Execution in an aggregator

Roel Oomen*

June, 2016

Abstract

An aggregator is a technology that consolidates liquidity – in the form of bid and ask prices and amounts – from multiple sources into a single unified order book to facilitate "best-price" execution. It is widely used by traders in financial markets, particularly those in the globally fragmented spot currency market. In this paper, I study the properties of execution in an aggregator where multiple liquidity providers compete for a trader's uninformed flow. There are two main contributions. Firstly, I formulate a model for the liquidity dynamics and contract formation process, and use this to characterise key trading metrics such as, the observed inside spread in the aggregator, the reject rate due to the so-called "last-look" trade acceptance process, the effective spread that the trader pays, as well as the market share and gross revenues of the liquidity providers. An important observation here is that aggregation induces adverse selection where the liquidity provider that receives the trader's deal request will suffer from the "Winner's curse", and this effect grows stronger when the trader increases the number of participants in the aggregator. To defend against this, the model allows liquidity providers to adjust the nominal spread they charge or alter the trade acceptance criteria. This interplay is a key determinant of transaction costs. Secondly, I analyse the properties of different execution styles. I show that when the trader splits her order across multiple liquidity providers, a single provider that has quick market access and for whom it is relatively expensive to internalise risk, can effectively force all other providers to join her in externalising the trader's flow thereby maximising the market impact and aggregate hedging costs. It is therefore not only the number, but also the type of liquidity provider and execution style adopted by the trader that determines transaction costs.

*Roel Oomen is employed as the global co-head of electronic FX spot trading at Deutsche Bank AG. The views and opinions rendered in this paper reflect the author's personal views about the subject and do not necessarily represent the views of Deutsche Bank AG or any part thereof. This article is necessarily general and is not intended to be comprehensive, nor does it constitute legal or financial advise in relation to any particular situation. Oomen would like to thank two anonymous referees, Natalia Fabra, Alex Gerko, Søren Johansen, Anthony Neuberger, Mark Podolskij, colleagues at Deutsche Bank, and the seminar participants at the London School of Economics, Erasmus University Rotterdam, and the "Microstructure of Foreign Exchange Markets" conference at the Cambridge-INET Institute for helpful comments.

1 Introduction

How do you secure a fair price in a fragmented market where the same product can be traded in different venues? If you contact a single dealer you may not get the best price available, whereas approaching all dealers may be impractical or too costly. A common approach, greatly facilitated by e-commerce developments over the past two decades, is to query a representative panel of dealers and then to transact with the one that provides the best price (e.g. car insurance, electronics, airline tickets, hotel rooms). Financial markets work in much the same way, particularly the over-the-counter markets where there is no centralised exchange. A prime example here is spot foreign exchange: the biggest financial market in the world (BIS, 2014), where a large and diverse set of liquidity providers (LPs) stand ready to buy and sell currencies on a bi-lateral and disclosed basis. To source liquidity, traders routinely put multiple LPs in competition and then transact with the one that offers the best price. To facilitate this process, aggregators are used: a technology that consolidates liquidity, in the form of bid- and offer-prices and amounts, from various sources into a single consolidated order book. But in a market where the terms of trade are privately negotiated and the liquidity provided is bespoke to the trader, deciding on a suitable aggregation setup is not a trivial task. For instance, how many LPs should the trader include into her aggregator? If there is heterogeneity across LPs, how to choose amongst them? Or perhaps, when the marginal costs are negligible, the trader should simply include them all? Once the setup is defined, the trader then needs to decide on how to execute within the aggregator. For large amounts, should she trade with the LP that provides the best price in that amount, or should she spread execution across multiple LPs trading only a portion of the order with each but perhaps at a tighter spread? And what is the impact of execution uncertainty on transaction costs when the LP rejects the trader's request to deal? In this paper I provide insights into these type of questions. On the basis of a model for the liquidity dynamics and contract formation process, I establish the determining factors of transaction costs associated with execution in an aggregator.

The model assumes a setup where multiple competing LPs provide liquidity for a specified security in a standard amount at a nominal spread centred around their best but imperfect estimates of the unobserved true or efficient price. The trader uses an aggregator to consolidate the liquidity provided and trades with the LP that shows the best price. She is assumed to be uninformed with respect to the future evolution of the price process and her liquidity demand is exogenously motivated, i.e. a "noise" trader in Kyle (1985) terminology. The LPs that participate in the aggregator each contribute a continuous stream of bid- and ask-prices without knowing what their competitors are showing (this is a key difference with exchange based trading where a market maker can observe the central limit order book prior to submitting an interest). The liquidity they provide is indicative: the prices and amounts are available to the trader for use in a deal request for consideration by the LP who will subsequently make an accept or

reject decision. As such, the contract formation process is one where finality of the deal resides with the LP. In practice, the trade acceptance process serves a critical role that lets the LP – for instance – check whether sufficient credit is available to satisfy the deal request, ensure trades are conducted at valid prices for allowable amounts, manage its exposures when it simultaneously provides liquidity to a large number of traders, and prevent uncontrolled trading over a system outage or market dislocation. Additionally, it may include what is often referred to as a "last-look" feature[1] (see, e.g. Bank of England, 2011) where the LP makes a trade acceptance decision based on pre-set criteria in light of its assessment of the market price at the moment of trade acceptance. The last-look feature can include taking a brief period of time – often referred to as a latency buffer – to update information sources and enable accurate decision making. In a globally fragmented market, and with information disseminating from a variety of venues each with bespoke publication protocols, the transmission and update times involved in gathering the required information necessitates the existence of this last-look feature, and, for certain traders including those with highly aggregated execution setups, an added latency buffer. The last-look trade acceptance decision is incorporated into the model setup via the specification of a tolerance level to price movements over the latency buffer that are adverse to the LP which, when exceeded, results in the rejection of a trader's deal request.

What determines execution costs in the above setup? Suppose there is only one LP in the aggregator. The trader pays the nominal (half-) spread on execution which the LP can fully retain as revenues because the flow is uninformed. With two or more LPs in the aggregator, this logic breaks down: when the LP wins a deal request he must have shown a better price than any of its competitors but with uncertainty of where the true price is, chances are that he mis-priced the deal. The mere act of aggregation induces adverse selection where the LP that secures the deal suffers from the "Winner's curse" (Thaler, 1988). The trader will observe tighter or even negative spreads in the aggregator while the LP will find that the post-deal price movement is more likely to go in the trader's favour than in his. To defend against this, the LP can enforce the last-look trade acceptance criteria and adjust its tolerance to adverse selection. I show that it is this interplay that determines transaction costs, i.e. the number of liquidity providers, the nominal spread, and the trade acceptance criteria translate into an effective spread which represents the true cost of execution in an aggregator.

This paper presents extensive results on the key metrics that characterise the properties of aggregation. For an arbitrary number of LPs with identical liquidity dynamics and trade acceptance criteria, I derive closed form expressions for the observed spread in the aggregator, the reject rate due to last-look, and the effective spread as a representative measure of actual transaction costs incurred by the trader taking into account the slippage result-

[1] In independent and concurrent work, Cartea and Jaimungal (2015) study how venue specific last look requirements influence the choice of where latency arbitrageurs operate.

ing from execution uncertainty. Next, I introduce heterogeneity of LPs by allowing each to have distinct dynamics, nominal spreads, and trade acceptance settings. This allows for an analysis of relative market share and gross revenues, along with effective spread and reject rate, which now vary across LPs. Finally, I investigate the impact of differences in the speed of price discovery across LPs and how that can be mitigated by use of the latency buffer.

How should the trader execute in the aggregator once its setup has been defined? For small amounts, she can simply trade on best price. For larger amounts, there may not be sufficient liquidity available at the inside spread and so she faces a choice: either aggress through the stack of bids or offers and simultaneously deal with multiple LPs for the combined amount required (i.e. "stack-sweep" execution) or trade with a single LP that offers her the best price in the full amount (i.e. "full-amount" execution). The usual argument for using stack-sweep is that each child-order is of smaller size and will therefore cross a tighter spread than what is charged for a single full amount order size. But this assumes of course that the LPs offer the same liquidity to a trader irrespective of execution style. To study this, I formulate a model where the LPs decide to either internalise trades by holding the risk until they find opposing interest from other traders, or to externalise trades by immediate one-for-one hedging on public venues thereby creating an instantaneous market impact that is proportional to the volume executed. Some LPs are quicker in accessing the market than others. All LPs aim to minimise cost of trading. In this setup, I characterise the equilibrium hedging strategies of the LPs and show that with stack execution, a single LP that has quick market access and for whom it is relatively expensive to internalise risk, can effectively force all other LPs to externalise the trader's flow thereby maximising the market impact and aggregate costs levied onto the LPs. The LPs are locked in a "Prisoner's dilemma" type equilibrium. This is unlikely to benefit the trader's effective spread. In fact, it is hard to imagine any scenario where a trader with uninformed flow will achieve lower trading costs with LPs that externalise than with those that internalise. The results here highlight the fragility of the aggregator setup in this regard, where the addition of a single new LP into an established well-functioning setup, can change the hedging behaviour of all participating LPs and radically increase execution costs for the trader.

The central message of the paper is then that execution costs associated with trading in an aggregator are not simply controlled by the nominal spreads the LPs charge or the inside spread observed in the aggregator, but are instead determined by a combination of factors including (i) the number of LPs participating in the aggregator, (ii) the type of LPs selected, (iii) the trader's execution style, (iv) the nominal spread charged by each LP, (v) the LPs' trade acceptance criteria as well as (vi) intrinsic characteristics of the LPs such as the quality of price discovery and (vii) market volatility. In practice, the first 5 factors are choice variables that can be negotiated between the trader and the LPs: the primary initiative on the first three generally sits with the trader whereas iv & v are then set by the LPs aiming to satisfy spread or fill-ratio requirements of the trader subject to commercial viability. The effective

spread measure proposed in the paper is a quantity that is representative of the all-in execution costs. For a given set of trades and the LP's post-deal price stream it is trivial to calculate it in practice, both for the trader and for the LP, as the required information is common to them (and only to them). Careful trade cost analysis and an open and informed dialogue between the trader and LPs on the nature of the liquidity provision and aggregator configuration is thus of fundamental importance.

The applicability of this paper is not restricted to the spot currency market. Aggregation of one form or another is taking an increasingly prominent role across a number of over-the-counter markets. For example, the Dodd-Frank act mandates that trading of vanilla interest rate and credit default swaps now takes place on Swap Execution Facilities where a minimum of three LPs are required to compete for a trader's flow (see Commodity Futures Trading Commission, 2013). Traders in the US Treasury and corporate bond markets adopt a similar approach where they request quotes from multiple LPs when they require liquidity. The contract formation process and execution style may vary across these markets but the basic mechanisms discussed in this paper still apply.

The remainder of the paper is organised as follows. Section 2 formulates the model which is then used to obtain the results for an arbitrary number of homogenous LPs in Section 3 and two heterogenous LPs in Section 4. The analysis of trader execution style and equilibrium hedging strategies is presented in Section 5. The appendices contain the proofs and some additional results.

2 The model setup

Let the unobserved true (logarithmic) price process of a specified security follow a random walk, i.e.

$$p_t^* = p_{t-1}^* + \varepsilon_t, \tag{1}$$

with $\varepsilon \sim$ i.i.d. $\mathcal{N}(0,\sigma^2)$. There are N competing liquidity providers (LPs) that offer liquidity in the security to a known counterpart or trader on a bi-lateral basis, i.e. they each post a bid-price (b) and an ask-price (a) at which they are willing to buy and sell a standard amount, at a spread $s = a - b$ centred around a mid-price p. I assume the dynamics of the observed bid- and ask-prices for LP-i, $i \in [1,2,\ldots,N]$, to be as follows:

$$b_t^{(i)} = p_t^{(i)} - \frac{s_i}{2} \qquad \text{and} \qquad a_t^{(i)} = p_t^{(i)} + \frac{s_i}{2}, \tag{2}$$

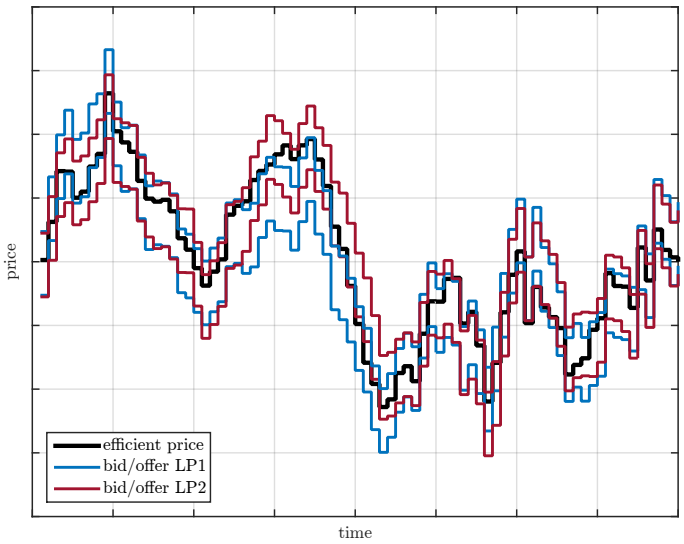
where

$$p_t^{(i)} = p_t^* + m_t^{(i)}, \tag{3}$$

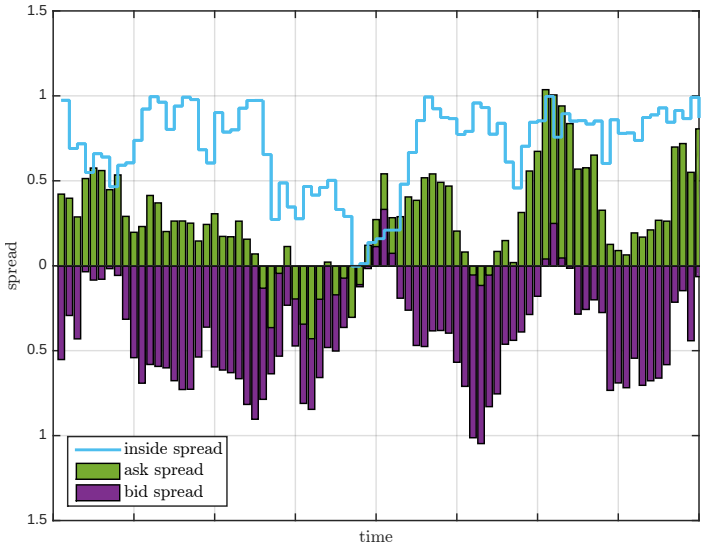
$$m_t^{(i)} = \beta_i m_{t-1}^{(i)} + \eta_t^{(i)}, \tag{4}$$

Figure 1: Aggregated liquidity dynamics

Panel A : sample price paths



Panel B : spread in the aggregator



Note. Panel A draws sample price paths of bids (b) and offers (a) of two ($N=2$) competing liquidity providers together with the unobserved "true" price (p^*). Parameters are equal for both LPs and set as $s=1, \sigma=0.5, \omega=0.35, \beta=0.9, \rho=0.5$. Panel B draws the inside spread observed in the aggregator decomposed into (unobservable) ask spread $\underline{a}-p^*$ and bid spread $p^*-\overline{b}$.

with $\eta^{(i)} \sim$ i.i.d. $\mathcal{N}(0,(1-\beta_i^2)\omega_i^2)$, $0 \leq \beta < 1$, and $\text{corr}(\eta_t^{(i)}, \eta_t^{(j)}) = \rho_{i,j}$ for $i \neq j$. The process m – also referred to below as p^*-deviation – allows for a dual interpretation. The LP may set the mid-price p equal to its best but imperfect estimate of the unobserved true price p^*. In this case m represents the measurement error of the estimator. Alternatively, one may assume the measurement error is negligible and m instead reflects a price "skew" that the LP uses to indicate its relative willingness to buy or sell the security. As part of an inventory risk management strategy, the LP may skew down its mid-price ($m<0$) to discourage further sell orders and actively solicit buy orders to reduce its long position, and vice versa. Whilst the LPs independently construct their mid-prices, in practice the information sets they use for the purposes of price discovery may be partially overlapping and this can lead to cross-sectional correlation in their measurement errors. Similarly, the set of counterparts the LPs provide liquidity to may be fully or partially overlapping, and this can lead to cross-sectional correlation in positions and thus in price skews. The parameter ρ captures this effect.

The trader's liquidity demand is assumed to be exogenously motivated, and independent of the future evolution of the price process, i.e. a "noise" trader in Kyle (1985) terminology. She uses an aggregator to consolidate the

liquidity provided by the LPs and deals in standard amounts on the best available price. That is, she submits an offer to buy at price $\underline{a}_t = \min_i a_t^{(i)}$ which the aggregator will route to LP $\underline{i}_t = \arg\min_i a_t^{(i)}$, and she submits an offer to sell at price $\overline{b}_t = \max_i b_t^{(i)}$ which the aggregator will route to LP $\overline{i}_t = \arg\max_i b_t^{(i)}$. As discussed in the introduction, the liquidity provided to the trader is indicative in nature and depending on the circumstances a trader's request to deal may be accepted or rejected by the LP: she is not guaranteed to transact at $\overline{b}_t$ or $\underline{a}_t$. From a transaction cost analysis perspective, it is thus important to distinguish between the observed inside spread in the aggregator (i.e. $\underline{a}_t - \overline{b}_t$) and the effective spread that incorporates any slippage costs introduced by the execution uncertainty. These quantities will be studied in detail below.

Figure 1 provides an illustration of the model setup.[2] Panel A draws a simulated sample path of two LPs' bid- and ask-prices around the unobserved true price, based on the model defined by Eqs. (1 – 4) . It highlights the alternating nature of the LP that has the best available price at any point in time. Panel B shows the time-varying dynamics of the inside spread decomposed into its bid and ask components.

The usual setup for a model of market making in the microstructure literature is one where a dealer faces a crowd of anonymous traders and either sets a single price or spread for all based on inventory considerations and the need to be compensated for absorbing risk or based on information considerations and the need to balance the costs of dealing with informed traders with the revenues made from uninformed traders (see, e.g. O'Hara, 1995, for an overview). The model presented here differs in a number of important ways, namely it focusses on the bi-lateral interactions between a single trader and the LPs competing for its flow, the relationship between trader and LP is disclosed, and this in turn allows the LP to provide liquidity and set prices that are bespoke to the trader. The setup is consistent with an over-the-counter market structure as opposed to anonymous exchange-based trading. And while the model in Eqs. (1 – 4) is of reduced form, it is not incompatible with the basic premise of information- and inventory-based models. For instance, despite the trader being uninformed, I will show below that important information effects arise: when the LP wins an offer to transact from the trader it knows that at that point its price was more aggressive than any of the other LPs competing in the aggregator and this information can be used by the

[2] The illustrations throughout the paper require a choice of specific model parameter values. Because statistical inference is beyond the scope of this paper, I normalise on σ and set $\omega \approx \sigma$ on the basis that the high-frequency data literature estimates the so-called "noise ratio", i.e. ω/σ in the setup here, to be around 0.5 for a range of liquid currencies and US equities(see, e.g., Christensen, Oomen, and Podolskij, 2014, Table 3). In standard market microstructure models, the spread s typically compensates the market maker for providing immediacy in a risky asset (with risk measured by σ) and/or to protect against adverse selection due to information asymmetry or mis-pricing (magnitude of this is measured by ω). It therefore seems reasonable to set the spread to a (small) multiple of the σ or ω. For the parameters β and ρ there is little guidance available so I set them to ad hoc values: $\rho = 0.5$ in a range $(0.5, 0.75)$ and β in a range $(0.5, 0.9)$ depending on the specific illustration.

LP to refine its estimate of the unobserved efficient price process. While this feedback mechanism is not explicitly modelled here, the mean-reverting nature of the m-process captures such dynamics. Similarly, when the market maker acquires a large position from traders' directional flow, inventory considerations would lead it to adjust prices in an attempt the balance the flow and reduce the position. Again, this is consistent with a mean-reverting process for m and the price skewing interpretation outlined above.

3 Homogenous liquidity providers

I start by analysing the case of $N \geq 1$ competing LPs with identical dynamics, i.e. $\omega_i = \omega, \beta_i = \beta, s_i = s$ and $\rho_{i,j} = \rho$ for all $i \neq j$. I derive properties of the observed spread and show that despite the uninformed nature of the trader's flow, strong adverse selection can be introduced by the act of aggregation. Using a simple rule to characterise the trade acceptance decision of the LP, I study the probability of a trader's deal request getting rejected and the factors that make this more or less likely. I then provide a characterisation of the effective spread: a representative measure of actual transaction costs incurred by the trader that incorporates any slippage costs associated with the execution uncertainty that the trade acceptance process introduces. Section 4 studies the same topics for heterogenous LPs with distinct liquidity dynamics and trade acceptance criteria.

3.1 Observed spread and adverse selection in an aggregator

Proposition 1 *For a panel of N homogenous liquidity providers, the expected observed spread in an aggregator is*

$$S \equiv E(\underline{a}_t - \overline{b}_t) = s - 2\omega\sqrt{1-\rho}\,\psi_N, \tag{5}$$

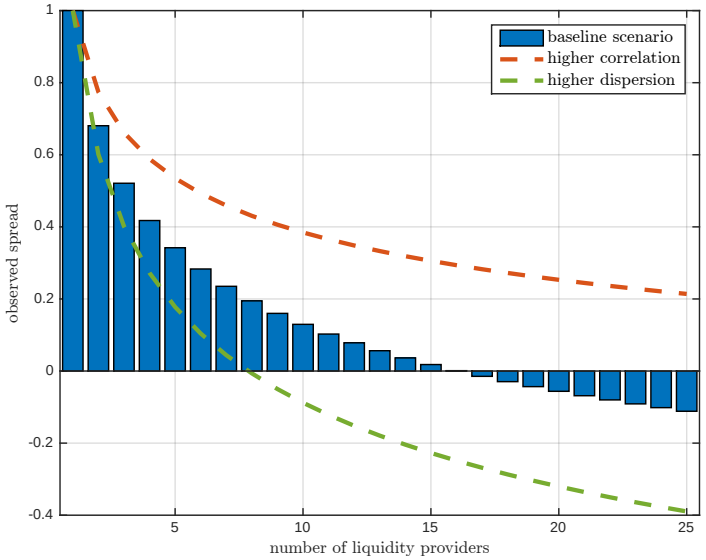
where $\psi_N = E(\max_i\{u_i\}_{i=1}^N)$ for $u_i \sim$ i.i.d. $\mathcal{N}(0,1)$. Note that $\psi_1 = 0, \psi_2 = 1/\sqrt{\pi}, \psi_3 = 3/\sqrt{4\pi}$, and $\psi_N \propto \sqrt{\log N}$ for large N.

Proof See Appendix B. ■

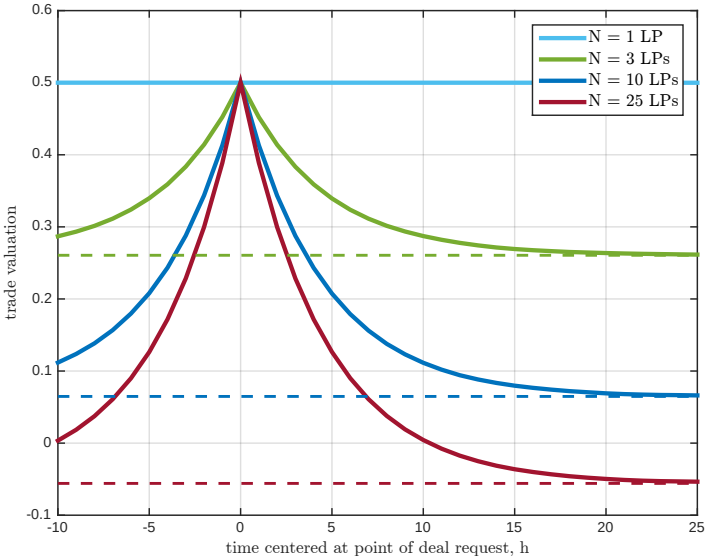
This result characterises a number of important properties of S. First, the observed spread decreases with an increase in ω (i.e. $\partial S/\partial\omega < 0$) or a decrease in ρ (i.e. $\partial S/\partial\rho > 0$). Intuitively, ω and ρ control the variability of the LPs prices relative to the common true price and relative to each other respectively. With increased variability or disagreements across the LPs prices, the higher the best bid and the lower the best ask will be and thus the tighter the observed spread. Note that when $\omega = 0$ or $\rho = 1$ the effective number of LPs is one (they all quote identical prices) and $S = s$. Second, the observed spread decreases with an increase in the number of liquidity providers N

Figure 2: Observed spread and adverse selection in an aggregator

Panel A : observed inside spread



Panel B : valuation of a trade



Note. Panel A draws the expected observed spread S as a function of the number of competing liquidity providers, N. For the baseline scenario the model parameters are set to $s = 1, \sigma = 0.5, \omega = 0.4, \beta = 0.8, \rho = 0.5$. This is varied to $\rho = 0.75$ for the higher correlation scenario and to $\omega = 0.5$ for the higher dispersion scenario. Panel B draws the marked-to-mid valuation of a trade V_h (solid lines for $h \geq 0$) as in Eq. (7) for different number of liquidity providers N using the baseline parameters as in Panel A. The pre-deal behaviour (solid lines for $h < 0$) and the long-term valuation V_∞ (dashed lines) are superimposed.

(i.e. $\partial S/\partial N < 0$, unless $\omega = 0$ and / or $\rho = 0$ in which case $\partial S/\partial N = 0$). With every addition of a new LP the best bid and ask can – ceteris paribus – only be improved and never worsened. However, because $\partial^2 S/\partial N^2 > 0$, the rate at which the spread contracts as new LPs are added decreases with N. Third, while the observed spread in the aggregator can never exceed the nominal spread s, there is nothing that prevents it from turning negative. In fact, for sufficiently large ω or N the expected observed spread can be arbitrarily negative. Note that a negative spread implies that the bid of at least one LP must exceed the ask of another and that this would present a guaranteed arbitrage opportunity were it not for the indicative nature of the liquidity and the associated trade acceptance process discussed below. Finally, the observed spread is invariant to the speed of mean reversion β (this is simply because the unconditional variance of m is assumed to be independent of β) and the efficient price volatility σ (it is deviations from the true price that affect the observed spread rather than the variability of the true price itself). Panel A of Figure 2 provides an illustration.

Additional insights into the properties of aggregation can be obtained by considering the value of a trade to the

liquidity provider (for the moment I assume that every deal request is accepted). At trade inception, the LP earns half the spread whilst the long-term expected valuation of the trade is established by adding any systematic (adverse or favourable) price movements. Specifically, the LP's valuation of a trade h periods after the point of execution is:

$$V_h \equiv E\left(\frac{s}{2}+(p^{(i)}_{t+h}-p^{(i)}_t) \mid b^{(i)}_t > b^{(\neq i)}_t\right) = E\left(\frac{s}{2}-(p^{(i)}_{t+h}-p^{(i)}_t) \mid a^{(i)}_t < a^{(\neq i)}_t\right) \qquad \text{for } h \geq 0. \tag{6}$$

Naively, one may expect due to the uninformedness of the trader's flow there to be no systematic price impacts post-deal and the value of a trade to equal half the spread charged. But this reasoning overlooks that to win the deal request in the first place, the LP needs to show a more attractive price than any of its competitors. With uncertainty of where the true price is, the LP that wins the trader's deal request to sell (buy) is likely the one that over- (under-) estimates the true value. This is the so-called Winner's curse (see Thaler, 1988). A type of adverse selection that was first discussed in the literature in the context of bidding in common value auctions (see, e.g., Capen, Clapp, and Campbell, 1971; Kagel and Levin, 1986) and to which the aggregator setup here bears a close resemblance. The below result formalises this intuition.

Proposition 2 *For a panel of N homogenous liquidity providers competing for a trader's uninformed flow, the LP's expected valuation of a trade marked-to-mid h periods post deal is:*

$$V_h = \frac{s}{2}-(1-\beta^h)\omega\sqrt{1-\rho}\,\psi_N \qquad \textit{for } h \geq 0, \tag{7}$$

where ψ_N is as defined in Proposition 1.

Proof See Appendix B. ■

When $N = 1$ the spread capture at deal inception is fully retained (i.e. $V_h = s/2$ for all $h > 0$): because there is no competition, the Winner's curse does not apply. When $N > 1$, adverse selection is introduced merely through the act of aggregation and this results in an erosion of initial spread capture and a valuation of $V_h < s/2$. Logic dictates that the effective half-spread the trader pays should equal the LP's long-term valuation of a trade. The effective spread, denoted by $\mathbb{S}$, can therefore be defined as:

$$\mathbb{S} \equiv 2\lim_{h\to\infty} V_h = s - 2\omega\sqrt{1-\rho}\,\psi_N. \tag{8}$$

Because every deal request is accepted by the LP, the observed spread equals the effective spread and its properties carry over one-for-one: the degree of adverse selection the LP is exposed to upon winning a deal request increases with an increase in ω and N and a decrease in ρ. The speed at which the LP's valuation converges to the effective

spread is determined by β. Panel B of Figure 2 provides an illustration. It emphasises the key point that even though the trader's flow is random and uninformed, each and every liquidity provider competing in the aggregator will perceive the flow as informed in that post-deal they will observe a systematic move in their mid-price p that favours the trader and is adverse to the LP.[3]

3.2 "Last look" trade acceptance

With adverse selection induced by aggregation, it is easy to arrive at a point where the effective spread paid by the trader is not commercially viable anymore for the LPs. At this point, one can proceed along a few different avenues. Firstly, individual LPs may decide to terminate their relationship with the trader by withdrawing liquidity provision. This leads to a reduction in N and, ceteris paribus, an increase in effective spread which may be sufficient to regain commercial viability for the remaining LPs. Secondly, the LPs may decide to widen the nominal spread they charge to a point where liquidity provision can be resumed on a sustained basis. The impact of both these options is quantified above, but neither may be desirable in practice. The LP may be reluctant to terminate its relationship especially when the liquidity provision is only one component of the overall service it provides to the trader, and equally the trader may require a certain number of LPs in her aggregator for redundancy purposes or to satisfy internal execution guidelines. Widening the spread by a single LP will lead to reduced market share and a likely increase in adverse selection which taken together may reduce the revenues for this LP (Section 4 studies this in more detail). A uniform widening of the spread across LPs would avoid this situation but the coordination required to achieve this is incompatible with the competitive nature of the market and the independent decision making by LPs. A third option exists to control the adverse selection and effective spread, and that is for the LPs to enforce trade acceptance criteria. To study this mechanism, I specify a simple rule where a trader's request to sell, submitted at time t, will be accepted by the LP at time $t+n$ if

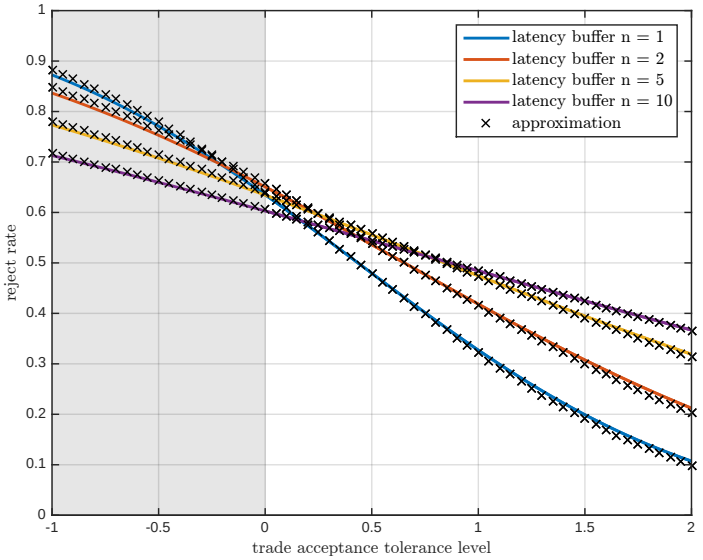
$$b^{(i)}_{t+n} - b^{(i)}_t > -\delta \qquad \text{for } n \geq 1, \delta \geq 0, \tag{9}$$

and rejected otherwise. Analogously, the LP will accept a trader's request to buy if $a^{(i)}_{t+n} - a^{(i)}_t < \delta$ and reject otherwise (in Oomen, 2016, I study a number of alternative last look specifications). The parameter n specifies the number of periods the LP takes to make a trade acceptance decision. In the spirit of the NIPS code (Bank of England, 2011), I refer to this as a "latency buffer". In a globally fragmented market, and with information disseminating from a variety

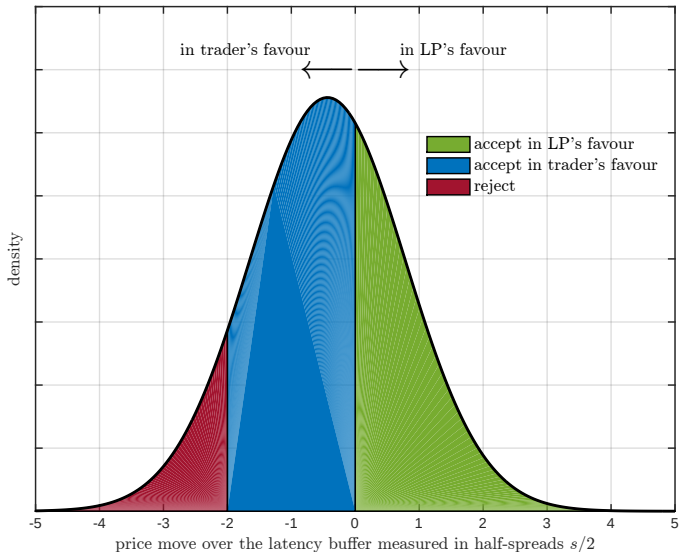
[3] As an aside, note that the chart also includes the pre-deal mid-price evolution (adjusted for trade direction) which shows an increasing aggression of the price in the run-up to winning a deal request. This pattern is specific to the setup here, and will look very different if for instance the trader's buy or sell decisions are triggered by the (true) price reaching specific levels, e.g. stop-loss or take-profit orders.

Figure 3: Trade acceptance and adverse selection in an aggregator

Panel A : reject rate for varying n and δ



Panel B : distribution of price move over latency buffer n



Note. Panel A draws the probability of a reject $\mathbb{R}$ as a function of tolerance level δ for varying latency buffer horizons n. The solid lines are based on numerical simulations, whereas the markers (×) indicate the analytical approximation given in Eq. (11). Panel B draws the price distribution over the latency buffer $n = 1$ with the shaded areas highlighting the range where deal requests are accepted or rejected for $\delta = s$. In both panels, the model parameters are set as $s = 1, \sigma = 0.5, \omega = 0.4, \beta = 0.5, \rho = 0.5, N = 10$.

of venues each with bespoke publication protocols, it takes time to gather all the relevant information required for accurate price discovery and to make an informed trade acceptance decision and this is precisely what the latency buffer provides (in practice, it is typically set to a fraction of a second). The parameter δ represents the maximum adverse price movement over the latency buffer that the LP is willing to tolerate short of rejecting the trade request. For instance, with $\delta = s/2$ the LP will reject the request to deal only if more than the full half-spread is lost to an adverse price movement within the latency buffer. Also note that negative values of δ are not permissible in the current model setup, i.e. I enforce that the LP will accept the deal request when the price doesn't move over the latency buffer.

The key quantity of interest in this discussion is of course the probability of a deal request getting rejected, i.e.

$$\mathbb{R} \equiv \Pr(b^{(i)}_{t+n} - b^{(i)}_t < -\delta \mid b^{(i)}_t > b^{(\neq i)}_t) = \Pr(a^{(i)}_{t+n} - a^{(i)}_t > \delta \mid a^{(i)}_t < a^{(\neq i)}_t). \tag{10}$$

Proposition 3 *For a panel of N homogenous liquidity providers competing for a trader's uninformed flow, and a trade*

acceptance rule as defined by Eq. (9)*, the probability of a deal request getting rejected is approximately:*

$$\mathbb{R} \approx \Phi\left(\frac{(1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N - \delta}{\sqrt{n\sigma^2 + (1-\beta^{2n})\omega^2}} \right), \tag{11}$$

where $\Phi(\cdot)$ *denotes the distribution of a standard normal random variable.*

Proof See Appendix B. ■

This result characterises the key properties of the reject rate in an aggregator setup with last-look trade acceptance. As expected, an increase in tolerance level δ lowers the reject rate (i.e. $\partial\mathbb{R}/\partial\delta < 0$) and drives it down to zero as $\delta \to \infty$. Note that when $\delta = 0$ the reject rate is at its highest and will always exceed 50%: this is a direct consequence of the adverse selection induced by aggregation that makes price moves in the trader's favour more likely than those in the LP's favour (provided that $N > 1, \omega > 0, \rho < 1$). An increase in the latency buffer n increases the reject rate (i.e. $\partial\mathbb{R}/\partial n > 0$). Two re-inforcing effects are at play here, namely (i) due to mean-reversion in m the adverse selection builds up over time (i.e. the second term in Eq. 7) so that with larger n the LP is better able to identify the effect and (ii) the natural variation of the efficient price process grows linearly over time and so a longer latency buffer makes it more likely for the price to exceed the tolerance level. Similarly, anything that elevates the variability or dispersion of LPs' prices leads to an increase in reject rate, i.e. $\partial\mathbb{R}/\partial\sigma > 0$, $\partial\mathbb{R}/\partial N > 0$, $\partial\mathbb{R}/\partial\omega > 0$, and $\partial\mathbb{R}/\partial\rho < 0$. With higher N, ω or lower ρ the adverse selection effect grows and this in turn heightens the chances of a price move to exceed the set tolerance level and generate a reject. Note that the same impact is observed with increases in the efficient price volatility σ. This is a somewhat undesirable yet unavoidable property of the last-look mechanism, as modelled here, in that efficient price moves are unrelated to the trader's actions or the Winner's curse and should therefore not affect the reject rate. But because p^* is unobservable, efficient price moves are indistinguishable and inseparable from adverse selection effects.[4] Finally, consider the impact of β on the reject rate. With a more persistent and less erratic measurement error, the price discovery is essentially slower and the ability of the LP to identify adverse selection is reduced, hence $\partial\mathbb{R}/\partial\beta < 0$. Likewise, with more persistent position skewing, the LP's prices are less volatile, reducing the probability of them exceeding the threshold.

Figure 3 further illustrates some of these points. Panel A draws the reject rates as a function of the tolerance level δ and for different values of the buffer n. It shows that the approximation in Eq. (11) is very accurate in a wide range of the parameter space. The shaded area highlights the impermissible range of negative tolerance levels where the

[4] In practice one may find the sensitivity of the reject rate to changes in σ to be limited because (i) empirically σ and s tend to move in tandem and (ii) whilst $\partial\mathbb{R}/\partial s = 0$ in the model here, a larger s does affords the LP with more room to loosen the tolerance level δ. Alternatively, it is of course possible to specify a trade acceptance rule where δ is an increasing function of σ or s.

reject rate continues to go up and eventually converges to 100%. Panel B draws the post-deal price distribution over the latency buffer to illustrate the adverse selection effect, i.e. price moves in favour of the trader are more likely than those in favour of the LP. There are three distinct regions, namely (i) the price moves in the LP's favour and the deal request is accepted (green area), (ii) the price moves against the LP but within the defined tolerance level δ and the deal request is accepted (blue area), and (iii) the price moves against the LP by more than δ and the deal request is rejected (red area).

How does the last-look mechanism impact the effective spread paid by the trader? Starting with the valuation of a trade, as before, the LP stands to earn the nominal half-spread at trade inception and its long-term valuation is obtained by adding any systematic post-deal price movements. The difference here is that instead of needing to only condition on winning the deal request, we now also need to condition on the request to successfully pass the trade acceptance rule in Eq. (9), i.e.

$$V_h = E\left(\frac{s}{2} + (p^{(i)}_{t+h} - p^{(i)}_t) \mid b^{(i)}_t > b^{(\neq i)}_t, b^{(i)}_{t+n} > b^{(i)}_t - \delta\right) = E\left(\frac{s}{2} - (p^{(i)}_{t+h} - p^{(i)}_t) \mid a^{(i)}_t < a^{(\neq i)}_t, a^{(i)}_{t+n} < a^{(i)}_t + \delta\right). \tag{12}$$

Following the same logic as above, the effective spread is defined as $\mathbb{S} = 2V_\infty$.

Proposition 4 *For a panel of N homogenous liquidity providers competing for a trader's uninformed flow, and a trade acceptance rule as defined by Eq. (9), a lower bound for the effective spread is:*

$$\mathbb{S} > s - 2\omega\sqrt{1-\rho}\,\psi_N + 2\frac{n\sigma^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} G((1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2). \tag{13}$$

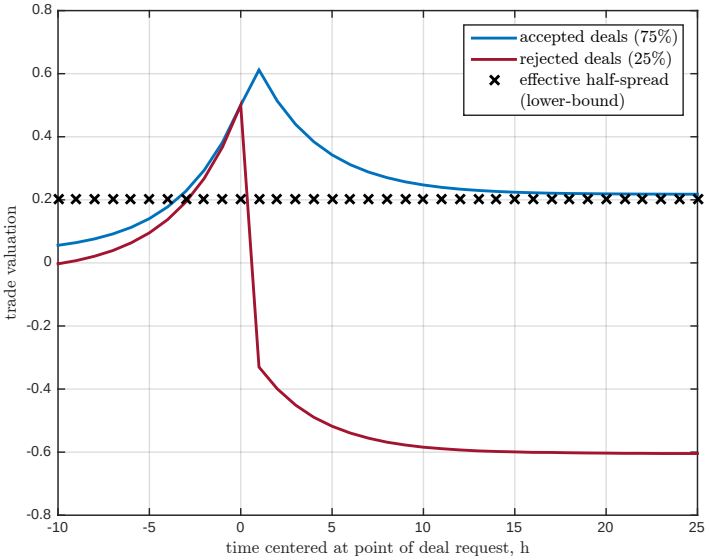
where $G(\mu_x, \sigma_x^2) = \sigma_x \phi(\mu_x/\sigma_x)/(1-\Phi(\mu_x/\sigma_x))$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution of a standard normal random variable.

Proof See Appendix B. ■

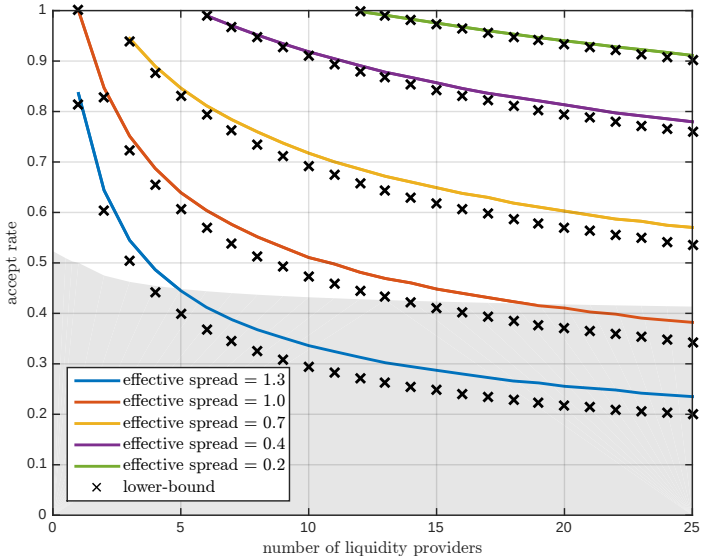
Eq. (13) shows that the effective spread can be decomposed into three separable and intuitive components, namely the nominal spread (+), the gross adverse selection costs (-), plus any recovered adverse selection costs via the trade acceptance rule (+). With similar intuition to the discussion of the reject rate, the effective spread increases with an increase in the latency buffer n (more protection for the LP), cross-sectional correlation of p^*-deviations ρ (less diversity amongst LPs), efficient price volatility σ (more likely to breach trade acceptance threshold), and of course the nominal spread s. The effective spread also increases with a decrease in the tolerance level δ (more protection for the LP), number of liquidity providers N (less competition amongst LPs), magnitude of p^*-deviations ω (more accurate pricing), and persistence of p^*-deviations β (quicker reversion to true price). It is worth noting

Figure 4: Trade valuation and effective spread with last-look trade acceptance

Panel A : trade valuation and effective spread



Panel B : indifference curves for effective spread



Note. Panel A draws the marked-to-mid valuations V_h of accepted and rejected deal requests, together with the effective half-spread approximation given by Eq. (13). The model parameters are set as $s = 1, \sigma = 0.5, \omega = 0.35, \beta = 0.75, \rho = 0.5, N = 25, n = 1, \delta = s/2$. Pre-deal dynamics are added to the chart for $h < 0$. Panel B draws a few indifference curves where the same effective spread is attained by different combinations of number of liquidity providers N (x-axis) and the accept rate $1-\mathbb{R}$ controlled by δ (y-axis). The grey area marks the region where the trade acceptance tolerance level δ is set to impermissible (negative) values.

that while the observed spread in the aggregator is invariant to the efficient price volatility σ, or the persistence of the measurement error β, the effective spread is impacted by these parameters : in turbulent markets characterised by elevated market volatility or erratic measurement error / price skews the effective spread paid by the trader increases despite her crossing the same spread in the aggregator.

Figure 4 provides an illustration of the trade valuation and effective spread with last-look. In the example of Panel A, the effective spread is composed of a nominal spread of $s = 1$ minus an adverse selection component of 0.97 plus recovered adverse selection costs via the trade acceptance process of 0.38 resulting in an effective half-spread of 0.20 (the crosses in the chart). Panel B draws indifference curves where the same effective spread is attained by different combinations of the number of LPs N and the accept rate $1-\mathbb{R}$ as controlled by δ. For instance, an effective spread of $\mathbb{S} = 0.7$ can be achieved – in this example – with three LPs accepting about 95% of the deal requests or with twenty LPs accepting about 60% of deal requests (note that the observed spread in the aggregator is 0.581 with three LPs and 0.076, or about a tenth of the effective spread, with twenty LPs). This underlines a fundamental point: for a trader

to fully understand the transaction costs associated with trading in a defined aggregator setup (i.e. N, ω, β, ρ and σ are fixed) she can't merely look at the nominal spread the LPs charge, or the observed spread that she crosses in the aggregator, but she needs to consider the triplet of choice variables (s, n, δ) and the effective spread that translates into.

To conclude the discussion, consider a trading setup without last-look. The equity market provides a good example, where so-called smart-order-routers (SORs) are routinely used to aggregate fragmented exchange liquidity and subsequently make decisions on where to route orders. A trader using an SOR to access this firm non-last-look exchange liquidity, however, still faces execution risk simply because of the physical distances and information transmission times between the competing venues and the trader location. For instance, by the time the SOR presents the trader with the latest liquidity available from a particular venue – or later still, by the time a resulting trader's order arrives at the venue attempting to access that liquidity – the price may have changed, the quote cancelled, or the liquidity removed by another trader.[5] The further the trader is located from the trading venues, or the greater the geographic dispersion of venues, the higher the execution risk will be. Co-location doesn't eliminate the issue either, because it can only get the trader close to one (or a few) venue but not all, and moreover, she will still need to compete with other traders in that same co-location. This last point highlights a key distinction between public exchange liquidity and the OTC liquidity: the former is available on a first-come-first-serve basis whereas the latter is typically available to many traders simultaneously and the liquidity demand of one trader doesn't necessarily impact the liquidity available to another trader.

4 Heterogenous liquidity providers

Up to now, the LPs each produce distinct prices but the processes that govern their liquidity dynamics and their trade acceptance settings are assumed to be identical. As a result, all LPs are exposed to the same degree of adverse selection, are equally likely to be top-of-book in the aggregator and to win a trader's request to deal, have the same market share, and each earn the same revenues. In this section I will study the properties of execution in an aggregator when the participating LPs have different characteristics. The case where $N = 2$ provides the key insights and retains analytic tractability so I'll limit the discussion to this.

[5] For example, see http://www.iextrading.com/insight/stats/, for monthly statistics on the fill ratios of the IEX SOR. For July 2015, it ranges from 69% for orders routed to CHI-X, to 88% for NYSE, to 99% for BATS.

4.1 Differences in liquidity characteristics and trade acceptance criteria

Proposition 5 *Consider two heterogenous liquidity providers competing for a trader's uninformed flow, and a trade acceptance rule as defined by Eq.* (9)*. The expected observed spread in the aggregator is:*

$$S = s_2 - (s_2 - s_1)\Phi\left(\frac{s_2 - s_1}{2\sigma_{\Delta m}}\right) - 2\sigma_{\Delta m}\phi\left(\frac{s_2 - s_1}{2\sigma_{\Delta m}}\right), \tag{14}$$

where $\sigma^2_{\Delta m} = \omega_1^2 + \omega_2^2 - 2\rho_m\omega_1\omega_2$*, and* $\rho_m = \rho\sqrt{(1-\beta_1^2)(1-\beta_2^2)}/(1-\beta_1\beta_2)$*. The probability of LP-*i *having the best price and winning a request to deal is:*

$$\mathbb{T}_i = \Pr(b_t^{(i)} > b_t^{(\neq i)}) = \Pr(a_t^{(i)} < a_t^{(\neq i)}) = \Phi\left(\frac{s_{\neq i} - s_i}{2\sigma_{\Delta m}}\right). \tag{15}$$

The probability of a deal request getting rejected by LP-i *is approximately:*

$$\mathbb{R}_i \approx \Phi\left(\frac{(1-\beta_i^{n_i})\frac{\omega_i^2 - \rho_m\omega_{\neq i}\omega_i}{\sigma^2_{\Delta m}} G(\frac{1}{2}(s_i - s_{\neq i}), \sigma^2_{\Delta m}) - \delta_i}{\sqrt{n_i\sigma^2 + (1-\beta_i^{2n_i})\omega_i^2}}\right). \tag{16}$$

A lower bound for the effective spread charged by LP-i *is:*

$$\begin{aligned}\mathbb{S}_i \quad > \quad & s_i - 2\frac{\omega_i^2 - \rho_m\omega_{\neq i}\omega_i}{\sigma^2_{\Delta m}} G\left(\frac{1}{2}(s_i - s_{\neq i}), \sigma^2_{\Delta m}\right), \\ & + 2\frac{n_i\sigma^2}{n_i\sigma^2 + (1-\beta_i^{2n_i})\omega_i^2} G\left((1-\beta_i^{n_i})\frac{\omega_i^2 - \rho_m\omega_{\neq i}\omega_i}{\sigma^2_{\Delta m}} G(\frac{1}{2}(s_i - s_{\neq i}), \sigma^2_{\Delta m}) - \delta_i, n_i\sigma^2 + (1-\beta_i^{2n_i})\omega_i^2\right).\end{aligned} \tag{17}$$

Proof See Appendix B. ■

There are a few additional execution metrics that can be derived from the above: the expected market share of LP-i, $\mathbb{M}_i = \mathbb{T}_i(1-\mathbb{R}_i)/\sum_j \mathbb{T}_j(1-\mathbb{R}_j)$, the expected gross revenues of LP-i, $\mathbb{W}_i = \frac{1}{2}\mathbb{S}_i\mathbb{M}_i$, and the effective spread the trader pays for execution within the aggregator, $\mathbb{S}_T = \sum_i \mathbb{M}_i\mathbb{S}_i$.

Differences in nominal spread To start, consider the scenario where both LPs have identical liquidity dynamics and trade acceptance settings except that LP-2 charges a different nominal spread to LP-1, i.e. $s_1 \neq s_2$ with all other parameters equal. In this case, the observed spread S is bounded by $\min(s_1, s_2)$ and the probability for LP-i to have the best price in the aggregator, $\mathbb{T}_i$, diminishes the wider the spread is in relation to that of its competitor. Turning to $\mathbb{R}_i$, note that the LP that widens its nominal spread will – ceteris paribus – reject a larger fraction of the deal requests but it also leads to a decrease in the reject rate of its competitor (Figure 5, Panel A). With a wider spread, the p^*-deviation needs to be stronger for the LP to win deal requests, but in those instances the adverse selection will also be stronger making it more likely for the LP to reject the deal request. The reject rate for the other LP

Table 1: Execution metrics with heterogenous liquidity providers

Two otherwise identical LPs with LP-2 incrementally …	observed spread	effective spread			reject rate		market share		gross revenues	
	S	$\mathbb{S}_1$	$\mathbb{S}_2$	$\mathbb{S}_T$	$\mathbb{R}_1$	$\mathbb{R}_2$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{W}_1$	$\mathbb{W}_2$
wider nominal spread (s)	↑	↑	↑	↑	↓	↑	↑	↓	↑	↓↑
more volatile p^*-deviations (ω)	↓	↑	↓	↓	↓	↑	↑	↓	↑	↓
more persistent p^*-deviations (β)	=	=	↓	↓	=	↓	↓	↑	↓	↓
more generous tolerance (δ)	=	=	↓	↓	=	↓	↓	↑	↓	↓
longer latency buffer (n)	=	=	↑	↑	=	↑	↑	↓	↑	↑

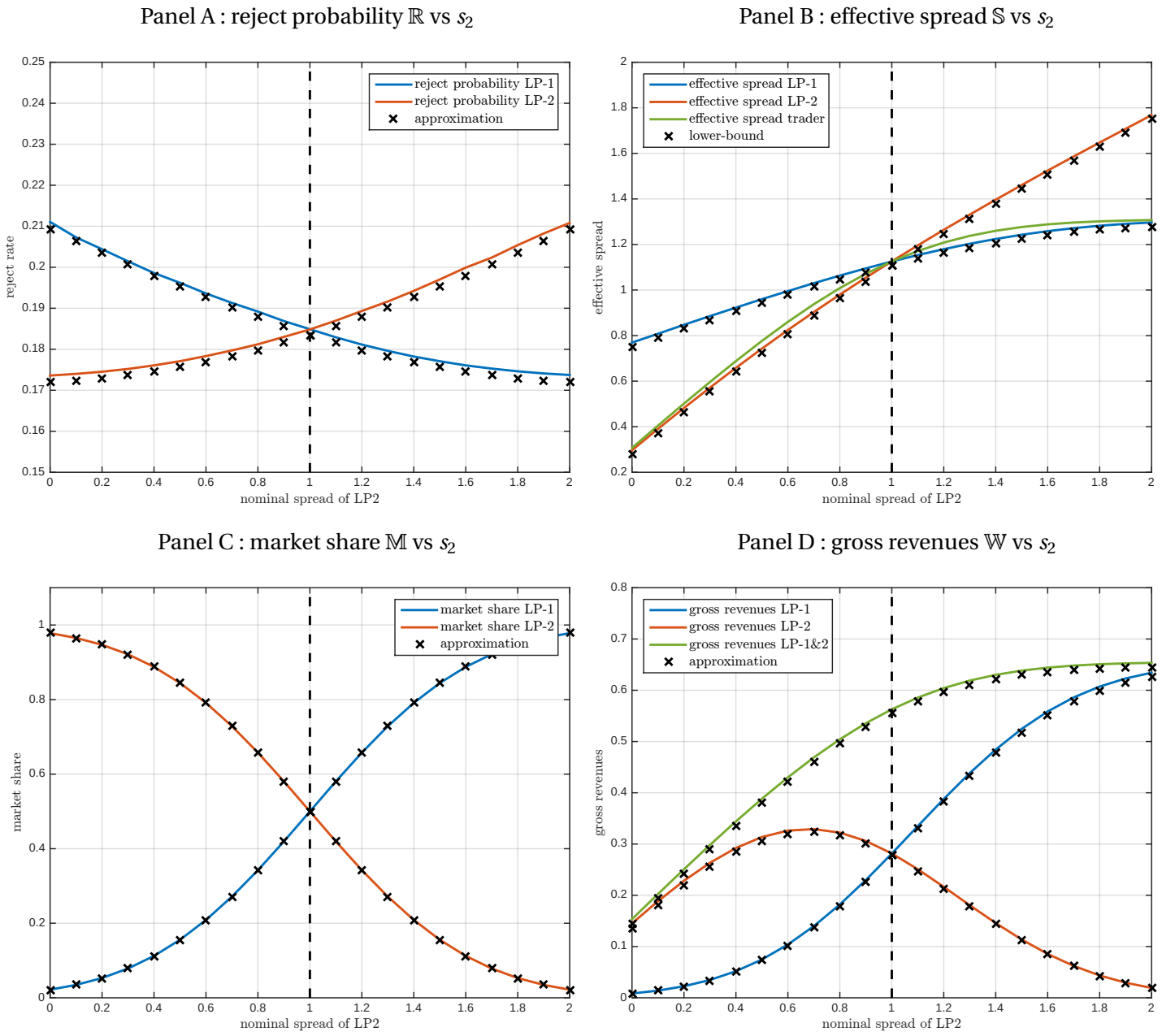
Note. For two identical LPs, this table reports how the various execution metrics change when LP-2 marginally increases one of the parameters governing the liquidity dynamics or trade acceptance process, e.g. for the impact of a widening of s_2 on effective spread for LP-2, the table reports the sign of $\partial \mathbb{S}_1 / \partial s_2 \mid_{s_2=s_1}$.

decreases because the diminishing competition for flow reduces the adverse selection. Next, the LP that widens its nominal spread, will increase its effective spread $\mathbb{S}_t$ but also raise the effective spread of its competitor (Figure 5, Panel B). Intuitively, the LP that leaves its spread unchanged will face less competition and receive deal requests on less aggressive prices, thereby increasing its effective spread. As expected, the market share decreases with a widening of the spread (Figure 5, Panel C) but note from the example that when $s_2 = 0$ (and $s_1 = 1$) the market share of LP-2 is still not 100%: on very strong p^*-deviation, LP-1 can still win deal requests despite the choice pricing of LP-2.

Perhaps the most interesting aspect here is that in a region of the parameter space, the LP that widens its spread will see its gross revenues $\mathbb{W}_i$ fall whilst its competitor will enjoy higher revenues (Figure 5, Panel C): a widening of nominal spread increases the effective spread, but this can be more than offset by a drop in market share leading to lower overall revenues. Correspondingly, in this situation, the LP can raise its revenues by tightening its nominal spread and undercutting its competitor. There is a limit to this: at some point, with further tightening, the reduction in effective spread is no longer compensated by larger gains in market share and the gross revenues will drop. In the limit, when $s_2 = 0$ (and $s_1 = 1$), LP-1's revenues dominate those of LP-2 although the latter are still positive because of the last-look mechanism.

The above discussion naturally leads to the question whether there is an equilibrium spread the LPs would charge. In an iterative process where conditional on the spread of one LP the other will set its spread to maximise gross revenues, will the spread converge, and if so to what value? This is a hard question to answer analytically, but

Figure 5: Execution metrics with two liquidity providers charging a different nominal spread



Note. This figure draws the effective spread (Panel A), reject rate (Panel B), market share (Panel C), and gross revenues (Panel D) as a function of LP-2 nominal spread s_2 with $s_1 = 1, \sigma = 0.5, \omega_1 = \omega_2 = 0.25, \beta_1 = \beta_2 = 0.75, \rho = 0.5, N = 2, n_1 = n_2 = 1, \delta_1 = \delta_2 = s/2$. The solid lines are based on simulations whereas the markers (×) are, or follow directly from, the analytical approximations given in Proposition 5.

using the expressions in Proposition 5 it is easy to consider a numerical example, see Figure 6. Panel A shows the nominal spread the LPs set in every round when the starting point is $s_1 = s_2 = 1$. The spread converges to $s_1 = s_2 \approx 0.3$

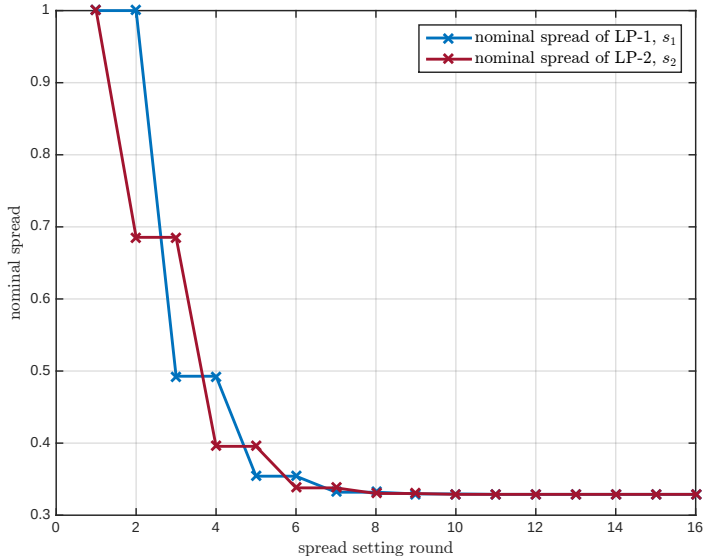
and Panel B shows that an equilibrium is attained because neither LP can increase its revenues by changing spreads. What happens if, for instance, the trade acceptance criteria differ between LPs? Panel C considers the case where $n_1 = 1$ and $n_2 = 2$: LP-2 imposes more stringent criteria than LP-1. Again the spreads converge, but now to different values, i.e. $s_1 \to 0.25$ and $s_2 \to 0.09$. Note from Panel D that while s_2 is less than half s_1 in steady state, the gross revenues that LP-2 earns are more than double those of LP-1 : the tighter spread earns LP-2 a larger market share while the stricter trade acceptance settings aid spread retention, resulting in higher revenues.

Differences in price dynamics and trade acceptance settings Table 1 summarises the trading metric properties by considering the impact of an incremental change in one of the liquidity or trade acceptance parameters, assuming all other model parameters are identical and unchanged (see also Figure 9 in the Appendix). Consider the parameters governing the price dynamics. A difference in ω between LPs can be interpreted as a difference in the quality of their price discovery. With higher ω, the LP's prices will be less accurate and more volatile, and this leads to a compression of the effective spread, an increase in the LP's reject rate, and a decrease in its market share and gross revenues. The lost revenues for this LP are effectively redistributed between the trader and competitor LP in some proportion defined by the exact model parameters: the trader will benefit from a tighter observed and effective spread and the competitor LP will be able to accept more deals, and enjoy higher market share and gross revenues. This emphasises that – despite the uninformedness of the trader's flow – there is a greater importance put on the quality of price discovery to meet the need for commercial viability when $N > 1$. A higher β implies more persistent skewing, or measurement errors that die down more slowly: the unconditional variance of m (and hence the observed spread S) is unchanged due to the scaling of the variance of η in Eq. 4, but over short horizons the process is less erratic. As a result, deal requests are less likely to be rejected as the variability of the LP's prices over the latency buffer is reduced. But because the adverse selection in the aggregator is of equal magnitude, with lower reject rate, the effective spread for the LP is also reduced. The competitor LP's reject rate and effective spread is unaffected by the increase in β, but because its market share drops, so do its gross revenues. Put simply, an increase in β increases competition in the aggregator, it leads to a higher market share for the LP with the higher β and the trader benefits from a reduction in effective spread.

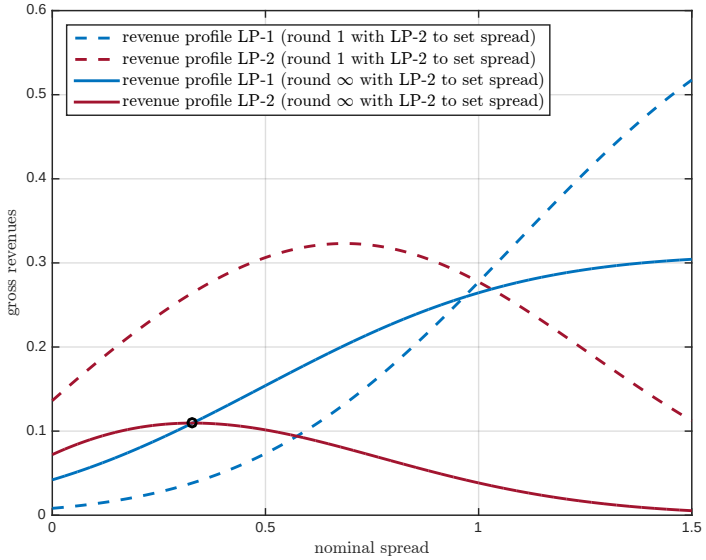
The impact of an increase in trade acceptance tolerance level δ is qualitatively equal to that of a decrease in latency buffer n and so I will limit discussion to δ. As expected, the effective spread and reject rate are reduced for the LP that loosens the tolerance levels whilst these metrics are unchanged for its competitor. The drop in reject rate leads to an increase in market share at the expense of the other LP but the increased competition leads to a reduction of revenues for both LPs. Intuitively, under-pricing of liquidity by one LP affects the viability of all other

Figure 6: Convergence to equilibrium spreads with homogenous and heterogenous liquidity providers

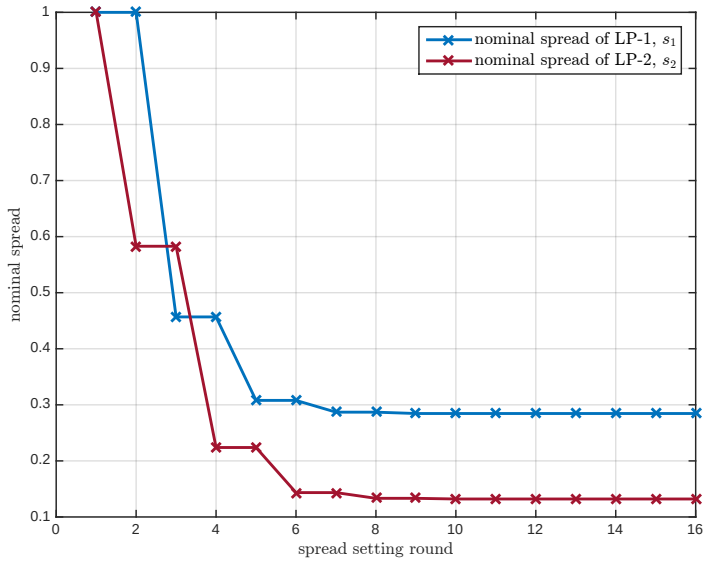
Panel A : spread convergence with homogenous LPs



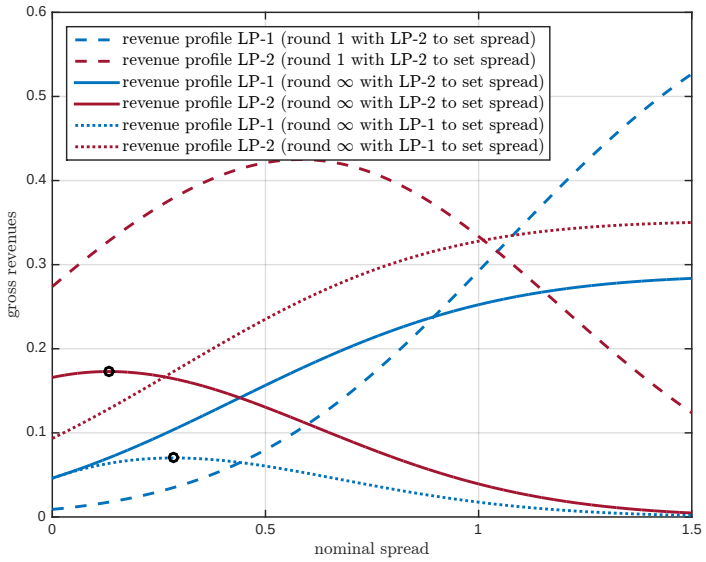
Panel B : revenue profiles across iterations



Panel C : spread convergence with heterogenous LPs



Panel D : revenue profiles across iterations



Note. This figure draws the spread convergence process in Panels A & C and the associated revenue profiles (i.e. $\mathbb{W}_i$ and $\mathbb{W}_{\neq i}$ as a function of s_i for fixed $s_{\neq i}$) in Panels B & D. In the top row, the LPs start the process as homogenous with $s_1 = s_2 = 1, n_1 = n_2 = 1$. In the bottom row, the LPs start the process as heterogenous with $s_1 = s_2 = 1, n_1 = 1, n_2 = 2$. The remaining model parameters are set at $\sigma = 0.5, \omega_1 = \omega_2 = 0.25, \beta_1 = \beta_2 = 0.75, \rho = 0.5, N = 2, \delta_1 = \delta_2 = s/2$.

LPs as they will need to price more aggressively to win the deal requests.

4.2 Differences in latency or speed of price discovery

One aspect not explicitly captured by the model in Eqs. (1 – 4) is that of latency differentials. In the globally fragmented FX market with its numerous price sources, trading locations, and news sources, it is a substantial challenge to obtain the relevant information on a continuous basis and in a timely fashion. Demand for fast data transmission has led to significant investments in state-of-the-art network links between the major financial centres (or perhaps the causality is the other way around). But in an aggregator setup, it is relative and not absolute speed that matters. So with differences in transmission latencies and speed of price discovery amongst LPs, what can be said about its impact on the execution metrics for both the trader and the LPs providing liquidity into the aggregator? To study this, I make a simple modification in the model and introduce a "slow" LP that prices at a time-lag of one period compared to a "fast" LP, i.e.

$$p_t^{\text{fast}} = p_t^* + m_t^{\text{fast}} \qquad \text{and} \qquad p_t^{\text{slow}} = p_{t-1}^* + m_{t-1}^{\text{slow}}. \tag{18}$$

Proposition 6 *Consider two liquidity providers – one fast and one slow as in Eq.* (18)*, but otherwise identical – competing for a trader's uninformed flow, and a trade acceptance rule as defined by Eq.* (9)*. The expected observed spread in an aggregator is*

$$S = s - 2\omega\sqrt{1-\rho\beta+\omega^{-2}\sigma^2/2}\,\psi_2. \tag{19}$$

The probability of LP-fast/slow having the best price and winning a request to deal is $\mathbb{T}_{fast} = \mathbb{T}_{slow} = \frac{1}{2}$*. The probability of a deal request getting rejected by LP-fast/slow for* $n > 1$ *is approximately:*

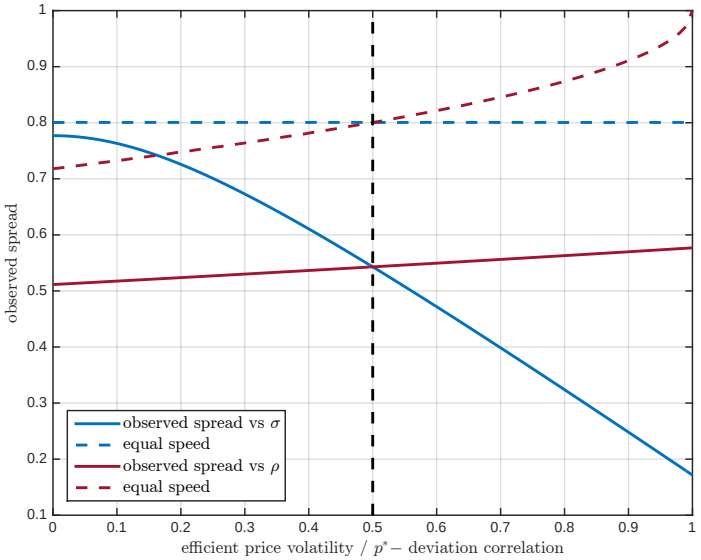
$$\mathbb{R}_{fast} \approx \Phi\left(\frac{\frac{(1-\beta^n)(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 - \delta}{\sqrt{n\sigma^2+\left(1-\beta^{2n}\right)\omega^2}}\right) \qquad \textit{and} \qquad \mathbb{R}_{slow} \approx \Phi\left(\frac{\frac{(1-\beta^n)(1-\rho\beta)\omega^2+\mu_{slow}}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 - \delta}{\sqrt{n\sigma^2+(1-\beta^{2n})\omega^2-\sigma^2_{slow}}}\right). \tag{20}$$

where $\mu_{slow} = \sigma^2 + \rho(\beta^{n-1} - \beta^{n+1})\omega^2$ *and* $\sigma^2_{slow} = \sigma^2 + \beta^{2n}(\beta^{-2}-1)\rho^2\omega^2$*. A lower bound for the effective spread charged by LP-fast/slow for* $n > 1$ *is:*

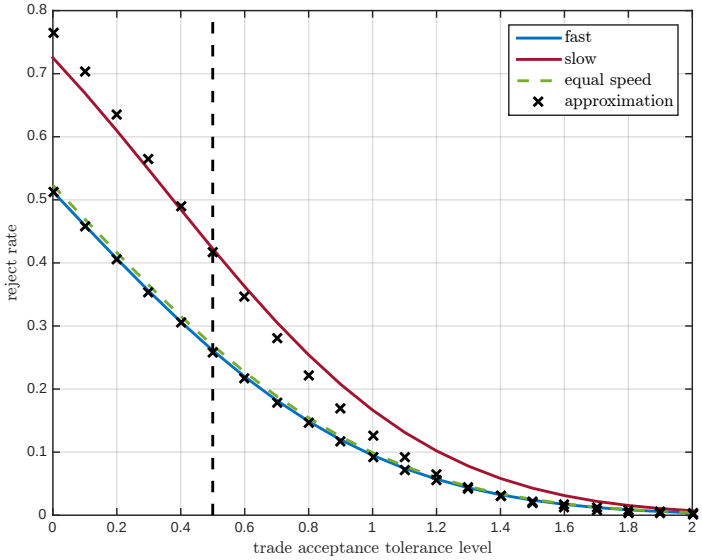
$$\begin{aligned}
\mathbb{S}_{fast} \;>\; & s - 2\frac{(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 \\
& + 2\frac{n\sigma^2}{n\sigma^2+(1-\beta^{2n})\omega^2}\,G\left(\frac{(1-\beta^n)(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 - \delta,\, n\sigma^2+(1-\beta^{2n})\omega^2\right), \qquad (21) \\
\mathbb{S}_{slow} \;>\; & s - 2\frac{(1-\rho\beta)\omega^2+\sigma^2}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 \\
& + 2\frac{(n-1)\sigma^2}{n\sigma^2+(1-\beta^{2n})\omega^2-\sigma^2_{slow}}\,G\left(\frac{(1-\beta^n)(1-\rho\beta)\omega^2+\mu_{slow}}{\sqrt{(1-\rho\beta)\omega^2+\sigma^2/2}}\psi_2 - \delta,\, n\sigma^2+(1-\beta^{2n})\omega^2-\sigma^2_{slow}\right). \qquad (22)
\end{aligned}$$

Figure 7: Execution metrics when speed of price discovery differs between LPs

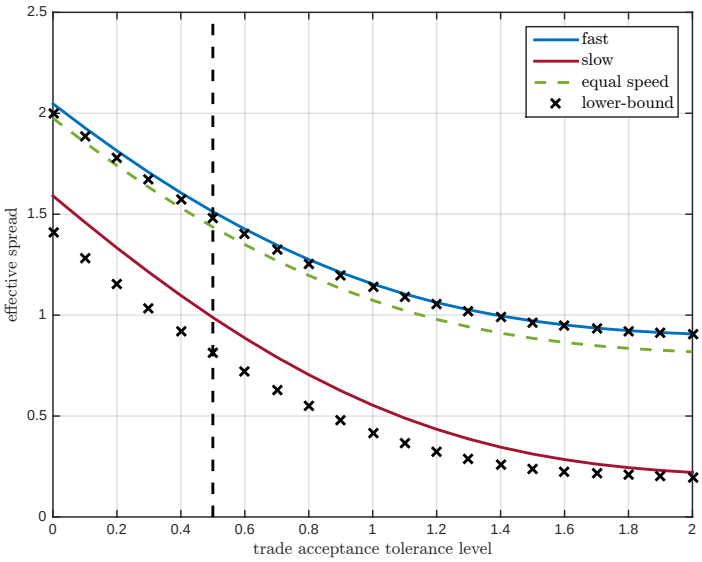
Panel A : observed spread S vs σ and ρ



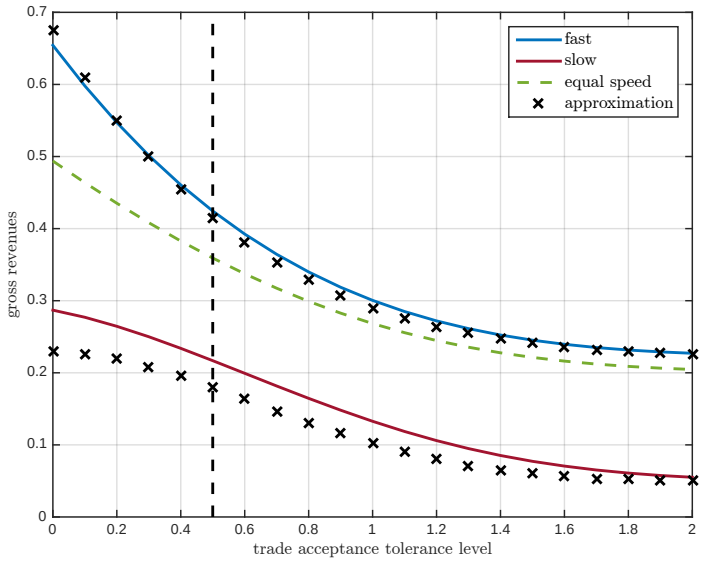
Panel B : reject probability $\mathbb{R}$ vs δ



Panel C : effective spread $\mathbb{S}$ vs δ



Panel D : gross revenues $\mathbb{W}$ vs δ



Note. Panel A draws the observed spread as a function of σ and ρ, and Panels B–D draw the reject rate, effective spread, and gross revenues as a function of tolerance level δ. The baseline model parameters are set at $s = 1, \sigma = 0.5, \omega = 0.25, \beta = 0.75, \rho = 0.5, N = 2, n = 2, \delta = s/2$. The solid lines are based on simulations whereas the markers (×) are, or follow directly from, the analytical approximations given in Proposition 6.

Proof See Appendix B. ■

Comparing Propositions 1 and 6 it is clear that the observed spread S is always tighter when there are latency differentials between LPs than when there are not, and this effect increases with the efficient price volatility (σ) and the persistence of p^*-deviations (β). LP-slow is lagging behind in price discovery and so in turbulent markets its pricing error will increase in magnitude and any corrections take longer with higher β. In isolation this would be inconsequential as the trader's flow is random, but when LP-slow is aggregated alongside LP-fast the discrepancies between their prices is larger than it would be without latency differentials, hence the stated impact on observed spread. As an aside, note that even when $\rho = 1$ and the LPs stream identical prices, the observed spread is still tighter than the nominal spread charged by the LPs because of the time lag in LP-slow's prices (Figure 7, Panel A). Recognising that LP-slow essentially has less accurate and more noisy prices than LP-fast, it is intuitive that its reject rate is higher, effective spread lower, and gross revenues suffer (Figure 7, Panels B-D).

Because the differences in trading metrics across LPs diminish as n grows, the latency buffer can be used to mitigate the "handicap" of LP-slow (trading costs need not be affected by this as δ and/or s can be adjusted accordingly). It is this kind of observation that has motivated the deliberate introduction of artificial latencies into numerous trading platforms (e.g. currency platforms ParFX, EBS, Reuters, and equity venue IEX) with the aim to level the playing field and to stop the technology arms race where participants seek to exploit technological anomalies of the platform to gain an advantage.

5 Full-amount versus stack-sweep execution

In the analysis above, the trader is assumed to require a standard amount of liquidity – i.e. the amount that each LP provides individually – and it is therefore feasible that she executes with a single LP that has the best price in the aggregator. I now consider the execution of larger amounts and focus in particular on how execution style and hedging strategy of the participating LPs impacts on transaction costs.

For larger amounts, there are two primary execution styles that the trader can adopt. She can divide up the order into a multiple of standard amounts and spread execution of these across as many providers at their best prices. This is so-called "stack-sweep" execution. Alternatively, she can request all LPs to provide additional liquidity up to the required amount and then execute the entire order with the single LP that provides the best price in the that amount. This is typically referred to as "full-amount" execution. Which of these two strategies is best? In practice that clearly depends on a variety of factors, including trader preferences and objective function, the agreed terms of liquidity provision, and specific pricing at the time of execution. Another important consideration is the impact

that execution style may have on the LPs' hedging behaviour. I will explore this last point in some detail using a stylised model for hedging costs, market impact, and speed-to-market.

Let $c_i = c_i^X - c_i^I$ denote the cost for LP-i to hedge an accepted trader's deal request via externalisation minus the cost of hedging via internalisation. For the purposes of this discussion, externalisation refers to the process of instantaneously hedging the trade one-for-one in external (inter-bank or public) markets. A (relatively) low cost of externalisation can be due to the LP's superior technology, speed to market, access to a diversified set of liquidity venues, smart order routing logic, and scale of trading operation for lower unit trading costs. Internalisation, on the other hand, refers to the process of absorbing the trade into the LP's inventory to then gradually reduce the risk position by attracting opposing interest from the LP's (private) client-base. Internalisation costs are driven by risk bearing capacity, scale of franchise operation for increased risk netting opportunities, and smart risk management logic.[6]

To simplify exposition, I do not explicitly link the LP's hedging costs to the liquidity it is able to offer, although in practice these are of course tightly coupled. Instead, I only consider the relative costs of hedging because this is what determines their hedging strategies. Specifically, if $c_i > 0$ then LP—i is naturally inclined to internalise and if $c_j < 0$ then LP-j will prefer to externalise. Note that it is possible for the LPs to have different relative costs (i.e. $c_i \neq c_j$) while the absolute costs of their preferred hedging strategy are the same (e.g. $c_i^I = c_j^X$) and both can offer equally competitive pricing to the trader.

Regarding market impact, I assume that externalisation creates an instantaneous and permanent price impact in the direction traded of θ per unit amount. Internalisation attracts negligible impact. This is a simplification of course: there is an extensive literature that distinguishes permanent from temporary impact, and similarly, it is natural to expect that internalisation will – over time – incur some market impact as the LP seeks to reduce its risk position. However, over short time scales and considering that the trader's liquidity demand is exogenously

[6]The use of internalisation as a risk management methodology is highlighted in Bank of England, H.M. Treasury, and Financial Conduct Authority (2014, p. 59): "This has led to an increase in "internalisation" in the spot FX markets where banks are able to match off client orders internally without having to go to the inter-dealer market to hedge their risk. Market participants have indicated that some dealers with large enough market share can now internalise up to 90% of their client orders in major currency pairs." In contrast, an example of a business model centred around externalisation is Virtu's: "Our strategies are also designed to lock in returns through precise and nearly instantaneous hedging, as we seek to eliminate the price risk in any positions held." (see Virtu Financial, Inc, 2014, p. 2). It is important to point out that the internalisation/externalisation classification is not simply one of bank versus non-bank LPs because there are banks that externalise significant portions of their flow and funds that actively internalise. In practice, the hedging approach adopted by any LP will lie somewhere along the spectrum from pure externaliser to pure internaliser, and may also vary by – for instance – market conditions and flow characteristics.

Table 2: Speed-to-market in a race amongst 3 LPs

	LP–1	LP–2	LP–3
1st to market	$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}$	$\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}$	$\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}$
2nd to market	$\frac{\lambda_1^2(\lambda_2+\lambda_3)+\lambda_1(\lambda_2^2+\lambda_3^2)}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_1+\lambda_2)(\lambda_1+\lambda_3)}$	$\frac{\lambda_2^2(\lambda_1+\lambda_3)+\lambda_2(\lambda_1^2+\lambda_3^2)}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_2+\lambda_3)(\lambda_1+\lambda_2)}$	$\frac{\lambda_3^2(\lambda_1+\lambda_2)+\lambda_3(\lambda_1^2+\lambda_2^2)}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}$
3rd to market	$\frac{\lambda_2\lambda_3(\lambda_2+\lambda_3)+2\lambda_1\lambda_2\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_1+\lambda_2)(\lambda_1+\lambda_3)}$	$\frac{\lambda_1\lambda_3(\lambda_1+\lambda_3)+2\lambda_1\lambda_2\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_2+\lambda_3)(\lambda_1+\lambda_2)}$	$\frac{\lambda_1\lambda_2(\lambda_1+\lambda_2)+2\lambda_1\lambda_2\lambda_3}{(\lambda_1+\lambda_2+\lambda_3)(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}$

Note. This table reports the probability of LP-i reaching the external market in first, second, or third place in a race amongst three competing liquidity providers. The distribution of speed-to-market is given by Eq. (23).

motivated, the assumption is justifiable.[7]

The third and final piece of the model specifies τ_i: the time it takes for LP-i to access the market when it decides to externalise. I assume this is an i.i.d. random variable with distribution:

$$\Pr(\tau_i < \tau) = 1 - e^{-\lambda_i \tau}. \tag{23}$$

The expected time-to-market for LP-i is $1/\lambda_i$ and so the higher the λ the quicker the LP. The purpose of this component of the model is to introduce a time-ordering amongst competing LPs when several want to externalise at the same time, for instance in response to a stack-sweep execution. It is only the relative values of τ_i (and λ_i) that matter, not their absolute values (in practice, differences in τ_i are typically of the order of milli-seconds, if not micro-seconds). The analysis below will require calculation of the probability that LP-i arrives to market in j^{th} place in a race to externalise with N LPs. I denote this probability by $P_{i,j}^{(N)}$ and Table 2 provides explicit expressions for the case where $N = 3$.

5.1 Equilibrium hedging strategy

The optimal hedging strategy for full-amount execution is trivial: the LP that wins the deal request knows that it is for the full amount and the cost-minimising strategy is to internalise if $c_i > 0$ and externalise otherwise. The

[7] Externalisation – as defined here – involves instantaneous hedging. Of course it is also possible to externalise via gradual hedging in a way that minimises market impact and makes it observationally indistinguishable from internalisation. Consequently, LPs that follow such a strategy should be considered internalisers in the context of this paper.

stack-sweep scenario is more interesting. To simplify exposition, I assume that $N = 3$ and $c_1 > 0, c_2 > 0, c_3 < 0$. This provides all the key insights and avoids any unnecessary complexity. The LPs need to decide on their preferred hedging approach and they do so simultaneously by aiming to minimise their costs while taking the actions of the other LPs as given. A Nash equilibrium is reached when none of the LPs have an incentive to – unilaterally – change their hedging decision.

Let's start by assuming that LP-3 will externalise (c_3 is negative after all). LP-1&2's intention is to internalise but they may change their approach conditional on LP-3's strategy. Individually, they will evaluate the following condition and decide to externalise if:

$$c_i + \theta P_{i,2}^{(2)} < \theta \qquad \text{for} \qquad i \in \{1,2\}. \tag{24}$$

The right-hand side of Eq. (24) measures the cost associated with an internalisation strategy: a guaranteed market impact cost of θ imposed on LP-i by the externalisation strategy of LP-3. The left-hand side measures the costs associated with switching to externalisation: LP-i incurs a cost of c_i but will now avoid the market impact cost with probability $1 - P_{i,2}^{(2)}$ when he reaches the external market ahead of LP-3. If neither LP-1 or LP-2 decide to externalise then an equilibrium is reached. If only one of LP-1 or LP-2 decide to externalise, then the other needs to re-consider his strategy as he now faces 2θ of market impact costs. Conditional on the other two LPs externalising, LP-i will now also externalise if the below condition is satisfied:

$$c_i + \theta P_{i,2}^{(3)} + 2\theta P_{i,3}^{(3)} < 2\theta \qquad \text{for} \qquad i = 1 \text{ or } 2. \tag{25}$$

See Appendix C for explicit expressions of Eqs. (24 – 25).

Using the above conditions, the equilibrium strategies can be mapped out in the parameter space of λ_1 and λ_2, for given $\{c_1, c_2, c_3, \lambda_3, \theta\}$. Panel A of Figure 8 and Table 3 provides an illustration, setting $c_1 = 0.100$, $c_2 = 0.075$, $c_3 = -0.250$, $\lambda_3 = 10$ and $\theta = 0.2$.

In scenario I of Table 3 (contained in the ■ region of Figure 8), both LP-1 and LP-2 are slower than LP-3 with $\lambda_1 = \lambda_2 = 5$: in a race with LP-3 alone LP-i will come second with probability $\lambda_3/(\lambda_i + \lambda_3) = \frac{2}{3}$. LP-1's choice is to (i) internalise and incur $\theta = 0.2$ of impact costs or (ii) pay $c_1 = 0.100$ to externalise and save θ with probability $\frac{1}{3}$. Because $0.2 < 0.1 + 0.2 \times \frac{2}{3} = 0.233$, LP-1 decides to internalise. The same applies to LP-2. Because LP-3 is better off by $-c_3$ compared to internalisation, no LP is inclined to change its hedging decision and equilibrium is reached.

In scenario II (contained in the ■ region of Figure 8), LP-1 is sufficiently quick to make him want to join the race to externalise: with a probability of 71% he'll reach the market before LP-3. The equilibrium is one where LP-1 reduces his impact costs to less than θ, LP-3 is still better off compared to internalisation. LP-2 now incurs 2θ of market impact but his costs would increase further were he to join the race to externalise.

Table 3: Incremental hedging costs associated with stack-sweep execution

	all	externalise				all	externalise			
	internalise	LP-3	LP-1/3	LP-2/3	all	internalise	LP-3	LP-1/3	LP-2/3	all
Scenario I: $\lambda_1 = 5, \lambda_2 = 5$ ($\lambda_3 = 10$)						*Scenario IV:* $\lambda_1 = 15, \lambda_2 = 10$ ($\lambda_3 = 10$)				
LP-1	0.000	**0.200**	0.233	0.400	0.333	0.000	0.200	0.180	0.400	**0.260**
LP-2	0.000	**0.200**	0.400	0.208	0.308	0.000	0.200	0.400	0.175	**0.295**
LP-3	0.250	**0.000**	0.067	0.067	0.133	0.250	0.000	0.120	0.100	**0.220**
Scenario II: $\lambda_1 = 25, \lambda_2 = 2$ ($\lambda_3 = 10$)						*Scenario V:* $\lambda_1 = 8, \lambda_2 = 10$ ($\lambda_3 = 10$)				
LP-1	0.000	0.200	**0.157**	0.400	0.172	0.000	0.200	0.211	0.400	**0.322**
LP-2	0.000	0.200	**0.400**	0.242	0.427	0.000	0.200	0.400	0.175	**0.264**
LP-3	0.250	0.000	**0.143**	0.033	0.176	0.250	0.000	0.089	0.100	**0.189**
Scenario III: $\lambda_1 = 2, \lambda_2 = 10$ ($\lambda_3 = 10$)						*Scenario VI:* $\lambda_1 = 20, \lambda_2 = 25$ ($\lambda_3 = 10$)				
LP-1	0.000	0.200	0.267	**0.400**	0.433	0.000	0.200	0.167	0.400	**0.278**
LP-2	0.000	0.200	0.400	**0.175**	0.208	0.000	0.200	0.400	0.132	**0.221**
LP-3	0.250	0.000	0.033	**0.100**	0.133	0.250	0.000	0.133	0.143	**0.276**

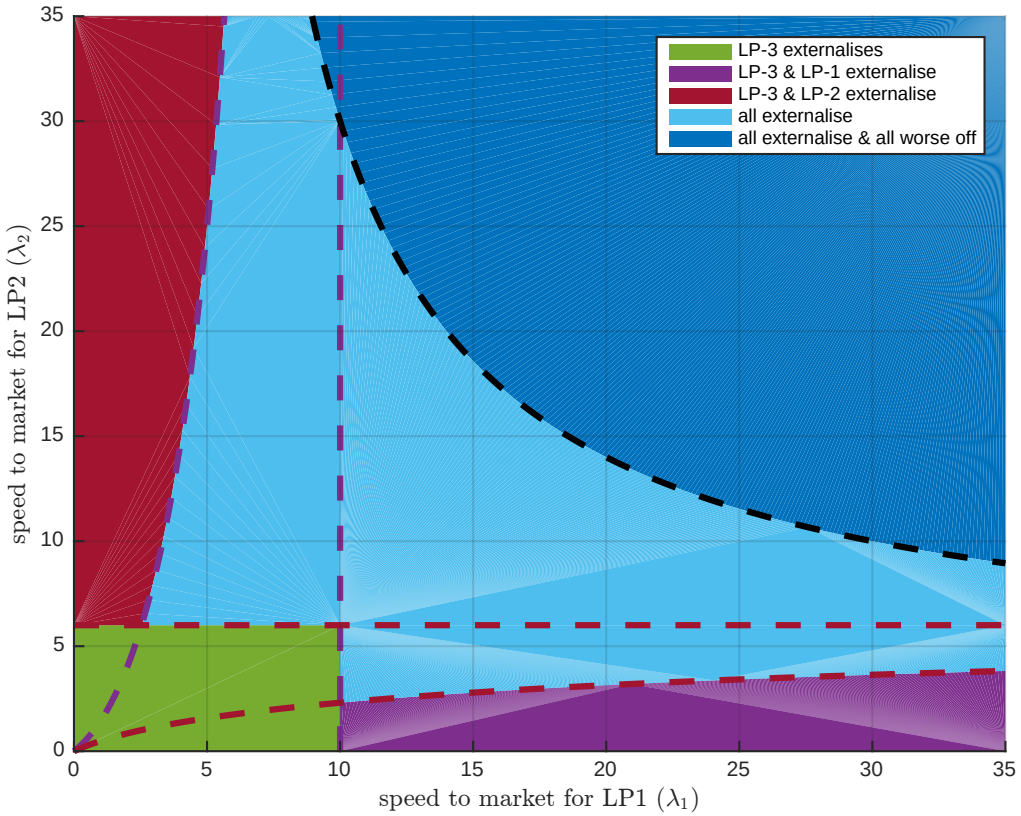
Note. This table reports the additional hedging costs incurred by the LPs when the trader uses stack-sweep execution and the LPs hedging decisions are as indicated in the table. The scenarios vary λ_1 and λ_2 while keeping fixed $\lambda_3 = 10$ and $c_1 = 0.100$, $c_2 = 0.075$, $c_3 = -0.250$, $\theta = 0.2$. The equilibrium states are highlighted in bold with colour coding consistent with Figure 8.

In scenario IV (contained in the ■ region of Figure 8), both LP-1 and LP-2 individually decide to join the race. The equilibrium is reached with all three LPs externalising the trader's flow. In the related scenario V, the condition in Eq. (24) is not satisfied for LP-1, i.e. the actions of LP-3 in isolation are not sufficient to make LP-1 change his default hedging strategy. However, the condition is satisfied for LP-2 which leads LP-1 to re-evaluate and externalise. This nicely illustrates the cascading nature of the hedging decisions: LP-2 only externalises because LP-3 does, and LP-1 only externalises because LP-2 and LP-3 do, ending up in a state where all LPs externalise when only one of them is naturally inclined to do so.

5.2 Discussion

Are the equilibrium hedging strategies described above "optimal"? The simple answer is that in some instances they are clearly not. Consider for instance Scenario IV where all LPs decide to externalise. Here the costs incurred

Figure 8: Equilibrium hedging strategies with stack-sweep execution



Note. This chart maps out the equilibrium hedging strategies as a function of λ_1 and λ_2 while keeping fixed $\lambda_3 = 10$ and $c_1 = 0.100, c_2 = 0.075, c_3 = -0.250, \theta = 0.2$.

by every LP are higher compared to the situation where only LP-3 externalises. Similarly, Scenario VI (contained in the ■ region of Figure 8) illustrates the case where LP-3 triggers a race to externalise amongst all participating LPs but given the speed of its competitors everyone ends up worse off compared to the situation where all internalise. The region defined by the equation below is one where LP-3 ends up bearing higher costs – due to LP-1 and LP-2 imposing significant market impact on him when they reach the market first – than if he had been able to commit to internalise and not forced LP-1&2 to externalise.

$$\theta P^{(3)}_{3,2} + 2\theta P^{(3)}_{3,3} > -c_3. \tag{26}$$

The equilibrium states described above, even those that are clearly inferior, are stable because no LP is incentivised to *unilaterally* change its decision. An agreement amongst the LPs to all internalise in Scenario VI won't hold because LP-3 can reduce its costs by non-compliance: he is incentivised to break the agreement short of any commitment device. This is the classical Prisoner's dilemma. The optimal "all-internalise" state in the ■ region in Figure 8 may

be reached in a repeated game with tit for tat strategies. This is beyond the scope of the current paper.

Turning to the choice of execution style, what is the trader advised to do? It is instructive to consider a baseline setup where the trader aggregates liquidity only from LP-1 and LP-2. In that case, the flow will be fully internalised irrespective of whether she executes on a full-amount basis or via stack-sweep. Now suppose the trader adds LP-3 into the aggregator. If she executes full amount, then LP-1 and LP-2 can continue to internalise and so any deals they win will attract minimal price impact. LP-3 will externalise and create market impact, but because the other LPs don't participate in the deal request they are unaffected. If, on the other hand, the trader executes via stack-sweep, after adding LP-3, a race amongst all participating LPs to externalise the flow may ensue – the highest aggregate cost equilibrium state (0.775 in the example in Table 3, compared to 0.250 for all-internalise or 0.400 for only LP-3 externalises). The trader therefore maximises its footprint and spreads are likely to widen due to the increased costs imposed on the LPs.

Taking this one step further, if the trader's flow is genuinely uninformed, it is unlikely she will want to interact with an externalising LP. Because trading in public venues is anonymous (the matching parties face a central clearing house), the externalising LP will pay a premium on any aggressive executions in the form of an adverse selection component embedded in the spread. Despite the LP hedging uninformed flow, the maker cannot distinguish between informed and uninformed aggressors due to the venue enforced anonymity of counterparts. The LP that internalises the flow avoids this premium and can reflect that in the nominal and effective spread charged. This reasoning then suggests that the trader should only include externalising LPs into the aggregator if her flow is informed and the value of the information content exceeds the adverse selection premium charged on-exchange. The LP becomes a route to market for the trader and the liquidity it offers is of no intrinsic value. Uninformed flow therefore gravitates towards over-the-counter markets with pre-deal counterparty transparency (where the trader can reveal it's "type") and flow that is sufficiently informed ends up on anonymous public markets.

Finally, consider the impact of one LP's hedging decision on the reject rate of another. If LP3 is the only one to externalise, then the instantaneous market impact θ created translates into an effective lowering of the other LPs' trade acceptance tolerance levels from δ to $\delta - \theta$. An example with conservative parameters illustrates the impact can be significant: with $s = 1, \omega = 0.25, \sigma = 0.50, \beta = 0.75, \rho = 0.50, N = 3, \delta = 0.5, n = 1$ the baseline reject rate is 19%, but with one externalising LP the reject rate of the other LPs will jump to 31%. If two LPs externalise and create a combined impact of 2θ, the reject rate of the other LP further increases to 45%.

6 Conclusion

This paper studies the properties of execution in an aggregator where multiple liquidity providers (LPs) compete on price for a trader's uninformed flow. Within the context of a simple model, I analyse the effective spread as a representative measure of the trader's all-in transaction costs and show how it is determined by a combination of factors, including (i) the number of LPs included in the aggregator, (ii) the type of LPs selected, (iii) the trader's execution style, (iv) the nominal spread charged by each LP, (v) the LPs' trade acceptance criteria as well as (vi) intrinsic characteristics of the LPs such as the quality of price discovery and (vii) market volatility. The results highlight intricate dependencies amongst the LPs' liquidity provision (e.g. a spread tightening by one LP can impact the reject rate and revenues of another) and the fragility of execution costs to the type of LPs included in the aggregator (e.g. the addition of a single LP reluctant to internalise the trader's flow, can lead to an equilibrium where all LPs externalise and market impact and collective hedging costs are maximised). The paper makes explicit that best-price execution doesn't necessarily lead to the lowest all-in transaction costs and provides traders with a framework to analyse and evaluate their aggregator design.

How do the theoretical predictions made in this paper translate into practice? A key message of the paper is that transaction costs are not necessarily lowered by increasing the number of LPs included in the aggregator: with many competing LPs the observed spread in the aggregator will certainly be tighter (and can even go negative) than with fewer LPs, but then the nominal spread and trade acceptance criteria will counterbalance this. So from a theoretical perspective, the number of LPs can be entirely inconsequential in that the same transaction costs can be achieved with many or with few LPs. In practice, there will be additional considerations. For instance, individual LPs may have particular strengths and weaknesses and the nature of their liquidity offering may vary (e.g. in terms of instrument coverage, service levels, platform functionality, amount of liquidity it can offer, etc) so combining a few can have benefits. Also, a trader may want a minimum number of LPs to participate in the aggregator to ensure resiliency or to satisfy internal execution guidelines. At the same time, a large number of LPs may be undesirable as relationship management can get costly, the economic incentives for individual LPs are reduced, and with the trader less reliant on an individual LP their liquidity provision may become less consistent.

The results presented throughout assume the trader to be entirely uninformed. Despite this, strong adverse selection effects can arise: competition for flow by the LPs combined with a best-price execution strategy on the trader's behalf means that the LP that wins the deal request will suffer from the Winner's curse. In practice, there will be traders that execute for exogenous liquidity reasons, but equally there will be those that act opportunistically based on short-term price predictions and those that seek to exploit temporary mis-pricing or pursue latency arbi-

trage opportunities. Depending on the nature of the trader's activity, the adverse selection effects described in the paper may be magnified and increase the importance of the last-look trade acceptance process as a defensive measure. The paper also shows that stack-sweep execution in an aggregator with participating LPs keen to externalise is unambiguously detrimental to transaction costs for an uninformed trader. For informed traders this conclusion may change, but such an analysis is beyond the scope of this paper and left for future research.

A Preliminaries

I first state some results on conditional expectations of normal random variables that will be used below. Let (x,y) be bi-variate normal with mean zero, variances σ_x^2,σ_y^2, and correlation ρ. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution function of a standard normal random variable. For $\sigma_x = 1$, we have

$$E(x \mid x>\delta) = \frac{E(X I_{(x>\delta)})}{E(I_{(x>\delta)})} = \frac{\int_\delta^\infty x\phi(x)dx}{1-\Phi(\delta)} = \frac{-\phi(x)\mid_\delta^\infty}{1-\Phi(\delta)} = \frac{\phi(\delta)}{1-\Phi(\delta)}.$$

And so for arbitrary σ_x we have

$$E(x \mid x>\delta) = \sigma_x E\left(x\sigma_x^{-1} \mid x\sigma_x^{-1}>\delta\sigma_x^{-1}\right) = \sigma_x \frac{\phi(\delta/\sigma_x)}{1-\Phi(\delta/\sigma_x)} \equiv G(\delta,\sigma_x^2), \tag{27}$$

where the function G is defined for notational convenience. Note that $G(0,\sigma^2) = \sigma\sqrt{2/\pi} = E(\mid x \mid)$.

$$E(\mid a+x \mid) = E(a+x \mid a+x>0)\Pr(a+x>0) - E(a+x \mid a+x<0)\Pr(a+x<0) = 2a\Phi(a/\sigma_x) + 2\sigma_x\phi(a/\sigma_x) - a. \tag{28}$$

For a bi-variate normal,

$$E(x \mid y>\delta) = E\left(E_y(x \mid y) \mid y>\delta\right) = \frac{\rho\sigma_x\sigma_y}{\sigma_y^2}E(y \mid y>\delta) = \frac{\rho\sigma_x\sigma_y}{\sigma_y}\frac{\phi(\delta/\sigma_y)}{1-\Phi(\delta/\sigma_y)}. \tag{29}$$

Using a change of variables and Eq. (29),

$$E(x \mid x>y) = E(x \mid z>0) = 2\phi(0)\frac{\sigma_x^2-\rho\sigma_x\sigma_y}{\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}} = \sqrt{2/\pi}\frac{\sigma_x^2-\rho\sigma_x\sigma_y}{\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}}, \tag{30}$$

where $z = x-y$ and noting that x,z are jointly normal, with $E(xz) = \sigma_x^2-\rho\sigma_x\sigma_y$ and $E(z^2) = \sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y$.

$$E(\max(x,y)) = E\left(\frac{x+y}{2}+\frac{\mid x-y \mid}{2}\right) = \frac{1}{2}E(\mid x-y \mid) = (2\pi)^{-1/2}\sqrt{\sigma_x^2+\sigma_y^2-2\rho\sigma_x\sigma_y}. \tag{31}$$

Also note that $E(\max(x,y)) = -E(\min(x,y))$ and so $E(\max(x,y)-\min(x,y)) = 2E(\max(x,y))$. For tri-variate normal (x,y,z) we have

$$\begin{aligned} E(\max(x,y,z)-\min(x,y,z)) &= \frac{1}{2}E\left(\mid x-y \mid + \mid x-z \mid + \mid y-z \mid\right), \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\sigma_x^2+\sigma_y^2-2\rho_{x,y}\sigma_x\sigma_y} + \frac{1}{\sqrt{2\pi}}\sqrt{\sigma_x^2+\sigma_z^2-2\rho_{x,z}\sigma_x\sigma_z} \\ &\quad + \frac{1}{\sqrt{2\pi}}\sqrt{\sigma_y^2+\sigma_z^2-2\rho_{y,z}\sigma_y\sigma_z}. \end{aligned} \tag{32}$$

B Proofs

Proof of Proposition 1. From $s_i = s$ it follows that $S = E(\underline{a}_t - \overline{b}_t) = s + E(\min_i\{m_t^{(i)}\}_{i=1}^N - \max_i\{m_t^{(i)}\}_{i=1}^N) = s - 2E(\max_i\{m_t^{(i)}\}_{i=1}^N)$. The unconditional variance of the measurement error is $V(m_t^{(i)}) = \omega^2$ and independent of β by specification of the process. Because $\beta_i = \beta$, the correlation amongst the measurement error processes is ρ and also independent of β. Now note that

$\{m_t^{(i)}\}_{i=1}^N \stackrel{d}{=} \{x_i\}_{i=1}^N$ where $x_i = \omega(\rho u_0 + \sqrt{1-\rho}\, u_i)$ and u are independent standard normal. Therefore, $E(\max_i\{m_t^{(i)}\}_{i=1}^N) = E(\max_i\{x_i\}_{i=1}^N) = \omega\sqrt{1-\rho}\, E(\max_i\{u_i\}_{i=1}^N)$. See Aksomaitis and Burauskaitė-Harju (2009) for more details. Berman (1964) shows that for large N, $\psi_N \propto \sqrt{\log N}$, from which it follows that $\partial^2\psi_N/\partial N^2 < 0$. ■

Proof of Proposition 2. For the model defined by Eqs. (1 – 4), it is easy to see that $V_h = s/2 + E(m_{t+h}^{(i)} - m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)})$. The expectation can be worked out as follows:

$$\begin{aligned} E(m_{t+h}^{(i)} - m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}) &= E((\beta^h - 1)m_t^{(i)} + \sum_{j=0}^{h-1}\beta^j \eta_{t+h-j} \mid m_t^{(i)} > m_t^{(\neq i)}), \\ &= (\beta^h - 1)E(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}), \\ &= (\beta^h - 1)E(\max_i\{m_t^{(i)}\}_{i=1}^N), \\ &= (\beta^h - 1)\omega\sqrt{1-\rho}\,\psi_N. \end{aligned} \tag{33}$$

■

Proof of Proposition 3. Starting with the definition of the reject rate in Eq. (10), note that:

$$\begin{aligned} \mathbb{R} &= \Pr(b_{t+n}^{(i)} - b_t^{(i)} < -\delta \mid b_t^{(i)} > b_t^{(\neq i)}), \\ &= \Pr\left(p_{t+n}^* - p_t^* + m_{t+n}^{(i)} - m_t^{(i)} < -\delta \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\ &= \Pr\left(\sum_{j=1}^{n}\varepsilon_{t+j} + \sum_{j=0}^{n-1}\beta^j \eta_{t+n-j}^{(i)} + (\beta^n - 1)m_t^{(i)} < -\delta \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\ &= \Phi\left(\frac{(1-\beta^n)m_t^{(i)} - \delta}{\sqrt{n\sigma^2 + (1-\beta^{2n})\omega^2}} \mid m_t^{(i)} > m_t^{(\neq i)}\right). \end{aligned} \tag{34}$$

To obtain the unconditional probability of a reject, one would need to integrate out the random variable $m_t^{(i)}$ conditioned on $m_t^{(i)} > m_t^{(\neq i)}$. This can be done numerically but it is not analytically tractable. An approximation can be obtained by replacing the measurement error by its conditional expectation, i.e.

$$\mathbb{R} \approx \Phi\left(\frac{(1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N - \delta}{\sqrt{n\sigma^2 + (1-\beta^{2n})\omega^2}}\right). \tag{35}$$

Because the function $\Phi(x)$ is convex for $x < 0$ and concave for $x > 0$, by Jensen's inequality, the above approximation will constitute an upper bound (lower bound) when the numerator is sufficiently positive (negative). ■

Proof of Proposition 4. I first derive an expression for V_h, defined in Eq. (12). The effective spread then trivially follows from $\mathbb{S} = 2V_\infty$.

$$\begin{aligned} \mathbb{V}_h &= \frac{s}{2} + E\left(p_{t+h}^{(i)} - p_t^{(i)} \mid b_t^{(i)} > b_t^{(\neq i)}, b_{t+n}^{(i)} > b_t^{(i)} - \delta\right), \\ &= \frac{s}{2} + E\left((\beta^h - 1)m_t^{(i)} + \sum_{j=0}^{h-1}\beta^j \eta_{t+h-j}^{(i)} + \sum_{j=1}^{h}\varepsilon_{t+j} \mid m_t^{(i)} > m_t^{(\neq i)}, b_{t+n}^{(i)} > b_t^{(i)} - \delta\right). \end{aligned} \tag{36}$$

The conditional expectation of $m_t^{(i)}$ in Eq. (36) can be expressed as:

$$E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}, (1-\beta^n)m_t^{(i)} < \delta + \sum_{j=0}^{n-1}\beta^j \eta_{t+n-j}^{(i)} + \sum_{j=1}^{n}\varepsilon_{t+j}\right). \tag{37}$$

This is of the form $E(x \mid y < x < z)$ and analytically intractable. In order to still obtain a lower bound on $\mathbb{S}$, an upper bound on the conditional expectation of $m_t^{(i)}$ is required (because $\beta^h - 1 < 0$ multiplying $m_t^{(i)}$ in Eq. 36):

$$E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}, b_{t+n}^{(i)} > b_t^{(i)} - \delta\right) = E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}, m_t^{(i)} < m_{t+n}^{(i)} + \delta\right) < E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}\right) = \omega\sqrt{1-\rho}\,\psi_N. \tag{38}$$

A lower bound on the second and third term in Eq. (36) for $h \geq n$ is given by:

$$\begin{aligned}
&E\left(E\left(\sum_{j=1}^{h}\varepsilon_{t+j} + \sum_{j=0}^{h-1}\beta^j\eta_{t+h-j}^{(i)} \mid \sum_{j=1}^{n}\varepsilon_{t+j} + \sum_{j=0}^{n-1}\beta^j\eta_{t+n-j}^{(i)} > (1-\beta^n)m_t^{(i)} - \delta\right) \mid m_t^{(i)} > m_t^{(\neq i)}\right) \\
&= E\left(E\left(\sum_{j=1}^{n}\varepsilon_{t+j} + \beta^{h-n}\sum_{j=0}^{n-1}\beta^j\eta_{t+n-j}^{(i)} \mid \sum_{j=1}^{n}\varepsilon_{t+j} + \sum_{j=0}^{n-1}\beta^j\eta_{t+n-j}^{(i)} > (1-\beta^n)m_t^{(i)} - \delta\right) \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\
&= \frac{n\sigma^2 + \beta^{h-n}(1-\beta^{2n})\omega^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} E\left(G((1-\beta^n)m_t^{(i)} - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2) \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\
&> \frac{n\sigma^2 + \beta^{h-n}(1-\beta^{2n})\omega^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} G((1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2).
\end{aligned} \tag{39}$$

In the first step, I use that the conditional expectation of ε_{t+j} and η_{t+j} is zero for $j > n$. In the second step I use the result in Eq. (29), and the final step follows from Jensen's inequality and the convexity of G. Collecting terms, yields a lower bound on $\mathbb{V}_h$ and the associated expression for $\mathbb{S}$. ■

Proof of Proposition 5. The unconditional variance of the measurement error noise is $V(m_t^{(i)}) = \omega_i^2$ and independent of β by specification of the process. For $N = 2$, the unconditional correlation between the measurement error processes is:

$$\begin{aligned}
\rho_m &\equiv \frac{1}{\omega_1\omega_2}\lim_{n\to\infty} E_0(m_n^{(1)}m_n^{(2)}), \\
&= \frac{1}{\omega_1\omega_2}\lim_{n\to\infty} E\left(\beta_1^n m_0^{(1)} + \beta_2^n m_0^{(2)} + \left(\sum_{j=0}^{n-1}\beta_1^j\eta_{n-j}^{(1)}\right)\left(\sum_{j=0}^{n-1}\beta_2^j\eta_{n-j}^{(2)}\right)\right), \\
&= \frac{1}{\omega_1\omega_2}\lim_{n\to\infty} E\left(\sum_{j=0}^{n-1}(\beta_1\beta_2)^j\eta_{n-j}^{(1)}\eta_{n-j}^{(2)}\right), \\
&= \rho\frac{\sqrt{(1-\beta_1^2)(1-\beta_2^2)}}{1-\beta_1\beta_2}.
\end{aligned} \tag{40}$$

The observed spread $S = -2E(\max(m_t^{(1)} - \frac{s_1}{2}, m_t^{(2)} - \frac{s_2}{2})) = \frac{s_1+s_2}{2} - E(\mid m_t^{(1)} - m_t^{(2)} + \frac{s_2-s_1}{2} \mid)$ from which it follows using Eq. (28) that

$$S = s_2 - (s_2 - s_1)\Phi\left(\frac{s_2 - s_1}{2\sigma_{\Delta m}}\right) - 2\sigma_{\Delta m}\phi\left(\frac{s_2 - s_1}{2\sigma_{\Delta m}}\right), \tag{41}$$

where $\sigma_{\Delta m}^2 \equiv E((m_t^{(2)} - m_t^{(1)})^2) = \omega_1^2 + \omega_2^2 - 2\rho_m\omega_1\omega_2$. To obtain the probability of reject for LP-2, I use the same approach as in the proof of Proposition 3 by noting that:

$$\mathbb{R}_2 = \Pr(b_{t+n_2}^{(2)} - b_t^{(2)} < -\delta_2 \mid b_t^{(2)} > b_t^{(1)}) = \Phi\left(\frac{(1-\beta_2^{n_2})m_t^{(2)} - \delta_2}{\sqrt{n_2\sigma^2 + (1-\beta_2^{2n_2})\omega_2^2}} \mid b_t^{(2)} > b_t^{(1)}\right). \tag{42}$$

From Eqs. (29) and (40) it follows that

$$E(m_t^{(2)} \mid b_t^{(2)} > b_t^{(1)}) = E(m_t^{(2)} \mid m_t^{(2)} - m_t^{(1)} > \frac{1}{2}(s_2 - s_1)) = \frac{\omega_2^2 - \rho_m \omega_1 \omega_2}{\sigma_{\Delta m}^2} G(\frac{1}{2}(s_2 - s_1), \sigma_{\Delta m}^2). \tag{43}$$

Replacing $m_t^{(2)}$ in Eq. (42) by the above expectation yields the required approximation. The reject rate for LP-1 follows by symmetry.

Following the same approach as in the proof of Proposition 4, the value of a completed trade to LP-2 can be expressed as:

$$\mathbb{V}_{h,2} = \frac{s}{2} + E\left((\beta_2^h - 1) m_t^{(2)} + \sum_{j=0}^{h-1} \beta_2^j \eta_{t+h-j}^{(2)} + \sum_{j=1}^{h} \varepsilon_{t+j} \mid b_t^{(2)} > b_t^{(1)}, b_{t+n_2}^{(2)} > b_t^{(2)} - \delta_2 \right). \tag{44}$$

An upper bound for the first term in Eq. (44) is:

$$E\left(m_t^{(2)} \mid b_t^{(2)} > b_t^{(1)}, b_{t+n_2}^{(2)} > b_t^{(2)} - \delta_2\right) < E\left(m_t^{(2)} \mid b_t^{(2)} > b_t^{(1)}\right) = \frac{\omega_2^2 - \rho_m \omega_1 \omega_2}{\sigma_{\Delta m}^2} G(\frac{1}{2}(s_2 - s_1), \sigma_{\Delta m}^2). \tag{45}$$

A lower bound on the second and third term in Eq. (44) for $h \geq n_2$ is given by:

$$\begin{aligned}
&E\left(E\left(\sum_{j=0}^{h-1} \beta_2^j \eta_{t+h-j}^{(2)} + \sum_{j=1}^{h} \varepsilon_{t+j} \mid \sum_{j=0}^{n_2-1} \beta_2^j \eta_{t+n_2-j}^{(2)} + \sum_{j=1}^{n_2} \varepsilon_{t+j} > (1-\beta_2^{n_2}) m_t^{(2)} - \delta_2 \right) \mid m_t^{(2)} > m_t^{(1)} + \frac{s_2 - s_1}{2} \right) \\
&= \frac{n_2\sigma^2 + \beta^{h-n_2}(1-\beta_2^{2n_2})\omega_2^2}{n_2\sigma^2 + (1-\beta_2^{2n_2})\omega_2^2} E\left(G((1-\beta_2^{n_2}) m_t^{(2)} - \delta_2, n_2\sigma^2 + (1-\beta_2^{2n_2})\omega_2^2) \mid m_t^{(2)} > m_t^{(1)} + \frac{s_2 - s_1}{2} \right), \\
&> \frac{n_2\sigma^2 + \beta^{h-n_2}(1-\beta_2^{2n_2})\omega_2^2}{n_2\sigma^2 + (1-\beta_2^{2n_2})\omega_2^2} G\left((1-\beta_2^{n_2}) \frac{\omega_2^2 - \rho_m \omega_1 \omega_2}{\sigma_{\Delta m}^2} G(\frac{1}{2}(s_2 - s_1), \sigma_{\Delta m}^2) - \delta_2, n_2\sigma^2 + (1-\beta_2^{2n_2})\omega_2^2 \right),
\end{aligned} \tag{46}$$

Collecting terms, yields a lower bound on $\mathbb{V}_{h,2}$ and the associated expression for $\mathbb{S}_2$ (and $\mathbb{S}_1$ by symmetry). ■

Proof of Proposition 6. For notational convenience, I assume throughout that LP-1 is "fast" and LP-2 is "slow". The expression for the observed spread directly follows from Eq. (31):

$$S = -2E(\max(b_t^{(1)}, b_t^{(2)}) - p_t^*) = s - E(\mid m_t^{(1)} - m_{t-1}^{(2)} + \varepsilon_t \mid) = s - E(\mid \beta m_{t-1}^{(1)} + \eta_t^{(1)} - m_{t-1}^{(2)} + \varepsilon_t \mid) = s - 2\psi_2 \sqrt{\omega^2(1-\rho\beta) + \sigma^2/2}. \tag{47}$$

An expression for the probability of reject for LP-slow for $n > 1$ is derived as follows:

$$\begin{aligned}
\mathbb{R}_{\text{slow}} &= \Pr\left(b_{t+n}^{(2)} - b_t^{(2)} < -\delta \mid b_t^{(2)} > b_t^{(1)}\right), \\
&= \Pr\left(p_{t+n-1}^* + m_{t+n-1}^{(2)} - p_{t-1}^* - m_{t-1}^{(2)} < -\delta \mid p_{t-1}^* + m_{t-1}^{(2)} > p_t^* + m_t^{(1)}\right), \\
&= \Pr\left(\sum_{j=0}^{n-1} \varepsilon_{t+j} + \sum_{j=0}^{n-1} \beta^j \eta_{t+n-1-j}^{(2)} + \left(\beta^n - 1\right) m_{t-1}^{(2)} < -\delta \mid m_{t-1}^{(2)} > \varepsilon_t + \beta m_{t-1}^{(1)} + \eta_t^{(1)} \right), \\
&= \Pr\left(\sum_{j=1}^{n-1} \varepsilon_{t+j} + \sum_{j=0}^{n-2} \beta^j \eta_{t+n-1-j}^{(2)} + \beta^{n-1}\sqrt{1-\rho^2}\eta_t^* < \left(1-\beta^n\right) m_{t-1}^{(2)} - \delta - \varepsilon_t - \beta^{n-1}\rho\eta_t^{(1)} \mid m_{t-1}^{(2)} > \varepsilon_t + \beta m_{t-1}^{(1)} + \eta_t^{(1)} \right), \\
&= \Phi\left(\frac{\left(1-\beta^n\right) m_{t-1}^{(2)} - \delta - \varepsilon_t - \beta^{n-1}\rho\eta_t^{(1)}}{\sqrt{(n-1)\sigma^2 + (1-\beta^{2n})\omega^2 - \beta^{2n}(\beta^{-2}-1)\rho^2\omega^2}} \mid m_{t-1}^{(2)} > \varepsilon_t + \beta m_{t-1}^{(1)} + \eta_t^{(1)} \right), \\
&\approx \Phi\left(\frac{\frac{\left(1-\beta^n\right)\left(1-\rho\beta\right)\omega^2 + \mu_{\text{slow}}}{\psi_2^{-1}\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}} - \delta}{\sqrt{n\sigma^2 + (1-\beta^{2n})\omega^2 - \sigma_{\text{slow}}^2}} \right),
\end{aligned} \tag{48}$$

where $\mu_{\text{slow}} = \sigma^2 + \rho(\beta^{n-1} - \beta^{n+1})\omega^2$ and $\sigma^2_{\text{slow}} = \sigma^2 + \beta^{2n}(\beta^{-2}-1)\rho^2\omega^2$. In the third step, I use that $\eta^{(2)}_t \overset{d}{=} \sqrt{1-\rho^2}\eta^*_t + \rho\eta^{(1)}_t$ with $\eta^*_t \sim$ i.i.d. $\mathcal{N}(0,(1-\beta^2)\omega^2)$. The final step uses Eq. (30) to obtain the expressions below:

$$\begin{aligned}
E\left(m^{(2)}_{t-1} \mid m^{(2)}_{t-1} > \varepsilon_t + \beta m^{(1)}_{t-1} + \eta^{(1)}_t\right) &= \frac{(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2, \\
E\left(\varepsilon_t \mid \varepsilon_t < m^{(2)}_{t-1} - \beta m^{(1)}_{t-1} - \eta^{(1)}_t\right) &= -\frac{\sigma^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2, \\
E\left(\eta^{(1)}_t \mid \eta^{(1)}_t < m^{(2)}_{t-1} - \varepsilon_t - \beta m^{(1)}_{t-1}\right) &= -\frac{\left(1-\beta^2\right)\omega^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2.
\end{aligned}$$

Similarly, an expression for the probability of reject for LP-fast is derived as follows:

$$\begin{aligned}
\mathbb{R}_{\text{fast}} &= \Pr\left(b^{(1)}_{t+n} - b^{(1)}_t < -\delta \mid b^{(1)}_t > b^{(2)}_t\right), \\
&= \Pr\left(p^*_{t+n+1} - p^*_{t+1} + m^{(1)}_{t+n+1} + m^{(1)}_{t+1} < -\delta \mid p^*_{t+1} + m^{(1)}_{t+1} > p^*_t + m^{(2)}_t\right), \\
&= \Pr\left(\sum_{j=1}^{n}\varepsilon_{t+j+1} + \sum_{j=0}^{n-1}\beta^j\eta^{(1)}_{t+n-j+1} < \left(1-\beta^n\right)m^{(1)}_{t+1} - \delta \mid \beta m^{(1)}_t > m^{(2)}_t - \varepsilon_{t+1} - \eta^{(1)}_{t+1}\right), \\
&= \Phi\left(\frac{\left(1-\beta^n\right)\left(\beta m^{(1)}_t + \eta^{(1)}_{t+1}\right) - \delta}{\sqrt{n\sigma^2 + \left(1-\beta^{2n}\right)\omega^2}} \mid \beta m^{(1)}_t + \eta^{(1)}_{t+1} > m^{(2)}_t - \varepsilon_{t+1}\right), \\
&\approx \Phi\left(\frac{\frac{\left(1-\beta^n\right)\left(1-\rho\beta\right)\omega^2}{\psi_2^{-1}\sqrt{\left(1-\rho\beta\right)\omega^2+\sigma^2/2}} - \delta}{\sqrt{n\sigma^2 + \left(1-\beta^{2n}\right)\omega^2}}\right). \qquad (49)
\end{aligned}$$

The value of a completed trade to LP-slow for $n > 1$ is:

$$\begin{aligned}
\mathbb{S}_{\text{slow}} &= 2\lim_{h\to\infty}\mathbb{V}_{h,2}, \\
&= s + 2\lim_{h\to\infty}E\left(p^{(2)}_{t+h} - p^{(2)}_t \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right), \\
&= s + 2\lim_{h\to\infty}E\left((\beta^h - 1)m^{(2)}_{t-1} + \sum_{j=0}^{h-1}\beta^j\eta^{(2)}_{t+h-j-1} + \sum_{j=1}^{h}\varepsilon_{t+j-1} \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right), \\
&= s + 2E\left(\sum_{j=2}^{n}\varepsilon_{t+j-1} \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right) + 2E\left(\varepsilon_t \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right) - 2E\left(m^{(2)}_{t-1} \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right).
\end{aligned}$$

Working through each term separately:

$$\begin{aligned}
&E\left(\sum_{j=2}^{n}\varepsilon_{t+j-1} \mid b^{(2)}_t > b^{(1)}_t, b^{(2)}_{t+n} > b^{(2)}_t - \delta\right) \\
&= E\left(E\left(\sum_{j=2}^{n}\varepsilon_{t+j-1} \mid \beta^{n-1}\sqrt{1-\rho^2}\eta^*_t + \sum_{j=0}^{n-2}\beta^j\eta^{(2)}_{t+n-j-1} + \sum_{j=2}^{n}\varepsilon_{t+j-1} > (1-\beta^n)m^{(2)}_{t-1} - \delta - \varepsilon_t - \rho\beta^{n-1}\eta^{(1)}_t\right) \mid m^{(2)}_{t-1} > \beta m^{(1)}_{t-1} + \eta^{(1)}_t + \varepsilon_t\right), \\
&= \frac{(n-1)\sigma^2}{n\sigma^2 + (1-\beta^{2n})\omega^2 - \sigma^2_{\text{slow}}}E\left(G((1-\beta^n)m^{(2)}_{t-1} - \delta - \varepsilon_t - \rho\beta^{n-1}\eta^{(1)}_t, n\sigma^2 + (1-\beta^{2n})\omega^2 - \sigma^2_{\text{slow}} \mid m^{(2)}_{t-1} > \beta m^{(1)}_{t-1} + \eta^{(1)}_t + \varepsilon_t\right), \\
&> \frac{(n-1)\sigma^2}{n\sigma^2 + (1-\beta^{2n})\omega^2 - \sigma^2_{\text{slow}}}G\left(\frac{(1-\beta^n)(1-\rho\beta)\omega^2 + \mu_{\text{slow}}}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2 - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2 - \sigma^2_{\text{slow}}\right).
\end{aligned}$$

Next,

$$E\left(\varepsilon_t \mid b_t^{(2)} > b_t^{(1)}, b_{t+n}^{(2)} > b_t^{(2)} - \delta\right) > E\left(\varepsilon_t \mid b_t^{(2)} > b_t^{(1)}\right) = E\left(\varepsilon_t \mid m_{t-1}^{(2)} > \beta\, m_{t-1}^{(1)} + \eta_t^{(1)} + \varepsilon_t\right) = -\frac{\sigma^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2.$$

Finally,

$$E\left(m_{t-1}^{(2)} \mid b_t^{(2)} > b_t^{(1)}, b_{t+n}^{(2)} > b_t^{(2)} - \delta\right) < E\left(m_{t-1}^{(2)} \mid m_{t-1}^{(2)} > \beta\, m_{t-1}^{(1)} + \eta_{t-1}^{(1)} - \varepsilon_t\right) = \frac{(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2.$$

To obtain the value of a completed trade for the fast LP:

$$\begin{aligned}
\mathbb{S}_{\text{fast}} &= 2 \lim_{h\to\infty} \mathbb{V}_{h,1}^{\text{fast}} \\
&= s + 2 \lim_{h\to\infty} E\left(b_{t+h}^{(1)} - b_t^{(1)} \mid b_t^{(1)} > b_t^{(2)}, b_{t+n}^{(1)} > b_t^{(1)} - \delta\right), \\
&= s + 2E\left(-m_{t+1}^{(1)} + \sum_{j=1}^{n} \varepsilon_{t+j} \mid b_t^{(1)} > b_t^{(2)}, b_{t+n}^{(1)} > b_t^{(1)} - \delta\right).
\end{aligned}$$

An upper bound for the first term is:

$$E\left(m_t^{(1)} \mid b_t^{(1)} > b_t^{(2)}, b_{t+n}^{(1)} > b_t^{(1)} - \delta\right) < E\left(m_t^{(1)} \mid m_t^{(1)} > m_{t-1}^{(2)} - \varepsilon_t\right) = \frac{(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2. \tag{50}$$

A lower bound for the second term is:

$$\begin{aligned}
&E\left(E\left(\sum_{j=1}^{n} \varepsilon_{t+j} \mid \sum_{j=0}^{n-1} \beta^j \eta_{t+n-j}^{(1)} + \sum_{j=1}^{n} \varepsilon_{t+j} > (1-\beta^n) m_t^{(1)} - \delta\right) \mid b_t^{(1)} > b_t^{(2)}\right) \\
&= \frac{n\sigma^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} E\left(G((1-\beta^n) m_t^{(1)} - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2) \mid m_t^{(1)} > m_{t-1}^{(2)} - \varepsilon_t\right), \\
&> \frac{n\sigma^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} G\left(\frac{(1-\beta^n)(1-\rho\beta)\omega^2}{\sqrt{(1-\rho\beta)\omega^2 + \sigma^2/2}}\psi_2 - \delta, n\sigma^2 + (1-\beta^{2n})\omega^2\right).
\end{aligned}$$

■

C Stack execution equilibrium boundaries

The condition in Eq. (24) where LP-i is better off externalising when LP-3 does so, can be expressed as:

$$\lambda_i > c_i \frac{\lambda_3}{\theta - c_i} \qquad \text{for} \qquad i \in \{1,2\}$$

The condition in Eq. (25) where LP-i is better off externalising when LP-$\neq i$ and LP-3 do so, can be expressed as:

$$\lambda_i > \frac{(c_i - \theta)\left(\lambda_3 + \lambda_{\neq i}\right) + \sqrt{c_i\left(\lambda_3 - \lambda_{\neq i}\right)^2 (c_i - 2\theta) + \theta^2\left(\lambda_3 + \lambda_{\neq i}\right)^2}}{2(2\theta - c_i)} \qquad \text{for} \qquad i \in \{1,2\},$$

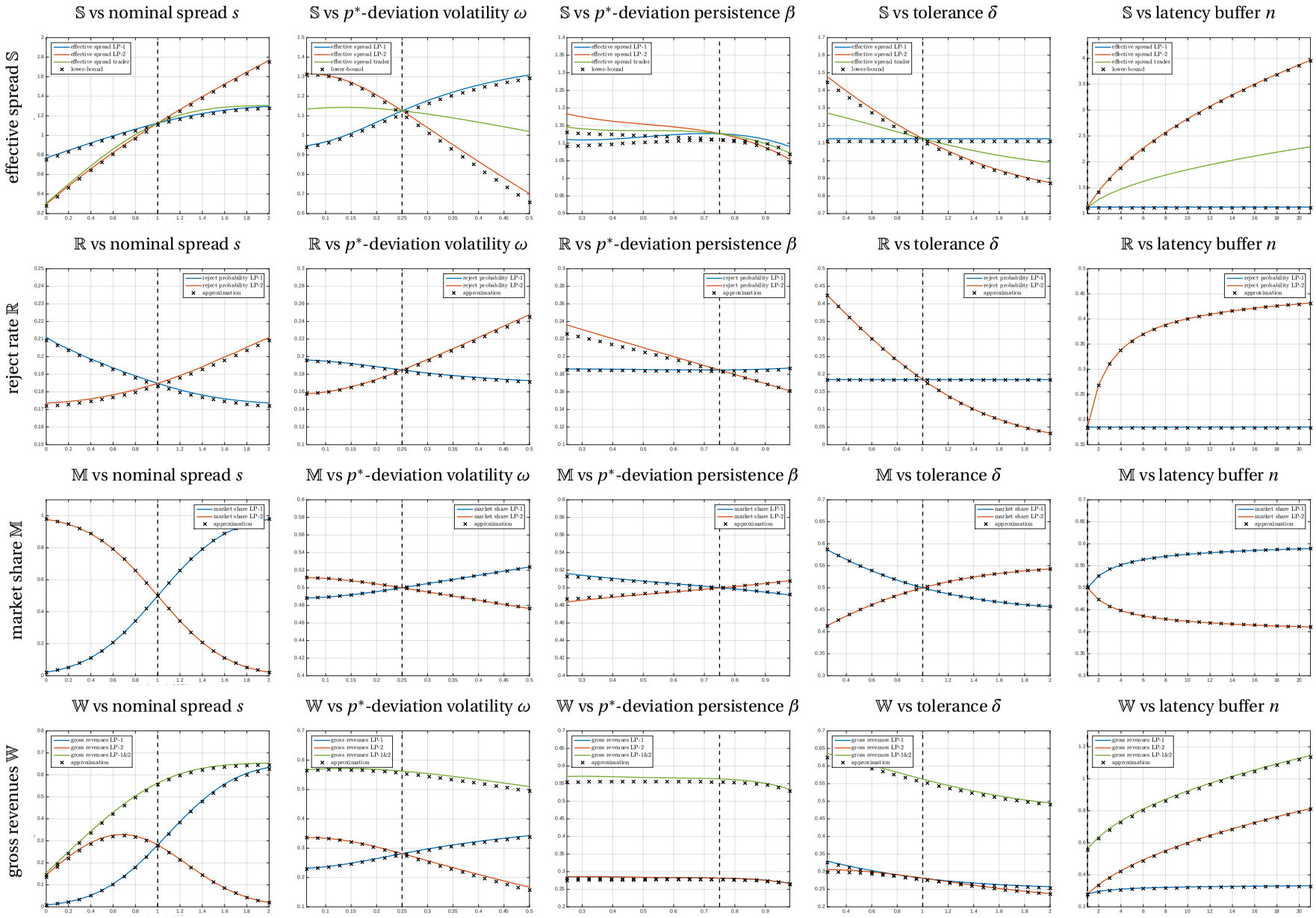
where $(\neq i) = 1$ if $i = 2$ and vice versa. Sometimes it is useful to translate the condition on λ_1 in terms of λ_2 into a condition on λ_2 for given λ_1:

$$\lambda_2 < -\lambda_1 \frac{c_1\lambda_1 - 2\theta\lambda_1 - \theta\lambda_3 + c_1\lambda_3}{c_1\lambda_1 + c_1\lambda_3 - \theta\lambda_1}$$

The no-regret condition for LP-3 in (26) can be expressed as:

$$\lambda_2 > -\lambda_3 \frac{c_3\lambda_1 + c_3\lambda_3 + \theta\lambda_1}{c_3\lambda_1 + 2\theta\lambda_1 + \theta\lambda_3 + c_3\lambda_3}$$

Figure 9: Execution metric sensitivities with heterogenous liquidity providers



Note. This figure draws the effective spread 𝕊, reject rate ℝ, market share 𝕄, and gross revenues 𝕎 (in rows 1 – 4 respectively) as a function of LP-2 nominal spread s_2, p^*-deviation volatility ω_2 and persistence β_2, trade acceptance tolerance δ_2, and latency buffer n_2. The LP-1 and remaining (i.e. those that are not varied) LP-2 parameters are fixed at $s = 1, \omega = 0.25, \beta = 0.75, \delta = s/2, n = 1$ and $\omega = 0.5, \rho = 0.5, N = 2$. The solid lines are based on simulations whereas the markers (×) are, or are derived from, the analytical approximations given in Proposition 5.

References

Aksomaitis, A., and A. Burauskaitė-Harju, 2009, "The moments of the maximum of normally distributed dependent values," *Information Technology and Control*, 38 (4), 301 – 302.

Bank of England, 2011, "The non-investment products code: for principals and broking firms in the wholesale markets," available at http://www.bankofengland.co.uk/markets/Documents/forex/fxjsc/nipscode1111.pdf.

Bank of England, H.M. Treasury, and Financial Conduct Authority, 2014, "Fair and effective markets review," available at http://www.bankofengland.co.uk/markets/Documents/femr/consultation271014.pdf.

Berman, S. M., 1964, "Limit theorems for the maximum term in stationary sequences," *Annals of Mathematical Statistics*, 35 (2), 502 – 516.

BIS, 2014, "Triennial central bank survey. Global foreign exchange market turnover in 2013," Monetary and Economic Department, Bank for International Settlements, available at http://www.bis.org/publ/rpfx13.htm.

Capen, E., R. Clapp, and W. Campbell, 1971, "Competitive bidding in high-risk situations," *Journal of Petroleum Technology*, 23, 641 – 653.

Cartea, A., and S. Jaimungal, 2015, "Foreign exchange markets with last look," working paper, University College London.

Christensen, K., R. C. Oomen, and M. Podolskij, 2014, "Fact or friction: jumps at ultra high frequency," *Journal of Financial Economics*, 114, 576 – 599.

Commodity Futures Trading Commission, 2013, "Core Principles and Other Requirements for Swap Execution Facilities," Billing Code 6351-01-P, RIN Number 3038-AD18 (Final rule), available at http://www.cftc.gov/ucm/groups/public/@newsroom/documents/file/federalregister051613b.pdf.

Kagel, J. H., and D. Levin, 1986, "The winner's curse and public information in common value auctions," *American Economic Review*, 76 (5), 894 – 920.

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

O'Hara, M., 1995, *Market microstructure theory*. Blackwell Publishers Ltd, Oxford, UK.

Oomen, R. C., 2016, "Last look," working paper.

Thaler, R. H., 1988, "Anomalies. The winner's curse," *Journal of Economic Perspectives*, 2 (1), 191 – 202.

Virtu Financial, Inc, 2014, "FORM S-1, REGISTRATION STATEMENT UNDER THE SECURITIES ACT OF 1933," Registration No. 333, available at https://www.sec.gov/Archives/edgar/data/1592386/000104746914002070/a2218589zs-1.htm.