. The second program uses the fat-tail prior to correct the errors of linear pricing strategy, just like the actions of arbitrageurs. This leads to the LASSO algorithm. Integrating both programs, market makers can keep using the linear pricing rule until their inventory exceeds the endogenous thresholds. At that point, they will switch to momentum trading and reduce excessive inventories. The no-trade zone in the second program is the ambiguity zone where they hesitate to correct uncertain pricing errors; this no-trade zone is also their comfortable zone to do market making. This new interpretation differs from conventional arguments that market makers' inventory limits are due to their high risk aversion or large inventory costs.

[23] See Jegadeesh and Titman (1993), Chan, Jegadeesh, and Lakonishok (1996), Carhart (1997), Hong and Stein (1999), Daniel, Hirshleifer, and Subrahmanyam (1998), Lee and Swaminathan (2000), among others.

2.5 Cartel Effect and Market Inefficiency

Arbitrageurs trade conservatively beyond the endogenous inaction zone. Their conservative trading facilitates their tacit collusion which mitigates their competition and impedes market efficiency. This has interesting implications for limits to arbitrage.

Proposition 2.5. *As $N \to \infty$, the total profit of arbitrageurs vanishes if they use the REE strategy. However, their total profit has a positive limit if they follow the robust strategy.*

Proof. If arbitrageurs all follow the optimal REE strategy $Z^o_{2,n}(s, y_1; \xi_v)$, they will compete away their total arbitrage profit when N goes to infinity:

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}\left[\sum_{n=1}^{N}(\tilde{v}-\tilde{p}_2)Z^o_{2,n}\right] &= \lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}\left[N\left(\tilde{v}-\lambda_1\tilde{y}_1-\lambda_2 X_2(\tilde{v},\tilde{y}_1)-N\lambda_2 Z^o_{2,n}\right)Z^o_{2,n}\right] \\
&= \lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}\left[\frac{(N+1)(\tilde{v}-\lambda_1\tilde{y}_1)-N(\hat{v}-\lambda_1\tilde{y}_1)}{2(N+1)}\cdot\frac{N(\hat{v}-\lambda_1\tilde{y}_1)}{2(N+1)\lambda_2}\right] \\
&= \lim_{N\to\infty} \frac{N}{4(N+1)^2\lambda_2}\mathrm{E}^{\mathcal{A}}[(\hat{v}(\tilde{y}_1)-\lambda_1\tilde{y}_1)^2] = 0, \qquad (22)
\end{aligned}$$

where in the above derivation we have used Eq. (11) and $\mathrm{E}^{\mathcal{A}}[\tilde{v}] = \mathrm{E}^{\mathcal{A}}[\mathrm{E}^{\mathcal{A}}[\tilde{v}|\tilde{y}_1]] = \mathrm{E}^{\mathcal{A}}[\hat{v}(\tilde{y}_1)]$.

In contrast, if arbitrageurs follow the robust strategy $Z_{2,n}(s, y_1; K^*)$, their total arbitrage profit will converge to a positive value, indicating a cartel effect:

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}\left[\sum_{n=1}^{N}(\tilde{v}-\tilde{p}_2)Z_{2,n}\right] &= \lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}\left[\frac{(N+1)(\tilde{v}-\lambda_1\tilde{y}_1)-N\hat{\theta}_{map}}{2(N+1)}\cdot\frac{N\hat{\theta}_{map}}{2(N+1)\lambda_2}\right] \\
&= \lim_{N\to\infty} \frac{\mathrm{E}^{\mathcal{A}}[N(N+1)(\hat{v}-\hat{v}_{map}+\hat{v}_{map}-\lambda_1\tilde{y}_1)\hat{\theta}_{map}-N^2\hat{\theta}^2_{map}]}{4(N+1)^2\lambda_2} \\
&= \frac{\mathrm{E}^{\mathcal{A}}[(\hat{v}-\hat{v}_{map})\hat{\theta}_{map}]}{4\lambda_2} > 0, \qquad (23)
\end{aligned}$$

where in the above derivation we have used Eq. (18) and $\mathrm{E}^{\mathcal{A}}[\tilde{v}] = \mathrm{E}^{\mathcal{A}}[\hat{v}(\tilde{y}_1)]$. The expression of the MAP estimate $\hat{\theta}_{map} \equiv (\hat{v}_{map}-\lambda_1\tilde{y}_1)\mathbf{1}_{|\tilde{y}_1|>K^*}$ implies $(\hat{v}_{map}-\lambda_1\tilde{y}_1)\hat{\theta}_{map} = \hat{\theta}^2_{map}$. The last expression is strictly positive because $(\hat{v}-\hat{v}_{map})$ and $\hat{\theta}_{map}$ has the same sign for $|\tilde{y}_1| > K^*$. □

Fig. 4(a) shows the total profit of a hundred arbitrageurs who follow the robust strategy, conditional on the observed order flow y_1. This profit profile (red curve) is proportional to the term $(\hat{v} - \hat{v}_{map}) \cdot \hat{\theta}_{map}$ in Eq. (23). It exhibits two spikes of profits beyond the trading thresholds (labeled by blue circles). These spikes indicate the major source of their extra

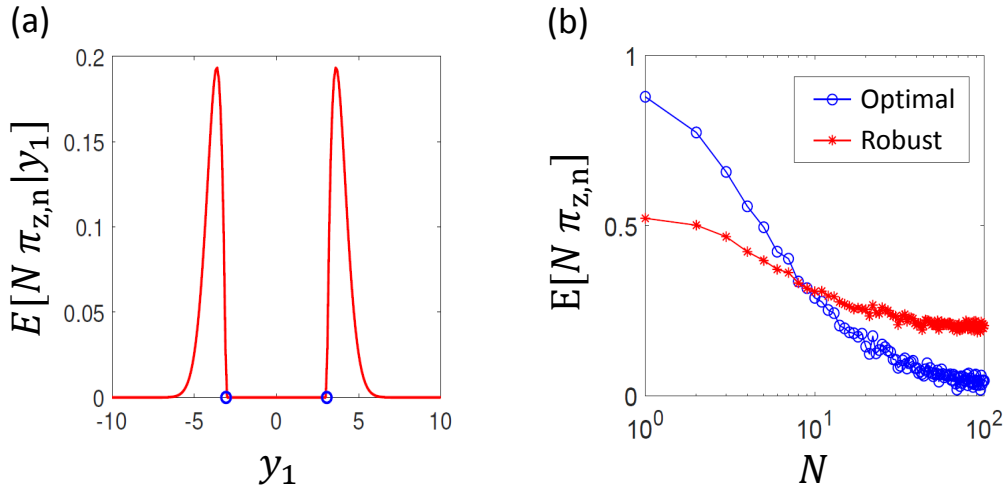


Figure 4. (a) The arbitrageurs' total profit under the robust strategy conditional on y_1. (b) The total arbitrage profit under the REE strategy vs. that under the robust strategy.

profits. Intuitively, arbitrageurs' under-trading is most prominent near the "kinks" of their robust strategy. Their non-competitive profits must be strongest there.

Fig. 4(b) compares the total payoffs to arbitrageurs when they follow different types of strategies. In the oligopolistic case (i.e., small N), the REE strategy allows them to earn higher profits, because the robust strategy ignores a wide range of profit opportunities. As N increases, the profitability of the REE strategy decays faster. In the competitive limit, arbitrageurs compete away their profits under REE and restore market efficiency at $t = 2$.

In contrast, arbitrageurs' total payoff converges to a positive value when they follow the robust strategy [Fig. 4(b)]. This confirms Proposition 2.5 and indicates a non-competitive effect. Their positive limiting payoff is attributable to the market power they amass beyond the inaction zone, where they trade less aggressively than they would do under REE [Fig. 3 and Fig. 4(a)]. This collusive behavior does not involve any communication device or explicit agreement. Their tacit collusion is not a result of financial constraints or trading frictions. It is due to traders' robust control for (non-Gaussian) model risk. Outside their inaction region, the cartel effect will prevent the market from being fully efficient .

Corollary 2.5. *In the limit $N \to \infty$, arbitrageurs will restore market efficiency when they follow the REE strategy, i.e.,* $\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}[P_2(\tilde{y}_1, \tilde{y}_2)|\tilde{y}_1] = \mathrm{E}^{\mathcal{A}}[\tilde{v}|\tilde{y}_1]$ *under* $Z^o_{2,n}(s, y_1; \xi_v)$ *for* $n = 1, ..., N$*; however, market efficiency is hindered when a finite fraction of arbitrageurs follow the robust strategy, i.e.,* $\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}[P_2(\tilde{y}_1, \tilde{y}_2)|\tilde{y}_1] \neq \mathrm{E}^{\mathcal{A}}[\tilde{v}|\tilde{y}_1]$ *under* $Z_{2,n}(s, y_1; K^*)$.

Proof. See Appendix A.5. □

By Corollary 2.5, it is difficult to restore market efficiency even if the economy hosts an infinite number of risk-neutral arbitrageurs. To restore price efficiency in the second period, it requires that (almost) every arbitrageur follows the REE strategy, that is, (almost) every arbitrageur knows *on average* the correct fat-tail prior and has no aversion to uncertainty. This is practically impossible because real-life arbitrageurs face different levels of model risks. Moreover, there exist both internal and external pressures that force them to manage such risks. Their robust control easily translates to their ambiguity aversion, which significantly limits their *willingness* to eliminate mispricings. As reviewed in Gromb and Vayanos (2010), existing studies mostly focus on different costs that limits arbitrageurs' *ability* in trading. Those frictions could be eased by injecting sufficient capital or removing certain constraints. The mechanism here is different. First, model risk is an intrinsic problem which may not be resolved easily. Second, arbitrageurs here are *able* to eliminate pricing errors; they hesitate to do so because of their aversion to uncertainty[24]. Third, arbitrageurs' hesitation in arbitrage has two characteristics: (1) the large inaction region tells them to leave money on the table; (2) their undertrading beyond the inaction region supports them as a "cartel". Consequently, even with an infinite number of risk-neutral arbitrageurs, a wide range of pricing errors can persist in this economy. This is an endogenous outcome of model risk.

Nowadays, financial markets have been largely occupied by algorithmic traders. The surge of quantitative modeling and machine-learning techniques can bring about hidden issues. The present paper demonstrates that statistical arbitrageurs can use machine-learning tools to combat model uncertainty and similar algorithmic "kinks" in their strategy can mitigate their competition at the expense of market efficiency. This is a general implication, given that many machine-learning algorithms have inaction regions and decision "kinks".

Equilibrium Condition. In the liquidity regime $\mu < 0$, an arbitrageur may find it profitable to trade in the first period and take advantage of the aggressive feedback trading of other arbitrageurs. One can verify Eq. (6) to see whether this unilateral deviation is profitable.

Corollary 2.6. *The conjectured equilibrium strategy profile may fail in the liquidity regime $\mu < \mu^*(N)$, where $\mu^*(N)$ is the largest root that solves $1 + \frac{N-1}{N+1} \cdot \frac{2}{1+\mu} = \frac{4}{\sqrt{1-\mu}}$. Given a large number of arbitrageurs using the same robust strategy, it can be profitable for an individual trader to deviate from the conjectured no-trade strategy in the first period. This deviation involves trading a large quantity $z_1 \gg K^*$ to trigger the other arbitrageurs and then unwinding the position at more favorable prices supported by the over-aggressive trading of others.*

Proof. See Appendix A.6 □

[24] Arbitrageurs are risk-neutral but ambiguity-averse in this setup. Their hesitation to perform arbitrage trading is not due to their risk aversion.

3 Model of Savvy Informed Trader

In this section, I extend the previous model to investigate how strategic interaction between the informed trader and the arbitrageurs affect equilibrium outcomes. This model extension can be interpreted as an institutional informed trader optimizes the dynamic order-execution algorithm by taking in account the responses of algorithmic arbitrageurs who use simple machine-learning strategies to exploit her trades. The extended model can be used, for example, to analyze controversial issues in algorithmic trading. It can shed light on hidden risks when algorithmic traders pervade financial markets. Such risks may account for market vulnerability and deserve more attention from regulators.

Let us consider a *savvy informed trader* who observes simultaneously the asset value $\tilde{v}$ and the distribution-type signal $\tilde{s}$ at the beginning. She anticipates the momentum trading of arbitrageurs and behaves strategically. In the Laplacian case, she will consider how her initial trading affects arbitrageurs' next responses. By backward induction, her expected total profit contains a nonlinear term reflecting her consideration of arbitrageurs' nonlinear inference. As a result, her first-period trading strategy is no longer linear and the *rational-expectations equilibrium* (REE) becomes intractable; more discussions are available in Appendix A.7.

To gain insights, the analysis in this section is devoted to a tractable model where strategic arbitrageurs only consider linear-triggering strategies that converge to the REE. This model keeps the basic structure (Table 1) elaborated in the previous section. I present a set of new assumptions to clarify traders' belief systems and information sets.

Assumption 3.1. *As common knowledge, this market has fixed linear pricing schedules,* $\tilde{p}_1 = P_1(\tilde{y}_1) = \lambda_1 \tilde{y}_1$ *and* $\tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2) = \lambda_1 \tilde{y}_1 + \lambda_2 \tilde{y}_2$*, that are exogenously given by Eq. (9).*

Assumption 3.2. *Arbitrageurs observe* $\tilde{s}$ *and have the correct priors:* $\mathcal{N}(0, \sigma_v^2)$ *at* $\tilde{s} = 0$ *and* $\mathcal{L}(0, \xi_v)$ *at* $\tilde{s} = 1$*. For simplicity, arbitrageurs only consider linear-triggering strategies of the form*[25]*:* $Z_{2,n}(s = 1, y_1; K_n) = Z^{\infty}(y_1, \xi_v)\mathbf{1}_{|Y_1|>K_n}$*, where* Z^{∞} *denotes the asymptotes of their REE strategy to be determined in the limit REE. Each arbitrageur chooses the optimal threshold, taking as given the best responses of other arbitrageurs and the informed trader.*

Assumption 3.3. *The risk-neutral informed trader observes both* $\tilde{v}$ *and* $\tilde{s}$ *at* $t = 0$*. This fact and Assumption 3.3 are held as common knowledge among the informed trader and arbitrageurs. In other words, the informed trader knows everything known by the arbitrageurs, including their prior belief and their adherence to linear-triggering strategies. Arbitrageurs also know everything known by the informed trader except the private information* $\tilde{v}$*.*

[25] Using linear-triggering strategies, arbitrageurs implicitly conjecture that the informed trader's strategy increases with her private signal. However, the Bayesian-rational strategy is not necessarily monotone.

The above assumptions put our focus on the strategic interplay between informed trader and arbitrageurs. The linear pricing rule in Assumption 3.1 can hold when market makers believe that they are living in the two-period Kyel model with the Gaussian prior $\tilde{v} \sim \mathcal{N}(0, \sigma_v^2)$. Arbitrageurs' adherence to linear-triggering strategies in Assumption 3.2 is motivated by the robust strategy discovered in Section 2. If traders worry about the complexity or overtrading of the REE strategy, they may favor such simple algorithms. The suggested linear-triggering strategies are determined by three parameters: slope, intercept, and threshold. These provide well-defined trading rules amenable for computerized executions. Assumption 3.3 explains the "savviness" of this informed trader who is Bayesian-rational, has correct knowledge about the information structure, and anticipates the strategy space of arbitrageurs.

The timeline of this model is identical to Table 1, except that the informed trader observes both $\tilde{v}$ and $\tilde{s}$ at $t = 0$. The strategies of informed trader and arbitrageurs are denoted by $\mathbf{X} = \langle X_1, X_2 \rangle$ and $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_N]$, where $\mathbf{Z}_n = \langle Z_{1,n}, Z_{2,n} \rangle$ is the n-th arbitrageur's strategy for $n = 1, ..., N$. The informed trader knows $\mathcal{I}_{1,x} = \{\tilde{v}, \tilde{s}\}$ before trading at $t = 1$ and $\mathcal{I}_{2,x} = \{\tilde{v}, \tilde{s}, \tilde{y}_1\}$ before trading at $t = 2$. We can write $\tilde{x}_1 = X_1(\tilde{v}, \tilde{s})$ and $\tilde{x}_2 = X_2(\tilde{v}, \tilde{s}, \tilde{y}_1)$. Given the information sets of arbitrageurs, $\mathcal{I}_{1,z} = \{\tilde{s}\}$ and $\mathcal{I}_{2,z} = \{\tilde{s}, \tilde{y}_1\}$, it is justified to write $\tilde{z}_{1,n} = Z_{1,n}(\tilde{s})$ and $\tilde{z}_{2,n} = Z_{2,n}(\tilde{s}, \tilde{y}_1)$ for $n = 1, ..., N$. Let $\tilde{\pi}_x = \sum_{t=1}^{2} (\tilde{v} - \tilde{p}_t)\tilde{x}_t$ be the informed trader's profit, and $\tilde{\pi}_{z,n} = \sum_{t=1}^{2} (\tilde{v} - \tilde{p}_t)\tilde{z}_{t,n}$ be the n-th arbitrageur's profit. It is common knowledge that the market-clearing prices are

$$\tilde{p}_1 = P_1(\tilde{y}_1) = \lambda_1 \tilde{y}_1 = \lambda_1 \left(X_1(\tilde{s}, \tilde{v}) + \sum_{n=1}^{N} Z_{1,n}(\tilde{s}) + \tilde{u}_1 \right), \tag{24}$$

$$\tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2) = \tilde{p}_1 + \lambda_2 \tilde{y}_2 = \lambda_1 \tilde{y}_1 + \lambda_2 \left(X_2(\tilde{s}, \tilde{v}, \tilde{y}_1) + \sum_{n=1}^{N} Z_{2,n}(\tilde{s}, \tilde{y}_1) + \tilde{u}_2 \right). \tag{25}$$

To stress the dependence of prices on the strategies of traders, we write $\tilde{p}_t = \tilde{p}_t(\mathbf{X}, \mathbf{Z})$ for $t = 1, 2$. We also write $\tilde{\pi}_x = \tilde{\pi}_x(\mathbf{X}, \mathbf{Z})$ and $\tilde{\pi}_{z,n} = \tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{Z})$ because the strategy of informed trader will affect the trading profits of arbitrageurs through direct competition and learning interference, and arbitrageurs' strategies also affect the informed trader's profits through competition and strategic interaction.

In this model, the informed trader and arbitrageurs have the same (consistent) belief system. In particular, they have correct common knowledge about the mixture distribution of $\tilde{v}$. Since $\tilde{s}$ is observed by all of them at $t = 0$, the informed trader is aware of the time at which arbitrageurs may trade. However, the informed trader cannot fool arbitrageurs into believing a different type of $\tilde{v}$. It is also common knowledge among them that every arbitrageur adheres to the linear-triggering strategy with only one choice variable: the trading threshold.

Definition of Equilibrium. The equilibrium here is defined as a pair of strategies $(\mathbf{X}, \mathbf{Z})$ such that, under the market-clearing prices Eq. (24) and Eq. (25), the following conditions hold:

1. For any alternative strategy $\mathbf{X}' = \langle X_1', X_2' \rangle$ differing from $\mathbf{X} = \langle X_1, X_2 \rangle$, the strategy $\mathbf{X}$ yields an expected total profit no less than $\mathbf{X}'$, and also X_2 yields an expected profit in the second period no less than any single deviation X_2':

$$\begin{aligned} \mathrm{E}[\tilde{\pi}_x(\mathbf{X}, \mathbf{Z})|\tilde{v}, \tilde{s}] &\geq \mathrm{E}[\tilde{\pi}_x(\mathbf{X}', \mathbf{Z})|\tilde{v}, \tilde{s}], & (26) \\ \mathrm{E}[(\tilde{v} - \tilde{p}_2(\langle X_1, X_2 \rangle, \mathbf{Z}))X_2|\tilde{v}, \tilde{s}, \tilde{y}_1] &\geq \mathrm{E}[(\tilde{v} - \tilde{p}_2(\langle X_1, X_2' \rangle, \mathbf{Z}))X_2'|\tilde{v}, \tilde{s}, \tilde{y}_1] & (27) \end{aligned}$$

2. For all $n = 1, ..., N$ and for any alternative strategy profile $\mathbf{Z}'$ differing from $\mathbf{Z}$ only in the n-th component $\mathbf{Z}_n' = \langle Z_{1,n}', Z_{2,n}' \rangle$, the strategy $\mathbf{Z}$ yields an expected profit no less than $\mathbf{Z}'$, and also $Z_{2,n}$ yields an expected profit in the second period no less than $Z_{2,n}'$:

$$\begin{aligned} \mathrm{E}[\tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{Z})|\tilde{s}] &\geq \mathrm{E}[\tilde{\pi}_{z,n}(\mathbf{X}, \mathbf{Z}')|\tilde{s}], & (28) \\ \mathrm{E}[(\tilde{v} - \tilde{p}_2(\cdot, Z_{2,n}))Z_{2,n}|\tilde{s}, \tilde{y}_1] &\geq \mathrm{E}[(\tilde{v} - \tilde{p}_2(\cdot, Z_{2,n}'))Z_{2,n}'|\tilde{s}, \tilde{y}_1]. & (29) \end{aligned}$$

   The strategy profile on the right hand side of Eq. (29) only differs from $(\mathbf{X}, \mathbf{Z})$ at $Z_{2,n}$.

In the Gaussian case, the informed trader's strategy remains the same as those in Proposition 2.1; arbitrageurs find no trading opportunity in this efficient market. To solve the equilibrium in the fat-tail case, it is useful to conjecture first and verify later that arbitrageurs will not trade in the first period. We first solve their second-period optimal strategy under this no-trade conjecture and then check if it is indeed unprofitable for any arbitrageur to trade in the first period. There is another implicit conjecture in the model development. To follow the linear-triggering strategies, arbitrageurs think that the informed trader plays a monotone strategy which increases with her private signal. This needs to be verified too.

3.1 Equilibrium with Linear-Triggering Strategies

In the fat-tail case, large order flows at $t = 1$ are mostly attributable to the informed trading. This simplifies the inference problem for arbitrageurs as they can conjecture that

$$X_1(s = 1, v) \to \frac{v}{\rho\lambda_1} + \mathrm{sign}(v)c\kappa\sigma_u, \tag{30}$$

where ρ and c are parameters to be determined in the limit equilibrium. The intercept term, $c\kappa\sigma_u$, reflects how the informed trader exploits her opponents' learning bias, $\kappa\sigma_u = \frac{\rho\lambda_1\sigma_u^2}{\xi_v}$. If

Eq. (30) holds, the arbitrageurs' estimate of $\tilde{v}$ will be asymptotically linear with the past order flow. In Appendix A.8, I solve the asymptotic $X_1(s=1,v)$ and derive two algebraic equations for ρ and c. Their solutions are given by

$$\rho(\mu,N)=\frac{2+5N+N^2+2\mu-N\mu-(N+2)\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}{2N(1-\mu)}, \tag{31}$$

$$c(\mu,N)=-\frac{3+N-\mu-\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}{1+N+\mu+\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}\cdot\frac{N}{2}. \tag{32}$$

Here, the parameter ρ decreases with μ and N, because poorer liquidity or higher competitive pressure tomorrow can stimulate more aggressive informed trading today. The parameter c increases (with μ) from -1 to 0, because poor future liquidity tends to discourage strategic actions; as shown in Appendix A.9, this parameter reflects the extent of how the informed trader strategically exploits the estimation bias of arbitrageurs. These two parameters can determine the REE asymptotes, Z^∞, which helps us to pin down the following equilibrium.

Proposition 3.1. *In the liquidity regime of $\mu>\mu_\epsilon$ where $\mu_\epsilon\approx 0.005$ according to numerical results, the following equilibrium $(\mathbf{X},\mathbf{Z})$ holds. First, arbitrageurs do not trade in the first period, i.e., $Z_{1,n}=0$ for $n=1,...,N$. Their optimal linear-triggering strategy at $t=2$ is*

$$Z_{2,n}(s,y_1;K^*)=sZ^\infty(y_1,\xi_v)\mathbf{1}_{|y_1|>K^*}=s\frac{(1-\mu)(\rho-1)}{N+2}\left[y_1-\text{sign}(y_1)\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]\mathbf{1}_{|y_1|>K^*}, \tag{33}$$

$$K^*(\mu,N)=\max\left[\kappa\sigma_u,\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]=\sigma_u\frac{2\sqrt{1+\mu}}{3+\mu}\max\left[\rho,\frac{\rho^2(1+c)}{\rho-1}\right]. \tag{34}$$

For the informed trader, the equilibrium strategy at $t=2$ is to trade

$$X_2(v,s,y_1;K^*)=(1-\mu)\frac{v-\lambda_1 y_1}{2\lambda_1}-s\frac{NZ^\infty(y_1)\mathbf{1}_{|y_1|>K^*}}{2}. \tag{35}$$

The strategy at $t=1$ is monotone with her signal and solved by Eq. (126) in Appendix A.10.

Proof. See Appendix A.10. □

Corollary 3.1. *The linear-triggering strategy Eq. (33) implies the heuristic learning rule, $\hat{\theta}_T=s\cdot(\hat{v}_T-\lambda_1 y_1)\mathbf{1}_{|y_1|>K^*}$, which estimates $\tilde{\theta}=\tilde{v}-p_1$, with*

$$\hat{v}_T(y_1;\xi_v)=\rho\lambda\left[y_1-\text{sign}(y_1)(1+c)\kappa\sigma_u\right]\mathbf{1}_{|y_1|>\kappa\sigma_u}. \tag{36}$$

Proof. See Appendix A.10 as well. □

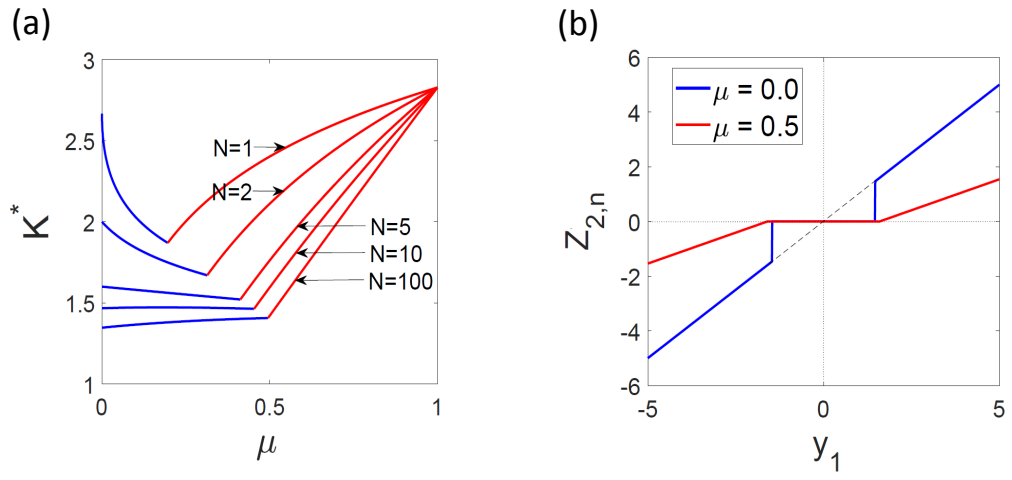


Figure 5. The threshold $K^*(\mu, N)$ and the strategy $Z_{2,n}(s, y_1; K^*)$ in two liquidity regimes.

The learning rule $\hat{v}_T$ looks similar to the MAP estimator $\hat{v}_{map}$ in Eq. (19), except that the horizontal intercept differs by a factor $(1 + c)$. The learning threshold, $\kappa\sigma_u \equiv \frac{\rho\lambda_1\sigma_u^2}{\xi_v}$, is independent of the parameter c, because parallel shifts of the informed trading strategy do not change the signal-to-noise ratio perceived by arbitrageurs. This learning threshold depends on the parameter ρ, because more aggressive informed trading (smaller ρ) can make arbitrageurs learn faster (smaller $\kappa\sigma_u$). The overall learning rule, $\hat{\theta}_T(y_1; K^*)$, is governed by the threshold K^*, which is the maximum of learning threshold $\kappa\sigma_u$ and strategic intercept term $\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}$. Since this intercept increases (with μ) from 0 to $2\kappa\sigma_u$, it must cross $\kappa\sigma_u$ at some intermediate value of μ. This indicates a kink in the equilibrium threshold:

Corollary 3.2. *There are two liquidity regimes separated by the critical liquidity value*

$$\mu_c(N) = \sqrt{N(N+2)^3} - N(N+3) - 1 \in \left[3\sqrt{3} - 5,\ \frac{1}{2}\right]. \tag{37}$$

For $\mu \in [0, \mu_c]$, $Z_{2,n}(s, y_1; K^*)$ is discontinuous at $|y_1| = K^* = \kappa\sigma_u$ which decreases with μ.
For $\mu \in [\mu_c, 1]$, $Z_{2,n}(s, y_1; K^*)$ is continuous and has $K^* = \frac{\rho(1+c)}{\rho-1}\kappa\sigma_u$ which increases with μ.

Proof. The critical liquidity μ_c is set by the crossover condition $1 = \frac{\rho(1+c)}{\rho-1}$ or $1 + \rho c = 0$. □

The rescaled threshold K^*/σ_u only depends on the liquidity level μ and the competition condition N (Fig. 5). Under good liquidity $\mu \in [0, \mu_c]$, the equilibrium threshold is set by the learning hurdle of $\hat{v}_T$, i.e., $K^* = \kappa\sigma_u$. Traders who use a threshold lower than $\kappa\sigma_u$ may engage in unjustified trading for a range of states where their estimated signal $\hat{v}_T$ is zero.

Under poor liquidity $\mu \in [\mu_c, 1]$, the equilibrium threshold is set by the horizontal intercept of $\hat{\theta}_T$, i.e., $K^* = \frac{\rho(1+c)}{\rho-1}\kappa\sigma_u$. Traders who use a threshold lower than this may do contrarian trading for a range of states where their estimated residual signal $\hat{\theta}_T$ is zero. Arbitrageurs will keep undercutting their thresholds as far as possible[26] until they hit the lower bound K^* in Eq. (34) which excludes contrarian trading or any unjustified trading.

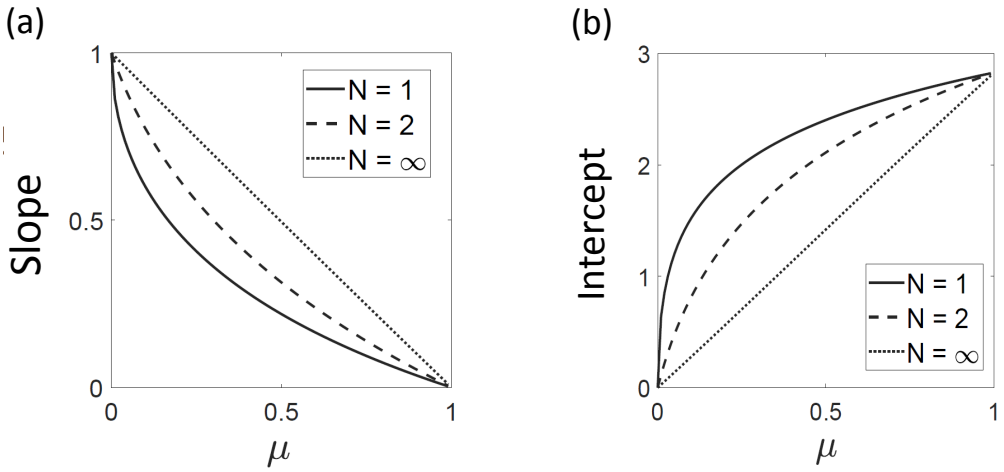


Figure 6. The slope and intercept of $Z_2(y_1) = \sum_{n=1}^{N} Z_{2,n}(s = 1, y_1; K^*)$ as a function of μ.

As shown in Fig. 6, the total arbitrage trading $Z_2(y_1) \equiv \sum_{n=1}^{N} Z_{2,n}(s = 1, y_1; K^*)$ has a slope, $\frac{N(1-\mu)(\rho-1)}{N+2}$, which decreases from 1 to 0 as μ varies from 0 to 1. Its horizontal intercept, $\frac{\rho(1+c)}{\rho-1}\kappa\sigma_u$, increases from 0 to $2\sqrt{2}\sigma_u$. At constant market depth, the total arbitrage trading collapses to the 45° line, $\lim_{\mu\to 0} Z_2 = y_1 \mathbf{1}_{|y_1|>\kappa\sigma_u}$, regardless of the number N. This is an "order-flow mimicking" strategy, since the total quantity traded by arbitrageurs exactly mimics the total order flow they observed earlier. Also, this is like a pool of stop-loss orders which get triggered to execute whenever the price change surpasses $\lambda_1\kappa\sigma_u = \frac{4\sqrt{2}}{9}\frac{N+1}{N}\sigma_v$ in either direction. A function of the form, $F(y) = y\mathbf{1}_{|y|>K}$, is often called "hard-thresholding" in machine learning. For $\mu > 0.5$, arbitrageurs always use the "soft-thresholding" strategy.

Let's look at the strategy of informed trader in different liquidity regimes. If market liquidity at $t = 2$ is good ($\mu < \mu_c$), her initial strategy $X_1(s = 1, v; K^*)$ is bended toward K^* to distort arbitrageurs' learning [Fig. 7(a)]. With $\tilde{x}_1 \approx K^*$ for a range of $\tilde{v}$, it will be difficult for arbitrageurs to infer the strength of $\tilde{v}$ from $\tilde{y}_1 = \tilde{x}_1 + \tilde{u}_1$. Their trading decisions are error-prone because they are largely influenced by noise trading $\tilde{u}_1$. The nonlinear pure

[26] As long as the informed trader's strategy monotonically increases with her signal, it will be profitable for arbitrageurs to undercut the threshold as much as possible.

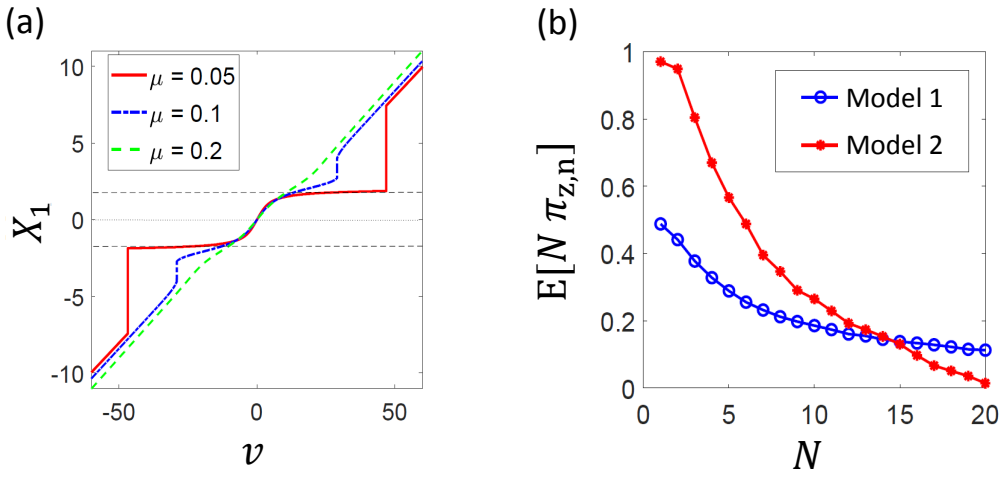


Figure 7. (a) the informed trader's strategy $X_1(s=1,v)$ under different μ. (b) the total payoffs to arbitrageurs in two models under respective linear-triggering strategies.

strategy allows the informed trader to hide her signal temporarily and inhibit the response of arbitrageurs. If future liquidity is poor ($\mu > \mu_c$), the informed trader will trade more at $t=1$ and play the game more honestly. Poor liquidity discourages arbitrage trading and reduces the incentive to distort their learning. Overall, the informed trader induces arbitrageurs to trade more competitively. This disrupts their market power and the cartel effect identified in the model of robust arbitrageurs. Facing the savvy informed trader, arbitrageurs can no longer sustain extra market power nor earn noncompetitive profits at large N [Fig. 7(b)].

3.2 Disruptive Strategies and Price Manipulations

In this trading game with linear-triggering strategies, there is an implicit belief in the arbitrageurs' minds that the informed trader will play a monotone strategy which increases with her signal. Numerically, this conjecture is found to hold in the liquidity regime where $\mu > \mu_\epsilon \approx 0.005$. However, the conjectured equilibrium becomes unstable when market depth is almost constant ($\mu \to 0$). If μ is arbitrarily close to 0, the total order flow from arbitrageurs will closely mimic the order flow y_1. This may invite the informed trader to trick them.

Corollary 3.3. *At $v=0$ and as $\mu \to 0$, the informed trader will first trade a sufficiently large x_1 to trigger arbitrageurs and then trade $x_2 = -y_1$ to offset their momentum trading, i.e., $\lim_{\mu\to 0} X_2(v=0,y_1) = -y_1 = -\lim_{\mu\to 0} Z_2(y_1)$. This Bayesian-rational strategy has a terminal position of $x_1+x_2=-u_1$ which is zero on average, with an expected profit of $\lambda_1\sigma_u^2$.*

Proof. This rational strategy follows from Eq. (33) and Eq. (35) by taking both limits $v \to 0$ and $\mu \to 0$. Detailed proof can be found in the Appendix A.11. □

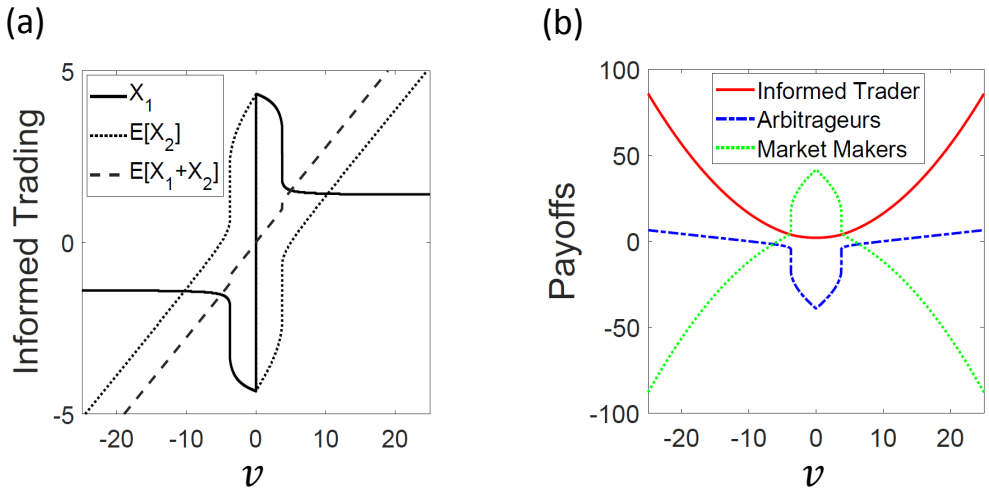


Figure 8. (a) the optimal strategy of informed trader under $\mu = 10^{-4}$, $N = 3$ and $\xi_v = 3$. (b) the total payoffs to different groups of traders.

As shown in Fig. 8(a), when the private signal v is very small, the informed trader places a large order $|x_1| \gg K^* = \kappa\sigma_u$ to trigger arbitrageurs whose trading at $t = 2$ closely mimics the total order flow observed at $t = 1$. This allows the informed trader to liquidate most of her inventory at more favorable prices at $t = 2$. The terminal position $\mathrm{E}[\tilde{x}_1 + \tilde{x}_2|\tilde{v} = v]$ is almost linear with her private signal v, but her strategy in each period is non-monotone with her signal. Fig. 8(b) shows the total payoffs to different groups of traders. Arbitrageurs incur dramatic losses near the origin as they have been fooled by the informed trader who earns a small profit on average. The losses of arbitrageurs mostly benefit market makers.

The non-monotone strategy seems disruptive and resembles controversial strategies in the real world, including *momentum ignition* and *stop-loss hunting*. These schemes are usually regarded as *trade-based price manipulations* by regulators. If such non-monotone strategies are prohibited (by regulators) in the model, the informed trader at the state $v = 0$ will not trade at $t = 1$. Instead, she will watch the market first and trade at $t = 2$ against either the noise-driven price changes or the order flows from arbitrageurs who are falsely triggered.

Kyle and Viswanathan (2008) recommend two economic criteria for regulators to define illegal price manipulations. These are pricing accuracy and market liquidity. Fig. 9 compares the (unconditional) probability distributions of prices when the non-monotone strategy is

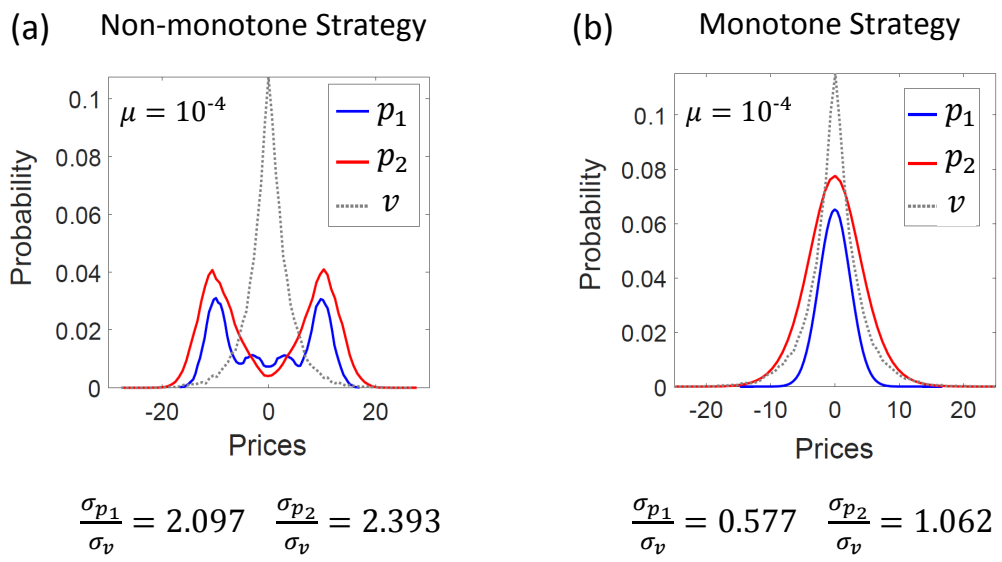


Figure 9. The unconditional probability distributions of the prices $\tilde{p}_1$ and $\tilde{p}_2$ under the non-monotone strategy versus the monotone strategy, given $\mu = 10^{-4}$, $N = 3$ and $\sigma_v = 3\sqrt{2}$.

allowed or banned. With the non-monotone strategy in Fig. 8(a), price distributions are bimodal in both periods [Fig. 9(a)]. Pricing accuracy is poor as prices do not reflect the fundamental value $\tilde{v}$ (with a unimodal distribution). Price volatilities are at least twice as large as the fundamental volatility σ_v. If a common investor arrives and trades this asset, she is likely to buy at a much higher ask price or sell at a much lower bid price. The bimodal price pattern reflects a much wider bid-ask spread for common investors. In contrast, if regulators set rules to ban such disruptive strategies, the price distributions become bell-shaped in both periods with reasonable price volatilities and pricing accuracy [Fig. 9(b)].

Regulators need to sort out the economic conditions for the *trade-base manipulations*. The results in this paper prescribe a list of conditions that could be necessary for the non-monotone disruptive strategy.

(1) Speculators think that market makers set inaccurate prices by using incorrect priors.
(2) Speculators have fat-tail priors about the fundamental value or trading opportunities.
(3) There is strategic interplay between the informed trader and those speculators.
(4) Market depth is not decreasing when the informed trader liquidates her inventory.
(5) Traders face no trading costs, no inventory costs, nor threat from regulators.
(6) There is no other informed trader who could interfere with the disruptive strategy.

The (non-monotone) disruptive strategy may fail if any of these conditions is not satisfied. It seems not easy at all, but the key condition is that the total feedback trading from

speculators has a slope no less than one. This could happen if speculators underestimate the actual number of speculators (N), since each speculator's demand is inversely proportional to the number of competitors (estimated by the speculator). This could also happen in the liquidity regime with $\mu < 0$, where the informed trader could dump her early inventory at a lower cost and speculators may trade more aggressively. In the conjectured equilibrium, the response slope of each speculator is given by $\frac{(1-\mu)(\rho-1)}{N+2}$. If all speculators keep using this strategy in the liquidity regime $\mu < 0$, the slope of their aggregate response will be greater than one: $N\frac{(1-\mu)(\rho-1)}{N+2} > 1$. Over-trading makes speculators susceptible to "disruptive attacks". For the informed trader, the profits of tricking speculators can be outweighed by the losses if she fails to liquidate the undesirable inventory in the second period.

4 Conclusion

This paper studies an equilibrium model of strategic arbitrage in the fat-tail environment. The presence of arbitrageurs is rationalized by applying random fat-tail shocks to the standard Kyle model where market makers adhere to Gaussian beliefs. If arbitrageurs are uncertain about the various of fat-tail shocks, their robust strategy under the max-min choice criteria is operationally equivalent to the LASSO algorithm in machine learning. For robustness, arbitrageurs choose to ignore a wide range of small (uncertain) mispricings and take actions only on large (certain) ones. This strategy is highly effective given its infrequent trading activity. As a result, many anomalies may be detected *ex post* by an external econometrician based on historical data in this economy. The econometrician may conclude that market inefficiency is due to arbitrageurs' behavioral bias as they overlook those anomalies. In fact, arbitrageurs are rational under their robust-control objective. They use Bayes rule to carefully evaluate all possible states over their multiple priors. Arbitrageurs can amass significant market power due to their under-trading beyond the kinks of robust strategy. This cartel effect allows them to earn noncompetitive profits which do not vanish even if their number goes to infinity. Therefore, price efficiency is further impaired.

If the informed trader strategically interacts with those arbitrageurs, she will try to distort their learning and induce them to trade more aggressively. Under certain market conditions, the informed trader may play a disruptive strategy that resembles real-life controversial practices (like momentum ignition). Such trading schemes can distort the informational content of prices and destabilize stock prices at the expense of common investors.

References

Abreu, D., and M. K. Brunnermeier. 2002. Synchronization risk and delayed arbitrage. *Journal of Financial Economics* 66:341–360.

Abreu, D., and M. K. Brunnermeier. 2003. Bubbles and crashes. *Econometrica* 71:173–204.

Allen, F., and D. Gale. 1992. Stock-price manipulation. *Review of Financial Studies* 5:503–529.

Back, K. 1992. Insider trading in continuous time. *Review of Financial Studies* 5:387–409.

Back, K., C. H. Cao, and G. A. Willard. 2000. Imperfect competition among informed traders. *Journal of Finance* 55:2117–2155.

Barberis, N., R. Greenwood, L. Jin, and A. Shleifer. 2015. X-CAPM: An extrapolative capital asset pricing model. *Journal of Financial Economics* 115:1–24.

Barberis, N., R. Greenwood, L. Jin, and A. Shleifer. 2018. Extrapolation and bubbles. *Journal of Financial Economics* .

Barberis, N., and R. Thaler. 2003. A survey of behavioral finance. *Handbook of the Economics of Finance* 1:1053–1128.

Bondarenko, O. 2003. Statistical arbitrage and securities prices. *Review of Financial Studies* 16:875–919.

Brunnermeier, M. K. 2005. Information leakage and market efficiency. *Review of Financial Studies* 18:417–457.

Brunnermeier, M. K., and L. H. Pedersen. 2005. Predatory trading. *Journal of Finance* 60:1825–1863.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay. 1997. *The Econometrics of Financial Markets*, vol. 2. princeton University press Princeton, NJ.

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Chakraborty, A., and B. Yılmaz. 2004a. Informed manipulation. *Journal of Economic Theory* 114:132.

Chakraborty, A., and B. Yılmaz. 2004b. Manipulation in market order models. *Journal of Financial Markets* 7:187–206.

Chakraborty, A., and B. Yılmaz. 2008. Microstructure bluffing with nested information. *American Economic Review* 98:280–84.

Chan, L. K., N. Jegadeesh, and J. Lakonishok. 1996. Momentum strategies. *Journal of Finance* 51:1681–1713.

Chinco, A. M., A. D. Clark-Joseph, and M. Ye. 2017. Sparse signals in the cross-section of returns. Tech. rep., National Bureau of Economic Research.

Collin-Dufresne, P., and V. Fos. 2016. Insider trading, stochastic liquidity, and equilibrium prices. *Econometrica* 84:1441–1475.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam. 1998. Investor psychology and security market under-and overreactions. *Journal of Finance* 53:1839–1885.

DeLong, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann. 1990. Positive feedback investment strategies and destabilizing rational speculation. *Journal of Finance* 45:379–395.

Dow, J., and S. R. d. C. Werlang. 1992. Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica* pp. 197–204.

Foster, F. D., and S. Viswanathan. 1994. Strategic trading with asymmetrically informed traders and long-lived information. *Journal of Financial and Quantitative Analysis* 29:499–518.

Foster, F. D., and S. Viswanathan. 1996. Strategic trading when agents forecast the forecasts of others. *Journal of Finance* 51:1437–1478.

Fox, M. B., L. R. Glosten, and G. V. Rauterberg. 2018. Stock Market Manipulation and Its Regulation. *Yale J. on Reg.* 35:67.

Freyberger, J., A. Neuhierl, and M. Weber. 2017. Dissecting characteristics nonparametrically. Tech. rep., National Bureau of Economic Research.

Gabaix, X. 2014. A sparsity-based model of bounded rationality. *Quarterly Journal of Economics* 129:1661–1710.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley. 2006. Institutional investors and stock market volatility. *Quarterly Journal of Economics* 121:461–504.

Gabaix, X., A. Krishnamurthy, and O. Vigneron. 2007. Limits of arbitrage: theory and evidence from the mortgage-backed securities market. *Journal of Finance* 62:557–595.

Gatev, E., W. N. Goetzmann, and K. G. Rouwenhorst. 2006. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies* 19:797–827.

Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with a unique set of priors. *Journal of Mathematical Economics* 18.

Goldstein, I., and A. Guembel. 2008. Manipulation and the allocational role of prices. *Review of Economic Studies* 75:133–164.

Grinold, R. C., and R. N. Kahn. 2000. Active Portfolio Management .

Gromb, D., and D. Vayanos. 2010. Limits of arbitrage. *Annual Review of Financial Economics* 2:251–275.

Han, J., and A. S. Kyle. 2017. Speculative Equilibrium with Differences in Higher-Order Beliefs. *Management Science* .

Hansen, L., and T. J. Sargent. 2001. Robust control and model uncertainty. *American Economic Review* 91:60–66.

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

Hogan, S., R. Jarrow, M. Teo, and M. Warachka. 2004. Testing market efficiency using statistical arbitrage with applications to momentum and value strategies. *Journal of Financial Economics* 73:525–565.

Holden, C. W., and A. Subrahmanyam. 1992. Long-lived private information and imperfect competition. *Journal of Finance* 47:247–270.

Hong, H., and J. C. Stein. 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance* 54:2143–2184.

Hong, H., and J. C. Stein. 2003. Differences of opinion, short-sales constraints, and market crashes. *Review of Financial Studies* 16:487–525.

Huang, J.-Z., and Z. Shi. 2011. Determinants of bond risk premia. Tech. rep., Citeseer.

Huberman, G., and W. Stanzl. 2004. Price manipulation and quasi-arbitrage. *Econometrica* 72:1247–1275.

Huddart, S., J. S. Hughes, and C. B. Levine. 2001. Public disclosure and dissimulation of insider trades. *Econometrica* 69:665–681.

Jarrow, R. A. 1992. Market manipulation, bubbles, corners, and short squeezes. *Journal of Financial and Quantitative Analysis* 27:311–336.

Jarrow, R. A. 2015. Asset price bubbles. *Annual Review of Financial Economics* 7:201–218.

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* 48:65–91.

Jiang, G., P. G. Mahoney, and J. Mei. 2005. Market manipulation: A comprehensive study of stock pools. *Journal of Financial Economics* 77:147–170.

Kahneman, D., J. L. Knetsch, and R. H. Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives* 5:193–206.

Khwaja, A. I., and A. Mian. 2005. Unchecked intermediaries: Price manipulation in an emerging stock market. *Journal of Financial Economics* 78:203–241.

Kirilenko, A., A. S. Kyle, M. Samadi, and T. Tuzun. 2017. The Flash Crash: High-frequency trading in an electronic market. *Journal of Finance* 72:967–998.

Kondor, P. 2009. Risk in dynamic arbitrage: the price effects of convergence trading. *Journal of Finance* 64:631–655.

Korajczyk, R. A., and D. Murphy. 2018. High frequency market making to large institutional trades. *Review of Financial Studies, Forthcoming* .

Kozak, S., S. Nagel, and S. Santosh. 2017. Shrinking the cross section. Tech. rep., National Bureau of Economic Research.

Kumar, P., and D. J. Seppi. 1992. Futures manipulation with "cash settlement". *Journal of Finance* 47:1485–1502.

Kyle, A. S. 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society* pp. 1315–1335.

Kyle, A. S., and A. A. Obizhaeva. 2016. Market microstructure invariance: Empirical hypotheses. *Econometrica* 84:1345–1404.

Kyle, A. S., and S. Viswanathan. 2008. How to define illegal price manipulation. *American Economic Review* 98:274–79.

Kyle, A. S., and W. Xiong. 2001. Contagion as a wealth effect. *Journal of Finance* 56:1401–1440.

Lee, C. M., and B. Swaminathan. 2000. Price momentum and trading volume. *Journal of Finance* 55:2017–2069.

Lehmann, B. N. 1990. Fads, martingales, and market efficiency. *Quarterly Journal of Economics* 105:1–28.

Lewis, M. 2014. *Flash Boys: a Wall Street Revolt*. WW Norton & Company.

Loeb, T. F. 1983. Trading cost: the critical link between investment information and results. *Financial Analysts Journal* pp. 39–44.

Samuelson, W., and R. Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1:7–59.

Scheinkman, J. A., and W. Xiong. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111:1183–1220.

Shleifer, A., and R. W. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52:35–55.

Silva, A. C., R. E. Prange, and V. M. Yakovenko. 2004. Exponential distribution of financial returns at mesoscopic time lags: a new stylized fact. *Physica A* 344:227–235.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

van Bommel, J. 2003. Rumors. *Journal of Finance* 58:1499–1520.

van Kervel, V., and A. J. Menkveld. 2017. High-frequency trading around large institutional orders. *Journal of Finance, Forthcoming* .

Vayanos, D. 1999. Strategic trading and welfare in a dynamic market. *Review of Economic Studies* 66:219–254.

Vayanos, D. 2001. Strategic trading in a dynamic noisy market. *Journal of Finance* 56:131–171.

Xiong, W. 2001. Convergence trading with wealth effects: an amplification mechanism in financial markets. *Journal of Financial Economics* 62:247–292.

Yang, L., and H. Zhu. 2017. Back-running: Seeking and hiding fundamental information in order flows. *Working paper, SSRN* .

A Appendix

A.1 Proof of Proposition 2.1

Under their common belief, the informed trader and market makers first conjecture that arbitrageurs do not trade if the market is efficient. As in the two-period Kyle (1985) model, they can seek a linear equilibrium $(\mathbf{X}, \mathbf{P})$, where $\mathbf{P} = \langle P_1, P_2 \rangle$ is the linear pricing strategy of market makers. Let $P_1(y_1) = \lambda_1 y_1$ and $P_2(y_1, y_2) = \lambda_1 y_1 + \lambda_2 y_2$. The information set of informed trader before trading at $t = 2$ is $\mathcal{I}_{2,x} = \{v, y_1\}$. After $t = 1$, she conjectures the price at $t = 2$ as

$$\tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2) = \lambda_1 y_1 + \lambda_2[X_2(v, y_1) + \tilde{u}_2], \quad \text{under} \quad \{\mathcal{I}_{2,x}, \mathcal{B}\}. \tag{38}$$

Her optimal strategy at $t = 2$ under the information set $\mathcal{I}_{2,x}$ and belief system $\mathcal{B}$ is

$$X_2(v, y_1) = \arg\max_{x_2} \mathrm{E}^{\mathcal{B}}\left[(v - \tilde{p}_2)x_2 | \mathcal{I}_{2,x}\right] = \frac{v - \lambda_1 y_1}{2\lambda_2}. \tag{39}$$

The informed trader conjectures the price at $t = 1$ to be $\tilde{p}_1 = \lambda_1[X_1(v) + \tilde{u}_1]$ under $\{\mathcal{I}_{1,x}, \mathcal{B}\}$. With this notion and $X_2(v, y_1)$, her subjective expected profit is a quadratic function of x_1:

$$\Pi_x(v, x_1) \quad = \quad x_1(v - \lambda_1 x_1) + \mathrm{E}^{\mathcal{B}}\left[\frac{(v - \lambda_1(x_1 + \tilde{u}_1))^2}{4\lambda_2}\middle| \mathcal{I}_{1,x} = \{v\}\right]. \tag{40}$$

The first order condition is $0 = v - 2\lambda_1 x_1 - \frac{v - \lambda_1 x_1}{2\delta}$, where $\delta \equiv \frac{\lambda_2}{\lambda_1}$. The optimal strategy is

$$X_1(v) = \frac{2\delta - 1}{4\delta - 1} \cdot \frac{v}{\lambda_1} = \frac{v}{\rho\lambda_1}, \tag{41}$$

where $\rho = \frac{4\delta - 1}{2\delta - 1}$. The above results constitute Eq. (11) in Proposition 2.1. Market makers hold the same Gaussian belief. As an extension of Proposition 1 in Huddart et al. (2001), it takes some similar calculations to derive that $\lambda_1 = \frac{\sqrt{2\delta(2\delta-1)}}{4\delta-1}\frac{\sigma_v}{\sigma_u}$, where the ratio δ is given by the largest root to the cubic equation:

$$8\gamma\delta^3 - 4\gamma\delta^2 - 4\delta + 1 = 0. \tag{42}$$

Here, $\gamma > 0$ is the ratio of noise trading volatilities over time. Under this pair of linear strategies $\mathbf{X}$ and $\mathbf{P}$, prices are conditional expectations of public information under market makers' belief $\mathcal{B}$. So the informed trader and market makers believe that if they play $\mathbf{X}$ and $\mathbf{P}$ no arbitrageurs would trade. This confirms the initial conjecture and completes the proof.

A.2 Proof of Proposition 2.2

Arbitrageurs know that they are not anticipated to trade by the informed trader and market makers. In the Gaussian case ($s = 0$), they have no informational advantage over market makers. The market is efficient under the subgame perfect equilibrium $(\mathbf{X}, \mathbf{P})$ in Proposition 2.1. Indeed, arbitrageurs will not trade when $s = 0$. In the Laplacian case ($s = 1$), they can exploit the pricing bias because market makers use the wrong prior. To solve the equilibrium, I conjecture first and verify later that arbitrageurs do not trade in the first period, i.e., $Z_{1,n} = 0$ for $n = 1, ..., N$. Under this conjecture, I solve their optimal strategy at $t = 2$. Arbitrageurs anticipate the informed trader's linear strategy and the market-clearing price,

$$P_2(\tilde{y}_1, \tilde{y}_2) = \lambda_1 \tilde{y}_1 + \lambda_2 \left(X_2(\tilde{v}, \tilde{y}_1) + \sum_{n=1}^{N} Z_{2,n}(\tilde{s}, \tilde{y}_1) + \tilde{u}_2 \right). \tag{43}$$

They estimate $\tilde{v}$ based on the observed y_1 and their Laplace prior $\mathcal{L}(0, \tilde{\xi})$. In the absence of model risk (i.e., $\tilde{\xi} = \xi$), the n-th arbitrageur solves her optimal strategy,

$$Z^o_{2,n}(s = 1, y_1; \xi) = \arg\max_{z_{2,n}} \mathrm{E}^{\mathcal{A}} \left[(\tilde{v} - \tilde{p}_2) z_{2,n} | \mathcal{I}_{2,z} \right], \tag{44}$$

under $\mathcal{I}_{2,z} \equiv \{s, y_1\}$ and the belief $\mathcal{A} = \{s, \xi\}$. Let $Z^o_{2,-n} = \sum_{m \neq n} Z^o_{2,m}$ be the their aggregate trading except the n-th arbitrageur's. The first order condition for $z_{2,n}$ is

$$\mathrm{E}^{\mathcal{A}}[\tilde{v}|\mathcal{I}_{2,z}] - \lambda_1 y_1 = \lambda_2 \left(\mathrm{E}^{\mathcal{A}}[X_2|\mathcal{I}_{2,z}] + 2 z_{2,n} + \mathrm{E}^{\mathcal{A}}[Z^o_{2,-n}|\mathcal{I}_{2,z}] \right). \tag{45}$$

Since $\mathrm{E}^{\mathcal{A}}[X_2|\mathcal{I}_{2,z}] = \frac{\hat{v} - \lambda_1 y_1}{2\delta\lambda_1}$ where $\hat{v} = \hat{v}(y_1; \xi) = \mathrm{E}^{\mathcal{A}}[\tilde{v}|\mathcal{I}_{2,z}]$, the solution is

$$Z^o_{2,n}(s = 1, y_1; \xi) = \frac{\hat{v} - \lambda_1 y_1}{2\delta\lambda_1} - \frac{\mathrm{E}^{\mathcal{A}}[X_2|\mathcal{I}_{2,z}] + \mathrm{E}^{\mathcal{A}}[Z^o_{2,-n}|\mathcal{I}_{2,z}]}{2} = \frac{\hat{v} - \lambda_1 y_1}{4\delta\lambda_1} - \frac{\mathrm{E}^{\mathcal{A}}[Z^o_{2,-n}|\mathcal{I}_{2,z}]}{2}. \tag{46}$$

The n-th arbitrageur conjectures that every other arbitrageur solves the same problem and trades $Z^o_{2,m} = \eta \cdot (\hat{v} - p_1)$ for any $m \neq n$, with a coefficient η to be solved. Eq. (46) becomes

$$Z^o_{2,n}(s = 1, y_1; \xi) = \frac{\hat{v} - \lambda_1 y_1}{4\delta\lambda_1} - (N - 1)\frac{\eta(\hat{v} - \lambda_1 y_1)}{2} = [\delta^{-1} - 2\lambda_1\eta(N - 1)]\frac{(\hat{v} - \lambda_1 y_1)}{4\lambda_1}. \tag{47}$$

Since each arbitrageur makes the same conjecture in a symmetric equilibrium, they find that $\eta = \frac{\delta^{-1} - 2\lambda_1\eta(N-1)}{4\lambda_1}$, which has a unique solution

$$\eta = \frac{1}{2\delta\lambda_1(N + 1)} > 0. \tag{48}$$

Without model risk, the optimal strategy of arbitrageurs under the Laplace prior $\mathcal{L}(0,\xi)$ is

$$Z^o_{2,n}(s,y_1;\xi)=\frac{\hat{v}(y_1;\xi)-\lambda_1 y_1}{2(N+1)\delta\lambda_1}=\frac{1-\mu}{N+1}\cdot\frac{\hat{\theta}(y_1;\xi)}{2\lambda_1},\qquad n=1,...,N. \tag{49}$$

Since $X_1(v)=\frac{v}{\rho\lambda_1}$, arbitrageurs have a Laplace prior for $\tilde{x}_1$, denoted $f_L(x_1)=\frac{\rho\lambda_1}{2\xi}\exp\left(-\frac{\rho\lambda_1|x_1|}{\xi}\right)$. By Bayes' rule, the posterior probability of the informed trading x_1 conditional on y_1 is

$$f(x_1|y_1)=\frac{f(y_1|x_1)f_L(x_1)}{f(y_1)}=\frac{\rho\lambda_1}{2\xi f(y_1)\sqrt{2\pi\sigma_u^2}}\exp\left[-\frac{(y_1-x_1)^2}{2\sigma_u^2}-\frac{\rho\lambda_1|x_1|}{\xi}\right]. \tag{50}$$

The probability density function of $\tilde{y}_1=\frac{\tilde{v}}{\rho\lambda_1}+\tilde{u}_1$ is found to be:

$$f(y_1)=\frac{\rho\lambda_1}{4\xi}\exp\left(\frac{\rho^2\lambda_1^2\sigma_u^2}{2\xi^2}\right)\left[e^{-\frac{\rho\lambda_1 y_1}{\xi}}\text{erfc}\left(\frac{\rho\lambda_1\sigma_u^2/\xi-y_1}{\sqrt{2}\sigma_u}\right)+e^{\frac{\rho\lambda_1 y_1}{\xi}}\text{erfc}\left(\frac{\rho\lambda_1\sigma_u^2/\xi+y_1}{\sqrt{2}\sigma_u}\right)\right]. \tag{51}$$

I define a dimensionless parameter $\kappa\equiv\frac{\rho\lambda_1\sigma_u}{\xi}$ and rewrite $f(y_1)$ in a dimensionless form

$$f(y_1=y'\sigma_u)=\frac{\kappa e^{\frac{\kappa^2}{2}}}{4\sigma_u}\left[e^{-\kappa y'}\text{erfc}\left(\frac{\kappa-y'}{\sqrt{2}}\right)+e^{\kappa y'}\text{erfc}\left(\frac{\kappa+y'}{\sqrt{2}}\right)\right], \tag{52}$$

which is symmetric and decays exponentially at large $|y'|$. Bayes' rule implies that

$$\text{E}^{\mathcal{A}}[\tilde{x}_1=x'\sigma_u|y_1=y'\sigma_u,\xi]=\sigma_u\int_{-\infty}^{\infty}xf(x|y)dx=\sigma_u\int_{-\infty}^{\infty}\frac{xf(y|x)f(x)}{f(y)}dx, \tag{53}$$

Given that $X_1(v)=\frac{v}{\rho\lambda_1}$, it is easy to derive the posterior expectation of $\tilde{v}$ explicitly:

$$\hat{v}=\text{E}^{\mathcal{A}}[\tilde{v}|y_1=y'\sigma_u,\xi]=\frac{\kappa\xi(y'-\kappa)\text{erfc}\left(\frac{\kappa-y'}{\sqrt{2}}\right)}{\text{erfc}\left(\frac{\kappa-y'}{\sqrt{2}}\right)+e^{2\kappa y'}\text{erfc}\left(\frac{\kappa+y'}{\sqrt{2}}\right)}+\frac{\kappa\xi(y'+\kappa)\text{erfc}\left(\frac{\kappa+y'}{\sqrt{2}}\right)}{\text{erfc}\left(\frac{\kappa+y'}{\sqrt{2}}\right)+e^{-2\kappa y'}\text{erfc}\left(\frac{\kappa-y'}{\sqrt{2}}\right)}. \tag{54}$$

The rescaled $\hat{v}/\xi$ is an increasing function of y' with a single shape parameter κ. Asymptotic linearity holds at $|y'|\gg\kappa$ that $\hat{v}\to\rho\lambda_1[y_1-\text{sign}(y_1)\kappa\sigma_u]$. All the second-order conditions are easy to check. The REE corresponds to the equilibrium where all arbitrageurs have the correct prior. Under REE, we have $\xi=\xi_v=\frac{\sigma_v}{\sqrt{2}}$ such that the shape parameter becomes

$$\kappa(\xi=\xi_v)=\frac{\rho\lambda_1\sigma_u}{\xi_v}=\frac{4\delta-1}{2\delta-1}\frac{\sqrt{4\delta(2\delta-1)}}{4\delta-1}=\frac{2}{\sqrt{1+\mu}}, \tag{55}$$

where $\mu\equiv1-\frac{1}{\delta}$ quantifies the percentage change of market depth in the second period.

To verify that no arbitrageurs would trade in the first period, I examine the condition Eq. (6). Suppose the n-th arbitrageur deviates from the conjectured strategy by trading a nonzero quantity $Z^{o,d}_{1,n} = z_1 \neq 0$ in the first period. Then the actual total order flow at $t = 1$ is $\tilde{y}'_1 = \tilde{x}_1 + \tilde{z}_1 + \tilde{u}_1$, instead of $\tilde{y}_1 = \tilde{x}_1 + \tilde{u}_1$ in the conjectured equilibrium. Taking $\mathbf{X}$, $\mathbf{P}$, and $Z^o_{2,m}(s, y'_1; \xi) = s\frac{\hat{v}(y'_1)-\lambda_1 y'_1}{2(N+1)\delta\lambda_1}$ for any $m \neq n$ as given, the n-th arbitrageur's optimal strategy at $t = 2$ conditional on the information set $\mathcal{I}'_{2,z} = \{s, y_1, z_1\}$ is

$$\begin{aligned}
Z^{o,d}_{2,n}(s, y'_1; \xi) &= s\frac{\hat{v}(y_1; \xi) - \lambda_1 y'_1}{2\delta\lambda_1} - s\frac{\mathrm{E}^{\mathcal{A}}[X_2(\tilde{v}, y'_1)|\mathcal{I}'_{2,z}] + \mathrm{E}^{\mathcal{A}}[Z^o_{2,-n}(s, y'_1; \xi)|\mathcal{I}'_{2,z}]}{2} \\
&= s\frac{\hat{v}(y_1; \xi) - \lambda_1 y'_1}{4\delta\lambda_1} - s\frac{Z^o_{2,-n}(s, y'_1; \xi)}{2} \\
&= s\frac{\hat{v}(y_1; \xi) - \lambda_1 y'_1}{4\delta\lambda_1} - s\frac{(N-1)[\hat{v}(y'_1; \xi) - \lambda_1 y'_1]}{4(N+1)\delta\lambda_1} \\
&= s\frac{\hat{v}(y_1; \xi) - \lambda_1 y'_1}{2(N+1)\lambda_2} + s\frac{N-1}{4(N+1)\lambda_2}[\hat{v}(y_1; \xi) - \hat{v}(y'_1; \xi)]. \qquad (56)
\end{aligned}$$

If the n-th trader does not deviate from the no-trade strategy in the first period, her optimal strategy should be $Z^o_{2,n}(s, y_1; \xi) = s\frac{\hat{v}(y_1;\xi)-\lambda_1 y_1}{2(N+1)\delta\lambda_1}$, where $y_1 = x_1 + u_1$. For convenience, we just need to consider the case $s = 1$. Let's add the notation that $\Delta P_1 \equiv \lambda_1(\tilde{y}'_1 - \tilde{y}_1) = \lambda_1 z_1$, $\Delta\hat{v} \equiv \hat{v}(\tilde{y}'_1; \xi) - \hat{v}(\tilde{y}_1; \xi)$, $\Delta Z \equiv Z^{o,d}_{2,n}(s, y'_1; \xi) - Z^o_{2,n}(s, y_1; \xi) = -\frac{\lambda_1 z_1}{2(N+1)\lambda_2} - \frac{N-1}{4(N+1)\lambda_2}\Delta\hat{v}$ and

$$\begin{aligned}
\Delta P_2 &\equiv P_2(\mathbf{X}, \mathbf{Z}') - P_2(\mathbf{X}, \mathbf{Z}) \\
&= \lambda_1 z_1 + \lambda_2[\Delta Z + X_2(\tilde{v}, \tilde{y}'_1) - X_2(\tilde{v}, \tilde{y}_1) + Z^o_{2,-n}(\tilde{y}'_1) - Z^o_{2,-n}(\tilde{y}_1)] \\
&= \lambda_1 z - \frac{\lambda_1 z}{2(N+1)} - \frac{N-1}{4(N+1)}\Delta\hat{v} - \frac{\lambda_1 z}{2} + \frac{(N-1)(\Delta\hat{v} - \lambda_1 z)}{2(N+1)} \\
&= \frac{\lambda_1 z_1}{2(N+1)} + \frac{N-1}{4(N+1)}\Delta\hat{v} = -\lambda_2\Delta Z, \qquad (57)
\end{aligned}$$

where $\mathbf{Z}'$ differs from $\mathbf{Z} \equiv [\langle 0, Z^o_{2,1}\rangle, ...\langle 0, Z^o_{2,N}\rangle]$ only in the n-th element $(\mathbf{Z}')_n = \langle z_1, Z^{o,d}_{2,n}\rangle$. Since $\tilde{y}_1 = X_1(\tilde{v}) + \tilde{u}_1$, we have $\mathrm{E}^{\mathcal{A}}[\tilde{y}_1 \cdot z_1] = 0$ and $\mathrm{E}^{\mathcal{A}}[\hat{v}(\tilde{y}_1) \cdot z_1] = 0$. The payoff difference is

$$\begin{aligned}
&\Delta\Pi^d_{z,n} = \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \tilde{p}_2(\mathbf{X}, \mathbf{Z}'))Z^{o,d}_{2,n} + (\tilde{v} - \tilde{p}_1(\mathbf{X}, \mathbf{Z}'))z_1 - (\tilde{v} - \tilde{p}_2(\mathbf{X}, \mathbf{Z}))Z^o_{2,n}|\tilde{s} = 1, \tilde{\xi} = \xi] \\
= \;& \mathrm{E}^{\mathcal{A}}[\tilde{v}z_1 - z_1\tilde{p}_1(\mathbf{X}, \mathbf{Z}') + \tilde{v}\Delta Z - \Delta P_2 \cdot Z^{o,d}_{2,n} - \tilde{p}_2(\mathbf{X}, \mathbf{Z}) \cdot \Delta Z|\tilde{s} = 1, \tilde{\xi} = \xi] \\
= \;& -\lambda_1 z_1^2 + \mathrm{E}^{\mathcal{A}}[\mathrm{E}^{\mathcal{A}}[(\tilde{v} - \tilde{p}_2(\mathbf{X}, \mathbf{Z}) + \lambda_2 Z^{o,d}_{2,n})\Delta Z|\tilde{y}_1]] \\
= \;& -\lambda_1 z_1^2 + \mathrm{E}^{\mathcal{A}}\left[\left(\frac{\hat{v}(\tilde{y}_1; \xi) - \lambda_1\tilde{y}_1}{N+1} + \lambda_2\Delta Z\right) \cdot \Delta Z\right] \\
= \;& -\lambda_1 z_1^2 + \frac{\mathrm{E}^{\mathcal{A}}[(\lambda_1 z_1 + \frac{1}{2}(N-1)\Delta\hat{v})^2]}{4(N+1)^2\lambda_2} - \frac{N-1}{4(N+1)^2\lambda_2}\mathrm{E}^{\mathcal{A}}\left[(\hat{v}(\tilde{y}_1; \xi) - \lambda_1\tilde{y}_1) \cdot \Delta\hat{v}\right]. \qquad (58)
\end{aligned}$$

One can rewrite Eq. (58) in a symmetric form with respect to z_1:

$$\begin{aligned}\Delta\Pi^d_{z,n} &= -\lambda_1 z_1^2 + \frac{\mathrm{E}^{\mathcal{A}}[(\lambda_1 z_1)^2 + \frac{1}{4}(N-1)^2(\Delta\hat{v})^2 - (N-1)[\hat{v}(\tilde{y}_1;\xi) - \lambda_1(\tilde{y}_1+z_1)]\Delta\hat{v}]}{4(N+1)^2\lambda_2} \\ &= -\lambda_1 z_1^2 + \frac{(\lambda_1 z_1)^2 + \mathrm{E}^{\mathcal{A}}[\frac{1}{4}(N-1)^2(\Delta\hat{v})^2 - (N-1)(\hat{\theta}(\tilde{y}_1+z_1;\xi) - \Delta\hat{v})\Delta\hat{v}]}{4(N+1)^2\lambda_2}. \end{aligned} \tag{59}$$

This is an even function of z_1 because one can use the symmetry of $\tilde{y}_1$ and $\hat{v}(\cdot)$ to prove

$$\begin{aligned} & \mathrm{E}^{\mathcal{A}}[\hat{\theta}(\tilde{y}_1 - z_1;\xi)\cdot\Delta\hat{v}(\tilde{y}_1,-z_1;\xi)] = \mathrm{E}^{\mathcal{A}}[\hat{\theta}(\tilde{y}_1 - z_1;\xi)(\hat{v}(\tilde{y}_1 - z_1;\xi) - \hat{v}(\tilde{y}_1;\xi))] \\ = \;& \mathrm{E}^{\mathcal{A}}[\hat{\theta}(-\tilde{y}_1 - z_1;\xi)(\hat{v}(-\tilde{y}_1 - z_1;\xi) - \hat{v}(-\tilde{y}_1;\xi))] \\ = \;& \mathrm{E}^{\mathcal{A}}[-\hat{\theta}(\tilde{y}_1 + z_1;\xi)(-\hat{v}(\tilde{y}_1 + z_1;\xi) + \hat{v}(\tilde{y}_1;\xi))] = \mathrm{E}^{\mathcal{A}}[\hat{\theta}(\tilde{y}_1 + z_1;\xi)\cdot\Delta\hat{v}(\tilde{y}_1,z_1;\xi)]. \end{aligned}$$

The first term of Eq. (59) is the average cost to play z_1 at $t=1$, whereas the second term represents the average profit from exploiting the biased response of other traders at $t=2$. The profit of this strategic exploitation has an upper limit which is achieved when all the arbitrageurs have the extreme fat-tail prior $\xi\to\infty$. In this limit, their response to the past order flow is the strongest and exactly linear with y_1: $\lim_{\xi\to\infty} Z^o_{2,n} = \frac{y_1}{(N+1)(2\delta-1)}$. Since $\Delta\Pi^d_{z,n}(-z_1) = \Delta\Pi^d_{z,n}(z_1)$, we only need to consider the positive deviation. For any $z_1 > 0$,

$$\Delta\hat{v}(\tilde{y}_1,z_1;\xi) \equiv \hat{v}(\tilde{y}_1+z_1;\xi) - \hat{v}(\tilde{y}_1;\xi) \le \rho\lambda_1(\tilde{y}'_1 - \tilde{y}_1) = \rho\lambda_1 z_1, \tag{60}$$

where the equality holds at $\xi\to\infty$. Given that $\lim_{\xi\to\infty}\hat{\theta}(\tilde{y}'_1;\xi) = \lambda_1(\rho-1)(\tilde{y}_1+z_1)$, I find

$$\begin{aligned}\Delta\Pi^d_{z,n} < \lim_{\xi\to\infty}\Delta\Pi^d_{z,n} &= -\lambda_1 z_1^2 + \lambda_1^2\frac{z_1^2 + \frac{1}{4}(N-1)^2\rho^2 z_1^2 - (N-1)\rho\mathrm{E}^{\mathcal{A}}[((\rho-1)(\tilde{y}_1+z_1) - \rho z_1)z_1]}{4(N+1)^2\lambda_2} \\ &= -\lambda_1 z_1^2 + (1-\mu)\lambda_1 z_1^2\frac{[(N-1)\rho+2]^2}{16(N+1)^2} \end{aligned} \tag{61}$$

The last expression of Eq. (61) is negative for any $\mu > \mu^*(N)$ where $\mu^*(N)$ is the largest root to the equation: $1 + \left(\frac{N-1}{N+1}\right)\frac{2}{1+\mu} = \frac{4}{\sqrt{1-\mu}}$. The maximum of $\mu^*(N)$ is found to be $\mu_\epsilon \equiv \lim_{N\to\infty}\mu^*(N) \approx -0.23191$, which is the largest root to the cubic equation:

$$\mu^3 + 21\mu^2 + 35\mu + 7 = 0. \tag{62}$$

In the liquidity regime of $\mu > \mu_\epsilon \approx -0.23191$, it is indeed unprofitable for any individual arbitrageur to trade in the first period, i.e., $\Delta\Pi^d_{z,n}(z_1) < 0$ for any $z_1 \neq 0$. This confirms the no-trade conjecture at $t=1$ and completes the proof of Proposition 2.2.

A.3 Proof of Proposition 2.3

All admissible strategies must lie in the area enclosed by $Z^o_{2,n}(y_1, \xi \to 0)$, $Z^o_{2,n}(y_1, \xi \to \infty)$, and the REE asymptotes $Z^\infty(y_1, K^*)$. Any strategy that runs outside this region will violate either the asymptotic requirement or the condition of convexity/concavity preservation. By symmetry, we just discuss the positive domain where the REE strategy is always convex. To satisfy the convexity-preservation rule, the first derivative of an admissible strategy, $\frac{\partial Z'_{2,n}}{\partial y_1}$, can never decrease in the domain of $y_1 > 0$. With a non-decreasing first derivative, the admissible strategy can never go beyond the asymptote $Z^\infty(y_1, K^*)$ and curve back to it.

For $y_1 \in [0, K^*]$, any selling decision located in the bottom triangle "a" would lose money in the worst-case scenario (i.e., if the highest prior ξ_H is true, under which one should buy). Similarly, any buying decision located in the up triangle "b" would lose money in the worst-case scenario (i.e., if the lowest prior ξ_L is true, under which one should buy). This argument indicates a no-trade strategy over $y_1 \in [0, K^*]$. For any $y_1 > K^*$, I will prove that any buying decision $Z'_{2,n}(y_1)$ located inside the area "c" may either lose more money or earn less money than the buying decision $Z^\infty(y_1, K^*)$ determined by the REE asymptotes. Let $Z_\Delta \equiv Z'_{2,n}(y_1) - Z^\infty(y_1, K^*)$. The difference of their payoffs under the lowest prior ξ_L is

$$\begin{aligned}
\mathrm{E}^{\mathcal{A}}[\Delta\tilde{\pi}_{z,n}|y_1, \tilde{\xi} = \xi_L] &= \mathrm{E}^{\mathcal{A}}\left[\left(\tilde{v} - \lambda_1 y_1 - \lambda_2(X_2 + Z'_{2,n} + Z_{2,-n} + \tilde{u}_2)\right) Z'_{2,n} \middle| y_1, \tilde{\xi} = \xi_L\right] \\
&\quad - \mathrm{E}^{\mathcal{A}}\left[(\tilde{v} - \lambda_1 y_1 - \lambda_2(X_2 + Z^\infty + Z_{2,-n} + \tilde{u}_2))\, Z^\infty \middle| y_1, \tilde{\xi} = \xi_L\right] \\
&= \mathrm{E}^{\mathcal{A}}\left[Z_\Delta\left[\frac{\tilde{\theta}}{2} - \lambda_2(Z^\infty + Z_{2,-n} + \tilde{u}_2)\right] \middle| y_1, \tilde{\xi} = \xi_L\right] - \lambda_2 Z'_{2,n} Z_\Delta \qquad (63)
\end{aligned}$$

The worst-case scenario is that ξ_L is true and every other arbitrageur trades $Z^\infty(y_1, K^*)$. Let $\hat{\theta}_L(y_1; \xi_L) \equiv \mathrm{E}^{\mathcal{A}}[\tilde{\theta}|y_1, \xi_L]$ and $Z^L \equiv \frac{\hat{\theta}_L}{2(N+1)\lambda_2}$. Obviously, $Z^L < Z^\infty < Z'_{2,n}$ and $Z_\Delta > 0$. It is not a profitable deviation for anyone to trade more than $Z^\infty(y_1, K^*)$, since

$$\begin{aligned}
\mathrm{E}^{\mathcal{A}}[\Delta\tilde{\pi}_{z,n}|y_1, \tilde{\xi} = \xi_L] &= \lambda_2 Z_\Delta[(N+1)Z^L - Z^\infty - (N-1)Z^\infty] - \lambda_2 Z'_{2,n} Z_\Delta \\
&= \lambda_2 Z_\Delta((N+1)Z^L - NZ^\infty - Z'_{2,n}) < 0. \qquad (64)
\end{aligned}$$

So the robust strategy is to follow the REE asymptote, $Z^\infty(y_1, K^*)$, for any $y_1 > K^*$.

By symmetry, the robust strategy is exactly Eq. (17). It remains to verify that no arbitrageur would find it profitable to trade in the first period, given that the other arbitrageurs only trade at $t = 2$ using the same robust strategy. The proof of no-trade condition Eq. (6) will be similar to the proof in Proposition 2.2; see Appendix A.6 for more details.

A.4 Proof of Proposition 2.4 and Corollary 2.4

Under the prior $\mathcal{L}(0, \xi_v)$, the Maximum a Posteriori (MAP) estimate of $\tilde{v}$ given y_1 is

$$\hat{v}_{map} = \arg\max_v \ f(v|y_1) = \arg\max_v \ f(y_1|v) f_L(v) = \arg\max_v \ \exp\left[-\frac{(y_1 - \frac{v}{\rho\lambda_1})^2}{2\sigma_u^2} - \frac{|v|}{\xi_v} \right], \tag{65}$$

We need to find the point of v that minimizes $(y_1 - \frac{v}{\rho\lambda_1})^2 + \frac{2\sigma_u^2|v|}{\xi_v}$ whose first order condition is $y_1 = \frac{v}{\rho\lambda_1} + \kappa\sigma_u \text{sign}(v)$. Graphically inverting this function $y_1(v)$ leads to the MAP estimator:

$$\hat{v}_{map}(y_1; \xi_v) = \text{sign}(y_1)\rho\lambda_1 \max[|y_1| - \kappa\sigma_u, 0] = \rho\lambda_1 \left[y_1 - \text{sign}(y_1)\kappa\sigma_u \right] \mathbf{1}_{|y_1| > \kappa\sigma_u}, \tag{66}$$

which has a learning threshold $\kappa\sigma_u = \frac{\rho\lambda\sigma_u^2}{\xi_v}$. Eq. (66) is also known as "soft-thresholding" in statistics. This gives a Bayesian interpretation for the LASSO algorithm. LASSO has a similar objective function that involves an l^1 penalty arising from the Laplace prior. The MAP estimate $\hat{v}_{map}$ is a continuous and piecewise-linear function of y_1. One can also apply the MAP learning procedure to directly estimate the residual signal $\tilde{\theta} = \tilde{v} - p_1$:

$$\hat{\theta}_{map} = \arg\max_\theta \ \exp\left[-\frac{(y_1 - \frac{\theta + \lambda_1 y_1}{\rho\lambda_1})^2}{2\sigma_u^2} - \frac{|\theta + \lambda_1 y_1|}{\xi_v} \right] = \arg\min_\theta \frac{(y_1 - \frac{\theta + \lambda_1 y_1}{\rho\lambda_1})^2}{2\sigma_u^2} + \frac{|\theta + \lambda_1 y_1|}{\xi_v}. \tag{67}$$

The first order condition of this objective leads to

$$y_1(\theta) = \frac{\theta}{\rho\lambda_1} + \text{sign}(\theta)\frac{\rho\kappa\sigma_u}{\rho - 1}. \tag{68}$$

Graphically inverting the function $y_1(\theta)$ yields the MAP estimator of $\tilde{\theta}$:

$$\hat{\theta}_{map} = (\rho - 1)\lambda_1 \left[y_1 - \text{sign}(y_1) K^* \right] \mathbf{1}_{|y_1| > K^*}, \quad \text{where } K^* = \frac{\rho\kappa\sigma_u}{\rho - 1} = \frac{\lambda_1 \rho^2 \sigma_u^2}{(\rho - 1)\xi_v} = \frac{\sqrt{2}\sigma_v}{\lambda_1}. \tag{69}$$

Since $K^* = \frac{\rho}{\rho - 1}\kappa\sigma_u > \kappa\sigma_u$, one can also write $\hat{\theta}_{map} = (\hat{v}_{map} - \lambda_1 y_1) \mathbf{1}_{|y_1| > K^*}$. This establishes an *observational equivalence* to the robust strategy, since we find the following identity

$$Z_{2,n}(s, y_1; K^*) = s Z^\infty(y_1, K^*) \mathbf{1}_{|y_1| > K^*} = s\frac{(\hat{v}_{map} - \lambda_1 y_1) \mathbf{1}_{|y_1| > K^*}}{2(N + 1)\lambda_2} = \frac{s \cdot \hat{\theta}_{map}}{2(N + 1)\lambda_2}. \tag{70}$$

Therefore, if arbitrageurs directly use the MAP rule to estimate the mispricing signal $\tilde{\theta}$, they will get the same strategy $Z_{2,n}(s, y_1; K^*)$. This MAP rule (posterior mode estimate) differs from the posterior mean $\hat{v}(y_1; \xi_v)$ which drives the REE strategy $Z^o_{2,n}(s, y_1; \xi_v)$.

Proof of Corollary 2.4: The MAP estimate for each asset value under the prior $\mathcal{L}(0,\xi_v)$ is:

$$\hat{v}_{i,map} = \arg\max_{v_i} f(v_i|y_{1,i}) = \exp\left[-\frac{(y_{1,i}-\frac{v_i}{\rho\lambda_1})^2}{2\sigma_u^2} - \frac{|v_i|}{\xi_v}\right] = \arg\min_{v_i} \left|p_{1,i} - \frac{v_i}{\rho}\right|^2 + \frac{2(\lambda_1\sigma_u)^2}{\xi_v}|v_i|, \tag{71}$$

which amounts to the LASSO objective in the Lagrangian form for $i \in \{1,...,M\}$. This leads to the trading algorithm below, which takes the price change $p_{1,i}$ for each stock as input:

$$Z_{2,n}(p_{1,i},\xi_v) = \frac{(\rho-1)\left[\lambda_1 y_{1,i} - \text{sign}(y_{1,i})\lambda_1 K^*\right]\mathbf{1}_{|\lambda_1 y_{1,i}|>\lambda_1 K^*}}{2(N+1)\lambda_2} = \frac{\rho-1}{N+1}\cdot\frac{p_{1,i}\pm 2\xi_v}{2\lambda_2}\mathbf{1}_{|p_{1,i}|>2\xi_v} \tag{72}$$

where we have used Eq. (16) to derive $\lambda_1 K^* = \sqrt{2}\sigma_v = 2\xi_v$ given $\xi_v = \sigma_v/\sqrt{2}$. Q.E.D.

What if arbitrageurs all adhere to the Gaussian prior? First, they will not trade if their Gaussian prior is identical to market makers' Gaussian prior because they will find out the market is efficient in the semi-strong sense. Arbitrageurs only trade when they have different prior beliefs. Let's model their Gaussian prior as $\tilde{v} \sim \mathcal{N}(0,\tilde{\zeta}^2)$, where $\tilde{\zeta}$ is a random variable reflecting the model uncertainty about the Gaussian prior dispersion. The assumption of prior distribution only changes how arbitrageurs learn from prices without affecting the informed trader's strategy by Assumption 2.1. For any specific value of $\tilde{\zeta} = \zeta$, the arbitrageurs' posterior belief about $\tilde{v}$ conditional on $\tilde{y}_1 = \frac{\tilde{v}}{\rho\lambda_1} + \tilde{u}_1$ is still Gaussian:

$$f(v|y_1) = \frac{f(y_1|v)f_G(v)}{f(y_1)} = \frac{1}{2\pi\zeta\sigma_u f(y_1)}\exp\left[-\frac{(y_1 - v/(\rho\lambda_1))^2}{2\sigma_u^2} - \frac{v^2}{2\zeta^2}\right]. \tag{73}$$

Under the Gaussian prior of $\tilde{v}$, arbitrageurs believe that $y_1 = \frac{\tilde{v}}{\rho\lambda_1} + \tilde{u}_1 \sim \mathcal{N}(0, \zeta^2/(\rho\lambda_1)^2 + \sigma_u^2)$ for a given value of ζ. By projection theorem, they obtain a linear estimator,

$$\hat{v}(y_1;\zeta) = \mathrm{E}^{\mathcal{N}}[\tilde{v}|y_1,\zeta] = \frac{\zeta^2/(\rho\lambda_1)}{\zeta^2/(\rho\lambda_1)^2 + \sigma_u^2}y_1 = \frac{\rho\lambda_1\zeta^2}{\zeta^2 + (\rho\lambda_1\sigma_u)^2}y_1. \tag{74}$$

The mean of a Gaussian distribution is the same as its mode. So the MAP estimate of $\tilde{v}$ coincides with the posterior mean, i.e., $\hat{v}_{map} = \hat{v}$ in this case. The rational strategy for arbitrageurs with Gaussian priors is always a linear function of the order flow y_1:

$$Z^o_{2,n}(y_1;\zeta) = \frac{1}{N+1}\frac{\hat{v}-\lambda_1 y_1}{2\delta\lambda_1} = \frac{(\rho-1)\zeta^2 - (\rho\lambda_1\sigma_u)^2}{\zeta^2 + (\rho\lambda_1\sigma_u)^2}\cdot\frac{y_1}{2(N+1)}, \quad \text{for } n = 1,...,N. \tag{75}$$

Any uncertainty about the prior ζ only changes the slope of this linear strategy. Therefore, the robust strategy must be linear under the max-min choice criteria.

A.5 Proof of Corollary 2.5

If arbitrageurs follow the REE strategy when $s = 1$, the price at $t = 2$ is

$$\tilde{p}_2 = \lambda_1 \tilde{y}_1 + \lambda_2 \left[X_2 + \sum_{n=1}^{N} Z^o_{2,n}(s, \tilde{y}_1; \xi_v) + \tilde{u}_2 \right] = \frac{\tilde{v} + \lambda_1 \tilde{y}_1}{2} + \frac{N}{N+1} \frac{\hat{v} - \lambda_1 \tilde{y}_1}{2} + \lambda_2 \tilde{u}_2. \quad (76)$$

As $N \to \infty$, the expectation of $\tilde{p}_2 = P_2(\tilde{y}_1, \tilde{y}_2)$ under arbitrageurs' information and belief is

$$\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}[\tilde{p}_2 | \mathcal{I}_{2,z}] = \frac{\hat{v} + \lambda_1 y_1}{2} + \frac{\hat{v} - \lambda_1 y_1}{2} = \hat{v} = \mathrm{E}^{\mathcal{A}}[\tilde{v} | \mathcal{I}_{2,z}]. \quad (77)$$

When arbitrageurs use the robust strategy, the price at $t = 2$ is

$$\tilde{p}_2 = \lambda_1 \tilde{y}_1 + \lambda_2 \left[X_2 + \sum_{n=1}^{N} Z_{2,n}(s, \tilde{y}_1; K^*) + \tilde{u}_2 \right] = \frac{\tilde{v} + \lambda_1 \tilde{y}_1}{2} + \frac{N}{N+1} \frac{\hat{v}_{map} - \lambda_1 \tilde{y}_1}{2} \mathbf{1}_{|\tilde{y}_1| > K^*} + \lambda_2 \tilde{u}_2. \quad (78)$$

The (ex ante) expected price under arbitrageurs' information and belief has a positive limit:

$$\lim_{N\to\infty} \mathrm{E}^{\mathcal{A}}[\tilde{p}_2 | \mathcal{I}_{2,z}] = \frac{\hat{v} + \lambda_1 y_1}{2} + \frac{\hat{v}_{map} - \lambda_1 y_1}{2} \mathbf{1}_{|y_1| > K^*} = \frac{\hat{v} + \hat{v}_{map}}{2} - \frac{\hat{v}_{map} - \lambda_1 y_1}{2} \mathbf{1}_{|y_1| > K^*} \neq \hat{v}, \quad (79)$$

indicating price inefficiency in the limit of $N \to \infty$.

A.6 Proof of Corollary 2.6

If arbitrageurs only trade at $t = 2$ and follow the robust strategy we derived, each of them may find that the total trading of other arbitrageurs has a response slope greater than one, i.e., $\frac{N-1}{N+1} \cdot \frac{1-\mu}{1+\mu} > 1$ if $-1 < \mu < 0$ and $N > -\frac{1}{\mu}$. It may become profitable for any arbitrageur to disrupt the equilibrium by trading a large quantity, $z_1 \gg K^*$, in the first period so that the other arbitrageurs will be triggered almost surely. If $z_1 > \frac{(N-1)(\mu-1)}{2(N\mu+1)} K^*$, the momentum trading of arbitrageurs at $t = 2$ can overwhelm the trade z_1. This may create opportunities for the initial instigator to unwind her position at favorable prices.

Suppose the n-th arbitrageur (instigator) secretly trades $z_1 \neq 0$ when $s = 1$ to trick other traders. Her objective at $t = 2$ is to maximize the minimum expected profit over all possible priors: $\max_{z'_{2,n} \in \mathcal{Z}} \min_{\xi \in \Omega} \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \lambda_1 \tilde{y}'_1 - \lambda_2 \tilde{y}'_2) z'_{2,n} | \mathcal{I}_{2,z}]$, where $\tilde{y}'_1 = X_1(\tilde{v}) + z_1 + \tilde{u}_1$ and $\tilde{y}'_2 = X_2(\tilde{v}, \tilde{y}'_1) + z'_{2,n} + Z_{2,-n}(\tilde{y}'_1, K^*) + \tilde{u}_2$. Here, $Z_{2,-n} = \sum_{m \neq n} Z_{2,m}(y'_1, K^*) = \frac{(N-1)\hat{\theta}_{map}(y'_1)}{2(N+1)\lambda_2}$ is the total quantity traded by the other arbitrageurs (excluding the n-th one) who form the estimate of $\tilde{\theta} = \tilde{v} - \lambda_1 y'_1$ based on y'_1 without knowing that y'_1 contains the secret trade z_1. The instigator's estimate, $\hat{\theta}_{map}(y_1) = [\hat{v}_{map}(y_1) - \lambda_1 y_1] \mathbf{1}_{|y_1| > K^*} = (\rho - 1)\lambda_1 [y_1 - \mathrm{sign}(y_1) K^*] \mathbf{1}_{|y_1| > K^*}$,

is however based on $y_1 = x_1 + u_1$ instead of y_1', because she is aware of the order flow z_1 secretly placed by herself. The strategy of this instigator in the second period reflects how she strategically exploits the other traders' overreaction due to her trade z_1:

$$\begin{aligned} Z'_{2,n}(y_1, z_1) &= \frac{\hat{v}_{map}(y_1) - \lambda_1 y_1'}{4\lambda_2}\mathbf{1}_{|y_1|>K^*} - \frac{N-1}{4(N+1)\lambda_2}[\hat{v}_{map}(y_1') - \lambda_1 y_1']\mathbf{1}_{|y_1'|>K^*} \\ &= \frac{\hat{\theta}_{map}(y_1)}{4\lambda_2} - \frac{z_1}{4\delta} - \frac{(N-1)(\rho-1)(y_1+z_1-K^*)}{4(N+1)\delta} \\ &= \frac{\hat{\theta}_{map}(y_1)}{2(N+1)\lambda_2} - \frac{(N+1)z_1 + (N-1)(\rho-1)[z_1 + (y_1-K^*)\mathbf{1}_{|y_1|<K^*}]}{4(N+1)\delta}, \quad (80) \end{aligned}$$

where we used the condition $z_1 \gg K^*$ so that $\mathbf{1}_{|y_1'=y_1+z|>K^*} = 1$ with probability arbitrarily close to 1. Her expected total profit is $\Pi^d_{z,n} = \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \lambda_1\tilde{y}_1')z_1 + (\tilde{v} - \lambda_1\tilde{y}_1' - \lambda_2\tilde{y}_2') \cdot Z'_{2,n}|\mathcal{I}_{1,z}]$ and the extra profit attributable to her unilateral deviation $(z_1, Z'_{2,n})$ is

$$\Delta\Pi^d_{z,n} = \Pi^d_{z,n} - \mathrm{E}^{\mathcal{A}}[(\tilde{v} - \lambda_1\tilde{y}_1 - \lambda_2\tilde{y}_2) \cdot Z_{2,n}|\tilde{s} = 1], \quad (81)$$

where $\tilde{y}_1 = X_1(\tilde{v}) + \tilde{u}_1$, $\tilde{y}_2 = X_2(\tilde{v}, \tilde{y}_1) + \sum_{n=1}^N Z_{2,n}(\tilde{y}_1, K^*) + \tilde{u}_2$, and $Z_{2,n}(\tilde{y}_1, K^*) = \frac{\hat{\theta}_{map}(\tilde{y}_1)}{2(N+1)\lambda_2}$. Using the results $\mathrm{E}^{\mathcal{A}}[\tilde{y}_1 \cdot z_1] = 0$, $\mathrm{E}^{\mathcal{A}}[\hat{\theta}_{map}(\tilde{y}_1) \cdot z_1] = 0$ and $\hat{\theta}_{map}\mathbf{1}_{|y_1|<K^*} = 0$, we derive that

$$\begin{aligned} \Delta\Pi^d_{z,n} &= -\lambda_1 z_1^2 + \lambda_2 \mathrm{E}^{\mathcal{A}}[(Z'_{2,n}(\tilde{y}_1, z_1))^2] - \lambda_2 \mathrm{E}^{\mathcal{A}}[(Z_{2,n}(\tilde{y}_1, K^*))^2] \\ &= -\lambda_1 z_1^2 + \lambda_2 \mathrm{E}^{\mathcal{A}}[(Z'_{2,n}(\tilde{y}_1, z_1) + Z_{2,n}(\tilde{y}_1, K^*))(Z'_{2,n}(\tilde{y}_1, z_1) - Z_{2,n}(\tilde{y}_1, K^*))] \\ &= -\lambda_1 z_1^2 + \lambda_2 \left(\frac{\lambda_1}{\lambda_2}\right)^2 \mathrm{E}^{\mathcal{A}}\left[\left(\frac{(N-1)\rho+2}{4(N+1)}z_1 + \frac{(N-1)(\rho-1)}{4(N+1)}(\tilde{y}_1 - K^*)\mathbf{1}_{|\tilde{y}_1|<K^*}\right)^2\right] \\ &= -\lambda_1 z_1^2 + (1-\mu)\lambda_1 z_1^2 \left[\frac{(N-1)\rho+2}{4(N+1)}\right]^2 + (1-\mu)\lambda_1 \frac{(N-1)^2(\rho-1)^2}{16(N+1)^2}\mathrm{E}^{\mathcal{A}}[(\tilde{y}_1 - K^*)^2\mathbf{1}_{|\tilde{y}_1|<K^*}]. \end{aligned}$$

Since $\delta = \frac{1}{1-\mu}$ and $\rho = \frac{3+\mu}{1+\mu}$ by definition, the above expression is positive if the coefficient in front of z_1^2 is positive. This is equivalent to the condition:

$$1 + \frac{N-1}{N+1} \cdot \frac{2}{1+\mu} > \frac{4}{\sqrt{1-\mu}}. \quad (82)$$

Given any $N > 1$, there exists a critical liquidity point $\mu^*(N)$ below which $\Delta\Pi^d_{z,n} > 0$. For example, $\mu^*(N=2) \approx -0.68037$, $\mu^*(N=3) \approx -0.54843$, $\mu^*(N=10) \approx -0.33525$, $\lim_{N\to\infty}\mu^* = \mu_\epsilon \approx -0.23191$. Thus, in the liquidity regime $\mu < \mu_\epsilon \approx -0.23191$, if the number of arbitrageurs is large enough, the conjectured equilibrium $\mathbf{Z} = [\langle 0, Z_{2,1}\rangle, ..., \langle 0, Z_{2,N}\rangle]$ may fail, because it may permit profitable deviations (or disruptive strategies) at $t = 1$.

A.7 Savvy Informed Trader: Rational-Expectations Equilibrium

For $s = 1$, we investigate the *rational-expectations equilibrium* (REE) in the model of savvy informed trader who anticipates arbitrageurs and strategically interacts with them. Based on $\mathcal{I}_{2,x} = \{v, s, y_1\}$, the informed trader conjectures her residual demand at $t = 2$ and solves

$$X_2(v, y_1) = \arg\max_{x_2} \mathrm{E}\left[(v - P_2(\tilde{y}_1, \tilde{y}_2))\, x_2 | \mathcal{I}_{2,x}\right] = (1-\mu)\frac{v - \lambda_1 y_1}{2\lambda_1} - \frac{\mathrm{E}[Z_2|\mathcal{I}_{2,x}]}{2}. \tag{83}$$

As the informed trader takes into account the price impact of all arbitrageurs, she will reduce her trading quantity by one half of the total arbitrage trading that she expects at $t = 2$. The information set of arbitrageurs right after $t = 1$ is $\mathcal{I}_{2,z} = \{s, y_1\}$, which is nested into the informed trader's information set $\mathcal{I}_{2,x} = \{v, s, y_1\}$. The n-th arbitrageur's objective is

$$\max_{z_{2,n}} \mathrm{E}\left[z_{2,n}\left(\tilde{v} - \lambda_1 \tilde{y}_1 - \lambda_2\left[X_2(\tilde{v}, \tilde{y}_1) + z_{2,n} + Z_{2,-n}(\tilde{y}_1) + \tilde{u}_2\right]\right) | \mathcal{I}_{2,z}\right], \tag{84}$$

from which she can solve the optimal strategy as below

$$Z_{2,n}(y_1) \;=\; (1-\mu)\frac{\hat{v} - \lambda_1 y_1}{4\lambda_1} - \frac{\mathrm{E}[Z_{2,-n}|\mathcal{I}_{2,z}]}{2} + \frac{\mathrm{E}[\mathrm{E}[Z_2|\mathcal{I}_{2,x}]|\mathcal{I}_{2,z}]}{4}. \tag{85}$$

Arbitrageurs are symmetric in terms of their information and objectives. The n-th arbitrageur conjectures that the other arbitrageurs will trade $Z_{2,m} = \eta \cdot (\hat{v} - \lambda_1 y_1)$ for $m = 1, ..., N$ and $m \neq n$, and she also conjectures the informed trader's conjecture that all arbitrageurs trade symmetrically $Z_{2,n} = \eta \cdot (\hat{v} - \lambda_1 y_1)$ for $n = 1, ..., N$. So her optimal strategy becomes

$$Z_{2,n}(y_1) = \left(\frac{1-\mu}{4\lambda_1} - \frac{(N-1)\eta}{2} + \frac{N\eta}{4}\right)(\hat{v} - \lambda_1 y_1). \tag{86}$$

In a symmetric equilibrium, every arbitrageur conjectures in the same way and solves the same problem. This symmetry requires $\eta = \frac{1-\mu}{4\lambda_1} - \frac{(N-1)\eta}{2} + \frac{N\eta}{4}$ that has a unique solution $\eta = \frac{1-\mu}{(N+2)\lambda_1}$. Thus the total order flow from arbitrageurs at $t = 2$ can be written as

$$Z_2 = \sum_{n=1}^{N} Z_{2,n} = N\eta \cdot (\hat{v} - \lambda_1 y_1) = \frac{N(\hat{v} - \lambda_1 y_1)}{(N+2)\lambda_2}. \tag{87}$$

One can prove a simple result that $Z_{2,n} = \mathrm{E}[X_2|\mathcal{I}_{2,z}]$, i.e., every arbitrageur expects that the informed trader on average trades the same quantity as she does. By Eq. (83) and (87),

$$\mathrm{E}[X_2(\tilde{v}, \tilde{y}_1)|\mathcal{I}_{2,z}] = \frac{\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] - \lambda_1 y_1}{2\lambda_2} - \frac{\mathrm{E}[\mathrm{E}[Z_2|\mathcal{I}_{2,x}]|\mathcal{I}_{2,z}]}{2} = \frac{\hat{v} - \lambda_1 y_1}{2\lambda_2} - \frac{Z_2}{2} = \frac{Z_2}{N} = Z_{2,n}. \tag{88}$$

As $\hat{v} = \mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}]$, we obtain the following

$$Z_{2,n}(y_1) = \eta(\hat{v} - \lambda_1 y_1) = \frac{1-\mu}{(N+2)\lambda_1}\left(\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] - \lambda_1 y_1\right) = \mathrm{E}[X_2|\mathcal{I}_{2,z}], \tag{89}$$

$$X_2(v, y_1) = \frac{v - \lambda_1 y_1}{2\lambda_2} - \frac{Z_2}{2} = \frac{v - \lambda_1 y_1}{2\lambda_2} - \frac{N}{N+2}\frac{\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] - \lambda_1 y_1}{2\lambda_2}. \tag{90}$$

One can rewrite the second-period informed trading strategy as

$$X_2(v, y_1) = \frac{v - \lambda_1 y_1}{(N+2)\lambda_2} + \frac{N}{N+2}\cdot\frac{v - \hat{v}}{2\lambda_2}, \tag{91}$$

where the first term is proportional to her informational advantage over market makers and the second term is proportional to her residual advantage over arbitrageurs. Let $\hat{v} = \mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] = g(y_1)$. The informed trader will conjecture the average price at $t = 2$ to be

$$\mathrm{E}[\tilde{p}_2|\mathcal{I}_{2,x}] = \mathrm{E}\left[\lambda_1\tilde{y}_1 + \lambda_2\left(X_2 + \sum_{n=1}^{N} Z_{2,n} + \tilde{u}_2\right)\middle|\mathcal{I}_{2,x}\right] = \frac{(N+2)v + Ng(y_1) + 2\lambda_1 y_1}{2(N+2)}, \tag{92}$$

The informed trader's expected profit from her second-period trading is

$$\Pi_{2,x}(v, y_1) = \mathrm{E}[x_2(v - \tilde{p}_2)|\mathcal{I}_{2,x}] = \frac{1}{\lambda_2}\left(\frac{(N+2)v - Ng(y_1) - 2\lambda_1 y_1}{2(N+2)}\right)^2. \tag{93}$$

The informed trader needs to choose x_1 that maximizes her total expected profits:

$$\begin{aligned}\Pi_x(v) &= \max_{x_1} \mathrm{E}\left[x_1(v - \lambda_1\tilde{y}_1) + \Pi_{2,x}(v, \tilde{y}_1)|\mathcal{I}_{1,x}\right] \\ &= \max_{x_1}\ x_1(v - \lambda_1 x_1) + \frac{1-\mu}{\lambda_1}\mathrm{E}\left[\left(\frac{(N+2)v - Ng(\tilde{y}_1) - 2\lambda_1\tilde{y}_1}{2(N+2)}\right)^2\middle|\mathcal{I}_{1,x}\right],\end{aligned} \tag{94}$$

where $\mathcal{I}_{1,x} = \{v, s = 1\}$. As regularity conditions permit, one can interchange expectation and differentiation operations to derive the first order condition (FOC) for $x_1 = X_1(v)$:

$$0 = v - 2\lambda_1 x_1 - \frac{1-\mu}{\lambda_1}\mathrm{E}\left[\frac{(N+2)v - Ng(x_1 + \tilde{u}_1) - 2\lambda_1(x_1 + \tilde{u}_1)}{2(N+2)}\cdot\frac{Ng'(x_1 + \tilde{u}_1) + 2\lambda_1}{N+2}\right]. \tag{95}$$

When $s = 1$, there does not exist a linear REE where the informed trader's strategy X_1 is a linear function of v. This is proved by contradiction: Suppose X_1 is a linear function of v, the posterior mean $g(y_1) = \mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}]$ will be a nonlinear function of y_1. With a nonlinear $g(y_1)$, the FOC Eq. (95) does not permit a linear solution to $X_1(v)$. Nonlinearity makes Eq. (95) and the REE intractable in general.

A.8 Savvy Informed Trader: Asymptotic Linearity

Based on the asymptotic conjecture of $X_1(v) \to \frac{v}{\rho\lambda_1} + c\kappa\sigma_u$ in the high signal regime, arbitrageurs will find that the posterior distribution of x_1 conditional on y_1 is asymptotically

$$f(x_1|y_1) \to \frac{\rho\lambda_1}{\xi_v f(y_1)\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{(y_1-x_1)^2}{2\sigma_u^2} - \frac{\rho\lambda_1(x_1 - c\kappa\sigma_u)}{\xi_v}\right]. \tag{96}$$

At large order flows, it is deduced that $\mathrm{E}[\tilde{x}_1|y_1] \to y_1 - \kappa\sigma_u$ and furthermore

$$\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] \to \rho\lambda_1[y_1 - (1+c)\kappa\sigma_u]. \tag{97}$$

This result makes the informed trader's FOC Eq. (95) for $x_1 = X_1(v)$ linear again:

$$0 = v - 2\lambda_1 x_1 - \frac{1}{\lambda_2}\mathrm{E}\left[\left(\frac{(N+2)v - N\rho\lambda_1[\tilde{y}_1 - (1+c)\kappa\sigma_u] - 2\lambda_1\tilde{y}_1}{2(N+2)}\right)\left(\frac{N\rho\lambda_1 + 2\lambda_1}{N+2}\right)\middle|\mathcal{I}_{1,x}\right]. \tag{98}$$

After some calculation with the notation $\delta \equiv \frac{\lambda_2}{\lambda_1} = \frac{1}{1-\mu}$, we get

$$0 = v - 2\lambda_1 x_1 - \frac{N\rho+2}{2\delta(N+2)^2}\left[(N+2)v - (N\rho+2)\lambda_1 x_1 + N(1+c)\kappa\rho\lambda_1\sigma_u\right]. \tag{99}$$

This FOC leads to a linear expression of x_1 which conforms to the original linear conjecture:

$$X_1(v) = \frac{(N+2)[2\delta(N+2) - N\rho - 2]}{4\delta(N+2)^2 - (N\rho+2)^2}\left(\frac{v}{\lambda_1}\right) - \frac{N\rho(N\rho+2)(1+c)\kappa}{4\delta(N+2)^2 - (N\rho+2)^2}\sigma_u. \tag{100}$$

Matching the first term leads to a quadratic equation for ρ:

$$-2(\rho-1)(N\rho+2) + 2\delta(\rho-2)(N+2)^2 = 0. \tag{101}$$

There are two roots to this equation but only one of them is sensible as it increases with δ:

$$\rho(\delta, N) = \frac{N + \delta(N+2)^2 - 2 - (N+2)\sqrt{\delta^2(N+2)^2 - 2\delta(3N+2) + 1}}{2N}. \tag{102}$$

Substituting $\delta = \frac{1}{1-\mu}$ into the above equation leads to

$$\rho(\mu, N) = \frac{2 + 5N + N^2 + 2\mu - N\mu - (N+2)\sqrt{N^2 + (1+\mu)^2 + 2N(3\mu-1)}}{2N(1-\mu)}. \tag{103}$$

For $N = 0$, we have $\rho = \frac{3+\mu}{1+\mu}$ which is identical to the parameter ρ in the previous model. There are two more useful limits: $\lim_{\mu\to 0}\rho = 2\left(1+\frac{1}{N}\right)$ and $\lim_{\mu\to 1}\rho = 2$. This equilibrium parameter ρ decreases with μ and N. It is bounded in the range $\left[2, \frac{2(N+1)}{N}\right]$. Now we match the intercept terms and utilize the slope-matching relation to obtain

$$c = -\frac{N(2+N\rho)}{2\delta(2+N)^2 - 2(2+N\rho)} = -\frac{3+N-\mu-\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}{1+N+\mu+\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}\cdot\frac{N}{2}. \quad (104)$$

In the competitive case, we have

$$\lim_{N\to\infty} c = \lim_{N\to\infty}\frac{-N(2+2N)}{2\delta(2+N)^2-2(2+2N)} = -\frac{1}{\delta} = -(1-\mu). \quad (105)$$

There are two more useful limits: $\lim_{\mu\to 0} c = -1$ and $\lim_{\mu\to 1} c = 0$.

Approximation to the rational equilibrium. The symmetry indicates that $X_1(-v) = -X_1(v)$. If $X_1(v)$ is monotone, it should cross the origin and be roughly linear in that neighborhood. With the linearized conjecture $X_1(v\to 0)\to\frac{v}{\alpha\lambda_1}$, one can use Taylor expansion of Eq. (14) at small y_1 to approximate $\mathrm{E}[\tilde{v}|y_1 \ll \kappa\sigma_u] \approx \alpha\beta\lambda_1 y_1$, where α and β are determined by

$$\beta N[\beta N-(N+2)]\alpha^2 + 2\left(\frac{(N+2)^2}{1-\mu}+2\beta N-(N+2)\right)\alpha - 4\left(\frac{(N+2)^2}{1-\mu}-1\right) = 0,$$

$$\beta = 1+\left(\frac{\alpha\lambda_1\sigma_u}{\xi}\right)^2 - \left(\frac{\alpha\lambda_1\sigma_u}{\xi}\right)\frac{e^{-\frac{(\alpha\lambda_1\sigma_u)^2}{2\xi^2}}\sqrt{\frac{2}{\pi}}}{\mathrm{erfc}\left(\frac{\alpha\lambda_1\sigma_u}{\sqrt{2}\xi}\right)}.$$

The first equation is derived from the FOC Eq. (95) and the second one is from the Taylor expansion of Eq. (14). Given $\{\mu, N, \xi\}$, one can numerically find a unique pair of positive solutions to α and β. With constant depth ($\mu = 0$), the first equation becomes $\alpha = \frac{2(N+3)}{N+2-N\beta}$ and the total demand from arbitrageurs becomes $\lim_{\mu\to 0} Z_2 \approx \lim_{\mu\to 0}\frac{N(\alpha\beta-1)}{N+2}y_1 = (\alpha-3)y_1$ for small y_1. The rational equilibrium is not tractable, but one can approximate the arbitrageurs' rational strategy by smoothly pasting the two regimes of asymptotic linearity. There are different methods to make a smooth transition between two linear segments; for example, any sigmoid functions that approach the Heaviside function may work. Here, I use $q(y) = \frac{1}{2}\mathrm{erfc}[a(\kappa\sigma_u - y)] + \frac{1}{2}\mathrm{erfc}[a(\kappa\sigma_u + y)]$, with a tunable parameter $a > 0$ and approximate the posterior mean estimate of $\tilde{v}$ by

$$\hat{v}_a(y_1) \approx [1-q(y_1)]\alpha\beta\lambda_1 y_1 + q(y_1)\rho\lambda_1[y_1 - \mathrm{sign}(y_1)(1+c)\kappa\sigma_u]. \quad (106)$$

Clearly, $\hat{v}_a \to \alpha\beta\lambda_1 y_1$ at $|y_1| \ll \kappa\sigma_u$ and $\hat{v}_a \to \rho\lambda_1[y_1 - \text{sign}(y_1)(1+c)\kappa\sigma_u]$ at $|y_1| \gg \kappa\sigma_u$. The figure below shows numerical approximations to the Bayesian-rational strategy $Z^o_{2,n}(s = 1, y_1; \xi)$ under different ξ, compared with the linear-triggering strategy $Z_{2,n}(s = 1, y_1; K^*)$.

Figure 10. Approximate rational strategies and the linear-triggering strategy (red line).

A.9 Learning Bias and Strategic Informed Trading

Corollary A.1. *Arbitrageurs tend to underestimate the private signal $\tilde{v}$ by a negative amount $-\rho\kappa\lambda_1\sigma_u < 0$. Anticipating this estimation bias, the informed trader in the high signal regime will strategically shift her demand downward by an amount of $c\kappa\sigma_u < 0$ at $t = 1$ and upward by an amount of $d\kappa\sigma_u > 0$ at $t = 2$ where the parameter $d(\mu, N)$ is given by Eq. (111). her average terminal position contains an informational component and a strategic component, that is, $E[X_1(v) + X_2(v, \tilde{u}_1)] \to X^*_{inf}(v) + X^*_{str}$, where $X^*_{str} = (c + d)\kappa\sigma_u$ and*

$$X^*_{inf} = \frac{N + 1 + \mu + \rho(1 - \mu)}{N + 2} \frac{v}{\rho\lambda_1}, \tag{107}$$

Given any $N > 0$, the maximum of $X^*_{inf}(v)$ is at $\mu_c(N) = \sqrt{N(N+2)^3} - N(N+3) - 1$.

Proof: In the asymptotic rational equilibrium we have shown $\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] \to \rho\lambda_1[y_1 - (1+c)\kappa\sigma_u]$ and $y_1 = X_1(v) + \tilde{u}_1 \to (\rho\lambda_1)^{-1}\tilde{v} + c\kappa\sigma_u + \tilde{u}_1$. Arbitrageurs tend to underestimate $\tilde{v}$,

$$\mathrm{E}[\tilde{v}|\mathcal{I}_{2,z}] - \tilde{v} = -\rho\lambda_1\kappa\sigma_u + \rho\lambda_1\tilde{u}_1 \sim \mathcal{N}[-\rho\lambda_1\kappa\sigma_u, (\rho\lambda_1\sigma_u)^2], \tag{108}$$

which has a negative mean $-\rho\lambda_1\kappa\sigma_u < 0$. This learning bias of arbitrageurs entices the informed trader to strategically exploit it. This can be seen from her asymptotic strategy:

$$X_2(v, y_1) \to (1 - \mu)\left[\frac{v - \lambda_1 y_1}{2\lambda_1} - \frac{N}{N + 2}\frac{(\rho - 1)y_1 - (1 + c)\kappa\rho\sigma_u}{2}\right], \tag{109}$$

whose average contains both an informational component and a strategic one:

$$\mathrm{E}[X_2|\tilde{v}=v] = \frac{(1-\mu)(1-\rho^{-1})}{\lambda_1(N+2)}v + \frac{(1-\mu)(N\rho-2c)}{2(N+2)}\kappa\sigma_u. \tag{110}$$

We define another parameter d for this strategic shift which decreases with μ and N:

$$d(\mu,N) = \frac{(1-\mu)(N\rho-2c)}{2(N+2)} = \frac{2N(1-\mu)}{1+N+\mu+\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}. \tag{111}$$

It has the following limit results: $\lim_{\mu\to 0} d = 1$, $\lim_{\mu\to 1} d = 0$, and $\lim_{N\to\infty} d = 1-\mu$. Thus, we have shown that $X_1 \to \frac{v}{\rho\lambda_1} + c\kappa\sigma_u$ where $c<0$ and $\mathrm{E}[X_2|v] \to \frac{(1-\mu)(1-\rho^{-1})}{\lambda_1(N+2)}v + d\kappa\sigma_u$ where $d>0$. This shows how the informed trader strategically exploit the arbitrageurs' bias $\kappa\sigma_u$.

The asymptotic terminal position of the informed trader can be decomposed into an informational term and a strategic term, that is, $\mathrm{E}[X_1(v)+X_2(v,\tilde{u}_1)] \to X^*_{inf}(v) + X^*_{str}$ where $X^*_{str} = (c+d)\kappa\sigma_u \geq 0$. The information-based target inventory is found to be

$$\begin{aligned} X^*_{inf}(v;\mu,N) &= \frac{v}{\rho\lambda_1} + \frac{(1-\mu)(1-\rho^{-1})}{\lambda_1(N+2)}v = \frac{N+1+\mu+\rho(1-\mu)}{N+2}\cdot\frac{v}{\rho\lambda_1} \\ &= \frac{1+3N+\mu-\sqrt{N^2+(1+\mu)^2+2N(3\mu-1)}}{N\rho}\cdot\frac{v}{2\lambda_1}. \end{aligned} \tag{112}$$

which is hump-shaped and reaches its maximum at

$$\mu_c(N) = \sqrt{N(N+2)^3} - N(N+3) - 1. \tag{113}$$

For example, X^*_{inf} has its maximum $0.5359\frac{v}{\lambda_1}$ at $N=1$ and $\mu_c(N=1) = 3\sqrt{3}-5 = 0.196152$. The informed trader manages to reach an informational target position roughly equal to $\frac{v}{2\lambda_1}$.

Figure 11. The information-based target inventory $X^*_{inf}(v)$ and the strategic position X^*_{str}.

A.10 Proof of Proposition 3.1

The candidate linear-triggering strategy for each arbitrageur along the REE asymptotes is

$$Z_{2,n}(s,y_1;K_n)=sZ^{\infty}(y_1,K^*)\mathbf{1}_{|y_1|>K_n}=s\frac{(1-\mu)(\rho-1)}{N+2}\left[y_1-\text{sign}(y_1)\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]\mathbf{1}_{|y_1|>K_n}. \tag{114}$$

For $s=1$, this can be rewritten as

$$\begin{aligned} Z_{2,n}(s,y_1;K_n) &= \frac{\rho\lambda_1[y_1-\text{sign}(y_1)(1+c)\kappa\sigma_u]-\lambda_1 y_1}{(N+2)\lambda_2}\mathbf{1}_{|y_1|>K_n} \\ &= \eta\cdot\left[\hat{v}_T(y_1;\xi_v)-\lambda_1 y_1\right]\mathbf{1}_{|y_1|>K_n}, \end{aligned} \tag{115}$$

where $\eta=\frac{1-\mu}{(N+2)\lambda_1}$ and the implied learning rule for $\tilde{v}$ is

$$\hat{v}_T(y_1;\xi_v)=\rho\lambda\left[y_1-\text{sign}(y_1)(1+c)\kappa\sigma_u\right]\mathbf{1}_{|y_1|>\kappa\sigma_u}. \tag{116}$$

The learning threshold $\kappa\sigma_u$ here ensures that $\hat{v}_T$ takes the same sign as y_1.

Now I prove that in equilibrium every arbitrageur will choose the same threshold

$$K^*=\max\left[\kappa\sigma_u,\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]. \tag{117}$$

Intuitively, any trader choosing K_n lower than the learning threshold $\kappa\sigma_u$ may take actions to trade over the states $|y_1|\in[K_n,\kappa\sigma_u]$ where she actually learns nothing under her learning rule, i.e., $\hat{v}_T=0$ for $|y_1|\in[K_n,\kappa\sigma_u]$. To exclude irrational trading when the inferred signal is zero, the equilibrium threshold must have a lower bound $\kappa\sigma_u$. On the other hand, any trader choosing K_n lower than the intercept $\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}$ may trade against the price trend (contrarian trading) over the states $|y_1|\in\left[K_n,\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]$. This may go against the true (fat-tail) signal and incur losses on average. Therefore, the condition $K_n\geq\max\left[\kappa\sigma_u,\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]$ could make arbitrageurs dedicate to the momentum trading strategy which is desirable in our fat-tail setup. When traders choose thresholds, they actually engage in Bertrand-type competition: each of them will keep undercutting the threshold as long as it is more profitable than the case she follows the common threshold used by other traders. Under this competition, the equilibrium threshold is the boundary K^* given by Eq. (117).

Let's first show that to use any threshold K' lower than K^* cannot be an equilibrium. It suffices to show that when everyone else uses $K_{-n}=K'<K^*$, it is a profitable deviation for

the n-th trader to choose $K_n = K^*$. We need to compare the difference of expected profits:

$$\begin{aligned}
& \mathrm{E}\left[\tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K^*, K_{-n} = K') - \tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K', K_{-n} = K')|\tilde{y}_1 = y_1\right] \\
= \quad & \mathrm{E}\left[\left\{-\eta\left(\frac{1}{2} - \frac{\lambda_1\eta(N-2)}{2(1-\mu)}\right)(\hat{v}_T - \lambda_1 y_1)^2 - \eta(\hat{v}_T - \lambda_1 y_1)\frac{\tilde{v} - \hat{v}_T}{2}\right\}\mathbf{1}_{K'<|y_1|<K^*}\middle| y_1\right] \\
= \quad & -\frac{\eta}{2}\left[\frac{4\Theta^2}{N+2} + (\hat{v} - \hat{v}_T)\Theta\right]\mathbf{1}_{K'<|y_1|<K^*}, \qquad (118)
\end{aligned}$$

where $\Theta = \hat{v}_T - \lambda_1 y_1$ is negative for $K' < y_1 < K^*$ and positive for $K' < -y_1 < K^*$. For the case $K^* = \kappa\sigma_u$, we have $\hat{v}_T = 0$ but $\hat{v} \geq 0$ for $|y_1| \in [K', \kappa\sigma_u]$. It means the last expression of Θ is a parabola that opens downward and crosses the origin. Since Θ takes the opposite sign of y_1 and $\hat{v}$ for $K' < |y_1| < K^*$, the last expression is strictly positive for $K' < |y_1| < K^*$. Similar arguments can be applied to the case $K^* = \frac{\rho(1+c)\kappa\sigma_u}{\rho-1}$. Therefore, $\mathrm{E}\left[\tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K^*, K_{-n} = K') - \tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K', K_{-n} = K')\right] > 0$ for $K' < K^*$, i.e., any threshold less than K^* cannot be an equilibrium threshold.

Similarly, any threshold K' larger than K^* cannot be an equilibrium threshold either. As before, it suffices to show that the deviation is profitable for any trader by just choosing $K_n = K^*$ less than K' used by others. The payoff difference given y_1 is positive as well:

$$\begin{aligned}
& \mathrm{E}\left[\tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K^*, K_{-n} = K') - \tilde{\pi}_{z,n}(\tilde{y}_1; K_n = K', K_{-n} = K')|\tilde{y}_1 = y_1\right] \\
= \quad & \mathrm{E}\left[\eta(\hat{v}_T - \lambda_1 y_1)\mathbf{1}_{K^*<|y_1|<K'}\left[\frac{\tilde{v} - \lambda_1 y_1}{2} - \frac{\lambda_1\eta(\hat{v}_T - \lambda_1 y_1)}{1-\mu}\right]\middle| y_1\right] \\
= \quad & \frac{\eta}{2}\left[\frac{N}{N+2}\Theta^2 + (\hat{v} - \hat{v}_T)\Theta\right]\mathbf{1}_{K^*<|y_1|<K'} > 0. \qquad (119)
\end{aligned}$$

It rules out any threshold larger than K^* to be an equilibrium. So the only possible equilibrium choice is K^*. When every trader uses the same threshold K^*, no one will deviate.

Now look at the informed trader in this algorithmic trading game. If arbitrageurs all use the same threshold K (which can be general), the informed trader at $t = 2$ will trade

$$X_2(v, y_1; K) = (1-\mu)\frac{v - \lambda_1 y_1}{2\lambda_1} - s\frac{N(1-\mu)(\rho-1)}{2(N+2)}\left[y_1 - \mathrm{sign}(y_1)\frac{\rho(1+c)\kappa\sigma_u}{\rho-1}\right]\mathbf{1}_{|y_1|>K}, \qquad (120)$$

and pick $X_1(v, K)$ that maximizes her total payoff. The price at $t = 2$ can be written as

$$\tilde{p}_2 = \lambda_1 y_1 + \lambda_2\left(x_2 + Z_2\mathbf{1}_{|y_1|>K} + \tilde{u}_2\right) = \begin{cases} \frac{(N+2)\tilde{v} + 2\lambda_1 y_1 + N\hat{v}_T}{2(N+2)} + \lambda_2\tilde{u}_2, & \text{if } |y_1| > K \\ \frac{\tilde{v} + \lambda_1 y_1}{2} + \lambda_2\tilde{u}_2, & \text{if } |y_1| < K, \end{cases} \qquad (121)$$

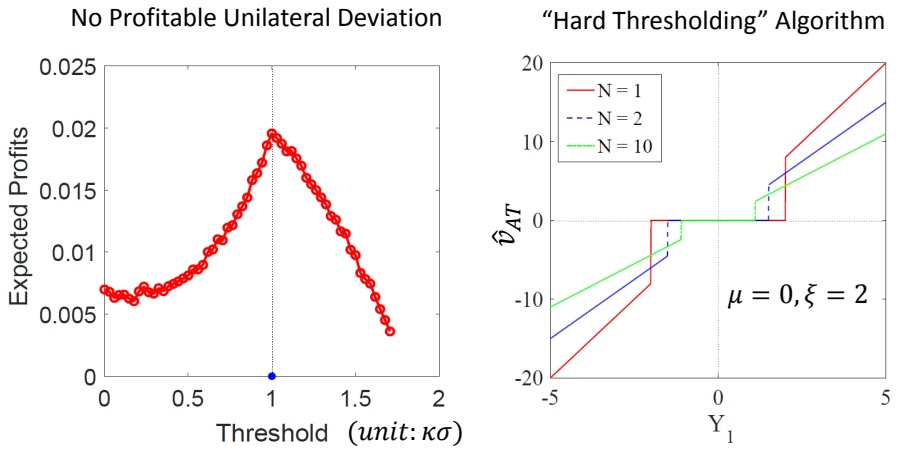


Figure 12. Left: Trading profits if someone deviates from K^*. Right: Learning rule $\hat{v}_T(y_1)$.

depending on whether the arbitrageurs are triggered. The informed trader's expected profit in the second period is also contingent on the state of arbitrageurs:

$$\Pi_{2,x}(v, y_1; K) = \mathrm{E}[x_2(v - \tilde{p}_2)|\mathcal{I}_{2,x}] = \begin{cases} \frac{1}{\lambda_2}\left[\frac{(N+2)v - N\hat{v}_T(y_1) - 2\lambda_1 y_1}{2(N+2)}\right]^2, & \text{if } |y_1| > K \\ \frac{1}{\lambda_2}\left(\frac{v-\lambda_1 y_1}{2}\right)^2, & \text{if } |y_1| < K. \end{cases} \tag{122}$$

Note that her expected profit in the second period is always positive because the informed trader fully anticipates the response of arbitrageurs. The informed trader needs to determine $x_1 = X_1(v; K)$ that maximizes the total expected profit from both periods. The calculation of her total profit, conditional on the private signal v, can be decomposed into three components:

$$\begin{aligned} \Pi_x(v, x_1; K) &= \max_{x_1} \mathrm{E}\left[\Pi_{1,x} + \Pi_{2,x}\mathbf{1}_{|\tilde{y}_1|<K} + \Pi_{2,x}\mathbf{1}_{|\tilde{y}_1|>K}|\mathcal{I}_{1,x}\right] \\ &= x_1(v - \lambda_1 x_1) + \mathrm{E}\left[\frac{(v - \lambda_1(x_1 + \tilde{u}_1))^2}{4\lambda_2}\mathbf{1}_{|x_1+\tilde{u}_1|<K}\middle|\mathcal{I}_{1,x}\right] \\ &\quad + \mathrm{E}\left[\frac{[(N+2)v - N\hat{v}_T(x_1 + \tilde{u}_1) - 2\lambda_1(x_1 + \tilde{u}_1)]^2}{4(N+2)^2\lambda_2}\mathbf{1}_{|x_1+\tilde{u}_1|>K}\middle|\mathcal{I}_{1,x}\right] \end{aligned} \tag{123}$$

On one hand, the informed trader may want to trade less to avoid triggering arbitrageurs and take full advantage of her information at $t = 2$. On the other hand, it is costly to hide her private signal if it is strong. This trade-off will reflect in the relative values of $\Pi^-_{2,x}$ and $\Pi^+_{2,x}$ which are defined below. Hereafter, I set $\sigma_u = 1$ for convenience. By direct integration,

one can derive their expressions:

$$\begin{aligned}
&\Pi^{-}_{2,x}(v,x_1;K) \equiv \mathrm{E}[\Pi_{2,x}\mathbf{1}_{|\tilde{y}_1|<K}|\mathcal{I}_{1,x}] = \mathrm{E}\left[\frac{(v-\lambda_1\tilde{y}_1)^2}{4\lambda_2}\mathbf{1}_{|\tilde{y}_1|<K}\middle|\mathcal{I}_{1,x}\right] \\
= \;& \frac{(1-\mu)[2v-\lambda_1(K+x_1)]\phi(K-x_1)}{4} - \frac{(1-\mu)[2v+\lambda_1(K-x_1)]\phi(K+x_1)}{4} \\
&+\frac{(1-\mu)[(v-\lambda_1 x_1)^2+\lambda_1^2]}{8\lambda_1}\left[\mathrm{erf}\left(\frac{K-x_1}{\sqrt{2}}\right)+\mathrm{erf}\left(\frac{K+x_1}{\sqrt{2}}\right)\right], \qquad (124)
\end{aligned}$$

$$\begin{aligned}
&\Pi^{+}_{2,x}(v,x_1;K) \equiv \mathrm{E}[\Pi_{2,x}\mathbf{1}_{|\tilde{y}_1|>K}|\mathcal{I}_{1,x}] \\
= \;& \mathrm{E}\left[\frac{[(N+2)v-N\hat{v}_T(\tilde{y}_1)-2\lambda_1\tilde{y}_1]^2}{4(N+2)^2\lambda_2}\mathbf{1}_{|\tilde{y}_1|>K}\middle|\mathcal{I}_{1,x}\right] \\
= \;& \frac{(1-\mu)(N\rho+2)}{4(N+2)^2}[-2Nw\rho\lambda_1-2(N+2)v+\lambda_1(N\rho+2)(K+x_1)]\phi(K-x_1) \\
&+\frac{(1-\mu)(N\rho+2)}{4(N+2)^2}[-2Nw\rho\lambda_1+2(N+2)v+\lambda_1(N\rho+2)(K-x_1)]\phi(K+x_1) \\
&+\frac{1-\mu}{8(N+2)^2\lambda_1}\left\{[(N+2)v+Nw\rho\lambda_1-(N\rho+2)\lambda_1x_1]^2+\lambda_1^2(N\rho+2)^2\right\}\mathrm{erfc}\left(\frac{K-x_1}{\sqrt{2}}\right) \\
&+\frac{1-\mu}{8(N+2)^2\lambda_1}\left\{[(N+2)v-Nw\rho\lambda_1-(N\rho+2)\lambda_1x_1]^2+\lambda_1^2(N\rho+2)^2\right\}\mathrm{erfc}\left(\frac{K+x_1}{\sqrt{2}}\right). \qquad (125)
\end{aligned}$$

where $w=(1+c)\kappa\sigma_u$ is the horizontal intercept of $\hat{v}_T(y_1)$ and where $\phi(K\pm x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(K\pm x)^2}{2}}$ denotes the probability density function of the standard normal distribution (with $\sigma_u=1$). Taking the first derivative, $\frac{d\Pi_x}{dx_1}=0$, one can find the FOC for $X_1(v;K)=x_1$:

$$\begin{aligned}
0 = \;& v-2\lambda_1x_1-\frac{(1-\mu)(v-\lambda_1x_1)}{4}\left[\mathrm{erf}\left(\frac{K-x_1}{\sqrt{2}}\right)+\mathrm{erf}\left(\frac{K+x_1}{\sqrt{2}}\right)\right] \\
&+\frac{(1-\mu)[(v-\lambda_1K)^2+2\lambda_1^2]}{4\lambda_1}[\phi(K+x_1)-\phi(K-x_1)]+(1-\mu)Kv\phi(K+x_1) \\
&+\frac{(1-\mu)\phi(K-x_1)}{4\lambda_1(N+2)^2}\{[K\lambda_1(N\rho+2)-(N+2)v]^2+2\lambda_1^2(N\rho+2)^2 \\
&\qquad +\lambda_1wN\rho[\lambda_1wN\rho-2K\lambda_1(N\rho+2)+2(N+2)v]\} \\
&-\frac{(1-\mu)\phi(K+x_1)}{4\lambda_1(N+2)^2}\{[K\lambda_1(N\rho+2)+(N+2)v]^2+2\lambda_1^2(N\rho+2)^2 \\
&\qquad +\lambda_1wN\rho[\lambda_1wN\rho-2K\lambda_1(N\rho+2)-2(N+2)v]\} \\
&-\frac{(1-\mu)(N\rho+2)}{4(N+2)^2}[(N+2)v+Nw\rho\lambda_1-(N\rho+2)\lambda_1x_1]\mathrm{erfc}\left(\frac{K-x_1}{\sqrt{2}}\right) \\
&-\frac{(1-\mu)(N\rho+2)}{4(N+2)^2}[(N+2)v-Nw\rho\lambda_1-(N\rho+2)\lambda_1x_1]\mathrm{erfc}\left(\frac{K+x_1}{\sqrt{2}}\right). \qquad (126)
\end{aligned}$$

This FOC equation defines the informed trader's optimal strategy $X_1 = x_1(v;K)$ at $t=1$. The unconditional expected total profit of all arbitrageurs is

$$\Pi_z^{tot} \equiv \mathrm{E}\left[\sum_{n=1}^{N} \tilde{\pi}_{z,n}(\tilde{v},\tilde{u}_1,\tilde{u}_2)\right] = \mathrm{E}[(\tilde{v}-\tilde{p}_2)Z_2\mathbf{1}_{|y_1|>K}]. \tag{127}$$

After solving $x_1 = X_1(v;K)$ given any v, one can compute the conditional expected profit:

$$\begin{aligned}
&\mathrm{E}\left[\sum_{n=1}^{N} \tilde{\pi}_{z,n}(\tilde{v},\tilde{u}_1,\tilde{u}_2)\middle|\tilde{v}=v\right] \\
=\;& \mathrm{E}\left[(\tilde{v}-\lambda_1\tilde{y}_1-\lambda_2\left(X_2(\tilde{v},\tilde{y}_1)+Z_2(\tilde{y}_1)\right))Z_2(\tilde{y}_1)\mathbf{1}_{|\tilde{y}_1|>K}|\tilde{v}=v\right] \\
=\;& \frac{N(1-\mu)}{2(N+2)^2}\left[w\lambda_1\rho(2N\rho+2-N)+(\rho-1)((N+2)v-\lambda_1(N\rho+2)(x_1+K))\right]\phi(K-x_1) \\
&+\frac{N(1-\mu)}{2(N+2)^2}\left[w\lambda_1\rho(2N\rho+2-N)-(\rho-1)((N+2)v-\lambda_1(N\rho+2)(x_1-K))\right]\phi(K+x_1) \\
&-\frac{N(1-\mu)}{4(N+2)^2}[(N+2)v(\rho(w-x_1)+x_1)+\lambda_1 N\rho^2(1+(w-x_1)^2) \\
&\qquad -\lambda_1\rho(N-2)(x_1^2-wx_1+1)-2\lambda_1(1+x_1^2)]\mathrm{erfc}\left(\frac{K-x_1}{\sqrt{2}}\right) \\
&+\frac{N(1-\mu)}{4(N+2)^2}[(N+2)v(\rho(w+x_1)-x_1)-\lambda_1 N\rho^2(1+(w+x_1)^2) \\
&\qquad +\lambda_1\rho(N-2)(x_1^2+wx_1+1)+2\lambda_1(1+x_1^2)]\mathrm{erfc}\left(\frac{K+x_1}{\sqrt{2}}\right),
\end{aligned} \tag{128}$$

where $w \equiv (1+c)\kappa\sigma_u$ and $\sigma_u = 1$. Finally, the unconditional total payoff to arbitrageurs is

$$\Pi_z^{tot} = \mathrm{E}\left[\sum_{n=1}^{N} \tilde{\pi}_{z,n}(\tilde{v},\tilde{u}_1,\tilde{u}_2)\right] = \int_{-\infty}^{+\infty} f_L(v)\mathrm{E}\left[\sum_{n=1}^{N} \tilde{\pi}_{z,n}(\tilde{v},\tilde{u}_1,\tilde{u}_2)\middle|\tilde{v}=v\right]dv. \tag{129}$$

A.11 Proof of Corollary 3.3

Since $\lim_{\mu\to 0} c = -1$ and $\lim_{\mu\to 0}\rho = 2+\frac{2}{N}$, one can derive that for the informed trader

$$\begin{aligned}
\lim_{v\to 0}\lim_{\mu\to 0}\Pi_x \;=\;& \lambda_1 - \frac{3\lambda(1+x_1^2)}{8}\left[\mathrm{erf}\left(\frac{K-x_1}{\sqrt{2}}\right)+\mathrm{erf}\left(\frac{K+x_1}{\sqrt{2}}\right)\right] \\
&+\frac{3\lambda_1}{4}\{[\phi(K-x_1)+\phi(K+x_1)]K+[\phi(K-x_1)-\phi(K+x_1)]x_1\},
\end{aligned} \tag{130}$$

which only depends on x_1, λ_1, and K. The FOC equation in this limiting case becomes

$$\frac{3\lambda}{4}\left[(2+K^2)\left[\phi(K-x_1)-\phi(K+x_1)\right]-x_1\left(\operatorname{erf}\left(\frac{K-x_1}{\sqrt{2}}\right)+\operatorname{erf}\left(\frac{K+x_1}{\sqrt{2}}\right)\right)\right]=0. \tag{131}$$

Using the equilibrium threshold $K^*(\mu=0)=\kappa$ with $\sigma_u=1$, we can rewrite the FOC as:

$$\frac{x_1}{2+\kappa^2}=\frac{\phi(\kappa-x_1)-\phi(\kappa+x_1)}{\operatorname{erf}\left(\frac{\kappa-x_1}{\sqrt{2}}\right)+\operatorname{erf}\left(\frac{\kappa+x_1}{\sqrt{2}}\right)}, \tag{132}$$

which may have multiple solutions: one is obviously $x_1=0$ and the other two are $\pm\infty$.

As long as the informed trader trades a sufficiently large quantity $x_1 \gg \kappa$ (instead of $\pm\infty$), the probability of triggering arbitrageurs to trade is arbitrarily close to one. In the second period, the informed trader's optimal strategy is found to be $\lim_{\mu\to 0} X_2(v=0,y_1)=-y_1$, which exactly offsets the total quantity traded by arbitrageurs $\lim_{\mu\to 0} Z_2(y_1)=y_1$. Thus, the terminal position of the informed trader is $x_1+x_2=-u_1$ which is zero on average. The expected profit from this disruptive strategy is found to be $\Pi_x \approx \lambda_1\sigma_u^2$, which is limited by the noise trading volatility in the first period.

Strategic Speed Choice by High-Frequency Traders under Speed Bumps*

Jun Aoyagi†

March 10, 2019

Abstract

We study how high-frequency traders (HFTs) strategically decide their speed level in a market with a random speed bump. If HFTs recognize the market impact of their speed decision, they perceive a wider bid-ask spread as an endogenous upward-sloping cost of being faster. We find that the speed elasticity of the bid-ask spread (slope of the endogenous cost function) negatively depends on the expected length of a speed bump since a longer delay makes market makers insensitive to HFTs' speed increment. Hence, speed bumps promote the investment of HFTs in high-speed technology by reducing the marginal cost of getting faster, undermining their intended purpose of protecting market makers. Depending on the expected length of a bump and exogenous cost of speed, an arms race among HFTs exhibits both complementarity and substitution. These findings explain the ambiguous empirical results regarding speed bumps and adverse selection for market makers.

Keywords: High-frequency trading, market structure, speed bumps, adverse selection, strategic speed decision.

JEL Classification: D40, D47, G10, G18, G20

*First draft: June 2018.

†Department of Economics, University of California at Berkeley. E-mail jun.aoyagi@berkeley.edu.

I appreciate the constructive comments from Michael Brolley, Nicolae Garleanu, Terry Hendershott, Ryo Horii, Ulrike Malmendier, Hayden Melton, Emi Nakamura, Christine Parlour, Yoko Shibuya, David Sraer, Yingge Yan, Haoxiang Zhu, and seminar participants at Osaka University (ISER), UC Berkeley and University of Tokyo. I am grateful to ClaireValgardson for copy-editing. This paper was awarded the 2018 Moriguchi Prize by the Institute of Social and Economic Research, Osaka University.

1 Introduction

> "Never before in human history have people gone to so much trouble and spent so much money to gain so little speed."—*Flash Boys: A Wall Street Revolt* by Lewis (2014).

The ever-increasing speed of electronic financial markets pushes traders to be lightning fast. They are obsessed with being the first to acquire information for trading purposes, spending significant amounts of money on high-speed technologies, such as custom-built fiber-optic cables and microwave/millimeter-wave transmissions. With these sophisticated tools, high-frequency traders (HFTs) can extract information from massive layers of signals at the speed of light.

Regulators are concerned about how quickly HFTs can access and act on information. It is argued that the informational advantage that HFTs obtain through increasing speed exposes market makers to the cost of adverse selection in the sense of Glosten and Milgrom (1985). That is, HFTs trade with market makers only if they receive news that is not yet publicly available and find market makers' orders outdated and mispriced.[1] By exploiting their speed advantage, HFTs "snipe" stale quotes provided by market makers (Budish et al., 2015).

The speed race by HFTs has prompted some exchange platforms to slow down HFT-involved transactions by introducing *speed bumps*. A speed bump imposes a delay on the arrival or execution of orders at a market, aiming to protect traders from exposure to the above-mentioned risks. For example, the Investors Exchange (IEX) adopts a 350-microsecond speed bump on incoming orders and outgoing information from the exchange. The Aequitas NEO Exchange and TMX Group, both Canadian exchanges, also apply a few milliseconds of random delay to non-cancellation orders.[2] Specifically, the speed bump in the latter markets aims to slow down only HFT-involved orders by classifying traders into high-frequency (latency-sensitive) and non-high-frequency categories.[3]

[1] One of the most frequently cited market benefits is liquidity provision by high-frequency market makers. However, extremum events, such as the May 2010 "flash crash," make regulators increasingly concerned that the liquidity provided by HFTs is likely to evaporate when it is most needed. See, Conrad et al. (2015) for the empirical study of this liquidity evaporation.

[2] See Appendix F of Baldauf and Mollner (2017) which provides a comprehensive summary of institutional details.

[3] Depending on the institutional details, the types of traders (or orders) to be protected may change. For

Table I: Top 5 Firms by Volume on BrokerTec

Firm	Volume($ millions)	Market Share
Jump Trading	2,291,000	28%
Citadel LLC	1,004,000	12%
Teza Technologies	905,000	11%
KCG	798,000	10%
JP Morgan	649,000	8%

Note: It tabulates shares in May-June, 2015. Data regarding top-10 HFTs is also available and indicates a similar result. Source: Risk.com, October 2015, Issue 10.

This paper analyzes the effect of speed bumps on speed decision of HFTs and on the adverse selection cost for market makers by focusing on a non-cancellation delay aimed at hampering sniping behavior of HFTs.[4] The key result is that a speed bump can *increase* the speed of HFTs and *worsen* adverse selection for market makers in contrast to its intended purpose. Specifically, once we allow HFTs to strategically choose their speed level (i.e, they are aware of the reaction of price setters), a speed bump increases the marginal benefit of being faster.

The strategic motive in the speed decision arises because major high-speed financial institutions have significant shares in the trading volume in markets.[5] For example, Table I shows the top five high-frequency financial institutions and their shares in the BrokerTec platform, through which more than half of the U.S. Treasury is traded. Typically, HFTs benefit from a huge number of small, short-lived transactions, and each trading decision does not impact the equilibrium price. However, when an institution decides a speed technology, she becomes aware of the market impact of her choice because a sizable number of transactions involve the same speed technology and affect the equilibrium price.

instance, the speed bump in the IEX is more likely to protect pegged orders from non-HFTs, but not market makers on a lit LOB, from being sniped by HFTs. The non-cancellation delay and the HFT-specific delay adopted by the two Canadian exchanges are more likely to save market makers from adverse selection cost.

[4]The term "speed" includes the choice of the geographical location of the firm's information server. For example, a spot in the mid-Atlantic ocean is the optimal point to exploit the price difference between the NYSExchange and the London Stock Exchange.

[5]There is anecdotal evidence for HFTs being aware of the market impact of their speed choice. For example, clients who purchase a speed device from a trade technology company often try to hide it by asking to peel corporate logos from shipments due to confidentiality clauses. See, for example, https://www.wsj.com/ and Lewis (2014).

As the literature points out, the faster the HFTs, the more severe adverse selection the market makers face. Thus, a higher speed puts positive pressure on the bid-ask spread and, in turn, reduces the sniping profit of HFTs. Therefore, the spread works as an *endogenous* upward-sloping cost of being faster. Importantly, it provides a channel through which a speed bump affects the speed decision of HFTs, because market makers' adverse selection risk, as well as the equilibrium spread, is affected by a speed bump. This channel is novel since an exogenous sunk cost of speed analyzed in the literature is independent of a bump.

First, we consider a simple benchmark structure to separate the key mechanism: homogeneous markets with a single HFT, having a random speed bump of a δ-period with $\lambda = E[\delta]$. If λ increases, market makers know that they are less likely to be picked off by the HFT. As a result, they do not care much about a marginal increase in the HFT's speed, and their pricing behavior, i.e., the bid-ask spread, becomes less responsive. This induces a lower endogenous marginal cost of speed investment for the HFT, providing her with a stronger incentive to be faster.

As an extension, we consider a speed competition among *multiple* HFTs—an "arms race"—and allow them to serve not only as snipers but also as high-frequency market makers. In the literature, such as Foucault et al. (2003), traders' speed levels interact with each other because one trader's speed affects other traders' probability of successful snipe, leading to the strategic substitution. In contrast, our endogenous cost of speed (bid-ask spread) provides a new channel for the interaction because the spread is an equilibrium variable. Specifically, depending on the relative significance of the exogenous and endogenous costs of speed, the arms race can exhibit both strategic complementarity and substitution.

If the exogenous sunk cost of speed is relatively small, an arms race creates strategic *complementarity*, because the speed-up by an HFT as a market maker reduces the sensitivity of her spread to other HFTs' speed-up. Also, a faster market maker decreases the sniping probability of snipers, making them care less about an adverse price movement caused by their speed increase. As a result, a faster market maker reduces snipers marginal cost of being faster and enhances their investment in speed.

In this situation, the introduction of a speed bump can backfire because a higher λ makes each HFT willing to be faster, triggering a fiercer speed competition and positive externality due to the complementarity. Although a speed bump protects market makers and mitigates adverse selection via its direct effect, the equilibrium speed increases substantially

and dominates the direct protection, worsening adverse selection risk.

Therefore, our strategic model with the endogenous cost of speed proposes opposite results to a traditional model with an exogenous sunk cost. We can think of these as two extreme cases: By introducing exogenous cost in our model and adjusting the exogenous cost parameter, for example, our model navigates between these extremes, leading to rich equilibrium behavior.

While our model is theoretical, we propose some testable implications and policy discussions. For example, our model implies that the SEC's policy in 2017 that approved the IEX (with a bump) as a National Securities Exchange can strengthen HFTs' demand for speed technologies, thereby allowing exchange platforms to charge higher fee for the direct data feed and colocation service. Our model indicates that this effect undermines the recent attempt of the SEC to block the exchange platforms from increasing the price for their data access.[6]

1.1 Literature Review

This paper contributes to the literature on high-frequency trading and market structure (see Jones, 2013; O'Hara, 2015; Menkveld, 2016 for reviews). Biais et al. (2015) analyze the effect of an arms race and show that a higher speed triggers more severe adverse selection for slow traders. Delaney (2018) describes the speed decision of HFTs as a model of irreversible investment with an optimal stopping time, while Bongaerts and Van Achter (2016) view it from a perspective of high-frequency market making.[7] However, the speed decision in these models is discrete (i.e., being fast or not), and they abstract away from addressing the implications of the equilibrium level of speed. Based on Foucault et al. (2003), Liu (2009) and Foucault et al. (2016) investigate a continuous choice of speed based on the monitoring intensity of traders.[8] However, traders decide on the speed level simultaneously with other

[6] As for the SEC's approval of the IEX, see Hu (2018). For the recent proposals regarding the increasing price of direct data feed charged by exchange platforms, see https://www.wsj.com/articles/nyse-nasdaq-take-it-on-the-chin-in-washington-1539941404.

[7] Aït-Sahalia and Saglam (2013), Hoffmann (2014), Foucault et al. (2016) construct models with HFTs to address the effect of high-frequency market making. See Conrad et al. (2015) for the empirical study of high-frequency quoting.

[8] Foucault et al. (2013) consider the optimal choice of the monitoring intensity by high-frequency snipers and market makers. It involves the exogenous cost but is not strategic. Both snipers and market makers

types of players (e.g., market makers), which requires them to focus on the exogenous cost of speed investment. Our model differs from theirs since the speed decision is continuous and bears an endogenous cost due to the strategic motive of HFTs. Our results are unique since these two modifications empower us to analyze how speed choice is affected by speed bumps.

As traders get faster, questions arise regarding the speed and frequency of executions by a trading platform. By altering the trading frequency of the Kyle-type model, Du and Zhu (2017) show that a low-frequency platform works better to reallocate assets, though it limits the ability to react to new information promptly. Pagnotta and Philippon (2018) also consider platforms' decisions regarding execution frequency and fees to attract speed-sensitive traders. Menkveld and Zoican (2017) also explore the effect of latency on HFTs' strategy and spread, citing risk aversion as a key to generating the result.[9] In their analyses, which pays little attention to the speed choice of HFTs, the frequency of transactions is determined at a market level and applies to all investors.

Our model shares the same interests as the studies on the impact of slow market structures, such as frequent batch auctions (Budish et al., 2015; Haas and Zoican, 2016) and speed bumps (Baldauf and Mollner, 2017; Brolley and Cimon, 2017; Aldrich and Friedman, 2018), on HFTs' behavior and adverse selection for market makers. However, they do not consider a continuous optimal speed decision by HFTs with a delay-sensitive endogenous cost. Thus, they conclude that these mechanisms mitigate adverse selection for market makers, an assertion that will be overturned in our model.[10]

The scope of the literature extends to other empirical findings regarding the HFT and the effect of bumps.[11] Hu (2018) analyzes the SEC approval of the IEX as a national securities

obtain a positive profit from trading due to heterogeneous private values of an asset, generating strategic complementarity in an arms race.

[9] Menkveld and Zoican (2017) obtain a hump-shaped equilibrium spread against a delay. This stems from the switch from the pure-strategy to mixed-strategy equilibrium, and it depends on the risk aversion parameter.

[10] Moreover, these models do not study the coexistence of slow and fast markets, which is analyzed in Appendix A.2. In independent work, Brolley and Cimon (2017) explore this coexistence and find a result consistent with ours, although it stems from a completely different mechanism.

[11] Hendershott and Moulton (2011) analyze the impact of the hybrid market at the NYSE and show that the faster market structure increases quoted and effective spreads and adverse selection cost. Riordan and Storkenmaier (2012) focus on the system upgrade in the Deutsche Boerse, Frino et al. (2014), Boehmer et al.

exchange, making traders route their orders under the "Order Protection Rule," and finds a net improvement in market quality measured by the spreads. Shkilko and Sokolov (2016) exploit interruptions of messaging via microwave communication caused by precipitation (i.e., rain or snow) to find a reduction in quoted spreads.[12] Chen et al. (2017) investigate the effect of a bump in the TMX Alpha, reporting an increase in quoted spreads. In our model, we can reconcile these results because, depending on the relative significance of the exogenous cost and the level of expected delay, speed bumps will affect a spread negatively, positively, or not at all.

2 The Benchmark Model

This section proposes a simple benchmark model to separate the main mechanism. Consider a one-shot exchange of an asset, in which a short-lived HFT tries to snipe stale limit orders. The asset has a stochastic liquidation value $v = \pm\sigma$ with equal probability. v is publicly announced at a stochastic time T, which occurs as a Poisson arrival with intensity γ. With the public announcement, the asset is liquidated. It is traded during $t < T$ due to liquidity needs or the arrival of private information, as in Glosten and Milgrom (1985) and Budish et al. (2015). Following the convention of market microstructure, we assume that each trader can hold only a unit position.

2.1 Traders

There is a continuum of competitive slow, uninformed market makers with a unit mass. At the beginning of the trading game ($t = 0$), all market makers submit a single-unit limit order with a half spread s to commit to trade at this price. The order will disappear from the limit order book if there is a taker or if the market maker cancels it based on public news. To focus on the short-horizon behavior, we assume that market makers do not return to the

(2015), and Brogaard et al. (2015) study the colocation as an example of latency reduction, and Hasbrouck and Saar (2013) construct a measure of low-latency in the NASDAQ to find subsequent shrinkages in spreads. On the other hand, Ye et al. (2013) analyze the importance of the tick-size constraint and report that latency declines at the NASDAQ did not significantly alter spreads (except for the smallest stocks).

[12] Although the interruption by precipitation may have a similar effect to a speed bump, they mentioned that this phenomenon is not paid much attention by financial institutions, while traders anticipate a speed bump and take it into their decision making.

market once they exit. Cancellation is immediate and incurs no cost.

There is one ($N = 1$) risk-neutral high-frequency trader (HFT). Before $t = 0$, the HFT invests in a technology that provides the speed ϕ.[13] Equipped with the speed device with ϕ, she can observe private news regarding v and react to it with a Poisson probability with intensity ϕ. We denote T_H as the arrival time of this Poisson news. Upon the arrival of the news, the HFT immediately submits market orders to "snipe" stale limit orders provided by the continuum of market makers.[14,15]

In addition, there is a continuum of liquidity traders who are exposed to a liquidity shock. The shock exogenously makes them submit buy or sell market orders with equal probability. We can think of them as noise traders, and trading against them conveys no information to market makers. Let T_L be the timing of the Poisson shock which arrives with intensity $\beta(\geq \gamma)$.

Finally, as in Haas and Zoican (2016) and Brolley and Cimon (2017), assume that trading information, including traders' identity, becomes public immediately after an order is executed, i.e., the market is perfectly transparent.

2.2 Market Structure

A continuous market imposes a random speed bump on incoming orders from the HFT. Specifically, an order submitted to the market at date t arrives at $t + \delta$, where δ is a random delay. Orders from liquidity traders and cancellation requests from market makers are executed promptly.[16] Thus, during $\tau \in (t, t + \delta)$, outstanding limit orders can be illusory

[13] In the benchmark model with $N = 1$, imposing a sunk cost on the speed investment does not change our result. In the extension with $N \geq 2$, we need a positive and convex cost to hamper the strategic complementarity and to derive an equilibrium.

[14] If we give an index $i \in [0, 1]$ to each market maker, the HFT submits marketable limit orders to obtain sniping profit $\sigma - s_i$ from each i. Since all the market makers quote a competitive homogeneous spread, the HFT's aggregate gain is $\int_0^1 (\sigma - s) di = \sigma - s$.

[15] We can show that the HFT does not intentionally delay the timing of the order submission: if she gets information at t, she immediately sends the order at t. Putting a time lag between obtaining the information and submitting the order can reduce a spread and increase sniping profit. However, without a commitment device, this cannot be an equilibrium since it is always optimal for the HFT at the information arrival time T_H to snipe immediately given the lag she announces at $t = 0$, i.e., there is a time inconsistency.

[16] For simplicity, we assume that there are no other sources for a latency, while the primitive parameters, such as β, can be seen as the potential latency that characterizes the speed of each type of trader.

for the HFT if liquidity traders trade against them or if market makers cancel due to public news.

For notational simplicity, we assume that δ follows an exponential distribution with a parameter b, and the expected length of a delay is denoted by $\lambda \equiv E[\delta] = b^{-1}$.[17]

Alternative Market Structures

As an extension, we analyze a situation with multiple HFTs, $N \geq 2$, in Section 3, in which each HFT serves not only as a sniper but also as a high-frequency market maker. This setting sheds light on a strategic property of an "arms race." Appendix A.1 considers a case where market makers can continuously update (cancel and resubmit) their limit orders. The coexistence of slow and fast markets is analyzed in Appendix A.2. This shows that a bump triggers a shift of adverse selection from slow markets with a bump to fast markets with no bumps, consistent with the empirical result by Chen et al. (2017).

2.3 Equilibrium

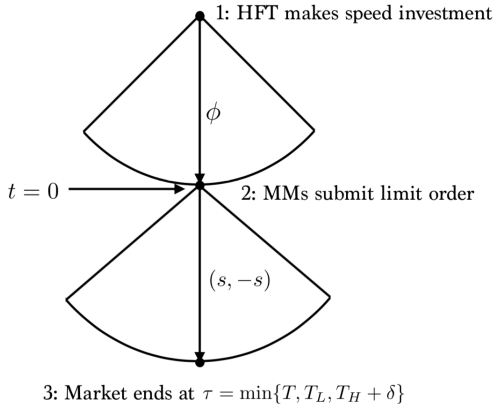
We conceptualize our model as a sequential game with two stages, as depicted in Figure I. In the first stage, the HFT decides the level of ϕ.[18] Given this, each market maker submits a competitive limit order, anticipating a confrontation with the informed HFT and liquidity traders. In the trading stage, the HFT looks for an opportunity to snipe.

The equilibrium concept is a subgame perfect equilibrium, and the HFT chooses the optimal level of speed ϕ in light of the optimal reaction of market makers. That is, the HFT knows the price impact of her *speed choice*, as the monopolist in Kyle (1985) knows the price impact of her trading behavior. In contrast to Kyle (1985), however, the trading stage in $t \in (0, T)$ is competitive, and the HFT behaves as if her trading strategy does not have a price impact. This is because she splits her orders and sends them to an infinitely large number of market makers given the outstanding limit orders. This follows the literature

[17] The randomness of δ does not significantly affect our result, while it makes the solution simpler. The case with a deterministic δ is available on request.

[18] Of course, the setting of the speed decision used in our model does not comprehensively reflect real-world conditions. As Dugast et al. (2014) suggest, some components of speed choice may occur simultaneously with market makers' behavior. However, we believe that the *ex-ante* strategic speed decision is still significant because HFTs would not invest in speed *ex-ante* if they did not exploit it in an *ex-post* trading game.

Figure I: Timeline



and captures the real-world behavior of HTFs, who send and cancel a massive number of small orders within a very short time frame.

2.3.1 Optimal Behavior of Market Makers

In a perfect competition, a limit order sent by a market maker yields zero expected profit, as in Glosten and Milgrom (1985). Without a loss of generality, let us consider how an ask price s is determined when $v = \sigma$.[19]

Given that a market order arrives at date t, it is possible that the taker is information or liquidity driven. As a result, the spread is set so that $s = E[v|\text{buy order at } t]$, where the expectation is over δ and the timing of the trade. The key effect of a bump is to reduce the probability of being picked off by the HFT or, put differently, to increase the probability that market makers observe public news to cancel their limit orders.

Suppose that a trade takes place at date t. If $t < \delta$, there is no fear of facing an information-driven HFT because of the speed bump. Put differently, the fastest possible arrival of the HFT occurs at δ. During this "safe interval," liquidity traders arrive before the public news with a density $\beta e^{-(\beta+\gamma)t}$. Otherwise, market makers can cancel their orders with density $\gamma e^{-(\beta+\gamma)t}$ at period t.

If a trade occurs at $t \geq \delta$, on the other hand, it bears an adverse selection cost: the HFT buys an asset only if the limit order is mispriced given the true information. The HFT gets to trade if she becomes informed at $t - \delta$, and there are no liquidity shocks or public news events during $(t - \delta, t)$. In this case, a market maker obtains $s - \sigma$. Market makers can also

[19] Results for the opposite case can be given by a symmetric argument.

trade with liquidity traders if there is a liquidity shock at t, and the HFT becomes informed after $t-\delta$. In this case, the trading profit is $s - E[v] = s$. Since δ is stochastic, the expected return for a market maker is

$$V = E_\delta \left[\int_0^\delta s\beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty (\beta s + \phi(s-\sigma)) e^{-\psi(t-\delta)} e^{-(\beta+\gamma)\delta} dt \right], \tag{1}$$

where the expectation relates to δ, and $\psi \equiv \phi + \beta + \gamma$. The first integral in (1) shows the trading profit in $t < \delta$, while the second describes the case with $t \geq \delta$.

This formulation is the result of the following probabilities: given δ,

$$\begin{aligned}
\Pr(\text{HFT arrives at } t) &= \phi e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t}, \\
\Pr(\text{Liq. traders arrive at } t) &= \beta e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t}, \\
\Pr(\text{cancellation at } t) &= \gamma e^{-\phi(t-\delta)} e^{-(\beta+\gamma)t},
\end{aligned}$$

which lead to the second term in (1). It is then possible to get the equilibrium spread from the break-even condition:

Proposition 1. *The equilibrium (half) spread is given by*

$$s = \frac{\phi E_\delta[e^{-(\beta+\gamma)\delta}]}{(\phi+\beta)E_\delta[e^{-(\beta+\gamma)\delta}] + \frac{\beta\psi}{\beta+\gamma}(1 - E_\delta[e^{-(\beta+\gamma)\delta}])} = \frac{\frac{\phi}{1+\lambda\psi}}{\frac{\phi}{1+\lambda\psi} + \beta}\sigma. \tag{2}$$

A few remarks on s are in order. First, a direct effect of the speed bump appears in the form of the discount on the arrival rate of the HFT, which is given by $(1+\lambda\psi)^{-1}$. This term mitigates adverse selection risk by generating a safe interval.

If ϕ is fixed, a lower λ induces a higher spread since the expected delay becomes shorter. Also, an infinitely small expected delay ($\lambda \to 0$) makes s converge to the traditional equilibrium spread of Glosten and Milgrom (1985). Therefore, as Budish et al. (2015) and Baldauf and Mollner (2017) point out, the direct effect of a bump mitigates the adverse selection cost for market makers.

This argument is built on the assumption that ϕ is fixed, i.e., the HFT's speed decision is not influenced by the bump. When the speed choice by the HFT is considered, the speed bump affects s via the fluctuation of the optimal speed as well. The existing models argue that the incentive to be faster diminishes as the bump gets longer, i.e., a higher λ reduces

s not only by the direct effect but also by making ϕ lower, while our model proposes the opposite effect.

The following properties of the spread are useful to understand the mechanism. First, note that the price impact of the speed is positive:

$$\frac{\partial s}{\partial \phi} > 0.$$

We call this derivative the "sensitivity" of a spread (price) to a speed-up by the HFT. It turns out that this represents the slope of the endogenous cost of being faster. At the same time, we have the following:

Lemma 1. *The sensitivity of the price to the speed is decreasing in λ, i.e.,*

$$\frac{\partial}{\partial \lambda}\left(\frac{\partial s}{\partial \phi}\right) < 0.$$

Therefore, the longer the expected delay, the less sensitive the spread becomes. A market with a higher λ is protected by a longer (expected) safe interval, and market makers behave as if the share (arrival rate) of the HFT is small. Hence, market makers care *less* about the speed investment by the HFT, making their pricing behavior less sensitive to ϕ.

2.3.2 Profit of the HFT

When the HFT becomes informed and submits market orders at t, they will be executed at $t+\delta$ if (i) there is no liquidity shock during $(t, t+\delta)$ and (ii) no public news arrives in the same interval. This happens with

$$\pi_t(\phi, \delta) \equiv \Pr(T_H = t, \min\{T, T_L\} > t+\delta) = \phi e^{-\psi t} e^{-(\beta+\gamma)\delta}. \tag{3}$$

Thus, if the random delay is δ, the profit from sniping at t is $\pi_t(\phi, \delta)(\sigma - s)$.

The first coefficient in (3) represents the probability that the HFT obtains the information at t. The sniping probability involves an additional exponential coefficient, $e^{-(\beta+\gamma)\delta}$, which shows the disadvantage of the HFT that stems from a speed bump, i.e., front-running by liquidity traders or cancellation by market makers due to the δ-delay. Therefore, a longer delay directly reduces the expected profit of the HFT in the second stage.

The objective function of the HFT in the first stage takes a simple form:

$$W(\phi) \equiv E_\delta \left[\int_0^\infty \pi_t(\phi, \delta)(\sigma - s) dt \right] = \frac{\phi}{\psi} \frac{1}{1 + (\beta + \gamma)\lambda} (\sigma - s). \tag{4}$$

Note that the HFT always submits unit orders since she exits the market once her orders are executed, i.e., she is a short-term investor.[20]

2.3.3 Optimal Speed

We move on to the speed choice by the HFT in the first stage. To obtain an interior solution, we make the expected length of the delay relatively short:

$$\lambda < \frac{1}{\sqrt{\beta(\beta + \gamma)}}. \tag{5}$$

The intuition behind this condition will be provided after offering our main propositions.

Under (5), the optimization problem of the HFT is

$$\begin{aligned} \max_\phi W(\phi) &\equiv E_\delta \left[\int_0^\infty \pi_t(\phi, \delta)(\sigma - s) dt \right], \\ s.t.\ s &= \frac{\frac{\phi}{1+\lambda\psi}}{\frac{\phi}{1+\lambda\psi} + \beta} \sigma. \end{aligned} \tag{6}$$

This indicates that the HFT decides ϕ knowing the price impact of her speed decision, i.e., she is strategic. In this case, being faster pushes up the price charged by market makers and saps her sniping profit. For this reason, we can think of the equilibrium spread as an *endogenous* cost of speed. Importantly, Lemma 1 suggests that the slope of this endogenous cost is affected by λ, and, in turn, affects the marginal cost of being faster for the HFT.

To analyze how λ alters the optimal decision, consider a marginal gain of being faster:

$$\begin{aligned} \frac{dW}{d\phi} &= E_\delta \left[\int_0^\infty \left\{ (\sigma - s(\phi)) \frac{d\pi_t(\phi, \delta)}{d\phi} + \pi_t(\phi, \delta) \frac{d}{d\phi} (\sigma - s(\phi)) \right\} dt \right]. \\ &= (\sigma - s(\phi)) E_\delta \left[\int_0^\infty \frac{d\pi_t(\phi, \delta)}{d\phi} dt \right] (1 - \varepsilon(\phi)), \end{aligned} \tag{7}$$

[20] See Appendix A.1 for a more general setting with continuous updating by market makers and time-dependent s_t.

where

$$\varepsilon \equiv -\frac{d\log(\sigma - s(\phi))}{d\log E_\delta\left[\int_0^\infty \pi_t(\phi,\delta)dt\right]} > 0.$$

ε is the sensitivity of the sniping profit $(\sigma - s)$ to a change in the expected sniping probability of the HFT $(E_\delta\left[\int_0^\infty \pi_t(\phi,\delta)dt\right])$. We call this the elasticity, and Appendix B.1 provides an explicit formula.

Note that obtaining a higher ϕ affects W through two competing channels and exhibits a price-liquidity tradeoff: it increases the sniping probability (the first term in [7]), while reducing the sniping profit via the adverse price movement (the second term in [7]).

When the equilibrium spread is more sensitive to ϕ than the sniping probability (i.e., $\varepsilon > 1$), being faster harms the profit of the HFT, and incentive to increase ϕ dwindles. On the other hand, if the HFT knows that the price impact of her speed choice is small, it is more likely that an improvement in sniping probability $(\frac{d\pi}{d\phi})$ dominates a decline in profit due to the wider spread $(\frac{d(\sigma-s)}{d\phi})$, luring her to be faster. In other words, for the strategic HFT, the marginal cost captured by the sensitivity of the spread matters considerably.

The following results guarantees the concavity of the problem (see Appendix B.1 for proofs):

Lemma 2. *The elasticity is increasing in the speed:* $d\varepsilon(\phi)/d\phi > 0$.

When the HFT is fast, market makers estimate that the economy is inhabited by a relatively large measure of the HFT in terms of the arrival rate. Therefore, a marginal increase in ϕ reduces market makers' expected profit, and they charge a wide spread to compensate for the expected loss. That is, as the HFT becomes faster, market makers grow more concerned about facing the HFT, and their pricing behavior is more sensitive to changes in speed. Thus, as ϕ increases, it is more likely that the steeper endogenous marginal cost of being faster will outweigh the higher marginal benefit from a higher π, making the objective function (6) concave.[21]

As a result, the optimal speed ϕ^* is derived by solving for the FOC.

[21] The condition in (5) is required to make Lemma 2 hold. If λ is sufficiently large, market makers become too insensitive to make $\frac{d\varepsilon}{d\phi} > 0$. Thus, the HFT can be infinitely fast, and we need an exogenous cost to make the problem well defined.

Proposition 2. *(i) The optimal speed is given by*

$$\phi^* = \frac{\sqrt{\beta+\gamma}(1+\lambda(\beta+\gamma))}{1-\lambda\sqrt{\beta(\beta+\gamma)}}. \tag{8}$$

(ii) ϕ^* is increasing in λ.

Proof. See Appendix B.1. □

In contrast to the traditional models, Proposition 2 demonstrates that, if the speed choice is strategic, a speed bump increases the equilibrium speed of the HFT. This modification of the speed decision is natural given that several high-frequency financial institutions control significant shares, as discussed in Section 1 and shown in Table I.

When the HFT knows how her speed investment affects the pricing behavior of market makers, an equilibrium spread generates an *endogenous* cost of being faster. This not only guarantees a bounded solution even without an exogenous cost of speed (Lemma 2), but also overturns the traditional result regarding speed bumps (Proposition 2).

The key mechanism is the negative impact of a speed bump on the sensitivity of the price in Lemma 1. Since a bump intends to slow down the HFT and protect market makers, the spread becomes insensitive to a change in the speed. That is, an intentional delay endogenously reduces the marginal cost of being faster. Hence, a speed bump does not prevent a speed race but promotes it, as Proposition 2 attests.

This surprising finding highlights the main difference of our results from the literature, such as Budish et al. (2015), Haas and Zoican (2016), and Baldauf and Mollner (2017). As in their models, if we assume that the HFT does not care about the effect of her speed choice on the spread, the second effect in (7) disappears. In this case, some exogenous costs of speed are required to make ϕ^* bounded, and the effect of λ on ϕ^* becomes reversed. We compare our model to the traditional ones in more detail in Subsection 3.4.

2.4 Effect on Adverse Selection

We are interested in how a speed bump affects adverse selection cost for market makers. It is straightforward that there are two competing effects. First, as the literature suggests, a speed bump reduces adverse selection cost because it dampens the probability for market makers of confronting the HFT. However, our strategic model adds an opposing channel:

a speed bump promotes speed investment by the HFT since it endogenously reduces the marginal cost of being faster. In the following, we take the equilibrium half spread s as a measure of adverse selection and investigate its equilibrium behavior.

Proposition 3. *The equilibrium spread is independent of the expected delay, i.e.,* $ds/d\lambda = 0$.

Proof. See Appendix B.2. □

This result shows that a speed bump cannot mitigate (or worsen) adverse selection for market makers.

With ϕ fixed, a speed bump reduces the profit of the HFT since the first-δ periods become safe intervals for market makers and the HFT cannot snipe. To compensate for this disadvantage, the strategic HFT gets faster. Anticipating the price impact of her speed investment, she chooses the level of ϕ that eliminates the cost from the speed bump, thereby muting its effect. As a result, the two competing consequences of a speed bump cancel each other out. Put differently, the speed-up due to a longer delay is an indirect effect of a change in the price sensitivity, $\frac{ds}{d\phi}$. Since the reduction of a spread by λ is a direct effect, the speed-up cannot predominate.

In the following sections, however, we show that this finding regarding adverse selection is specific to the benchmark setting and significantly changes if we consider more general market structures.

3 Multiple HFTs and High-Frequency Market Making

In the real world, HFTs serve not only as takers (snipers) but also as liquidity providers. We modify the benchmark model to capture this fact.

Assume that there are two HFTs ($i = 1, 2$), both of whom provide limit orders at $t = 0$ as market makers at a competitive price. At a random date, $T_i \sim \exp(\phi_i)$, HFT i obtains private news about v. When the news arrives, it is optimal for HFT i to immediately send market orders to snipe the stale limit orders of her opponent (HFT j) and to simultaneously cancel her limit orders.

The behavior of liquidity traders is the same as in the benchmark, but we ignore public news at $T \sim \exp(\gamma)$, since it only adds complexity. The other structures of the game stay the same as in the benchmark. Note that the results with $N = 2$ can be easily extended to

$N \geq 3$ with an additional parameter N that measures the (inverse of) market power. For technical reasons, assume that $\beta \geq 1$ and focus on the symmetric equilibrium.

3.1 Optimal Behavior of HFTs

Consider HFT j as a high-frequency market maker (HFMM). Her behavior is the same as that of ordinary market makers in the benchmark model except that she can cancel her limit orders at $T_j \sim \exp(\phi_j)$. Thus, the break-even condition provides the following equilibrium spread:

$$s_j = \frac{\phi_i}{\phi_i + \beta(1 + \lambda(\phi_i + \phi_j + \beta))}\sigma.$$

This spread has the same structure as s in (2): it reflects an expected value of v conditional on the trade. Note that the symmetric equilibrium makes the spreads set by both HFTs the same; $s = s_1 = s_2$.

We turn to the optimal speed decision of HFT i as a sniper. Since the competition drives her total profit from market making to zero, her gains come only from the sniping profit. Thus, the optimization problem is analogous to (4):

$$\max_{\phi_i} W_i(\phi) = \frac{1}{1 + \lambda(\beta + \phi_j)} \frac{\phi_i}{\phi_i + \phi_j + \beta}(\sigma - s_j),$$
$$\text{s.t., } s_j = \frac{\phi_i}{\phi_i + \beta(1 + \lambda(\phi_i + \phi_j + \beta))}.$$

Since this is exactly the same as the benchmark case if we substitute ϕ_j for γ, the best response function of HFT i is a modified (8):

$$BR_i(\phi_j) = \frac{\sqrt{\beta + \phi_j}[1 + \lambda(\beta + \phi_j)]}{1 - \lambda\sqrt{\beta(\beta + \phi_j)}}, \tag{9}$$

as long as $1 > \lambda\sqrt{\beta(\beta + \phi_j)}$. Otherwise, $\phi_i = \infty$ is the best response. In this section, we focus on bounded responses, while Subsection 3.2 analyzes all possible symmetric equilibria.[22] The following property of the best response function helps explain the mechanism:

Proposition 4. *The best response function exhibits strategic complementarity, i.e.,* $\frac{dBR_i(\phi_j)}{d\phi_j} > 0$.

[22] Technically, we can avoid the unbounded equilibrium if we introduce a positive exogenous sunk cost for the speed.

The intuition behind this proposition should be clear if we analyze the marginal gain of being faster for HFT i:

$$w_i \equiv \frac{\partial W_i}{\partial \phi_i} = (\sigma - s_j)\frac{\partial \pi_i}{\partial \phi_i} + \pi_i \frac{\partial(\sigma - s_j)}{\partial \phi_i}, \tag{10}$$

where

$$\pi_i \equiv \frac{\phi_i}{[1 + \lambda(\beta + \phi_j)]\psi}$$

represents her sniping probability.

The first term is the marginal improvement in the sniping probability, and the second stands for a decline in the profit. These terms can be seen as the marginal benefit and cost of being faster. The structure of the marginal gain of increasing ϕ_i is the same as in the benchmark case, while it depends on the speed of the competitor.

Furthermore, we need a cross derivative to obtain the reaction of BR_i.

$$\frac{\partial w_i}{\partial \phi_j} = \left[(\sigma - s_j)\frac{\partial^2 \pi_i}{\partial \phi_j \partial \phi_i} + \frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s_j)}{\partial \phi_j}\right] + \overbrace{\left[\frac{\partial \pi_i}{\partial \phi_j}\frac{\partial(\sigma - s_j)}{\partial \phi_i} + \pi_i \frac{\partial^2(\sigma - s_j)}{\partial \phi_j \partial \phi_i}\right]}^{\oplus}. \tag{11}$$

When the opponent (HFT j) increases her speed, it affects both the marginal benefit and cost of being faster for HFT i. The first component of (11) is a change in the marginal benefit that stems from a marginal improvement in π_i. A faster opponent (i) increases or decreases the marginal improvement in the sniping probability and (ii) raises the sniping profit, making it more worthwhile to have a higher π_i. These are the first and second terms in the first brackets in (11). At the same time, a faster opponent reduces the (endogenous) marginal cost of being faster for HFT i. Intuitively, (iii) since a faster opponent makes HFT i less likely to snipe, she does not need to care much about the adverse price movement of being faster. Moreover, (iv) a faster opponent becomes more insensitive to HFT i's speed-up due to the same logic that states that a higher λ makes s less sensitive to ϕ in the benchmark case. As a result, the second term is positive, as is the total effect of ϕ_j on w_i, i.e., a faster opponent renders speeding up more profitable for HFT i.

Moreover, "tit for tat" due to complementarity can be strong enough when the opponent is sufficiently fast.

Lemma 3. *There is a unique $\phi_j = \phi_0$ such that*

$$\frac{d^2 BR_i(\phi_j)}{d\phi_j^2} > 0 \Leftrightarrow \phi_j > \phi_0. \tag{12}$$

Proof. See Appendix B.3. □

If ϕ_j is sufficiently high, the negative effect of ϕ_i on the expected profit becomes minimal. This is because a very fast opponent makes it extremely difficult for HFT i to snipe. Thus, she barely cares about the negative impact of ϕ_i on the price. In addition, a fast opponent as a market maker tends to be highly insensitive to a change in ϕ_i because she estimates that being sniped by HFT i is not likely to happen. Both of these effects strongly prompt an incentive of HFT i to be faster, making the best response function convex.

3.2 Equilibrium Speed

To see if (9) has a symmetric solution, we first observe that $BR_i(0) > 0$, i.e., facing a zero-speed opponent, HFT i still maintains a positive speed. This is because $\phi_i > 0$ yields a positive profit, while $\phi_i = 0$ keeps it at zero. Together with (12), this implies that multiple equilibria can arise. We focus on the symmetric equilibria.

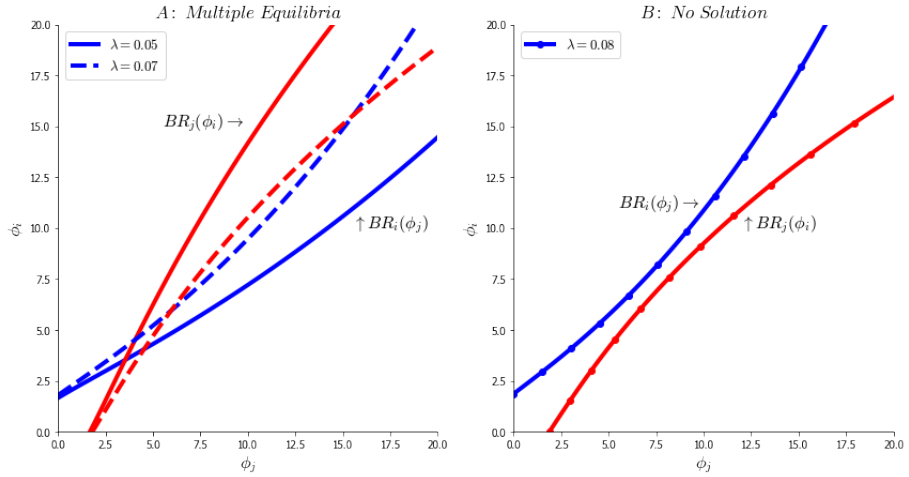
Proposition 5. *(i) There is a unique $\lambda = \lambda_0$. If $\lambda > \lambda_0$, no bounded solution exists. If $\lambda \leq \lambda_0$, there are two bounded solutions to $BR(\phi) = \phi$. The low-ϕ solution is stable and the high-ϕ solution is unstable.*
(ii) In the stable equilibrium, $\phi^* \equiv \phi_i = \phi_j$ is increasing in λ.

Proof. See Appendix B.3. □

Note that a higher λ has the same implication as a higher ϕ_j for the sensitivity of s to ϕ_i and for the improvement in the sniping probability π_i. Thus, due to the same logic as in Lemma 3, a sufficiently high λ makes the complementarity strong enough to eliminate a bounded solution, i.e., $\phi = \infty$ is always optimal. On the other hand, when λ is small, we obtain bounded symmetric equilibria.

Following the convention (Hendershott and Mendelson, 2000; Zhu, 2014), we use stability as an equilibrium selection criterion. The unstable equilibrium is not robust to a small

Figure II: Best Response Functions



perturbation in a parameter value, whereas the stable one does not diverge even if a parameter changes slightly. Thus, our focus is on the small-ϕ solution.

Figure II provides the best response functions for different values of λ. In multiple equilibria, a small-ϕ solution is stable, while a higher ϕ makes "tit for tat" strong and the solution can explode.

As Figure II indicates, a longer delay has the same effect on the optimal speed as in the benchmark, i.e., it increases the marginal benefit of being faster. Thus, the best response function shifts upward, leading to a higher speed in the stable equilibrium.

3.3 Effect of Speed Bumps on the Spread

The effect of a bump on the spread and adverse selection can be derived analytically. Due to "fast market making," adverse selection risk is mitigated by fast market makers, and it helps λ protect them. However, in the symmetric equilibrium, the increase in market makers' speed occurs identically for snipers. Then, the above-mentioned effect of fast market making is offset by the increase in the snipers' speed.

Since the strategic complementarity is sufficiently strong, this arms race outweighs the direct protection of the speed bump, expanding the spread.

Proposition 6. *A longer speed bump widens the spread;* $\frac{ds}{d\lambda} > 0$.

Proof. See Appendix B.4. □

The introduction of a speed bump or a longer expected delay can backfire not only in terms of speed but also of adverse selection. In the benchmark model, we have $\frac{ds}{d\lambda} = 0$ because the speed-up by the single HFT is an indirect consequence of the speed bump and cannot offset the direct protection of market makers.

By contrast, multiple HFTs generate a positive externality through strategic complementarity (Proposition 4). In this situation, an increase in λ indirectly affects the best response functions of both HFTs, shifting them upward, as shown in Figure II. This triggers an arms race with positive externality, amplifying the first indirect effect. As a result, the speed-up in the symmetric equilibrium dominates the direct protection of market makers, leading to more severe adverse selection.

3.4 Comparison with Traditional Models with an Exogenous Cost

The results in the previous subsection run counter to traditional models with an exogenous cost of speeding up. To illustrate this, consider a model with non-strategic HFTs; Instead of an endogenous cost, we introduce an exogenous sunk cost of being faster denoted by $C(\phi_i) = \frac{c}{2}\phi_i^2$, as in Foucault et al. (2016). To make the comparison clearer, we call our model in Subsection 3.1 the *strategic model*.

If the strategic motive is absent, the FOC in (10) and cross derivative in (11) are modified as follows:

$$\begin{aligned} w_i \equiv \frac{\partial W_i}{\partial \phi_i} &= (\sigma - s_j)\frac{\partial \pi_i}{\partial \phi_i} - c\phi_i, \\ \frac{\partial w_i}{\partial \phi_j} &= (\sigma - s_j)\frac{\partial^2 \pi_i}{\partial \phi_j \partial \phi_i} + \frac{\partial \pi_i}{\partial \phi_i}\frac{\partial(\sigma - s_j)}{\partial \phi_j}. \end{aligned} \tag{13}$$

The second term of (10) that represents an endogenous marginal cost is replaced by the exogenous marginal cost, $c\phi_i$, and the effect of the opponent's speed via the strategic motive, denoted by the second set of brackets in (11), disappears from (13).

We focus on a symmetric equilibrium and obtain the following results.

Proposition 7. *(i) Around the symmetric equilibrium, the best response function exhibits strategic substitution;* $\frac{dBR_i(\phi_j)}{d\phi_j} < 0$.
(ii) The equilibrium speed and spread are decreasing in λ.

Proof. See Appendix B.5. □

The third column (panels A3, B3, and C3) of Figure III shows these results. As we have established, if HFTs are strategic, HFT j's speed-up improves w_i through the second brackets in (11), while this effect is absent in the traditional models.

To understand this intuition, note that the profit function of an HF sniper is roughly given by

$$V_i = \max_{\phi_i} \pi_i(\phi_i, \phi_j, \lambda)(\sigma - s) - C(\phi_i),$$

where π_i is the sniping probability of HFT i. In this formulation, the interaction between HFT i and j occurs only through π_i, i.e., probability of successful sniping.

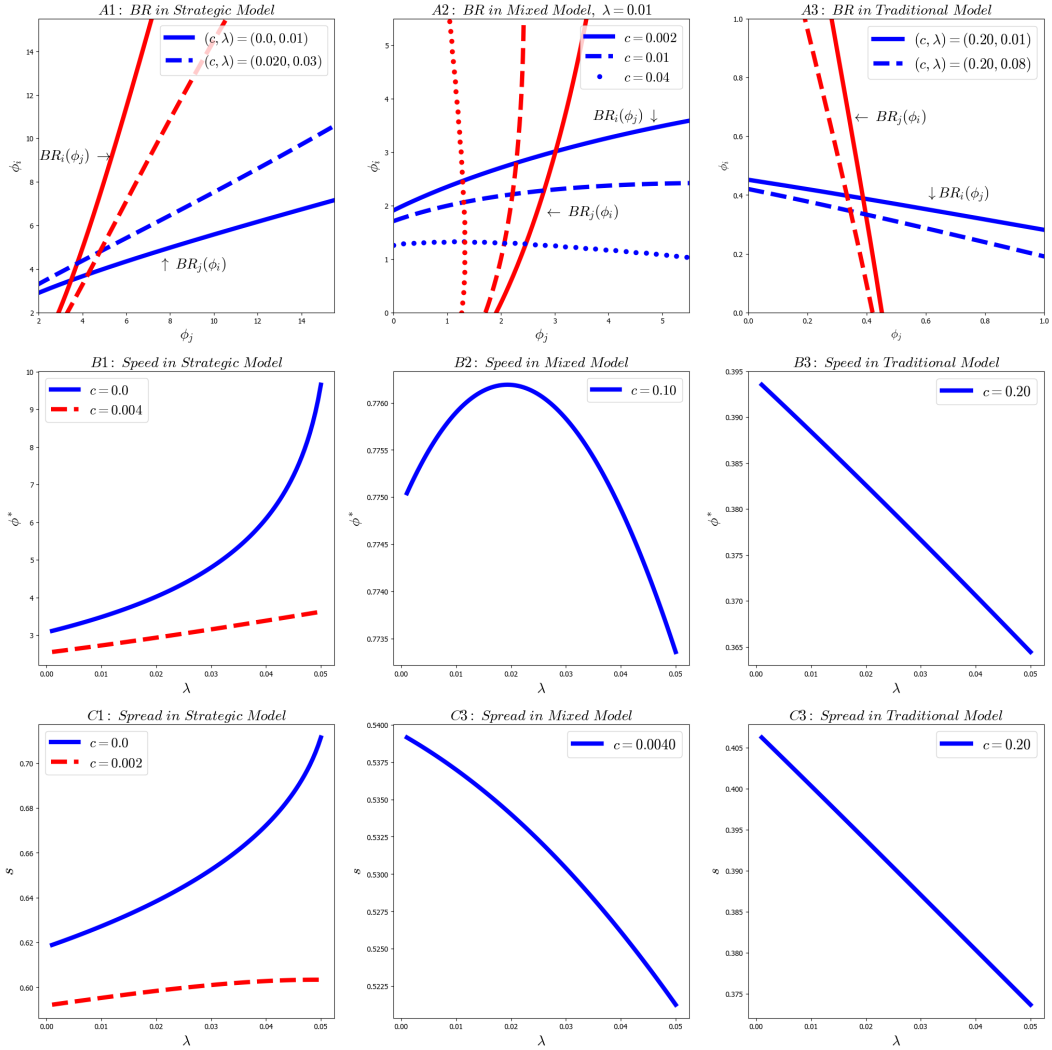
A marginally faster opponent in a traditional model affects HFT-i's decision by making sniping more difficult, i.e., ϕ_j changes V_i only by reducing π_i. Since the exogenous cost is sunk, HFT i must pay it anyway. By contrast, her speed investment pays out only if her sniping attempt is fulfilled. Therefore, if HFT i thinks she is less likely to snipe due to a faster opponent (a lower π_i), the exogenous cost becomes more salient (C/π_i increases), hampering her speed investment. This logic applies to a speed bump as well: a bump makes sniping less likely, leaving HFTs reluctant to pay the sunk cost. This is why traditional models conclude that a speed bump is effective to slow HFTs down and mitigate adverse selection, which is replicated by Proposition 7.

Once HFTs become aware of the price impact of their speed choice, they perceive the price as $s = s(\phi_i, \phi_j, \lambda)$, generating a new channel through which ϕ_j affects V_i. In the previous subsection, we showed that this modification overturns the result of the benchmark. This is because the pricing behavior of market makers becomes insensitive to an increase in ϕ_i (the endogenous marginal cost declines) when the opponent as a market maker becomes faster or a bump is expected to be longer. In our strategic model, this endogenous cost channel works against the traditional exogenous cost.

3.5 Strategic Complementarity or Substitution

As Proposition 7 and the following discussion suggest, an exogenous sunk cost tends to make an arms race exhibit strategic substitution, while an endogenous cost in our model promotes complementarity (Proposition 4). These results imply that introducing an exogenous sunk cost in our strategic model allows us to explain both complementarity and substitution in an arms race, as well as the positive and negative reaction of the spread to a

Figure III: Effect of Exogenous Cost



Note: The first row plots the best response functions with different $(c\lambda)$, the second and third rows plot the optimal speed and the equilibrium spread, respectively, against the expected length of a speed bump with different values of c. In each row, $c = 0$ corresponds to the strategic model with no exogenous costs. The figures in the third column replicate the traditional results.

bump, by changing the parameter values.

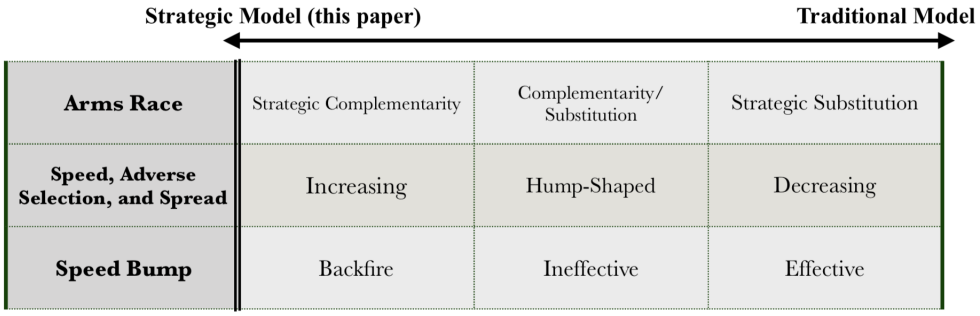
Consider a model that is the same as in Subsection 3.1, except that HFT i solves

$$\max_{\phi_i} W_i(\phi) = \frac{1}{1+\lambda(\beta+\phi_j)} \frac{\phi_i}{\psi}(\sigma - s_j) - C(\phi_i),$$
$$\text{s.t., } s_j = \frac{\phi_i}{\phi_i + \beta(1+\lambda(\phi_i+\phi_j+\beta))},$$

where $C(\phi) = \frac{c}{2}\phi^2$ is the exogenous sunk cost of speed.

The best response functions, equilibrium speed, and spread are provided in Figure III.

Figure IV: Summary



The strategic model in Subsection 3.1 corresponds to $c = 0$. As expected, a small exogenous cost (panels A1 and A2 with $c = 0.002$ and 0.01) offers strategic complementarity and an increasing ϕ^* against λ (panel B1). This leads to the increasing s against λ (panel C1).

As long as $c > 0$, ϕ^* becomes flatter as λ rises and can be hump shaped (panels B1 and B2). This is because a higher λ makes C more salient. That is, a longer delay makes sniping more difficult, and HFTs become more reluctant to pay the sunk cost. In this situation, an arms race exhibits both complementarity and substitution (Panel A2 with $c = 0.04$), i.e., the BR curves become hump shaped. When λ is small enough, the model is close to our strategic model with complementarity, while a large λ makes it similar to the traditional model with substitution. As a result, the direct protection of market makers by the bump dominates the speed increase in the high-λ region, and s slopes downward, as shown in Panels C1 with $c = 0.002$ and C2.

Finally, if the exogenous cost is sufficiently large, the economy reverts to the traditional world with no strategic speed choice (panels A3, B3, and C3). The competition exhibits global substitution, and the equilibrium speed and the spread are downward sloping against λ, i.e., a speed bump mitigates adverse selection.

Overall, we can think of the strategic model with complementarity and the traditional models with substitution as two extremes. By changing the significance of the exogenous cost, c, we can explain intermediate cases. Moreover, we have established that a speed bump, λ, also works to adjust the relative significance of the exogenous cost because a longer delay makes it more salient. This premise is summarized in Figure IV and has some empirical implications.

4 Empirical Implications and Policy Discussion

Our model provides some testable implications for the strategic nature of an arms race among HFTs and for the effect of speed bumps on a spread and adverse selection for market makers.

As Subsection 3.5 demonstrates, an arms race exhibits strategic complementarity or substitution depending on the exogenous cost of speed investment and the expected length of a speed bump. As summarized in Figure IV, a long (resp. short) expected speed bump or a significant (resp. slight) exogenous cost of speed creates strategic substitution (resp. complementarity), and the introduction of a bump and a marginally longer delay would effectively reduce (resp. widen) a spread.

In the real world, the speed cost for HFTs involves two factors. The first is a large sunk cost to develop a high-speed communication technology or investigate the optimal location of an information servers to exploit an arbitrage between segmented markets. For example, it is well known that Spread Networks LLC invested about $300 million to *re*construct a fiber-optic network between the Chicago and New York exchanges. The second cost can be a relatively small subscription fee to access these technologies developed by communication service companies or to colocate an information server to an exchange platform.[23]

Depending on the condition of the financial markets, our model suggests different implications. If HFTs take the first investment approach, the traditional model fits better: an arms race involves strategic substitution and a bump is effective. In contrast, if the exogenous cost is relatively small compared to the total profit, our strategic model is more appropriate, suggesting the complementarity and detrimental effect of a bump. This comparison can also be applied to the market power of a high-frequency financial institution because the endogenous cost stems from the strategic motives of HFTs. In other words, when N is large, the adverse price impact of increasing ϕ_i diminishes, leaving the model close to the traditional one.

Also, we can vary the length of a speed bump or the distribution of a random delay to test the implications of λ on a spread. If a bump is expected to be long, a marginally longer delay can be effective, slowing HFTs down due to the substitution in the traditional model. On the other hand, if a platform tries to avoid a delay cost by keeping the length

[23] For example, monthly payments to plug into NSYSE and NASDAQ amount to $10,000∼$22,000.

of a bump minimal, the introduction of a bump can aggravate adverse selection since HFTs become faster. Although we can compare the length of bumps in each platforms (i.e., bumps in the ANEO and TMX are longer than those in the IEX or Chicago Stock Exchange), our model shows that the *level* of λ also matters. Deriving the critical (c, λ) involves numerical calculation in our model, and estimating it requires further data on the cost of the speed technologies, speed levels, profit, and the market power of high-frequency financial institutions.

4.1 Policy Implications

Recently, exchange platforms have experienced declines in their revenue from transaction fees. Instead, they charge an increasingly high price to the fast access to their data, such as direct data feed and colocation of data servers. Namely, one of the suppliers of the speed technology, which we did not specify in the model, can be the exchange platforms.

The SEC is concerned about this skyrocketing price of fast data feed and issued a proposal to block the exchanges to raise the fee in March 2018. Another HFT-related policy that SEC adopted is the approval of the IEX with a speed bump as a National Securities Exchange (NSE) in 2017. Due to the Reg. NMS and Order Protection Rule, this approval effectively makes all orders being affected by the speed bump. In a nutshell, the SEC tries to curb the price of speed technologies supplied by the exchange platforms, while they prompt the introduction of a speed bump.

Importantly, our model suggests that the introduction of a delay does not necessarily conflict with a provision of the expensive speed technologies by an exchange platform. That is, a speed bump can increase HFTs' demand for fast information access, allowing an exchange platform to charge a higher price. Thus, aforementioned policies adopted by the SEC can be self negating.

In this situation, we can analyze whether "the market will fix the market," as Budish et al. (2018) put forth. They argue that a platform does not have an incentive to introduce FBA, as long as competing platforms can copy the innovation with a low cost. In our model, a bump does not always mitigate adverse selection, while an exchange platform may have an incentive to introduce a bump to obtain the higher demand for the speed technology. More detailed analyses can be our future research topics, but aforementioned mechanism can provide another explanation for why "the market cannot fix the market."

5 Conclusion

A speed bump, which seeks to mitigate adverse selection for market makers, can backfire. When HFTs strategically choose their speed level by considering the impact of their speed decision on market behavior, a bid-ask spread charged by market makers works not only as a trading cost but also as an endogenous cost of speeding up, since it widens as HFTs get faster. This endogenous cost tends to be insensitive to an increase in speed by HFTs when a speed bump is expected to be longer. This is because market makers behave as if the share of HFTs is small under a longer expected speed bump and do not care much about a speed increase. Then, if a bump is introduced or the length of a delay becomes longer, the marginal cost of speed diminishes, leading to a higher equilibrium speed of HFTs.

We also consider an arms race among multiple HFTs who serve both as snipers and as market makers. When the significance of the exogenous speed cost is not high compared to the endogenous cost, or when the expected length of a delay is relatively short, a speed competition can show strategic complementarity. In this situation, a longer speed bump triggers a positive externality among HFTs, leading to a very fast equilibrium speed. As a result, the increase in HFTs' trading speed dominates a direct reduction of an arrival rate of HFTs by a speed bump. That is to say, a longer bump exacerbates adverse selection and widens the equilibrium spread. The opposite holds in the case of substitution.

Thus, our model, which incorporates strategic speed choice and the endogenous cost (i.e., the equilibrium spread), generates results that are opposite to the traditional models, in which an arms race exhibits strategic substitution, the speed and adverse selection decrease with the length of the delay, and a bump is effective. By adding a traditional exogenous cost to our strategic model, we can derive a rich description of the characteristics of an arms race among HFTs and can explain the somewhat ambiguous effects of speed bumps on spreads and adverse selection.

References

Aït-Sahalia, Yacine and Mehmet Saglam, "High frequency traders: Taking advantage of speed," Technical Report, National Bureau of Economic Research 2013.

Aldrich, Eric M and Daniel Friedman, "Order protection through delayed messaging," Technical Report, WZB Discussion Paper 2018.

Baldauf, Markus and Joshua Mollner, "High-frequency trading and market performance," *SSRN Electronic Journal*, 2017.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, "Equilibrium fast trading," *Journal of Financial Economics*, 2015, *116* (2), 292–313.

Boehmer, Ekkehart, Kingsley Fong, and Julie Wu, "International evidence on algorithmic trading," *SSRN Electronic Journal*, 2015.

Bongaerts, Dion and Mark Van Achter, "High-frequency trading and market stability," *SSRN Electronic Journal*, 2016.

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan, "Trading fast and slow: Colocation and liquidity," *The Review of Financial Studies*, 2015, *28* (12), 3407–3443.

Brolley, Michael and David A Cimon, "Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays," *SSRN Electronic Journal*, 2017.

Budish, Eric, Peter Cramton, and John Shim, "The high-frequency trading arms race: Frequent batch auctions as a market design response," *The Quarterly Journal of Economics*, 2015, *130* (4), 1547–1621.

_ , Robin Lee, and John Shim, "Will the Market Fix the Market? A Theory of Stock Exchange Competition and Innovation," *Manuscript in Preparation*, 2018.

Chen, Haoming, Sean Foley, Michael Goldstein, and Thomas Ruf, "The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets," *SSRN Electronic Journal*, 2017.

Conrad, Jennifer, Sunil Wahal, and Jin Xiang, "High-frequency quoting, trading, and the efficiency of prices," *Journal of Financial Economics*, 2015, *116* (2), 271–291.

Delaney, Laura, "Investment in high-frequency trading technology: A real options approach," *European Journal of Operational Research*, 2018, *270* (1), 375–385.

Du, Songzi and Haoxiang Zhu, "What is the optimal trading frequency in financial markets?," *The Review of Economic Studies*, 2017, *84* (4), 1606–1651.

Dugast, Jérôme, Thierry Foucault et al., "False news, informational efficiency, and price reversals," *Banque de France, Working Paper*, 2014, (513).

Foucault, Thierry, Ailsa Roell, and Patrik Sandas, "Market making with costly monitoring: An analysis of the SOES controversy," *The Review of Financial Studies*, 2003, *16* (2), 345–384.

_ , Ohad Kadan, and Eugene Kandel, "Liquidity cycles and make/take fees in electronic markets," *The Journal of Finance*, 2013, *68* (1), 299–341.

_ , Roman Kozhan, and Wing Wah Tham, "Toxic arbitrage," *The Review of Financial Studies*, 2016, *30* (4), 1053–1094.

Frino, Alex, Vito Mollica, and Robert I Webb, "The impact of co-location of securities exchanges' and traders' computer servers on market liquidity," *Journal of Futures Markets*, 2014, *34* (1), 20–33.

Glosten, Lawrence R and Paul R Milgrom, "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of financial economics*, 1985, *14* (1), 71–100.

Haas, Marlene and Marius Zoican, "Beyond the frequency wall: Speed and liquidity on batch auction markets," *SSRN Electronic Journal*, 2016.

Hasbrouck, Joel and Gideon Saar, "Low-latency trading," *Journal of Financial Markets*, 2013, *16* (4), 646 – 679.

Hendershott, Terrence and Haim Mendelson, "Crossing networks and dealer markets: Competition and performance," *The Journal of Finance*, 2000, *55* (5), 2071–2115.

__ **and Pamela C Moulton**, "Automation, speed, and stock market quality: The NYSE's hybrid," *Journal of Financial Markets*, 2011, *14* (4), 568–604.

Hoffmann, Peter, "A dynamic limit order market with fast and slow traders," *Journal of Financial Economics*, 2014, *113* (1), 156–169.

Hu, Edwin, "Intentional access delays, market quality, and price discovery: Evidence from IEX becoming an exchange," *SEC Working Paper*, 2018.

Jones, Charles M, "What do we know about high-frequency trading?," *Columbia University Working Paper*, 2013.

Kyle, S Albert, "Continuous Auctions and Insider Trading," *Econometrica*, 1985, *53* (6), 1315–1335.

Lewis, Michael, *Flash boys: a Wall Street revolt*, WW Norton & Company, 2014.

Liu, Wai-Man, "Monitoring and limit order submission risks," *Journal of Financial Markets*, 2009, *12* (1), 107–141.

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

__ **and Marius A Zoican**, "Need for speed? Exchange latency and liquidity," *The Review of Financial Studies*, 2017, *30* (4), 1188–1228.

O'Hara, Maureen, "High frequency market microstructure," *Journal of Financial Economics*, 2015, *116* (2), 257–270.

Pagnotta, Emiliano S and Thomas Philippon, "Competing on speed," *Econometrica*, 2018, *86* (3), 1067–1115.

Riordan, Ryan and Andreas Storkenmaier, "Latency, liquidity and price discovery," *Journal of Financial Markets*, 2012, *15* (4), 416–437.

Shkilko, Andriy and Konstantin Sokolov, "Every cloud has a silver lining: Fast trading, microwave connectivity and trading costs," *SSRN Electronic Journal*, 2016.

Ye, Mao, Chen Yao, and Jiading Gai, "The externalities of high frequency trading," *SSRN Electronic Journal*, 2013.

Zhu, Haoxiang, "Do dark pools harm price discovery?," *The Review of Financial Studies*, 2014, *27* (3), 747–789.

A Different Market Structures

A.1 Continuous Updating by Market Makers

We modify the benchmark case by allowing each market maker to update (cancel and resubmit) limit orders continuously before HFTs move. Other structures of the model is the same as the benchmark.[24]

Consider a market maker who updates her limit order by resubmitting competitive s_t. The competition drives $s_t = E[v|\text{trade at } t]$. Since δ is stochastic,

$$\begin{aligned} s_t &= \int_0^\infty be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta \\ &= \int_t^\infty be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta + \int_0^t be^{-b\delta} E[v|\text{trade at } t,\delta]d\delta \qquad (14) \\ &= 0 \times \int_t^\infty be^{-b\delta}d\delta + \int_0^t be^{-b\delta}\frac{\phi}{\phi+\beta}\sigma d\delta \qquad (15) \\ &= (1-e^{-bt})\frac{\phi}{\phi+\beta}\sigma. \end{aligned}$$

In (14), the first term represents the case that t is in the "safe interval," i.e., $0 < t < \delta$. Conditional on trade occurs, the market maker expects that $E[v] = 0$ because the trade must be against a liquidity trader, and it does not convey any information. This is why the first term in (15) bears 0. The second is the case that t is outside of the "safe interval," leading the conditional expected return to be the probability of the HFT arrival (times σ) in (15).

A speed bump has the same effect on the endogenous marginal cost as in the benchmark (the proof is omitted as it is straightforward):

Proposition 8. *(i)* $\frac{\partial s_t}{\partial b} > 0$, *and (ii)* $\frac{\partial}{\partial b}\left(\frac{\partial s_t}{\partial \phi}\right) > 0$.

Since $b = \lambda^{-1}$ represents the inverse of the expected length, a longer delay (i) directly mitigates adverse selection for market makers, but (ii) it makes the marginal cost (spread) less sensitive to speed-up by the HFT.

We impose an exogenous sunk cost to make the model well-defined. The optimization problem of the HFT regarding the speed is given by

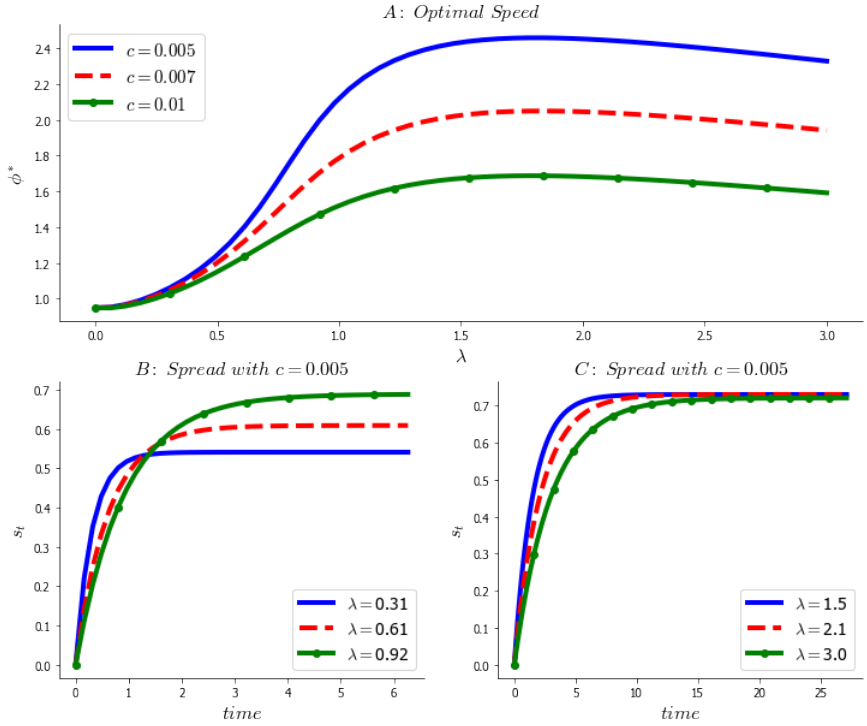
$$\begin{aligned} \max_\phi W(\phi) &= E_\delta\left[\int_0^\infty \phi e^{-\psi t}e^{-\eta\delta}(\sigma - s_{t+\delta})dt\right] - \frac{c}{2}\phi^2, \\ \text{s.t., } s_t &= (1-e^{-bt})\frac{\phi}{\phi+\beta}\sigma. \end{aligned}$$

The sniping profit from sending market orders at t is given by $\sigma - s_{t+\delta}$ since they possibly arrive and executed at $t+\delta$. Note that, given that the trade occur, there is no price uncertainty since s_t is deterministic.

The behavior of the optimal speed and spread is hard to show analytically. However, the numerical solutions in Figure V can be discussed by using the ingredients we have already analyzed. In the single HFT economy, a speed bump positively affects the optimal speed ϕ^* through (i) decline in the marginal cost and (ii) increase in the sniping profit, while (iii) it reduces ϕ^* by magnifying the exogenous sunk cost. When λ is small, the execution risk is relatively small, leaving effect (iii) less significant compared to (i) and (ii), while a longer delay in the high-λ region makes (iii) more salient.

[24] We can eliminate the possibility that an informed HFT splits her orders across the time because executions in a part of the markets let other market makers know the arrival of the HFT and true information. This event triggers the cancellation of outstanding limit orders. We also abstract away from the possibility of a mixed strategy since each order from the HFT does not have price impact. As mentioned in Footnote 15, we can show that the mixed strategy is not an equilibrium because waiting is not credible.

Figure V: Effect of λ on ϕ^* and s_t



Note: The panel A plots the optimal speed against λ with different values of c. The panels B and C plot the dynamics of the spread, s_t with different values of λ. The panel B is the case with the increasing optimal speed, $\frac{d\phi^*}{d\lambda} > 0$, and the panel C is the decreasing optimal speed, $\frac{d\phi^*}{d\lambda} < 0$.

As a result, ϕ^* takes the hump-shaped curve as depicted by Panel A in Figure V. Given λ, a larger cost slows HFT down as we can see from a parallel shift in the curves.

As (14) suggests, the spread is time dependent and increasing in t. This is because the market makers expect that they are less likely to be in the safe interval as t increases. The effect of $\lambda = b^{-1}$ is

$$\begin{aligned}\frac{ds_t}{d\lambda} &= -\frac{1}{\lambda^2}\frac{\partial s_t}{\partial b} + \frac{d\phi}{d\lambda}\frac{\partial s_t}{\partial \phi} \\ &= -\frac{1}{\lambda^2} te^{-bt}\frac{\phi}{\beta+\phi} + \frac{d\phi}{d\lambda}(1-e^{-bt})\frac{\beta}{(\phi+\beta)^2}. \end{aligned} \tag{16}$$

First, a longer delay (higher λ) reduces the spread since market makers are directly protected by the longer safe interval. This is represented by the first term in (16). However, as a higher λ may push ϕ up or down, it increases or decreases the spread, as the second term in (16) suggests. When $\frac{d\phi}{d\lambda} > 0$, the second effect competes with the first effect: the safe interval gets longer, while the HFT becomes faster. The result is provided by Panel B in Figure V.

Whether the first effect dominates the second effect depends on t. From (16), the first negative effect is increasing in $t \leq b^{-1}$ and then starts decreasing. On the other hand, the second one is monotonically increasing in t and concave. There is a unique $t = \tau$, such that $\frac{ds_t}{d\lambda} > 0$ if and only if $t > \tau$, i.e., a longer speed bump increases the spread and worsens the adverse selection problem in the long-run.

The intuition is straightforward. When the current period t is $t > \tau$, the probability that t is in the safe interval is relatively small. Then, market makers think that an increase in λ has only a small effect to mitigate the adverse selection, while it increases the speed of the HFT.

When $\frac{d\phi}{d\lambda} < 0$, the result is given by Panel C in Figure V. Since a bump reduces the speed in this case, the second effect helps the first effect reduce the spread and adverse selection cost. Once again, whether or not a speed bump increases the equilibrium speed depends on the exogenous cost, c, competing with the endogenous cost effect. Hence, the effect on the spread is governed by the market structure c and the time frame t.

A.2 Coexistence of Fast and Slow Markets

In the real economy, the major market structure is still the continuous limit market with no speed bumps. Thus, the introduction of a speed bump inevitably makes these market structures coexist. However, analyses provided by the literature deal only with homogeneous markets. We extend our model to relax this limitation.

A.2.1 Environment

Consider the benchmark economy in Section 2. $q \in (0,1)$ fraction of market makers are in the market with a delay $\delta > 0$, which is stochastic, and the rest of them are in the market with no delay, $\delta = 0$. We call the first market the *slow market* and the latter one the traditional *fast market*. Each market is competitive. Market makers in the slow market submit limit orders with the (half) spread s_λ, while those who in the traditional fast market provide s_0. As in the benchmark, λ is the expected length of the delay. In contrast to the literature (Biais et al., 2015), we impose no restrictions on the venue choice by the HFT. Transactions information in each market becomes public right after an order execution, i.e., the markets are perfectly transparent.

A.2.2 Strategies of HFT

Consider a strategy of the HFT who becomes informed of $v = \sigma$ at date t. There are two possible (pure) trading strategies for the HFT. First, if she submits orders into the fast market, they are immediately executed, fulfilling $1 - q$ of her total buying attempts. This market activity is publicly observable, allowing all market makers to realize that the transactions are information driven.[25] Based on this premise, market makers in the slow market can cancel their limit orders in the interval $\tau \in (t, t+\delta)$ which is protected by the speed bump. We call this the "strategy one" and denote it by $A = 1$.

Second, the HFT who becomes informed at t can immediately send market orders for q shares into the slow market, anticipating the execution with the δ-delay. She refrains from sending orders to the fast market at t and waits until the orders sent to the slow market are executed. By observing the execution in the slow market, she sends orders to the traditional fast market at $t+\delta$, which incur no delay by the construction. In this case, all of her orders arrive at the markets at the same time, $(t+\delta)$, and she can conceal her identity. Hence, none of the market makers can cancel their quotes. This "wait-and-grasp-all" strategy is denoted as $A = 2$.[26] Overall, taking $A = 2$ bears the execution risk, though the return from it is larger than $A = 1$ if accomplished.

The mixed strategy is the probability distribution over the set of actions $A \in \mathcal{A} = \{1, 2\}$, and let $\theta_t \in [0,1]$ be the probability that the HFT takes the action $A = 2$. For $A \in \mathcal{A}$, let $w_A(t)$ be the expected profit from taking $A \in \mathcal{A}$ when the information arrives at date t. Figures VI and VII illustrate the timing of the executions when the HFT becomes informed at t.

First, $A = 1$ does not bear the execution risk because the HFT can immediately snipe limit orders in the fast market. However, this behavior becomes public immediately, allowing market makers in

[25] This is because orders from liquidity traders will be fulfilled at the slow and the fast markets simultaneously.

[26] Note that making other lengths of strategic time lag is not optimal for the HFT, as any other intentional delay than δ tells that the orders are not from liquidity traders but from the HFT.

Figure VI: Strategy 1

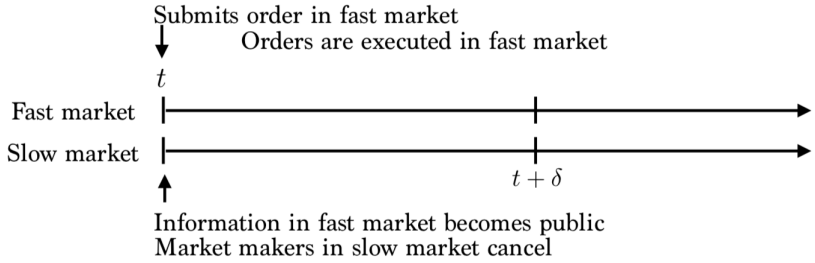
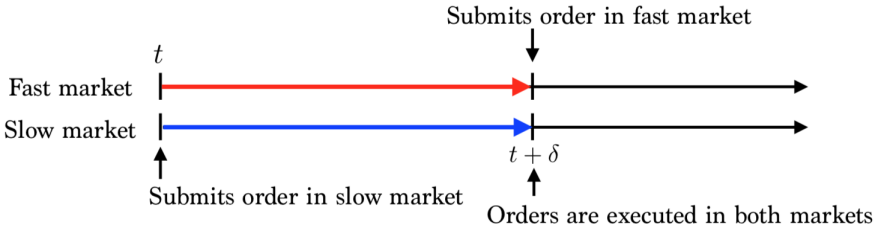


Figure VII: Strategy 2



the slow market to cancel their orders. Thus, q fraction of quotes disappear, and the expected profit is given by

$$w_1(t) = (1-q)(\sigma - s_0).$$

On the other hand, $A = 2$ can snipe all outstanding liquidity at the same time, while it bears the execution risk that stems from the δ-delay. If there is a liquidity shock or the public news during $(t, t+\delta)$, the HFT cannot exploit her information and speed. Given that the HFT gets informed at date t, she obtains the profit with probability $\Pr(T_L > \delta, T > \delta) = e^{-(\beta+\gamma)\delta}$. Moreover, since the HFT is a price taker regarding her trading behavior, her expected return is

$$\begin{aligned} w_2(t) &= \int_0^\infty b e^{-(b+\beta+\gamma)\delta} \left[q(\sigma - s_\lambda) + (1-q)(\sigma - s_0)\right] d\delta, \\ &= \frac{q(\sigma - s_\lambda) + (1-q)(\sigma - s_0)}{1 + \lambda(\beta+\gamma)}. \end{aligned}$$

Note that both of $\{w_j(t)\}_{j\in\mathcal{A}}$ are time independent due to the memoryless property of the exponential distribution. This implies that the optimal decision of $A \in \mathcal{A}$ is also time independent. No matter when the HFT becomes informed, a timing of private news does not matter—only the delay can be her concern.

A.2.3 Behavior of Market Makers

We take q an exogenous parameter in our model and assume that each market maker is randomly assigned the structure of the market. Given an assigned market, each market maker earn zero profit and does not have an incentive to move to another market with a different structure.

In the Fast Market

The market makers in the fast market suffer from adverse selection cost no matter what strategy the HFT takes. In other words, They are not given any chances to cancel orders by observing market-

based information before the HFT snipes them. The expected return is

$$\begin{aligned} V_0 =& E_\delta \left[\theta \left(\int_0^\delta s_0 \beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta} (\beta s_0 + \phi(s_0 - \sigma)) dt \right) \right. \\ & \left. + (1-\theta) \int_0^\infty e^{-\psi t} (\beta s_0 + \phi(s_0 - \sigma)) dt \right]. \end{aligned} \tag{17}$$

The first line is the case that the HFT takes $A = 2$. In this case, the fast market is under the protection of the speed bump even though the speed bump is not applied to the fast market.[27] This is because the HFT takes "wait-and-grasp-all" strategy, and she does not snipe the fast market until she accomplishes her trading attempt in the slow market. The expected profit, in this case, is identical to those in the benchmark with homogeneous markets. In the second line, the possibility of $A = 1$ is characterized. In this case, the speed bump does not matter for the fast market, and the conditional expected return is the same as a model with $\lambda = 0$.

In the Slow Market

In contrast to the fast market, the strategy of the HFT determines whether or not market makers in the slow market are protected. If $\theta = 0$, the slow market is perfectly protected by the speed bump: they can cancel their limit orders to avoid the HFT for sure. On the other hand, if $\theta \neq 0$, it is possible that the HFT arrives at the slow market to trade.

The expected return is

$$\begin{aligned} V_\lambda =& E \left[\theta \left(\int_0^\delta s_0 \beta e^{-(\beta+\gamma)t} dt + \int_\delta^\infty e^{-\psi(t-\delta)-(\beta+\gamma)\delta} (\beta s_0 + \phi(s_0 - \sigma)) dt \right) \right. \\ & \left. + (1-\theta) \int_0^\infty s \beta e^{-\psi t} dt \right]. \end{aligned} \tag{18}$$

The intuitions behind the first line are the same as those in (17). As mentioned above, when the HFT takes $A = 1$, there is no chance for the HFT to snipe in the slow market. On the other hand, liquidity traders arrive at the slow market at t if $T_L = t$, $T_H > t$, and $T > t$, which gives the integrand in the second line.

A.2.4 Equilibrium in the Trading Stage

Let $Q \equiv (1-q)/q$. We first solve for the equilibrium spread given θ:

Proposition 9. *The equilibrium spread in fast and slow markets are given by*

$$s_j = \begin{cases} \dfrac{\phi}{\phi + \beta + \beta\psi\theta \frac{\lambda}{1+\lambda(\beta+\gamma)(1-\theta)}} & \text{for } j = 0, \\ \dfrac{\phi\theta}{\beta + \phi\theta + \lambda\beta(\beta+\gamma)\left(1 + \frac{\phi\theta}{\beta+\gamma}\right)} & \text{for } j = \lambda. \end{cases} \tag{19}$$

Proof. Solving $V_j = 0$ yields the result. □

These formulae show the following:

[27] Liquidity traders arrive at t if (i) the HFT becomes informed after t or (ii) the HFT becomes informed before t and takes $A = 2$. Given that the quote remains alive at t_-, $\Pr(T_H < t, A_{T_H} = 2 | \text{quote is alive at } t_-) = \Pr(T_H < t)$. Since, otherwise, the HFT arrives immediately at T_h and snipes the stale quotes. Therefore, $\Pr(\text{Liq. trade at } t) = \beta e^{-\psi t} + \beta e^{-(\beta+\gamma)t} \int_0^t \phi e^{-\phi\tau} d\tau = \beta e^{-(\beta+\gamma)t}$.

Corollary 1. *θ affects s_0 and s_λ in an opposite way, that is,*

$$\frac{\partial s_0}{\partial \theta} < 0, \frac{\partial s_\lambda}{\partial \theta} > 0.$$

With ϕ fixed, s_0 is decreasing in θ, while s_λ is increasing in θ. For market makers in the fast market, a higher θ (i.e., the probability of $A = 2$) implies that the fast market is more likely to be protected by the speed bump in the slow market due to the HFT's "wait-and-grasp-all" strategy. This mitigates the adverse selection risk for the market makers in the fast market, making s_0 lower.

On the other hand, a higher θ (or $\phi\theta$) has a negative impact on market makers in the slow market. This is because a higher probability of $A = 2$ reduces the chance for market makers to observe sniping activity in the fast market to cancel the quote. Thus, a higher θ exposes the slow market to more severe adverse selection and pushes the spread s_λ up.

Now, the (mixed) strategy is characterized by the following, in which θ must satisfy the indifference condition, $w_1 = w_2$.

Proposition 10. *The optimal trading strategy for the HFT is*

$$\theta = \begin{cases} 0 & \text{if } \lambda Q(\beta+\gamma) > \frac{\phi+\beta}{\beta}, \\ \theta^* \in [0,1] & \text{if } \lambda Q(\beta+\gamma) \in [1, \frac{\phi+\beta}{\beta}], \\ 1 & \text{if } \lambda Q(\beta+\gamma) < 1, \end{cases} \tag{20}$$

with

$$\theta^* = \frac{(\phi+\beta) - \beta(\beta+\gamma)Q\lambda}{1+\lambda Q(\beta+\gamma) - \beta\lambda(1-\lambda\eta Q)} \frac{1+\lambda\eta}{\phi}. \tag{21}$$

Proof. See Appendix B.6. □

With ϕ fixed, the strategy θ of the HFT in the second stage game crucially depends on (i) the expected length of delay λ and (ii) the share of the slow market q. As proposed by (20), a higher λ and smaller q make the HFT reluctant to take $A = 2$ because both negatively impact the expected profit of $A = 2$ by imposing a higher execution risk and lower profit in the slow market, respectively. Thus, as λ or Q increases, θ^* declines and converges to 0. On the other hand, the HFT sticks to the strategy $A = 2$ (i.e., $\theta \to 1$) when λ or Q is sufficiently small.

A higher speed ϕ has two effects on the behavior of θ^*. First, it is straightforward that a higher ϕ widens the region for $\theta = \theta^* \in (0,1)$. Also, under the mixed strategy, we have the following:

Corollary 2. *Ceteris paribus, $\frac{\partial \theta^*}{\partial \phi} > 0$.*

When the HFT becomes faster, the spreads in the fast and slow markets are differently affected. Since the slow market is more likely to be protected by the speed bump, a higher ϕ has a stronger effect on s_0 than s_λ. Moreover, as mentioned earlier, the slow market will face the HFT only if she takes $A = 2$ with probability θ. Hence, as we can see from (19), the effect of ϕ on s_λ is discounted by θ (i.e., ϕ affects s_λ via $\phi\theta$). Thus, when ϕ is high, the profit from the fast market shrinks more compared to the profit from the slow market. This induces the stronger incentive for the HFT to shift her priority towards the gain from the slow markets. Therefore, she tends to refrain from taking $A = 1$, and θ increases.

A.2.5 Strategic Speed Choice

Given the equilibrium in the trading stage, the HFT decides her speed level. Her objective function is denoted by

$$W(\phi) = \int_0^\infty \phi e^{-\psi t} w_A(\phi) dt$$

subject to

$$w_A(\phi) = \begin{cases} w_1(t) = (1-q)(\sigma - s_0(\phi,\theta)) & \text{if } \theta \in [0,1) \\ w_2(t) = E_\delta\left[e^{-(\beta+\gamma)\delta}(\sigma - s_\lambda(\phi,\theta))\right] & \text{if } \theta = 1, \end{cases} \tag{22}$$

the spreads in (19) as functions of (ϕ,θ), and the equilibrium strategy θ given by (20). Note that the mixed strategy $\theta^* \in (0,1)$ makes it indifferent for the HFT to take $A = 1$ and $A = 2$, leading to the first line in (22). Furthermore, when $\theta = 1$, the economy converges to the benchmark case since the effect of the speed bump in the slow market encompasses the fast market too. Thus, $s_0 = s_\lambda$, and we obtain w_2 in (22).

A.2.6 Short expected delay

As we have established in (20), a sufficiently short expected delay, such that $\lambda < (\beta+\gamma)^{-1}Q^{-1}$, does not hamper the incentive of the HFT to take $A = 2$. She always takes "wait-and-grasp-all" strategy, resulting in $\theta = 1$. This makes the economy, as well as the equilibrium results, same as the benchmark case in Section 2. Therefore, a longer expected speed bump increases the speed level ϕ^*, which completely offsets the reduction in the adverse selection cost due to the longer safe interval (Proposition 3). This region is depicted by the left region of the shaded area in Figure IX.

A.2.7 Long expected delay

When a delay is sufficiently long, the HFT becomes reluctant to take $A = 2$ because of the higher execution risk. She starts adopting the mixed strategy ($\theta = \theta^*$) or immediately snipes in the fast market ($\theta = 0$). The switch between these two cases occurs at

$$\hat{\phi} \equiv \beta\left[\lambda(\beta+\gamma)Q - 1\right]. \tag{23}$$

When $\phi < \hat{\phi}$ (resp. $\phi > \hat{\phi}$), the strategy of the HFT is $\theta = 0$ (resp. $\theta = \theta^*$). As we have discussed, this threshold is increasing in λ and decreasing in q since both of them reduce the expected profit from sniping in the slow market.

Lemma 4. *When $\theta = 0$, the optimal speed level is given by $\phi_0^* = \sqrt{\beta(\beta+\gamma)}$. The speed and the spread are independent of the expected length of the speed bump λ.*

Proof. Plugging $\theta = 0$ into (22) and taking derivative immediately derive the result. □

Since the objective function W switches at $\hat{\phi}$, we have a couple of candidates for ϕ^* depending on $\hat{\phi} \gtrless \phi_0^*$, and this is crucially affected by the values of λ and Q.

Figure VIII plots the objective function W against ϕ with various parameter values for λ, in which the effect of ϕ on θ is taken into account. This function is not smooth due to the switch at $\phi = \hat{\phi}$. When λ is relatively small, we have $\theta = \theta^* \in (0,1)$, and the optimal ϕ is higher than $\hat{\phi}$. Then the speed is positively affected by λ, i.e., the longer the expected delay, the faster the HFT. As shown by Corollary 2, this pushes θ^* up. However, because the longer expected delay escalates the execution risk and the expected return starts waining, $W(\theta^*)$ dips below $W(\theta = 0)$ at some λ. Thus, there is a λ that makes $\theta = \theta^*$ and $\theta = 0$ indifferent.

As a result, the optimal speed plummets as λ increases when θ switches from $\theta = \theta^*$ to $\theta = 0$. Intuitively, the optimal speed must incorporate the execution risk by the speed bump only if the HFT snipes in the slow market with a strictly positive probability. Otherwise (if $\theta = 0$), the speed bump has nothing to do with the HFT's expected profit. The speed decision cares only about the endogenous cost at the fast market, which is more sensitive to the change in ϕ compared to the endogenous cost that stems from the slow market. As a result, the optimal speed with $\theta = \theta^*$ is too fast if $\theta = 0$ is the optimal strategy, leading to a dive of ϕ^* at the switch. See Figure IX for the visual illustration of the effect of λ.

Figure VIII: $W(\phi)$ with different λ

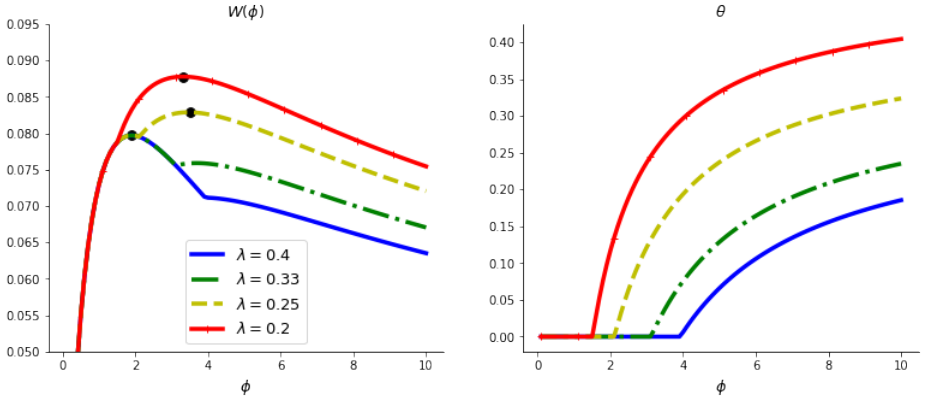
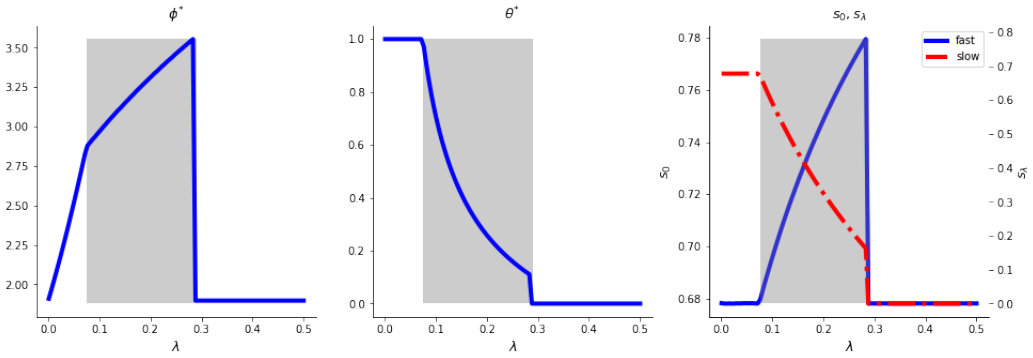


Figure IX: Effect of λ



A.2.8 Adverse Selection Cost

Figure IX shows the level of the optimal speed ϕ^*, the mixed strategy θ^*, and the spreads in both markets as functions of λ. First, if the delay is sufficiently small, so that $\lambda Q(\beta+\gamma) < 1$, the execution risk for the HFT is sufficiently low, and she takes $A = 2$ for sure ($\theta^* = 1$). The result is the same as Section 2, and the optimal speed is increasing in λ, while the adverse selection cost, measured by the half spread, is constant. Note that there is no difference between the slow and the fast markets.

Second, if $\lambda Q(\beta+\gamma) > 1$ but λ is intermediate, the HFT finds it not attractive to take $A = 2$ with 100% probability because of the relatively high execution risk. Thus, she starts to mix $A = 2$ with $A = 1$, so that she stochastically snipes at the timing of information revelation. This is represented by the shaded area in Figure IX. In this case, if she keeps ϕ constant, the welfare declines as λ increases. However, a longer expected delay makes the endogenous cost (i.e., the spread) insensitive to an increase in the speed because market makers set s as if the HFT arrives with a lower probability. This promotes the investment in the speed. In contrast to the speed, the probability of taking $A = 2$ declines, although a higher ϕ^* has a positive impact on θ^*. This is due to the dominating negative effect of λ on θ^*: a longer expected delay makes $A = 2$ a less attractive choice for the HFT.

Finally, if λ becomes sufficiently large, as in the right region of the shaded area, the HFT no longer figures that taking $A = 2$ pays out because the execution risk becomes sufficiently high. Thus, she switches to taking $A = 1$ with 100% probability, making the optimal level of the speed insensitive to λ. The optimal level ϕ^* jumps down from the middle region of λ as mentioned in the previous subsection.

Regarding the adverse selection cost in both markets, the first region with a small λ provides the

constant and same level of s. This is natural because both markets face the same risk of the HFT arrival. The intermediate region of λ makes them behave differently. A higher λ directly mitigates adverse selection for market makers, while it increases the optimal speed and the probability of the HFT-confrontation. The fast market bears the risk of the HFT no matter what strategy the HFT takes, though $A = 2$ is discounted by the delay λ. On the other hand, the slow markets are exposed to the HFT only if she takes $A = 2$, and this is protected by λ. As shown by Corollary 1, this asymmetry makes s_λ decreasing and s_0 increasing—a longer speed bump protects the slow markets at the expense of the traditional fast markets.

Once the delay becomes sufficiently long (right side of the shaded area), the risk of HFT completely diminishes in the slow market because the HFT takes $A = 1$ for sure. Hence $s_\lambda = 0$. In the fast market, the spread drops as well because the speed of the HFT is humbled. The fast market still bears the risk of the HFT and keeps the spread strictly positive.

B Appendix: Proofs

B.1 Proof of Lemma 2 and Proposition 2

Let $\eta \equiv \beta + \gamma$. The explicit formula for W is given by

$$W(\phi) = \frac{1}{1+\lambda\eta}\frac{\phi}{\psi}\frac{\beta(1+\lambda\psi)}{\phi+\beta(1+\lambda\psi)},$$

where

$$p \equiv E_\delta\left[\int_0^\infty \pi_t(\phi,\delta)dt\right] = \frac{1}{1+\lambda\eta}\frac{\phi}{\psi},$$
$$\sigma - s = \frac{\beta(1+\lambda\psi)}{\phi+\beta(1+\lambda\psi)}.$$

Therefore,

$$\begin{aligned} W'(\phi) &= p'(\sigma-s) + p(\sigma-s)' \\ &= p'(\sigma-s)(1-\varepsilon) \end{aligned}$$

with

$$\varepsilon \equiv -\frac{p}{p'}\frac{(\sigma-s)'}{\sigma-s} = \frac{(1+\lambda\eta)\psi}{\eta(1+\lambda\psi)}\frac{\phi}{\phi+\beta(1+\lambda\psi)}.$$

It is obvious that $d\varepsilon/d\phi > 0$. This implies that the optimization problem satisfies the SOC.

The solution is derived by solving the FOC, which is reduced to

$$1 = \varepsilon(\phi).$$

Note that $\varepsilon(0) = 0, \varepsilon'(\phi) > 0$, and

$$\lim_{\phi\to\infty}\varepsilon(\phi) = \frac{1+\lambda\eta}{\eta\lambda(1+\beta)}.$$

Thus, as long as $\lim_{\phi\to\infty}\varepsilon(\phi) > 1$, there is a unique solution. We can easily check that this condition is expressed as (5). If this is not satisfied, we have $\phi^* = \infty$.

When (5) holds, the $\phi^* > 0$ solves $1 = \varepsilon(\phi)$, and some tedious calculations show that the solution is given by (8). The second statement is obvious from (8).

B.2 Proof of Proposition 3

By taking a derivative, we have

$$\frac{ds}{d\lambda} \sim -\phi\psi + \frac{d\phi}{d\lambda}(1+\lambda\eta). \tag{24}$$

Moreover, by the implicit function theorem,

$$\frac{d\phi}{d\lambda} = \frac{\eta g^2 - \phi^2\gamma}{\eta g^2 + \beta\psi^2(1+\eta\lambda)^2} \tag{25}$$

where $g \equiv \phi + \beta(1+\lambda\psi)$. By substituting (25) for the one in (24),

$$\begin{aligned}\frac{ds}{d\lambda} &\sim \psi\beta(1+\lambda\eta)(1+\lambda\psi) - \phi g \\ &= \eta\beta(1+\lambda\psi)^2 - \phi^2.\end{aligned}$$

Therefore, at the optimal speed $\phi^* = \frac{\sqrt{\eta}(1+\lambda\eta)}{1-\lambda\sqrt{\beta\eta}}$, we have $ds/d\lambda = 0$.

B.3 Proof of Lemma 3 and Proposition 5

Let $r \equiv \sqrt{\beta+\phi_j}$. The second order derivative of BR_i is

$$\frac{d^2BR_i(\phi_j)}{d\phi_j^2} = \frac{dr}{d\phi_j}\frac{R}{2r^2}\left(r\frac{R'}{R} - 1\right),$$

with

$$R \equiv \frac{1+\lambda r^2}{(1-\lambda\sqrt{\beta}r)^2} + \frac{2\lambda r}{1-\lambda\sqrt{\beta}r}.$$

We can check that $r\frac{R'}{R} > 1$ is identical to $Z(r) < 0$ with

$$Z(r) \equiv 2\beta\lambda^3 r^3 - \lambda(1+\lambda\sqrt{\beta})r^2 - \lambda\sqrt{\beta}(3+2\lambda)r + 1.$$

Note that we are focusing on the bounded solution, that is $1 > \lambda\sqrt{\beta}r$. Since $Z(\frac{1}{\lambda\sqrt{\beta}}) < 0$ and $Z(0) > 0$, there is a unique r^* such that $r > r^* \Leftrightarrow Z(r) < 0$. Then, we can define ϕ_0 be the solution of $r = r^*$ and obtain the result.

The symmetric equilibrium is given by solving $\phi = BR(\phi)$, which is rewritten as $X(r,\lambda) = 0$ with $r \equiv \sqrt{\beta+\phi}$ and

$$X(r,\lambda) = \lambda(1+\sqrt{\beta})r^3 - r^2 + (1-\lambda\beta\sqrt{\beta})r + \beta.$$

This function has the following properties:

$$\begin{aligned}&\frac{\partial X(r,\lambda)}{\partial\lambda} > 0, \ \forall r > 0, \\ &X(r,0) = -r^2 + r + \beta, \ \lim_{\lambda\to\infty} X(r,\lambda) = \infty.\end{aligned}$$

Therefore, as λ increases, X shifts up from $X(r,0)$ and eventually explodes for all r. At $\lambda = 0$, $X = 0$ has a unique solution in the positive r region. By the continuity of X regarding λ, if $\lambda \searrow 0$, then $X = 0$ attains three solutions, two in the positive region (a larger one can be greater than $\frac{1}{\lambda\sqrt{\beta}}$). Let r^+ and r_- be these two solutions. Since $\frac{\partial X(r^+,\lambda)}{\partial r} > 0$ and $\frac{\partial X(r_-,\lambda)}{\partial r} < 0$, the implicit function

theorem implies $\frac{dr^+}{d\lambda} < 0$ and $\frac{dr_-}{d\lambda} > 0$, which means that the stable solution is increasing in λ. By the monotonicity of X regarding λ, there is a unique $\lambda = \lambda_0$ such that $r_-(\lambda_0) = r^+(\lambda_0)$, and $X(r, \lambda) > 0$ for all r if $\lambda > \lambda_0$, i.e., there are no solutions.

B.4 Proof of Proposition 6

First, by letting $\psi \equiv \sum_i \phi_i + \beta$ and $\eta \equiv \phi_j + \beta$, the FOC for HFT i can be expressed as

$$1 = \frac{\phi_i}{\phi_i + \beta(1+\lambda\psi)} \frac{\Upsilon(\psi)}{\Upsilon(\eta)},$$

with

$$\Upsilon(x) = \frac{x}{1+\lambda x}.$$

Under the symmetric equilibrium, it reduces to

$$1 = s(\phi, \lambda) \frac{\Upsilon(\psi)}{\Upsilon(\eta)},$$

with $\psi \equiv 2\phi + \beta, \eta \equiv \phi + \beta$, and ϕ is the equilibrium speed. We can check that

$$\frac{ds}{d\lambda} \sim \phi\psi[\psi(1+\lambda\psi) - 2\eta(1+\lambda\eta)] - \lambda\eta\phi\psi(1+\lambda\beta). \tag{26}$$

Since the symmetric equilibrium solves

$$\phi^2 = \beta\eta(1+\lambda\psi)^2,$$

we know that $\phi = \sqrt{\beta\eta}(1+\lambda\psi) \geq \beta \geq 1$. By using these conditions, we can check that the RHS of 26 is positive.

B.5 Proof of Proposition 7

First of all, the traditional model satisfies the SOC: By letting $\Gamma \equiv \phi_i + \beta(1+\lambda\psi)$,

$$\begin{aligned}\frac{dw_i}{d\phi_i} &= (\sigma - s)\frac{\partial^2 \pi_i}{\partial \phi_i^2} + \frac{\partial(\sigma - s)}{\partial \phi_i}\frac{\partial \pi_i}{\partial \phi_i} \\ &\sim -\Gamma\beta(1+\lambda\psi) - \psi(\Gamma - \phi_i(1+\beta\lambda)) \\ &< 0.\end{aligned}$$

Then, the FOC to solve is

$$c\phi_i = \frac{\beta}{\psi\Gamma}\frac{\Upsilon(\eta)}{\Upsilon(\psi)} \equiv K(\phi_i, \phi_j, \lambda), \tag{27}$$

with $\psi \equiv \sum_i \phi_i + \beta$ and $\eta \equiv \phi_j + \beta$. We can easily check that the RHS of (27) is decreasing in ϕ_i. Since K is decreasing in ϕ_j and λ around the symmetric equilibrium, we can prove that $\frac{dBR_i}{d\phi_j} < 0$ and $\frac{d\phi}{d\lambda} < 0$. Since the form of the equilibrium spread is identical to the strategic model, the opposite effect of $\frac{d\phi}{d\lambda}$ in Proposition 6 shows that $\frac{ds}{d\lambda} < 0$.

B.6 Proof of Proposition 10

The comparison is

$$w_1 \gtrless w_2 \Leftrightarrow \lambda\eta Q(\sigma - s_0) \gtrless \sigma - s_\lambda.$$

By plugging the formulae for the equilibrium spreads into the inequality above,

$$L(\theta) \equiv \lambda\eta Q(\sigma - s_0) = \lambda\eta Q \frac{K(\theta)}{\phi + \beta K(\theta)},$$
$$R(\theta) \equiv \sigma - s_\lambda = \frac{J(\theta)}{\phi + \beta J(\theta)},$$

with

$$K(\theta) = 1 + \frac{\psi}{\eta}\theta\frac{\lambda\eta}{1 + (1-\theta)\lambda\eta}, J(\theta) = 1 + \lambda\eta\left(1 - \theta + \theta\frac{\psi}{\eta}\right).$$

These functions have the following properties:

$$\frac{dL}{d\theta} > 0, L(0) = \frac{\lambda\eta Q}{\phi + \beta}, L(1) = \frac{\lambda\eta Q(1+\lambda\psi)}{\phi + \beta(1+\lambda\psi)},$$
$$\frac{dR}{d\theta} < 0, R(0) = \beta^{-1}, R(1) = \frac{1+\lambda\psi}{\phi + \beta(1+\lambda\psi)}.$$

Thus, if $\lambda\eta Q < 1$, we have $L(1) < R(1)$, indicating that $R > L$ for all $\theta \in [0,1]$. Therefore, $\theta^* = 1$ is the optimal. When $\lambda\eta Q \geq 1$, the result depends on $L(0) \gtrless R(0)$. If $\lambda\eta Q < (\beta + \phi)/\beta$, then $R(0) > L(0)$, which implies that there is a unique interior solution θ^* that solves the indifference condition. The solution solves $L(\theta) = R(\theta)$, and tedious calculation gives (21). Finally, if $\lambda\eta Q > (\beta + \phi)/\beta$, we have $R < L$ for all $\theta \in [0,1]$. Thus, $\theta = 1$ is the optimal strategy.