Uncertain Market Making*

Bart Zhou Yueshen†

current version: February 1, 2016

* This paper tremendously benefits from discussions with Shmuel Baruch, Matthijs Breugem, Adrian Buss, Jintao Du, Darrell Duffie, Bernard Dumas, Andrew Ellul, Pekka Hietala, Vincent van Kervel, John Kuong, Massimo Massa, Albert Menkveld, Joël Peress, Farzad Saidi, Vikrant Vig, Lucy White, Liyan Yang, and Haoxiang Zhu. In addition, comments and feedbacks from participants at the 2015 NBER Market Microstructure Meeting are greatly appreciated. The numerical estimation of the structural model receives great help from discussions with Marcin Zamojski. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

† Bart Zhou Yueshen is affiliated to INSEAD. Address: Boulevard de Constance, 77300 Fontainebleau, France. Phone: +33 1 60 72 42 34. E-mail: b@yueshen.me.

Uncertain Market Making

Abstract

This paper argues that market makers' presence is uncertain over any short time interval, as their operations are subject to shocks and constraints of, e.g., capital, technology, and attention. Such uncertain market making implies a random pricing equilibrium in a noise rational expectations framework. Implications for risk, liquidity, and efficiency are discussed. A structural model captures from data the predicted dispersion of random pricing. In 2014, the estimated dispersion is around 10 times of the average price impact, compared to only 2 times in the early 2000s. The evidence suggests deteriorated short-run order flow pricing efficiency in the U.S. equity market.

Keywords: market making, price dispersion, risk, liquidity, price efficiency
JEL code: G10, D40

(There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.)

1 Introduction

Market makers play a central role in financial securities markets. They provide liquidity to investors and, in the meantime, update prices to reflect new information. The canonical assumption of competitive market making—typical in frameworks pioneered by Kyle (1985) and Glosten and Milgrom (1985)—requires market makers perfectly monitor and timely react to all news and events. Only this way does the market achieve price efficiency and maximum liquidity via an implicit Bertrand competition. Do market makers always manage to achieve such perfection?

This paper argues no. In today's fast and fragmented market, effectively monitoring market conditions in real time is very challenging, despite the tremendous advancement in technology and speed. This is because speed is a double-edged sword. On the one hand, traders can process information faster and more efficiently. On the other, they generate more market events to be processed per unit of time—trades, order submission and revision, etc. The net effect is arguably unclear. However, unlike the former, the latter effect is compounded by the number of traders gaining speed. For example, if each of m traders is now able to submit one more order per unit of time, then every trader in the market needs to parse m—not just one—more events. Thus, the more traders gaining speed, the heavier is the burden of monitoring the market for everyone.

Further, product complexity and market fragmentation also add to the cost of monitoring. O'Hara (2015) cites Berman (2014) for an example of strenuous market making across 14 exchange-traded products linked to gold, involving 91 distinct pairs of arbitrage relationships that must be monitored continuously in time. It seems onerous to effectively monitor and react to, in real time, all events like quotes, trades, and news flashing on all these manifold marketplaces.

In view of such challenge in monitoring the market, this paper argues that market making is hardly perfect and, specifically, is subject to two disturbances. First, market makers face time constraints. Given the high speed at which the market moves, it is hard to imagine that market makers will always have sufficient time to fully agree on the price for each and every trade (e.g. by

undercutting each other's quotes in a Bertrand fashion). Second, relatedly, within the short time interval, the presence of (sufficiently many) market makers is uncertain. Some might be unable to respond due to technology-related search frictions, some might be confined by capital constraints, and some others might be intentionally focusing on other markets because of limited attention.[1]

Taken together, the above two frictions—1) the time constraint and 2) the uncertain presence of competing market makers—imply that there is "uncertain market making". That is, no individual market participant knows the exact number of market makers who are actively providing liquidity at any given point of time.

This paper first develops a theory of uncertain market making and examines the implications on market quality. A structural model is then built to incarnate the theoretical concept of uncertain market making into a concrete statistic measure. Applied to real-world trading data, the structural model finds evidence consistent with the theory. It is shown that market making uncertainty is an order of magnitude larger in recent years than was one decade ago.

The theory builds on the informed trading model of Kyle (1985) by blending in the price dispersion model of Burdett and Judd (1983). Relaxing the perfectly competitive market makers, the model assumes that they independently bid for the aggregate order flow once and only once (time constraint) and that, when bidding, they do not know against how many others they are competing (uncertain presence). A noise rational expectations equilibrium exists in which, conditioning on the order flow size, the trading price is always random. The random pricing feeds back to investors' optimal trading behavior via their rational expectations, affecting market quality in all aspects.

More concretely, the random pricing for a given order flow involves two components: 1) an efficient price impact, which is deterministic given the order size, as is standard in Kyle-type models; and 2) a novel *random markup*, due to market makers' strategic behavior in view of uncertain market making. Such a decomposition of trading price contributes to the understanding of market quality

[1] The notion of "market makers" in this paper is meant to be broad. It can refer to designated market makers registered with the exchange or any individual who supplies liquidity (e.g. by posting limit orders).

in three aspects. First, price efficiency is adversely affected by uncertain market making in two different ways: Directly, there is more (transitory) noise due to market makers' random markup. Indirectly, informed trader trades less aggressively in anticipation of the costly random markup. While the indirect channel suggests reduced order flow informativeness, which hurts the long-run price efficiency, the direct channel negatively affects the immediate (short-run) pricing efficiency of the order flow.

Second, the model distinguishes the notions of "trading cost", "(il)liquidity", and "adverse selection", which are largely treated as synonyms in models with perfect market making. Specifically, investors' trading cost is jointly determined by 1) the market makers' adverse selection cost (Kyle's λ) and 2) the random markup, both of which affect market liquidity but reflect different economics. The former is the marginal cost of providing liquidity (like firms' manufacturing cost), while the latter reflects the price premium charged by market makers (like firms' market power from lack of competition). By distinguishing these related notions, the model predicts scenarios of very illiquid market with little informed trading—e.g., few oligopolistic market makers might be charging high markups—and vice versa.

Third, the random markup affects the price return in several ways: 1) It multiplicatively scales up, rather than simply adding to, the price return volatility. The more severe is the market making uncertainty, the stronger is the volatility amplification by the random markup. 2) The random markup is shown to have a positively skewed supported bounded from below by zero (so that market makers make profit from competitors' uncertain presence). The price return inherits such skewness. That is, the model predicts positive (negative) skewness in price returns following buys (sells), *conditional on* the order flow sign and size. 3) The random markup fattens the tails of the price return distribution (amplified kurtosis), increasing the frequency of observing extreme price movements. These aspects highlight the importance of market structure on the risk aspect of financial securities.

The model also finds that limited attention endogenously gives rise to uncertain market making.

In an extension, market makers are offered a menu of marketplaces—a key characteristic of the fragmented financial market nowadays—but constrained by limited attention, they can only choose a subset of the venues to provide liquidity (in a short period of time). It is shown that there is a symmetric mixed-strategy equilibrium in which all market makers split their attention across marketplaces. As each marketplace only gets a fraction of market makers' attention, market making uncertainty aggravates across the board. The model extension thus provides a formal equilibrium argument for the "limited attention hypothesis" analyzed by Corwin and Coughenour (2008) and the model predictions echo their empirical evidence on the NYSE specialists.

A structural model is then developed to capture market making uncertainty from real-world trading data as a measurable statistic. The structural model is most closely related to the state space models in, e.g., Menkveld (2013) and Brogaard, Hendershott, and Riordan (2014). As inspired by the theory, a price impact dispersion term *that scales with the order flow* is added to the pricing error equation. That is, instead of being restricted to a constant parameter, the price impact is allowed to deviate from its (long-run) mean momentarily. While the conventional measure captures the (long-run) *average* effect of order flows on price, market making uncertainty affects its temporary *dispersion*. Economically, the dispersion can be understood as the (in)efficiency of how order flows are priced in the short-run.[2]

The structural model is then applied to intraday trading data of 500 randomly picked stocks from S&P 1500 index over a one-year sample period (January to December 2014). The estimates suggest that an order flow of size $10,000 on average generates a permanent price impact of 1.47 basis points. However, this price impact is accompanied by a large dispersion of 22.51 basis points (per $10,000) in the short-run, about 15 times as large as the long-run effect. The magnitude of the dispersion decreases along stock sizes. On average, a small stock sees a contemporaneous price impact dispersion about 19 times as large as the mean, while it is roughly 12 times for a large stock.

[2] For example, on average the price impact might be 2 basis points per trade, but the immediate, trade-to-trade effect might be sometimes 1 basis point, sometimes 3 or 4 basis points. If market making is certain, the dispersion will be virtually zero, implying that order flows are priced very efficiently; and vice versa.

This dispersion reflects the uncertain market making proposed by the theory: The data suggests positive correlation between the dispersion and the lack of competition of quoting activity, measured by a number of Herfindahl indices, following Hasbrouck (2015). That is, the higher Herfindahl indices of quoting activity (lack of competition), the larger is the dispersion.

The same structural model is applied to a 15-year sample (2000 to 2014) of the component stocks of Dow Jones Industrial Average. The long time series demonstrates how the dispersion of contemporaneous price impact—market making uncertainty—evolved over the one and half decade. At the beginning of the century, the dispersion was around two to three times. Starting from year 2007, it began to rise, gradually to the level of about 10 times by 2014. In the meantime, the conventional *average* price impact measure is found to decrease over the same period, suggesting improved (long-run) price efficiency. However, the increase of the *dispersion* suggests that the market has become less efficient in pricing an order flow in the short-run over the years.

The rest of the paper is organized as follows. Section 2 reviews related literature. Section 3 sets the benchmark model of uncertain market making and discusses the implications. Section 4 then builds an extension with market makers' limited attention. A structural model is developed in section 5 and is applied to real-world trading data in section 6. Section 7 then concludes.

2 Related literature

The model developed in this paper is largely inspired by Baruch and Glosten (2013), who show that random quoting can be optimal in a dynamic limit order market environment. Similarly, Dennert (1993) studies an equilibrium in the context of a dealership market where dealers also quote according to a mixed-strategy. Apart from the market structure difference, the equilibrium random pricing (or quoting) in this paper originates from the *unknown* number of active market makers (attributable to their limited attention), while in both above papers, the population of market makers is known and fixed. Such uncertain market making is a friction prevalent in the current

financial markets and less well understood in the literature.

Jovanovic and Menkveld (2015) analyze non-zero entry costs in a first-price auction. They show that there is a unique equilibrium where bidders endogenously choose to randomly enter the auction and, if enters, bid according to a symmetric mixed-strategy. They then apply the model to high-frequency bidding for S&P500 stocks and identify the model parameters from realized quote price dispersion in the limit order book. Their work and the current paper complement mainly in two ways. First, while Jovanovic and Menkveld (2015) focuses on prices, the current paper, relying on the canonical framework of Kyle (1985), encompasses strategic order flows submission via rational expectations. Second, whereas Jovanovic and Menkveld (2015) generate their probabilistic participation from an unspecified, exogenous participation cost, this paper suggests that market makers' limited attention can contribute to such cost, as those who choose to bid in one marketplace forgo the opportunity of bidding in some other marketplaces.

There is a large body of literature studying equilibrium price dispersion in different contexts. Notable examples include Hausch and Li (1993) and Cao and Shi (2001) in auction; Butters (1977) and Burdett and Judd (1983) in search; Salop and Stiglitz (1977) and Varian (1980) in industrial organization. More recently, Duffie, Dworczak, and Zhu (2015) also establish a similar mixed-strategy equilibrium for dealers' price offers to study benchmarks in over-the-counter markets.

More broadly, this paper contributes to the literature studying market makers' competition. The uninformed traders in Kyle (1989) serve the role of market making and their strategic behavior affects the equilibrium pricing and market quality. Bondarenko (2001) studies a dynamic extension of Kyle (1985) with finite strategic market makers. Moving to the quote-driven market setting, competition among dealers or limit order traders is studied by, to name a few, Glosten (1989), Biais, Martimort, and Rochet (2000, 2013), and Back and Baruch (2013). This paper adds to this literature the new element of market makers' uncertain presence.

There is a recent strand of literature studies how uncertainty, in addition to asset payoffs, affects trading and market. While this literatue (to name a few, Banerjee and Green, 2015; Rossi and Tinn,

2014; Back, Crotty, and Li, 2014; Yang, 2015) mostly focus on the uncertainty on the investor side, this paper studies the uncertainty on the market making side.

The structural model and its estimation relate to the empirical literature studying price impact, trading cost, and market liquidity (see, e.g., Glosten and Harris, 1988; Hasbrouck, 1991, 1993; Brennan and Subrahmanyam, 1996; Sadka, 2006). The model specification is seen in Hasbrouck (2007) and the state space model treatment is pioneered by Menkveld, Koopman, and Lucas (2007), with recent applications like Menkveld (2013); Hendershott and Menkveld (2014); and Brogaard, Hendershott, and Riordan (2014). The innovation of this paper lies in the empirical identification of price impact dispersion, which is argued to represent the short-run order flow pricing efficiency. In addition, the state space model is estimated with a novel generalized method of moments (GMM). In particular, the GMM approach does not require disturbances be Gaussian and identifies the skewness (and other higher moments, if needed) of the price impact dispersion.

In the current paper, the estimates from component stocks of Dow Jones Industrial Average suggest the short-run pricing efficiency of order flows began to deteriorate around 2007. A few recent studies also document a notable market structure "breakpoint" around year 2007 in the U.S. equity market. Skjeltorp, Sojli, and Tham (2013) demonstrate an increasing trend of quote-to-trade ratio from 1999 to 2012 for a large cross-section of stocks. While a mild slope is seen from 1999 to early 2006, a abrupt rise is saliently seen in 2007 and most noticeable for large-cap stocks. Lyle, Naughton, and Weller (2015) find a turning point around 2007, since when the reduction of bid-ask spread, presumably associated with the rise of algorithmic trading, has stagnated. These observations are consistent with the argument that too much market events (e.g. quoting activity) impair participants' ability to parse information timely, raising concerns of uncertain market making.

3 A model of uncertain market making

This section develops a model of uncertain market making. The model builds on the static model of Kyle (1985) but enriches the market making process with just necessary frictions that lead to uncertain market making.

3.1 Model setup

As a notation convention in this paper, capital letters indicate random variables while their corresponding lower case letters indicate the realizations.

Assets. There is a risky asset and a risk-free numéraire (money). Each unit of the risky asset will pay off a random amount of V units of the numéraire. Without loss of generality and to simplify notation, $\mathbb{E}V$ is normalized to 0.

Investors. There is one risk-neutral informed trader, the "insider", who privately observes the realization of V and chooses his optimal order size $X = x(V)$—as a function of V—to maximize his expected profit from trading. There is one noise trader who submits an order of size U (e.g. due to unmodeled liquidity demand). V and U are referred to as the fundamentals of the asset. They are independently normally distributed, with zero means and respective variances σ_V^2 and σ_U^2.

Market makers. There are in total m (possibly infinite) risk-neutral market makers indexed by $i \in \{1, ..., m\}$. However, not all of them are "active" and ready to absorb the incoming order flow. Denote the event that market maker i is active by a successful Bernoulli draw of $\mathbb{1}_i = 1$.[3] The joint distribution of $\{\mathbb{1}_i\}_{i=1}^m$ is common knowledge and, in particular, it is independent of the fundamentals V or U. The total number of active market makers, $M := \sum_{i=1}^m \mathbb{1}_i$, is also random

[3] This structure of market makers' activeness is assumed to be exogenously given in this section. Jovanovic and Menkveld (2015) show, in a more general auction setting, that the Bernoulli arrival draws are a unique equilibrium outcome when market makers (limit order traders) endogenously choose whether to pay a non-zero entry cost. Section 4 below endogenizes this participation cost via market makers' limited attention.

and the realization is not known to any agent in the economy. Later in the analysis, a specific distribution will be adopted to ensure the tractability of the model. Only these M active market makers can participate in the trading process described below.

Trading process. The insider and the noise trader first submit orders independently and their orders pool into $Y = X + U$. Then, just like in Kyle (1985), the active market makers observe the pooled order flow realization y and bid to absorb it. However, rather than perfect competition, the bidding process is specifically modeled: Each active market maker i independently submits a uniform price p_i, agreeing to absorb the order flow y at cost of $p_i y$. Denote by $\mathcal{P} = \{p_i | \mathbb{1}_i = 1\}$ the set of all submitted prices. The winning price is the lowest (highest) price if the pooled flow is a buy when $y > 0$ (sell when $y < 0$):

$$P := \arg\min_{p \in \mathcal{P}} py. \tag{1}$$

The order flow then executes at this winning price against the market maker who bids it. After the trade, the risky asset pays off and all players consume. Such a trading process requires at least $M \geq 1$ active market makers. In the case of $M = 0$, it is assumed that the submitted order flows are nullified and no transaction takes place.

Equilibrium. There are two types of strategic players under this setup: the insider, choosing order size $x(V)$, and active market makers, bidding $p_i(Y)$ upon seeing the aggregate flow $Y = x(V) + U$. This section focuses on an equilibrium in which all active market makers choose a symmetric strategy. Therefore, the equilibrium is a pair $\{x(\cdot), p_i(\cdot)\}$, such that the informed trader maximizes her expected profit

$$x \in \arg\max_{x} \mathbb{E}[(V - P)x | V], \tag{2}$$

where P is the winning price defined in equation (1), and that each active market maker maximizes his expected profit

$$p_i \in \arg\max_{p_i} \mathbb{P}(p_i = P)\mathbb{E}\big[(p_i - V)Y \,|\, Y = x(V) + U, p_i = P\big]. \tag{3}$$

Note that in equation (3), market maker i earns zero with probability $\mathbb{P}(p_i \neq P)$ as the trade does not occur to him in this case. (As will be seen later, in equilibrium, ties almost surely do not occur.)

Some remarks of the model setup

Compared with the original Kyle (1985) model, the model setup presented here differs only in the price setting behavior by market makers. Note that equilibrium condition (3) states an optimality condition, which differs from the original price efficiency condition saying $P = \mathbb{E}[V|Y = x(V)+U]$. In this model, trading price P is efficient only when market makers' optimized profit is forced to zero (e.g. by competition). To see this, set the right-hand side of condition (3) to zero (due to competition) and take expectation (over P) to get $\mathbb{E}\big[(p_i - V)Y \,|\, Y = x(V) + U\big] = 0$, which implies that every market maker bids the efficient price $p_i = \mathbb{E}[V \,|\, Y = x(V) + U]$. Without the zero-profit condition, the trading price is, in general, different from the (semi-strong) efficient price $\mathbb{E}[V \,|\, Y]$.

The trading process is modeled as a first-price, sealed-bid auction (with unknown number of rivals). Such an auction implies that the winner takes all, a key simplifying assumption that keeps the tractability of the model. Effectively, the insider (and the noise trader) always trades with one and only one market maker. This makes sure that the realized trading price is not "contaminated"—in the sense of analytic tractability—by the uncertainty from the random number of market makers. To compare, consider an alternative setup a la Kyle (1989) but with random number of uninformed traders, i.e. market makers. In such a rational expectations equilibrium, the trading price implied by the market clearing condition would also reflect the realized number of market makers. The resulting equilibrium strategy turns nonlinear through Bayesian updating and has limited analytic tractability. It should be emphasized that the winner-takes-all simplifying assumption is not the

driver of all the results to be derived below. Rather, it is the uncertain number of market makers.

The setup recalls the price dispersion model of Burdett and Judd (1983). Indeed, one can think of the informed and the noise traders in the model as liquidity demanders (consumers) and the market makers as liquidity suppliers (firms). Then an equivalent interpretation of the model can be as follows: Let the demand side be given the initiative. The liquidity consumer searches and finds a *random* number of firms, who independently quote for the liquidity consumer's order flow, and then trades with the firm quoting the cheapest price. A large body of literature has extended the original price dispersion model by Burdett and Judd (1983). This section adds to the literature the rational expectations perspective of firms (market makers) who learn the adverse selection cost from consumers' demand (order flow), while consumers anticipate their demand to have price impact. Very relatedly, Jovanovic and Menkveld (2015) in an auction setting endogenize the distribution of the random number of bidders (i.e. firms), which is exogenously given in Burdett and Judd (1983). A similar endogenization is deferred to section 4 where a friction of market makers' limited attention is introduced.

3.2 Equilibrium analysis

3.2.1 The insider

The insider privately knows the risky asset's true value v. Based on this, she chooses her order size $x(v)$ to maximize her expected profit from trading:

$$\max_x \ \mathbb{E}[(V - P)x \,|\, V = v\,],$$

where $P = p(x + U; \mathbb{1}_1, ..., \mathbb{1}_m)$ is the winning bid price as defined in equation (1). Note that the insider is unsure of the execution price P of her order, for two reasons: 1) her order is going to be pooled with the noise flow, U; and 2) the number of active market makers, M, is unknown. This second reason is borrowed from the "noisy search" in Burdett and Judd (1983) but is new to the original Kyle (1985) framework. To facilitate the analysis, conjecture the following for the insider

(recall the normalization of $\mathbb{E}V = 0$):

Conjecture (Kyle, 1985)**.** *The insider's optimal order size is*

$$x(v) = \beta v \tag{4}$$

for some aggressiveness parameter $\beta > 0$*.*

This conjecture will be verified later in subsection 3.2.3.

3.2.2 Market makers

To begin with, the following heuristic argument (similar to the case of Burdett and Judd, 1983) is useful to understand why there is no pure-strategy equilibrium for market makers. Given the conjectured linear strategy $x(v) = \beta v$ by the insider, a Bertrand competitive price of $\mathbb{E}[V|x(V)+U = y]$ is known to all active market makers. However, due to uncertain market making, this Bertrand price is not an equilibrium. If knowing that all others were to bid this Bertrand price, an active market maker i will deviate: He will want to bid a higher price of $\mathbb{E}[V|x(V) + U = y]$ + markup so that with probability $\mathbb{P}(M = 1|\, \mathbb{1}_i = 1)$ he is the only present market maker. Because of uncertain market making, the above probability is not zero. In expectation, hence, the market maker earns $\mathbb{P}(M = 1|\, \mathbb{1}_i = 1)$markup, which is strictly larger than zero profit implied by the Bertrand price. Note, however, that the markup cannot be deterministic, because then the standard Bertrand argument would apply to drive down the price to the perfectly competitive level, which is not an equilibrium per the above argument. The analysis below formally establishes a mixed-strategy equilibrium for market makers.

Let the realization of the pooled order flow be y, which is observed by all active market makers. Suppose market maker i is active, i.e. $\mathbb{1}_i = 1$. Then he knows the probability of competing against other $0 \leq k \leq m-1$ active market makers is $\mathbb{P}(M = k + \mathbb{1}_i|\, \mathbb{1}_i = 1)$. Suppose further that all other $M - 1$ active market makers follow a symmetric linear strategy of bidding $p_j(y) = \Lambda_j y$, for all $j \neq i$ and $\mathbb{1}_j = 1$, where Λ_j is i.i.d. from some known distribution (possibly degenerate). The active

market maker i realizes that the probability of winning with a price p_i is

$$\begin{aligned}\mathbb{P}(p_i = P|\,\mathbb{1}_i = 1) &= \sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i|\,\mathbb{1}_i = 1)\mathbb{P}\Big(p_i y < \min\Big\{(\Lambda_j y)y\Big\}_{j\neq i}|\,M = k + \mathbb{1}_i, \mathbb{1}_i = 1\Big) \\ &= \sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i|\,\mathbb{1}_i = 1)\mathbb{P}\Big(\lambda_i < \min\Big\{\Lambda_j\Big\}_{j\neq i}|\,M = k + \mathbb{1}_i, \mathbb{1}_i = 1\Big),\end{aligned}$$

where the second equality rewrites $p_i = \lambda_i y$. Note that this is merely a rewriting and it does not restrict the strategy space market maker i enjoys, since no constraint on λ_i is imposed. Therefore, market maker i also knows his expected profit with price $p_i = \lambda_i y$ is

$$\begin{aligned}&\sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i|\,\mathbb{1}_i = 1)\mathbb{P}\Big(\lambda_i < \min\Big\{\Lambda_j\Big\}_{j\neq i}|\,M = k + \mathbb{1}_i\Big)(\lambda_i y - \mathbb{E}[V|\,Y = y\,])y \\ =&\sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i|\,\mathbb{1}_i = 1)\mathbb{P}\Big(\lambda_i < \min\Big\{\Lambda_j\Big\}_{j\neq i}|\,M = k + \mathbb{1}_i\Big)(\lambda_i - \lambda)y^2, \qquad (5)\end{aligned}$$

where, exactly as in Kyle (1985),

$$\lambda := \mathbb{E}[V|\,Y = y\,]/y = \frac{\beta\sigma_V^2}{\beta^2\sigma_V^2 + \sigma_U^2} \qquad (6)$$

follows the conditional expectation of bivariate normal distribution. That is, λ is the *efficient price impact factor* (under the conjecture that the insider's aggressiveness is β).

Equivalently, each market maker can be viewed as choosing a random markup that adds to the efficient price impact λ by Z_j times:

$$\Lambda_j = \Big(1 + Z_j\Big)\lambda \qquad (7)$$

where Z_j is i.i.d. from some distribution $F(\cdot)$ to be determined. Similarly, by choosing $p_i = \lambda_i y$, the market maker i is equivalently choosing his markup $z_i = \lambda_i/\lambda - 1$. Then market maker i's expected profit simplifies to[4]

$$\lambda y^2 \sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i|\,\mathbb{1}_i = 1)(1 - F(z_i))^k z_i. \qquad (8)$$

[4] The simplification uses the minimum statistics: For a sample of k observations i.i.d. from $F(\cdot)$, the minimum of the sample has c.d.f. $(1 - F(\cdot))^k$.

Is market maker i also going to use a mixed-strategy where he randomizes the markup z_i? If so, he must be indifferent across all z_i on the relevant support. The partial derivative of the above expected profit with respect to z_i must satisfy[5]

$$\cancel{\lambda y^2} \sum_{k=0}^{m-1} \mathbb{P}(M = k + \mathbb{1}_i | \, \mathbb{1}_i = 1)(1 - F(z_i))^{k-1} \left[1 - F(z_i) - k z_i \dot{F}(z_i)\right] = 0, \tag{9}$$

where the notation $\dot{F}(\cdot)$ is the probability density function of z_i. The result is a first-order ordinary differential equation of $F(z_i)$, which does not depend on order flow y but only on the joint distribution of $\{\mathbb{1}_j\}$. Therefore, an equilibrium exists if all active market makers play the linear mixed-strategy of $p_j = \Lambda_j y$ by marking up λ with Z_j drawn from $F(\cdot)$.

It remains to solve $F(\cdot)$ (and its support) from the differential equation (9). To ensure analytic solution and to develop further implications from the model, assume the following specific joint distribution for market makers' activeness:

***Assumption* 1** *(Market makers' activeness).* Each market maker's activeness is determined through i.i.d. Bernoulli draws with success probability $\theta \in (0, 1)$.

Under assumption 1, the parameter θ can be interpreted as an average market maker's activeness. It measures how likely a market maker is able to react to market events (within some limited amount of time). Among others, such activeness can be affected by the market maker's technology, liquidity constraints, or his strategic allocation of attention. Section 4 studies how θ can be affected by market makers' endogenous allocation of their limited attention.

One more technical assumption is needed. Observe that the profit expression (8) at $m = 1$ (i.e. monopoly) becomes $\lambda y^2 z_i$, which is strictly monotone increasing in the markup z_i. That is, lacking sufficient force of competition, a monopolist will want to charge infinite price to clear the order flow. To rule out such unrealistic infinite profit, a cap on z_i is needed.

***Assumption* 2** *(Maximum markup).* The support of the markup z_i is confined to be a subset of $[0, a]$, where $0 < a < \infty$.

[5] The analysis here restricts to the search only for differentiable $F(\cdot)$.

The cap can be motivated by the maximum possible bid-ask spread that market makers are required to quote by the exchange. When the spread exceeds this threshold, a "circuit breaker" will be triggered and trading will halt, as was the case in the Flash Crash on May 6, 2010. In their auction setting, Jovanovic and Menkveld (2015) endogenize the cap by assuming a fixed cost for bidders to pay before participating in the auction.[6]

The differential equation 9 can now be solved.

Lemma 1 (Solution to differential equation 9). *Under assumption 1, the solution to* $F(\cdot)$ *is*

$$F(z;\theta,m) = \frac{1}{\theta} - \frac{1-\theta}{\theta}\left(\frac{a}{z}\right)^{\frac{1}{m-1}} \quad \text{for } z \in \left[(1-\theta)^{m-1}a, a\right] \text{ and some } a \in \mathbb{R}_{++}. \tag{10}$$

The solution in equation (10) describes the distribution and support according to which the market makers will randomize their price impact markup Z_i. The markup is positive because, intuitively, there is non-zero probability that i is actually the only active market maker and he would like to profit from such probabilistic monopoly power. To see this, evaluating a market maker's expected profit, *conditional on he being active and on the observed flow* y, gives:

$$\pi(y;\theta,m) = \underbrace{(1-\theta)^{m-1}}_{\text{probability of being the monopolist}} \overbrace{a\lambda y^2}^{\text{monopoly profit}}. \tag{11}$$

It can be seen that the expected profit is simply the monoply profit scaled by the probability of being the monopolist. This observation also motivates the necessity of assumption 2.

3.2.3 The equilibrium

With the optimal strategy of market makers (lemma 1), it remains to verify the conjecture about the insider's linear strategy as stated on page 12. Since all active market makers use the symmetric

[6] Very helpful discussion with Shmuel Baruch is acknowledged in understanding the importance of such an upper bound a for the existence of the equilibrium. The discussion also reveals that the exogenous cap a on the markup implies an endogenous cap on the price impact: $\Lambda_j = (1+Z_j)\lambda \leq (1+a)\lambda < \infty$. An alternative of directly imposing an exogenous cap of $\Lambda_j \leq \bar{\lambda}$, similar to the consumers' reservation price $\tilde{p}$ in Burdett and Judd (1983), achieves the same purpose. The exogenous $\bar{\lambda}$, however, complicates the algebra behind comparative static analysis (section 3.3) without affecting the qualitative predictions. As such, the analysis here sticks to the cap a on z_i.

mixed-strategy of $p_i = \Lambda_i y = (1 + Z_j)\lambda y$, the winning price (equation 1) simplifies to

$$P = (1 + Z)\lambda y, \text{ with } Z := \min\{Z_i | \mathbb{1}_i = 1\}. \tag{12}$$

Denote $\mathbb{E}Z = \zeta$. Hence, the insider's optimization problem becomes

$$\max_x \mathbb{E}[(V - (1 + Z)\lambda \cdot (x + U))x | V = v] \quad \Longrightarrow \quad \max_x (v - (1 + \zeta)\lambda x)x$$

which has a unique linear solution of $x(v) = v \cdot (2(1 + \zeta)\lambda)^{-1}$, confirming the conjecture on page 12 with $\beta = (2(1 + \zeta)\lambda)^{-1}$, under the second-order condition of $\beta > 0$. Using lemma 1, the distribution of Z can be found to evaluate ζ. The following lemma summarizes the equilibrium.

Lemma 2. *There is a linear strategy equilibrium, in which the insider submits $x(v) = \beta v$ and all active market makers follow a symmetric mixed-strategy of $p_i(y) = (1 + Z_i)\lambda y$ by drawing Z_i independently from $F(\cdot)$ stated in lemma 1. The equilibrium aggressiveness β and the efficient price impact λ are given by*

$$\beta = \frac{1}{\sqrt{1 + 2\zeta}} \frac{\sigma_U}{\sigma_V} \text{ and } \lambda = \frac{\sqrt{1 + 2\zeta}}{2(1 + \zeta)} \frac{\sigma_V}{\sigma_U}$$

where $\zeta = \mathbb{E} \min\left\{ Z_j \middle| \mathbb{1}_j = 1 \right\} = (1 - \theta)^{m-1}\theta ma/(1 - (1 - \theta)^m)$.

Compared with the original Kyle (1985) equilibrium, the novel feature is the random markup Z, which is the result of uncertain market making. In the context of Burdett and Judd (1983), Z is just a rewriting of the random prices offered by the (liquidity) suppliers. Yet, the fact that the equilibrium is in rational expectations allows a rich set of implications for market quality due to the strategic interaction between the insider (the liquidity consumer) and the market makers (firms "producing" liquidity). These implications are explored next in section 3.3.

3.3 Properties and implications of the equilibrium

The key feature of the equilibrium sought above is that each market maker marks up the efficient price impact factor λ by a random proportion z_i. As only the minimum random markup Z matters

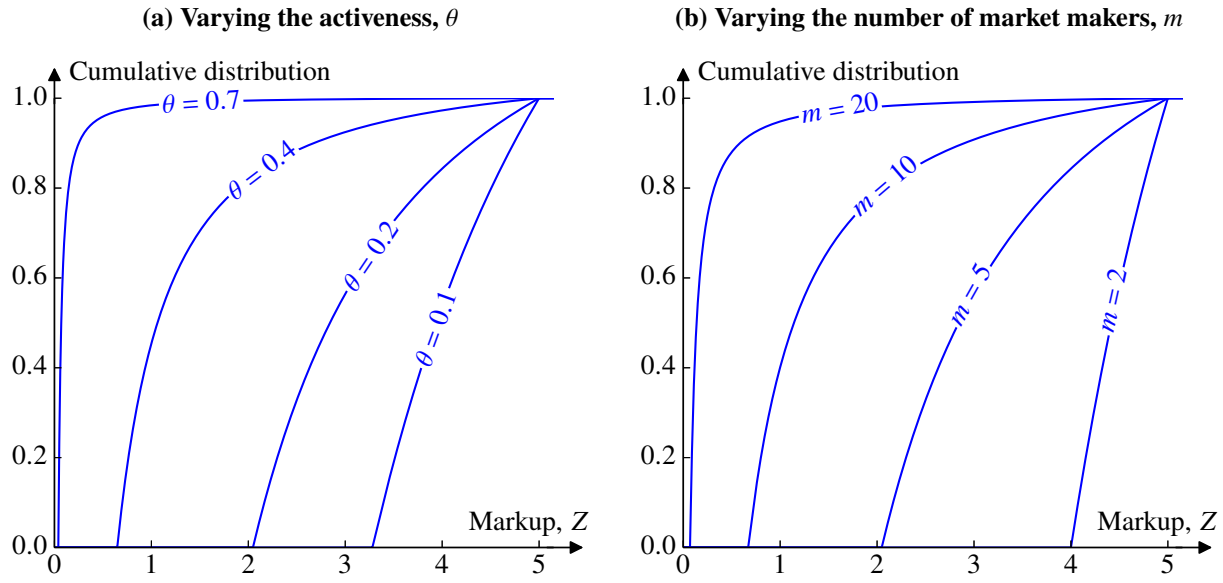


Figure 1: Distribution of price impact markup Z. Panel (a) plots the cumulative distribution function of Z under various activeness parameter, $\theta \in \{0.1, 0.2, 0.4, 0.7\}$, fixing $m = 5$. Panel (b) plots the cumulative distribution function of Z under various number of market makers, $m \in \{2, 5, 10, 20\}$, fixing $\theta = 0.2$. The other primitive parameters are $\sigma_V^2 = 1$, $\sigma_U^2 = 2$, and $a = 5$.

for realized trades, this section focuses on its properties and implications.

3.3.1 Uncertain market making, competition, and the random markup

The random markup Z on the efficient price impact factor λ adds to market makers' expected profit. As either market makers' activeness θ or the number of market makers m increases, the profit reduces to zero (see equation 11). This suggests market makers become more competitive and the market making uncertainty reduces accordingly:

> **Lemma 3 (Stochastic dominance).** *Denote by $G(z; \theta, m)$ the cumulative distribution function for the minimum price impact markup Z. Then for $\theta' > \theta$ and $m' > m$, $G(\cdot; \theta, m)$ first-order stochastically dominates both $G(\cdot; \theta', m)$ and $G(\cdot; \theta, m')$.*

Figure 1 illustrates how the distribution of Z changes as θ or m increases. The more active are the market makers, the less likely the random markup will be large. The effect is similar when there

are increasingly more market makers. More probability mass is allocated to the left tail, closer to 0, hence making the minimum markup Z more likely to be small (closer to zero).

In the limit of $\theta \to 1$ (or $m \to \infty$), Z degenerates to 0 almost surely and hence, $\mathbb{E}Z = \zeta = 0$. In this extreme case, the equilibrium converges to the original static Kyle (1985) (as can be easily verified via the expressions in lemma 2). Uncertain market making and (lack of) market makers' competition are synonyms in this context. Since both θ and m have similar effects, the rest of this subsection focuses on θ to avoid repetitiveness. For convenience, θ as market makers' activeness measure will also be discussed as the (inverse of) market making uncertainty.

3.3.2 Order flow aggressiveness, price efficiency, and trading cost

The random markup makes trading more costly for both the insider and the noise traders. Knowing this, hence, the strategic insider bids less aggressively: $\beta = 1/(2(1+\zeta)\lambda) < 1/(2\lambda)$; c.f. Kyle (1985). Panel (a) of figure 2 shows this effect. The horizontal axis, σ_u/σ_v, is the noise-to-information ratio in the pooled order flow. The dashed line plots the insider's aggressiveness in a perfectly competitive market making environment, where $\beta = \sigma_u/\sigma_v$ as in Kyle (1985). The solid lines plot the insider's aggressiveness under the current environment where there is uncertain market making (following lemma 2).

The insider's reduced aggressiveness, in turn, lowers the price efficiency, simply because there is a lower amount of information compounded into the order flow. Panel (b) of figure 2 shows the the efficient price adjustment λ is lower than the case of no market making uncertainty (the dashed line). As the market making uncertainty increases (θ reduces), the inefficiency is also elevated.

From investors' point of view, the "total (marginal) cost" of their order flows is measured as $\Lambda = (1+Z)\lambda$, which is the marginal price to pay for the additional unit of order. In this total cost, λ reflects the efficient price impact cost and Z reflects the (random) cost of uncertain market making. Panel (c) of figure 2 plots the expected value of this total cost. As can be seen, contrary to the reduction of price efficiency λ, the total trading cost *increases* with market making uncertainty

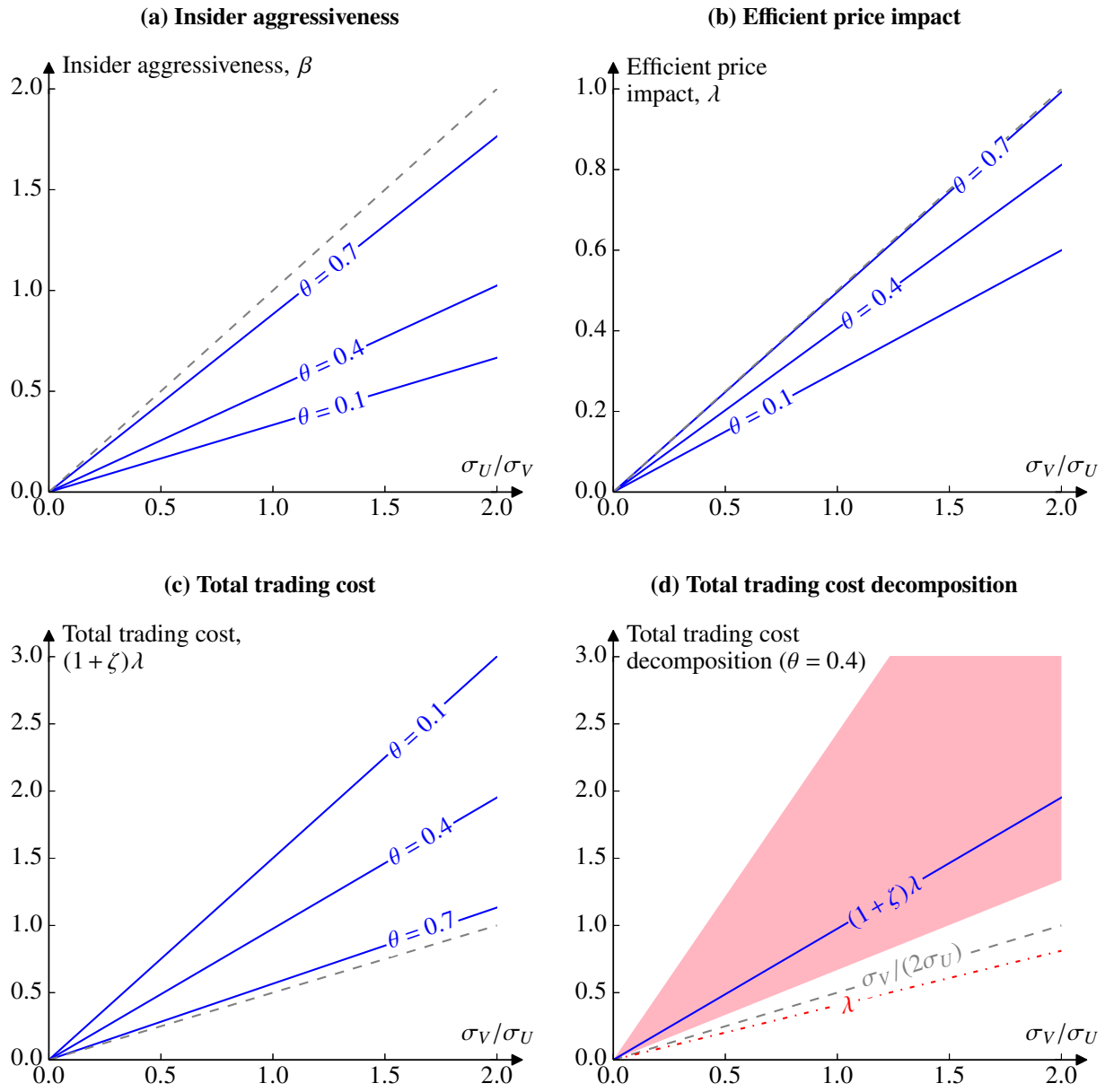


Figure 2: Price impact and insider aggressiveness. Panel (a) and (b) plot the insider's aggressiveness β and the efficient price impact λ, respectively, for different levels of market maker activeness, i.e. $\theta \in \{0.1, 0.4, 0.7\}$. Panel (c) plots how the total trading cost, as the sum of the efficient price impact and the expected minimum markup, varies across different θ values. Panel (d) decomposes investors' total trading cost into λ and Z, fixing $\theta = 0.4$. The shaded area shows the possible range of variation of $(1 + Z)\lambda$. The top solid line is the expected price impact, $(1 + \zeta)\lambda$. The middle dashed line is the competitive efficient price impact, i.e. $\sigma_v/(2\sigma_u)$ as in Kyle (1985). The bottom dot-dashed line shows the efficient price impact λ. For all panels, the remaining primitive parameters are set at $m = 5$ and $a = 5$.

(lower θ). This suggests that it is the cost of uncertain market making ζ, instead of the efficient price impact λ, that dominates investors' trading cost. Similarly, panel (d) visually decomposes this total marginal cost. Notably, as shown by the shaded area, the variation of the total trading cost can be very large due to uncertain market making. The following proposition summarizes the above discussion.

Proposition 1 (Uncertain market making and market quality). *As market making uncertainty increases (smaller θ), the insider trades less aggressively (smaller β), the efficient price impact reduces (smaller λ), and the average total trading cost increases (larger $(1+\zeta)\lambda$).*

Another way of interpreting the total trading cost Λ is that it is the total price impact per unit of the order flow (see equation 13 below). The total price impact thus has two components: λ and $Z\lambda$. The former is the long-run efficient price impact, as it reflects the permanent price change per unit of the order flow. The later $Z\lambda$ is a measure of short-run order flow pricing (in)efficiency, which will die out in the long-run in a dynamic setting. The structural model developed later in section 5 builds on such a decomposition of Λ.

3.3.3 Market (il)liquidity

In this type of models, market illiquidity is often measured by Kyle's λ, which is a measure of the asset price elasticity to the order flow. The larger is the sensitivity, the less liquid is the asset—often interpreted as reduced order book depth. Kyle's λ is essentially a cost measure: It exactly reflects market makers' competitive marginal price for executing an additional unit of order. In the same spirit, with uncertain market making, market illiquidity can be measured by the total trading cost $(1+Z)\lambda$. When compared to the canonical expression λ, the new expression sheds light on the understanding of market liquidity.

First, the random markup Z indicates a new source of market illiquidity. Conventionally, market illiquidity is only attributed to the information asymmetry in the order flow, in which case $\lambda = \sigma_V/(2\sigma_U)$. The new component of $Z\lambda$ arises only from uncertain market making, a

feature related to the market structure but independent of the asset fundamentals like σ_V^2 or σ_U^2. From market makers' point of view, the information asymmetry component λ is the necessary cost to provide liquidity, just like the production cost for a manufacturer. The market making uncertainty component λZ, instead, reflects the suppliers' competitiveness. This component of market illiquidity does not exist in the classic framework where perfectly competitive market makers are typically assumed. Because of the multiplicative structure, the effect of uncertain market making on market illiquidity is expected to be economically sizable. Section 5 introduces a structural model to estimate the magnitude of such a component.

Second, the effects of uncertain market making (reflected by Z) and information asymmetry (reflected on λ) can go in opposite ways. In particular, proposition 1 suggests that a thin market (e.g. low depth, wide spread) is not necessarily related to information asymmetry. The lack of liquidity can simply be due to uncertain market making. Similarly, more informed trading and increased market liquidity can coexist: If market makers are very active in a stock's trading, the average market making uncertainty cost $\mathbb{E}Z = \zeta$ reduces and informed trader will trade more aggressively, raising λ. The net effect, as can be seen in panel (c) of figure 2, is improved market liquidity (lower trading cost). This is because the insider always optimally chooses her aggressiveness so that the increase in the efficient price impact λ does not exceed the reduction in the expected markup ζ—overall, she pays a lower trading cost.

Finally, because the markup Z is random, it also affects the variation of market illiquidity, i.e. the liquidity risk: $\text{var}[(1 + Z)\lambda] > 0$. Following the stochastic dominance property (lemma 3), higher liquidity risk is expected when market makers are relatively inactive.

3.3.4 Price volatility, skewness, and tail risk

The random price impact markup, Z, has a novel scaling effect on asset price volatility. To see this, write the equilibrium price return as[7]

$$\Delta p = (1 + Z)\lambda Y = \Lambda \cdot (\beta V + U). \tag{13}$$

where $\Lambda := (1 + Z)\lambda$ as in equation (7) and $Y = x(V) + U = \beta V + U$. Then the variance of the return becomes

$$\text{var}[\Delta p] = \mathbb{E}\Lambda^2 \cdot \left(\beta^2\sigma_V^2 + \sigma_U^2\right) = \left[\text{var}\Lambda + (\mathbb{E}\Lambda)^2\right]\left(\beta^2\sigma_V^2 + \sigma_U^2\right)$$

The term $\left(\beta^2\sigma_V^2 + \sigma_U^2\right)$ represents the two fundamental sources of the price return volatility, the innovation V and the noise demand U. It can be seen via $\text{var}\Lambda = \text{var}(1 + Z)\lambda^2 > 0$ that market making uncertainty *multiplicatively scales up*, rather than adding to, the fundamental return volatility components. This scaling effect is only present due to the randomness in the markup Z.

Panel (a) of figure 3 illustrates the effect. Notably, when market maker's activeness is relatively low, the scaling effect on the return volatility can be sizable. When activeness is high, the market makers' competition becomes more fierce and the markup Z diminishes (in a stochastic dominance sense; see lemma 3). Eventually, the return volatility converges to the perfectly competitive case.

The higher moments of the markup Z also affect the price return distribution.

Proposition 2 (Conditional positive skewness). *Conditional on the order flow, the price return exhibits positive (negative) skewness following a buy (sell).*

To emphasize, proposition 2 is a conditional statement saying there is still skewness in price return after controlling the potential skewness in the order flow, if any.

[7] For exposition clarity, the return equation (13) assumes that prior to this order flow, the (traded) asset price is at its (semi-strong) efficient price, $\mathbb{E}V$ (which is normalized to zero).

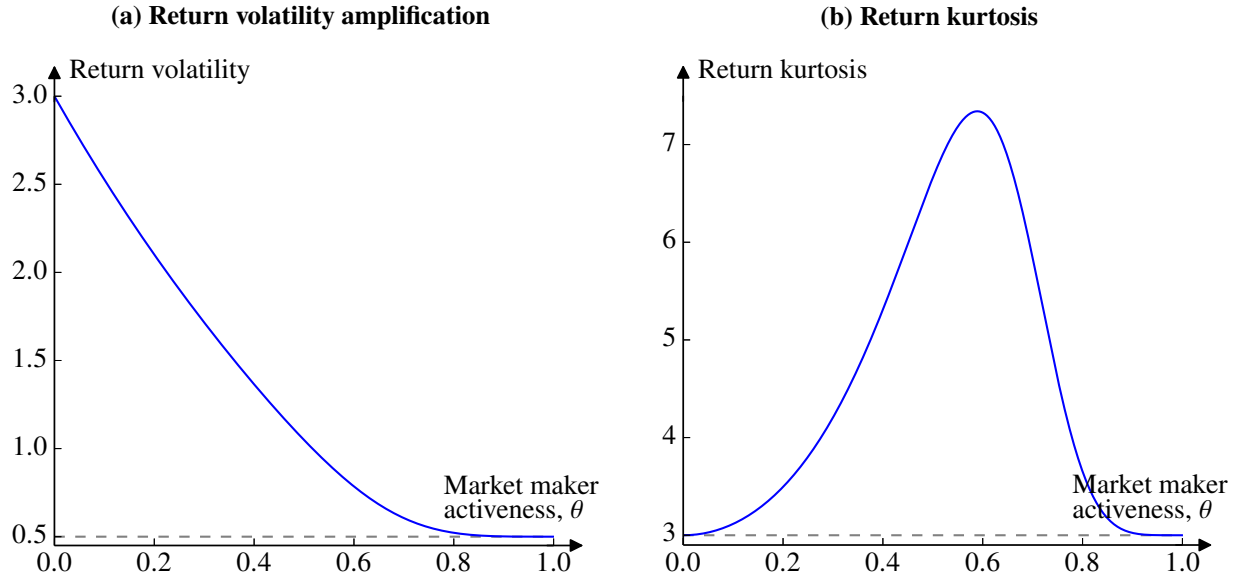


Figure 3: Price return volatility and kurtosis. Panel (a) and (b) plot, respectively, the volatility and kurtosis of the price return (Δp) against market maker activeness θ, i.e. the inverse of market making uncertainty. In both panels, the dashed lines show the asymptotic values as $\theta \to 1$, i.e. the limit result of Kyle (1985). The other primitive parameters are set at $m = 5$, $a = 5$, $\sigma_V^2 = 1$, and $\sigma_U^2 = 2$.

Proposition 3 (Tails of price return distribution). *Uncertain market making fattens the tail of the price return distribution. Statistically, the kurtosis of the asset return is larger when there is uncertain market making than when there is not.*

Panel (b) of figure 3 illustrates this inflated kurtosis and finds that it is non-monotone in the market making activeness: With moderately large θ, the tail of the asset return appears to be the fattest (the kurtosis is the largest).

4 Limited attention and uncertain market making

There are many reasons why a market maker might not be always active. He might be constrained by his technology accessing the market; he might be just given a margin call and need time to raise fund for further operation; his analyst team might need more time to study a company's financial report.

This section explores how one particular type of constraint, market makers' limited attention, affects uncertain market making through a model with endogenous attention allocation.

4.1 Setting

The setting extends from section 3.1. There are $n \geq 2$ marketplaces, indexed by $j \in \{1, ..., n\}$, each trading one risky security with random payoff V_j. In each marketplace, there is one informed trader who privately knows the realization of V_j and strategically submits informed order flow $x_j(V_j)$. As before, there is noise flow U_j in each marketplace as well. Only the pooled order flow $y_j = U_j + x_j(V_j)$ is observable by market makers who are active in market place j. The random variables $\left\{V_j\right\}$ and $\left\{U_j\right\}$ are assumed to follow a joint normal distribution with zero mean. The noise flow U_j is assumed to be independent of the payoff V_j. The variances are denoted by σ^2_{Vj} and σ^2_{Uj}. No other restriction is imposed on how these n risky assets are related. For example, the security payoffs can be identical (only the trading venues are fragmented) and the cross-sectional noise flows can be correlated.

Trading is facilitated by m (≥ 2) identical, risk-neutral market makers. At the beginning of the game ("attention allocation period"), each active market maker chooses in which market(s) he wants to be active. Specifically, a market maker i chooses whether $\mathbb{1}_{i,j} = 1$ for all venues $j \in \{1, ..., n\}$. Limited attention is modeled as a constraint $\bar{n}$ that for each market maker i, such that

$$\sum_{j \in \{1,...,n\}} \mathbb{1}_{i,j} < \bar{n} < n.$$

That is, no market maker can be active in all marketplaces. For ease of exposition, this paper only considers the special case of $\bar{n} = 1$. In particular, mixed-strategy is allowed: A market maker i can choose to operate in a marketplace j with some probability $\mathbb{P}\left(\mathbb{1}_{i,j} = 1\right) = \theta_{i,j}$, subject to $0 \leq \sum_{j=1}^{n} \theta_{i,j} \leq 1$. The choice variable $\theta_{i,j}$ is market maker i's attention paid to marketplace j.

After all market makers have allocated their attention $\left\{\theta_{i,j}\right\}$, $\forall (i, j) \in \{1, ..., m\} \times \{1, ..., n\}$, trades take place as described in section 3.1 ("trading period"). Figure 4 sketches the time line.

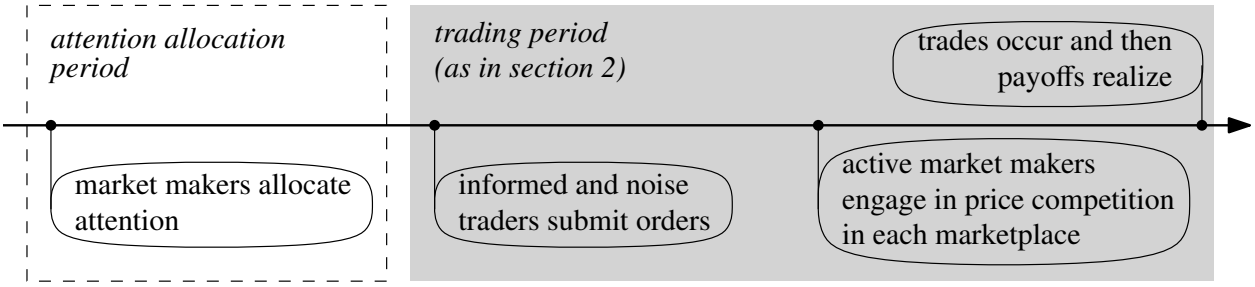


Figure 4: Time line of the attention allocation game. This figure illustrates the time line of the extended game presented in section 4. The shaded part of the time line corresponds to the trading game presented in section 3. The dashed-line box indicates that the extension of market makers' attention allocation.

4.2 Equilibrium attention allocation

The equilibrium is found backwardly by first solving for a market maker i's expected profit from the trading period and then by imposing an indifference condition in the attention allocation period.

4.2.1 Market makers' expected profit from trading

Consider an arbitrary market maker i, who happens to be active in marketplace j, i.e. $\mathbb{1}_{i,j} = 1$. Suppose that he knows that all other market makers' attention choice for this marketplace is the same $\theta_j \in [0,1]$ (to be determined). By assumption 1, $M_{-i,j} := \sum_{h \neq i} \mathbb{1}_{i,j}$ follows a Binomial distribution with $m-1$ Bernoulli draws of success rate θ_j.

Suppose further that the informed trader in marketplace j believes that all market makers' attention choices for marketplace j are the same θ_j. Lemma 1 and lemma 2 then apply to each individual marketplace. In particular, given $\mathbb{1}_{i,j} = 1$, market maker i's expected profit from marketplace j is (c.f. equation 11)

$$\pi_j(\theta_j, m) := \mathbb{E}\pi_j(Y_j; \theta_j, m) = \left(1-\theta_j\right)^{m-1} a\lambda_j \mathbb{E}Y_j^2 = \left(1-\theta_j\right)^{m-1} a\beta_j \sigma_{Vj}^2 \tag{14}$$

where the last equality follows the expressions from lemma 2.

The expression (14) explains how the average attention choice θ_j affects a market maker i's

profit, in two ways: 1) The "competition channel" reduces the profit: The probability of being the monopolist in marketplace j, $(1-\theta_j)^{m-1}$, is reducing in θ_j. As more attention is paid to marketplace j by all market makers, the monopoly probability reduces for each of them and hence lower rent remains. 2) The "participation channel" increases the profit: Note that the informed trader's strategy is affected by market makers' activeness, θ_j, via her aggressiveness β_j. This is because, intuitively, the more attention paid to the marketplace, the less market making uncertainty remains and the reduced illiquidity (proposition 1) makes the informed trader willing to trade more, increasing the order flow size. Since market makers' profit is scaling on the order flow size $\mathbb{E}Y_j^2$, a higher rent, therefore, follows. In the proof of lemma 4, it is shown that the net effect of the two is negative, i.e. the competition channel dominates.

4.2.2 Attention allocation

A critical assumption underlying the expected profit expression (14) is that all market makers pay the same attention to a marketplace. If there is such an equilibrium, the resulting expected profit from any marketplace equate each other. Otherwise, the market makers will deviate to paying more attention to the marketplace that generates a higher expected profit.

More rigorously, a market maker i's optimization problem is to allocate his attention $\left\{\hat{\theta}_j\right\}$:

$$\max_{0\le\hat{\theta}_j\le 1}\sum_{j=1}^{n}\hat{\theta}_j\pi_j\left(\theta_j,m\right),\ \text{s.t.}\ \sum_{j=1}^{n}\hat{\theta}_j\le 1,$$

given that all other market makers are choosing $\left\{\theta_j\right\}$ *and* that each informed trader in marketplace j expects the same attention θ_j from all market makers. This linear optimization problem, in general, will have a bang-bang solution unless the other market makers' attention allocation satisfies $\pi_j(\theta_j,m)=\pi_h(\theta_h,m)>0$ for all marketplaces j and h. If such an indifference condition holds, no market maker will have incentive to deviate from the same allocation of $\left\{\theta_j\right\}$. Such an equilibrium always exists.

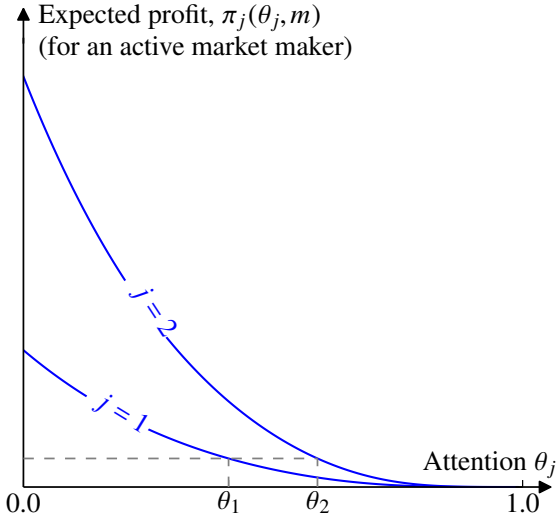


Figure 5: Equilibrium attention allocation. This figure illustrates the market makers' equilibrium attention allocation for $n = 2$ marketplaces. Each solid curve corresponds to the expected profit function $\pi_j(\theta_j)$ for marketplace j. The equilibrium is found for θ_1 and θ_2 such that $\theta_1 + \theta_2 = 1$ and that the profits equate each other: $\pi_1 = \pi_2$.

Lemma 4 (Equilibrium attention allocation). *There exists a unique, symmetric-strategy equilibrium in which all market makers mix to pay attention* $0 \leq \theta_j < 1$ *to marketplace j such that* $\sum_{j=1}^{n} \theta_j = 1$.

Figure 5 illustrates this equilibrium with $n = 2$ marketplaces. The two marketplaces have different characteristics in terms of σ^2_{Vj} and σ^2_{Uj}. Depending on these characteristics, the shapes of π_j differ but the proof of lemma 4 establishes that they are always strictly downward sloping and positive on $\theta \in [0, 1]$. In equilibrium, all market makers derive the same expected profit from each marketplace. The equilibrium is found for some $\{\theta_1, \theta_2\}$ such that $\pi_1(\theta_1, m) = \pi_2(\theta_2, m)$ and $\theta_1 + \theta_2 = 1$.

4.3 Implications of equilibrium attention allocation

Market makers' limited attention forces them to strategically allocate their attention across the marketplaces. Which marketplace gets more attention and hence is subject to less market making uncertainty? The answer lies in the drivers behind an active market maker's expected profit. Expanding the expression (14) by substituting the expression of β_j from lemma 2 gives

$$\pi_j(\theta_j, m) = \frac{\left(1-\theta_j\right)^{m-1} a}{\sqrt{1+2\zeta_j}} \sigma_{V,j}\sigma_{U,j}.$$

It can be seen that when either $\sigma_{U,j}$ or $\sigma_{V,j}$ is higher, market makers derive higher expected profit from marketplace j and will want to allocate more attention there. The two fundamental parameters reflect, respectively, the liquidity demands by the noise traders (e.g. hedging needs) and by the informed traders (relative magnitude of non-public information).[8]

As such, a "liquidity-beget-liquidity" phenomenon occurs: Marketplaces with higher liquidity demand—larger $\sigma_{U,j}$ or $\sigma_{V,j}$—tend to attract more market makers' attention, lowering the market making uncertainty, which, in turn, reduces the total trading cost in that marketplace (recall proposition 1). The following comparative static result formally states this result.

Proposition 4 (Comparative statics of attention allocation). *The equilibrium attention θ_j paid to a marketplace j is monotone increasing in insider's information σ^2_{Vj} and in noise trading σ^2_{Uj}.*

Proposition 4 can be alternatively interpreted as a way to understand how market makers reallocate attention when there is a shock in market fundamentals. Consider two marketplaces 1 and 2. Suppose a shock strikes marketplace 1 and either σ^2_{V1} or σ^2_{U1} increases, then market makers will all want to allocate more attention to marketplace 1. This suggests that while marketplace 1 gets more attention and sees less market making uncertainty and higher liquidity, marketplace 2 will experience more market making uncertainty and lower liquidity. Therefore, from the point view of market quality, limited attention can lead to illiquidity spillover.

[8] The "liquidity demand" in this context broadly refers to investors' general interest for trading. It encompasses demands from both informed traders and noise traders. The terminology could be misleading as, in a narrower sense, "liquidity demand" sometimes specifically refers to noise traders' order flows only.

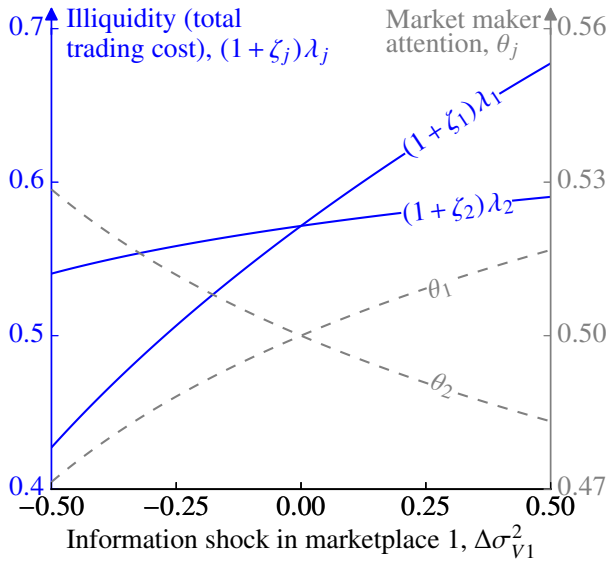


Figure 6: Illiquidity spillover. This figure illustrates how illiquidity can propagate from one marketplace to another without interconnected fundamentals. An information shock on σ^2_{V1} hits marketplace 1, where the pre-shock value is $\sigma^2_{V1} = 1.0$. The shock size, ranging from −0.5 to 0.5, is indicated on the horizontal axis. Other characteristics of the marketplaces do not change: $\sigma^2_{U1} = \sigma^2_{U2} = 2.0$ and $\sigma^2_{V2} = 1.0$. The solid curves show the post-shock market illiquidity (total trading cost), $\lambda_j + \zeta_j$ for $j \in \{1, 2\}$, on the left axis. The dashed curves plot the post-shock equilibrium attention allocation of θ_1 and θ_2 by market makers, shown on the right axis. The other primitive parameters are set at $m = 5$ and $a = 5$.

Corollary 1 (Illiquidity spillover). *Positive shocks in the fundamentals of marketplace j—either insider's information σ^2_{Vj} or noise trading σ^2_{Uj}—reduce market makers' attention in the other marketplaces, resulting in higher trading cost (illiquidity) there.*

Figure 6 illustrates such an illiquidity spillover from marketplace 1 to marketplace 2, due to an idiosyncratic information shock specific to the asset traded on marketplace 1. The shock size $\Delta\sigma^2_{V1}$ is shown on the horizontal axis. It can be seen that a positive shock drives up the illiquidity (total trading cost) $(1 + \zeta)\lambda$ in *both* marketplaces. The reasons, however, are different. In marketplace 1, the increased level of information asymmetry σ^2_{V1} determines that market makers require higher compensation for adverse selection, λ_1. (Although more market makers' attention is drawn to marketplace 1, the reduction in the market making uncertainty ζ_1 is still dominated by the increase

in λ_1.) The increase in σ^2_{V1} drives up the expected trading volume in marketplace 1 and more market makers' attention is drawn from marketplace 2. The reduced θ_2 implies a marketplace 2 with higher market making uncertainty, implying higher ζ_2—the main source of higher illiquidity in marketplace 2.[9]

The predictions above echo the empirical findings in Corwin and Coughenour (2008), who examine NYSE specialists' portfolio choice and the liquidity provision and document higher trading cost for stocks whose specialist has turned attention elsewhere. Consistent with proposition 4, they also find such phenomenon is most evident for least active stocks. The theory developed on market makers' limited attention and the effects on uncertain market making formalizes the "limited attention hypothesis" proposed by Corwin and Coughenour (2008) as an equilibrium outcome.

To emphasize, such a spillover arises even the shock in the "originating" marketplace is completely independent of the assets traded in the "infected" marketplaces. In fact, in the example illustrated in figure 6, there is no fundamentals changed in marketplace 2. The only reason that market illiquidity rises there is because market makers have turned more attention to marketplace 1 and market making becomes more uncertain. The illiquidity spillover here has a very different mechanism from the learning channel proposed by Cespa and Foucault (2014).

5 A structural model

This section develops a structural model that based on the theory developed in section 3. After laying out the model structure, the interpretation of parameters as well as estimation techniques are discussed. Appendix A shows how the structural specification relates to price return heteroskedasticity.

[9] The implicit assumption underlying such illiquidity spillover is that the total number of market makers is fixed, i.e. there is no free entry, not at least in the short time frame considered. This is perhaps a reasonable characterization for the high-frequency trading world as traders' entry decisions are typically decided off trading hours: They need to pay an arguably high fixed cost to setup the equipment and to develop algorithms.

5.1 Model structure and assumptions

Consider the following generic structure of an asset's unobservable efficient price m_t, the unobservable pricing error s_t, and the observed price p_t:

$$m_t = m_{t-1} + w_t$$

$$p_t = m_t + s_t$$

where w_t is the permanent increment of the efficient price at time t. This specification is seen in Hasbrouck (2007) and the state space model treatment is pioneered by Menkveld, Koopman, and Lucas (2007). More recent applications include Menkveld (2013), Hendershott and Menkveld (2014), and Brogaard, Hendershott, and Riordan (2014). Specifically, the efficient price innovation w_t and the pricing error s_t are affected by the order flow y_t and its surprise y_t^* in the following standard way:

$$w_t = \lambda y_t^* + \mu_t$$

$$(1 - \phi(L))s_t = \psi(L)y_t + \nu_t$$

where $\phi(L)$ and $\psi(L)$ are some arbitrary lag polynomials. In words, the efficient price increment w_t features two components, 1) λy_t^* from the surprise in the order flow and 2) μ_t from information unrelated to trading (e.g. public news). The pricing error has an autoregressive structure on the possibly lagged order flows $\psi(L)y_t$ and some disturbance ν_t.

To ensure the stationarity of the pricing error, it is assumed that all roots to the polynomial $1 - \phi(L)$ fall strictly inside the unit circle. For example, when $\phi(L) = 0$, i.e. there is no lasting effect of pricing error s_t on subsequent prices, the structural specifications above imply a price return of $\Delta p_t = \lambda y_t^* + \mu_t + (1 - L)\big[\psi(L)y_t + \nu_t\big]$, which is reminiscent of the structural models laid down in, for example, Brennan and Subrahmanyam (1996) and Sadka (2006). When $\phi(L) = \phi L$ (for $|\phi| < 1$), the pricing error becomes $s_t = \phi s_{t-1} + \psi(L)y_t + \nu_t$, i.e. the same specification in Menkveld (2013) and Brogaard, Hendershott, and Riordan (2014) with $\psi(L) = \psi$.

This paper further imposes a structure on the disturbance term of the pricing error:

$$\nu_t = \varepsilon_t + \lambda z_t y_t^*.$$

Hence, the price pressure process becomes

$$(1 - \phi(L))s_t = \psi(L)y_t + \varepsilon_t + \lambda z_t y_t^*. \tag{15}$$

This structure suggests three components contributing to the pricing error. The first two are standard: $\psi(L)y_t$ reflects the order flows' price impact and ε_t is the part unrelated to trading. The third term $\lambda z_t y_t^*$ is new and is inspired by the model developed in section 3. It reflects the idea that uncertain market making generates a random markup of size $\lambda z_t y_t^*$, where the process z_t is designed to capture the random multiplicative scaling effect. Of course, all three components only contribute to the pricing error s_t and will die out in the long run.

To gauge market making uncertainty, the main econometric objective is to estimate sample moments of the random markup z_t. The following assumption is needed for identification:

***Assumption* 3** *(White noise disturbances).* Conditional on order flow series $\{y_\tau\}_{\tau \le t}$ (and, hence, also $\{y_\tau^*\}_{\tau \le t}$), the disturbances $\{\mu_t\}$, $\{\varepsilon_t\}$, and $\{z_t\}$ are independent white noises, whose first $\bar{k}$ (≥ 2) moments exist.

The white noise assumption about z_t perhaps requires a bit more elaboration. In particular, a white noise process is zero-mean and, yet, the theory from section 3 predicts that the random markup z_t has strictly positive mean of ζ (lemma 2). Here in the structural model, assuming a zero-mean $\{z_t\}$ process is, in fact, necessary. To see this, suppose the true random markup process is $\tilde{z}_t$ with unconditional mean ζ, i.e. $\tilde{z}_t = \zeta + z_t$. Consider a general specification of s_t akin to equation (15):

$$(1 - \phi(L))s_t = b(L)y_t + \varepsilon_t + \lambda \tilde{z}_t y_t^* = \underbrace{\left(b(L)y_t + \lambda\zeta y_t^*\right)}_{:=\psi(L)} + \varepsilon_t + \lambda z_t y_t^*. \tag{16}$$

To the extent that the innovation y_t^* can be written as some $c(L)y_t$, the above is equivalent to

equation (15) by writing $\psi(L) := b(L) + \lambda\zeta c(L)$.

The insight by comparing the structures (15) and (16) is that the unconditional mean of the random markup, ζ, is unfortunately not identifiable, unless additional condition exists for the structure $b(L)$. In a sense, ζ is already "absorbed" in the lag polynomial $\psi(L)$ and the $\{z_t\}$ process should be interpreted as the *demeaned* random markup. In view of the above concerns, this paper turns to the second moment of the random markup, denoted by $\sigma_Z^2 := \mathbb{E}z_t^2$, for measuring market making uncertainty.

Using σ_Z^2 as a measure for market making uncertainty has intuitive economic interpretation. Consider a shock y_t^* in the order flow. From the structure model above, this surprising order flow generates a "net contemporaneous price impulse response", denoted by ξ_t, of size

$$\xi_t = \lambda y_t^* + (\lambda z_t + \psi(0)) y_t^*$$

where λy_t^* is the permanent price impact and $\psi(0)y_t^* + \lambda z_t y_t^*$ is the transitory impact. This impulse response ξ_t is still *random* conditional on y_t^*, due to the white noise markup z_t. Its conditional variance is $\text{var}[\xi_t | y_t^*] = \lambda^2\sigma_Z^2(y_t^*)^2$. Hence, σ_Z can be written as

$$\sigma_Z = \frac{\sqrt{\text{var}\left[\xi_t \middle| y_t^*\right]}}{\left|\lambda y_t^*\right|}. \tag{17}$$

This expression measures how disperse the contemporaneous price response is relative to the long-term price impact, per unit of order flow surprise y_t^*. The wider is the dispersion, the more volatile is the random markup, reflecting higher degree of market making uncertainty and worse short-run order flow pricing efficiency.

5.2 Parameter estimation

This subsection discusses how to estimate the above structural model from a dataset of observed prices $\{p_t\}$ and order flows $\{y_t\}$ (hence, also $\{y_t^*\}$). The parameters of interest are the market making uncertainty measure σ_Z^2, the (permanent) price impact λ, how order flows affect pricing errors as

described by $\psi(L)$, and the autoregressive structure $\phi(L)$ for the pricing error.

This paper proposes to estimate these parameters using generalized method of moments (GMM). GMM is deemed advantageous over conventional maximum likelihood estimation (MLE) for this particular model. Notably, GMM requires no additional assumption on the joint distribution of μ_t, η_t, and z_t apart from they being independent white noises (unlike MLE, which typically would need Gaussianity). This is particularly relevant because, as will be seen shortly, the empirical estimates suggest positive skewness in z_t, rejecting Gaussianity, as the theory predicts.

Sufficiently many moment conditions are needed in order to apply GMM. To begin with, observe that the time t price returns can be written as (assuming $1 - \phi(L)$ is invertible)

$$\begin{aligned}\Delta p_t &= (1-L)p_t = (1-L)(m_t + s_t) = w_t + (1-L)s_t \\ &= (\lambda y_t^* + \mu_t) + \frac{1-L}{1-\phi(L)}(\psi(L)y_t + \varepsilon_t + \lambda z_t y_t^*) \\ &= \lambda y_t^* + \hat{y}_t + \underbrace{\mu_t + \frac{1-L}{1-\phi(L)}(\varepsilon_t + \lambda z_t y_t^*)}_{\text{the "remainder"}} \end{aligned} \tag{18}$$

where, in the third line, $\hat{y}_t$ is defined as $(1 - \phi(L))\hat{y}_t := (1 - L)\psi(L)y_t$. (A feasible way of constructing $\hat{y}_t$ is discussed in the next subsection.) Note from above that by removing λy_t^* and $\hat{y}_t$ from Δp_t, the remainder part has mean zero and is uncorrelated with all lags of y_t, thanks to the white noise assumption 3. The above allows to generate as many moment conditions as needed for identifying the lag polynomials $\phi(L)$ and $\psi(L)$ as well as λ:

$$\begin{aligned}&\mathbb{E}_y[\Delta p_t - \lambda y_t^* - \hat{y}_t] = 0 \\ &\mathbb{E}_y[(\Delta p_t - \lambda y_t^* - \hat{y}_t) y_{t-k}] = 0, \forall k \in \{0, 1, \ldots\}\end{aligned} \tag{19}$$

The subscript of the expectation operators above are understood as being conditional on all order flows $\{y_\tau, y_\tau^*\}_{\tau \le t}$.

To identify the second (and higher) moments of z_t, higher moments of the remainder are

exploited. Observe that the remainder can always be restructured into

$$
\begin{aligned}
r_t &= (1-\phi(L))(\Delta p_t - \lambda y_t^* - \hat{y}_t) = (1-\phi(L))\mu_t + (1-L)\varepsilon_t + (1-L)\lambda z_t y_t^* \\
&= \eta_t + \lambda \cdot \left(z_t y_t^* - z_{t-1} y_{t-1}^*\right),
\end{aligned}
$$

where the last equality rewrites $\eta_t := (1-\phi(L))\mu_t + (1-L)\varepsilon_t$, which is a collection of disturbances unrelated to trading (the "non-trade residual"). Note that $\{\eta_t\}$ has zero mean and its first $\bar{k}$ moments exist. Then the following second moment conditions holds:

$$
\begin{aligned}
&\mathbb{E}_y\left[r_t^2 - \sigma_\eta^2 - \lambda^2\sigma_Z^2 \cdot \left((y_t^*)^2 + \left(y_{t-1}^*\right)^2\right)\right] = 0 \\
&\mathbb{E}_y\left[\left(r_t^2 - \sigma_\eta^2 - \lambda^2\sigma_Z^2 \cdot \left((y_t^*)^2 + \left(y_{t-1}^*\right)^2\right)\right)(y_t^*)^2\right] = 0,
\end{aligned} \tag{20}
$$

which are just enough to identify σ_η^2 and σ_Z^2 The third moment of z_t can be identified in a similar way (requiring $\bar{k} \geq 3$ in assumption 3).

The above moment conditions demonstrate the strength of the GMM approach. There is no need to assume Gaussian disturbances, a much stronger assumption than assumption 3. The GMM approach can identify as many moments about z_t so long they exist. In section 6, the structure model is brought to data and both the second and the third moments of z_t are estimated. In particular, the estimates suggest positive skewness in z_t.

The disadvantage of this GMM approach is that it does not immediately identify moments of μ_t and ε_t separately. Instead, only the moments of $\eta_t = (1-\phi(L))\mu_t + (1-L)\varepsilon_t$ is identified. This is, however, not an important concern in the current paper where the focus is on measuring market making uncertainty σ_Z^2. Nevertheless, as shown in Chapter 8 of Hasbrouck (2007), for structural models of this type, the variance of w_t—the efficient price increment—can always be identified. That is, there exists an estimator for $\sigma_w^2 = \lambda^2 \mathrm{var}[y_t^*] + \sigma_\mu^2$, which can then be used together with the estimate of λ from GMM and $\mathrm{var}[y_t^*]$ from data to determine σ_μ^2. Further, from the second moment of η_t, σ_η^2 can then also be recovered. Hence, all parameters governing the data generating process can be properly estimated and, when Gaussian disturbances are valid, be applied to Kalman

filter to predict, filter, and smooth the unobservable states and disturbances without maximizing the likelihood.

5.3 Implementation

To utilize the GMM approach above, the series $\{y_t^*\}$, $\{\hat{y}_t\}$, and $\{r_t\}$ must be constructed from the raw data which only contains $\{y_t, p_t\}$. The order flow innovation $\{y_t^*\}$ can be readily estimated from an autoregressive regression on y_t. For example, Brogaard, Hendershott, and Riordan (2014) include 10 lags in their estimation. To construct $\{\hat{y}_t\}$ and r_t, the lag polynomials must be fixed first. For exposition clarity, consider the simple example of $\phi(L) = \phi L$ with $|\phi| < 1$ and $\psi(L) = \psi$. Through a Taylor expansion, $\hat{y}_t$ can be written as

$$\hat{y}_t = \frac{1-L}{1-\phi L}\psi y_t = \psi y_t - (1-\phi)\psi \sum_{k=1}^{\infty} \phi^{k-1} y_{t-k}.$$

Because $|\phi| < 1$, the above infinite sum converges quickly. Truncating it at some large number of lags, $\hat{y}_t$ can be constructed as a linear combination of the observable order flow series $\{y_t\}$ to arbitrary accuracy. Then r_t can be constructed by $\varepsilon_t = (1-\phi L)(\Delta p_t - \lambda y_t^* - \hat{y}_t)$. The same method applies to any other $\phi(L)$ and $\psi(L)$ structure.

It is perhaps useful to point out that because of the way $\hat{y}_t$ is constructed, the moment conditions are non-linear in the $\phi(L)$ coefficients. Numerical optimization is needed. For estimation efficiency, the applications in this paper adopt the continuously-updating GMM.

6 Empirical findings

This section applies the GMM approach to estimate the structural model, using real-world trading data. The preliminary empirical findings provide evidence supporting the theory of uncertain market making.

6.1 Data

While the theory speaks to general financial securities trading, the empirical analysis has to be specific. This section studies the U.S. equity market. To provide a representative and up-to-date overview, a sample of 500 stocks is randomly chosen from the S&P 1500 index for a one-year period from January to December 2014. The intraday trading data of these stocks are collected from the Monthly Trade and Quote (TAQ) database, accessed via Wharton Research Data Service (WRDS). The algorithm developed by Holden and Jacobsen (2014) is applied in order to alleviate the potential data issues associated with Monthly TAQ data. The authors' SAS script (dated June 26, 2014; retrieved from authors' websites) is directly adopted.

Daily stock information is collected from the Center for Research in Security Price (CRSP) for these 500 stocks from October to December 2013, i.e. three months before the sample period. These 500 stocks are then sorted into three groups—small, medium, and large—according to their daily average dollar trading volume during the three months. The large and medium stock groups have 150 stocks each, and the small stock group has the rest. Sorting on dollar volume is inspired by the theory prediction that market makers pay more attention to securities that have more activity (proposition 4). The grouping would be almost the same if instead the stocks are sorted on average daily market capitalization, which is the conventional choice in portfolio formation.

The structural model is estimated for each stock, each day. In each stock-day, series of $\{p_t\}$ and $\{y_t\}$ are constructed by sampling the 23,400 one-second snapshots of the trading hours. To avoid bid-ask bounces, the national best bid-offer (NBBO) midquote price at the end of each second is used for $\{p_t\}$. Each trade is signed using the algorithm proposed by Ellis, Michaely, and O'Hara (2000) (the results are robust to alternative signing algorithms). To facilitate comparison across stocks, the net order flow in each second $\{y_t\}$ is measured per \$10,000, following Brogaard, Hendershott, and Riordan (2014). The clock-time sampling follows the motivation of uncertain market making: In a short time interval, market makers' presence is uncertain. Instead, the time interval between two adjacent trades (event-time sampling) is endogenous of market conditions

like information, liquidity, etc., which might complicate the interpretation. The estimate results are qualitatively robust to other frequencies of clock-time snapshots (e.g., 5-second) and to share-based order flows.

6.2 Estimation result

The structural model developed in section 5 is estimated with lag polynomials of $\phi(L) = \phi L$ and $\psi(L) = \psi L$ in the pricing error equation (15). This parsimonious specification is consistent with Menkveld (2013) and Brogaard, Hendershott, and Riordan (2014) and, hence, allows comparison with their results. In constructing $\hat{y}_t$, 10 lags of y_t are used.

There are 248 trading days in the one-year sample period, which, together with the 500-stock cross-section, yields a maximum of 248×500 stock-day observations. The number of valid estimates falls short of this maximum for two reasons. First, if a stock-day has fewer than 1,000 trades or 20 intraday price changes, it is excluded from estimation. Second, even for stock-days passing this threshold, the numerical optimization does not always converge. The overall convergence rate is about 97.3%, with a breakdown of 98.6%, 98.1%, and 95.1% respectively for large, medium, and small stocks. The converged estimates yield a panel of 109,897 stock-day observations. Table 1 reports the estimation results. Figure 7 plots the time series of σ_Z for small, medium, and large stocks in the sample.

The estimates show that the order flows of the sample stocks, on average, generate a permanent price impact of $\lambda = 1.47$ basis points per \$10,000 in 2014. However, this permanent impact has a dispersion of $\sigma_Z = 15.31$ times large, or approximately 22.51 basis points (per \$10,000 order flow). Across the size terciles, the magnitude of dispersion—the standard deviation of the random markup z_t—is decreasing from 19.22 times for small stocks to 11.65 times for large stocks. Consistent with the theory, the cross-sectional variation of σ_Z suggests uncertain market making is more severe in less active stocks (where market makers are less attentive).

Such a large dispersion of permanent price impact is one source of the contemporaneous pricing

						Percentile in the full sample				
	Unit	All	Large	Medium	Small	5%	25%	50%	75%	95%
λ	bps/\$10,000	1.47**	0.04**	0.20**	10.23**	0.01	0.03	0.10	0.37	3.04
		(12.84)	(18.34)	(15.55)	(12.01)					
ψ	bps/\$10,000	-0.69**	-0.02**	-0.10**	-3.79**	-1.58	-0.15	-0.03	-0.01	0.01
		(-12.58)	(-15.56)	(-12.87)	(-12.06)					
σ_Z	times	15.31**	11.65**	15.56**	19.22**	3.75	6.41	9.08	14.08	43.46
		(60.28)	(43.71)	(39.10)	(45.42)					
σ_η	bps	0.78**	0.57**	0.68**	1.08**	0.32	0.50	0.67	0.95	1.59
		(45.99)	(41.59)	(38.77)	(43.25)					
skew[z_t]		2.94**	2.39**	2.93**	3.57**	-4.81	0.31	2.49	4.96	11.35
		(61.63)	(39.00)	(43.99)	(38.53)					
skew[η_t]		0.05	0.04	0.09	-0.01	-7.80	-1.66	-0.06	1.52	8.21
		(1.05)	(1.12)	(1.29)	(-0.18)					
ϕ		0.18**	0.15**	0.19**	0.19**	-0.39	-0.09	0.10	0.53	0.82
		(21.86)	(10.13)	(12.94)	(17.95)					
count		109,897	37,020	36,928	35,949					

Table 1: Estimates of the structural model. This table reports the summary statistics of the estimated parameters of the structural model (see more detailed discussion in section 5):

$$\begin{aligned}
\text{observed price:} \quad & p_t = m_t + s_t \\
\text{efficient price:} \quad & m_t = m_{t-1} + \lambda y_t^* + \mu_t \\
\text{pricing error:} \quad & s_t = \phi s_{t-1} + \psi y_t + \varepsilon_t + \lambda z_t y_t^* \\
\text{non-trade residual:} \quad & \eta_t = \mu_t - \phi \mu_{t-1} + \varepsilon_t - \varepsilon_{t-1}
\end{aligned}$$

The reported estimates include 1) the permanent price impact λ, 2) the contemporaneous pricing error correction ψ, 3) the volatility σ_Z of the random markup z_t, 4) the volatility σ_η of the non-trade residual η_t, 5) the skewness of the random markup z_t and of the non-trade residual η_t, and 6) the autoregressive coefficients ϕ for the pricing error. The sample averages of the estimates are separately reported for the full sample as well as the size terciles. The t-statistics, reported in brackets, are calculated using two-way clustered (stock and day) standard errors. In addition, selected percentiles of the estimates in the full sample are reported. The superscript ** indicates 1% statistical significance based on two-sided t-test.

error, which eventually reverts to zero for two reasons: 1) It is corrected by order flows at $\psi = -0.69$ basis points (per \$10,000 order flow); and 2) it quickly decays at a rate of $\phi = 0.18$ (per second).

Also reported in table 1 are the estimated skewness of the dispersion z_t and of the non-trade residual η_t. A statistically significant skewness is found for z_t (but not for η_t), consistent with the theory prediction (proposition 2). Also consistent with the friction of attention allocation, large stocks are found to have lower skewness in the markup than small stocks. Intuitively, the skewness exists because uncertain market making incentivizes market makers to mark up, but not down, the competitive price impact λ by a positive amount. Hasbrouck (2015) also finds more skewed price returns when there is more high-frequency quoting activity. The difference lies in that the skewness here specifically originates from a component of pricing error $\lambda z_t y_t^*$, which is *uncorrelated* with order flows. This finding also justifies the GMM approach in estimating the parameters, as the positive skewness rejects Gaussianity in z_t.

6.3 Uncertain market making and competition in quoting

Econometrically, the structural model in section 5 uses σ_Z to capture the transitory dispersion of the efficient price change λy_t^* (equation 17). Does this dispersion indeed reflect the economic force—market makers' competitiveness—as argued by the theory? A tentative answer is provided in this subsection by showing statistical significant correlation between σ_Z and some competition measures.

Following Hasbrouck (2015), four Herfindahl indices are constructed to measure the competitiveness in quoting activities. The quote files from TAQ data do not identify individual order submitters but document the reporting exchanges. The Herfindahl indices are thus calculated based on the exchange identifiers. As such, the result in this subsection should be cautiously read as indicative but not conclusive. Specifically, for each stock-day, let τ_j^{Ask} denote exchange j's 1) milliseconds at the best ask price, 2) milliseconds at the best ask price alone, 3) milliseconds multiplying the depth (in shares) at the best quotes, or 4) number of price improvements (reducing

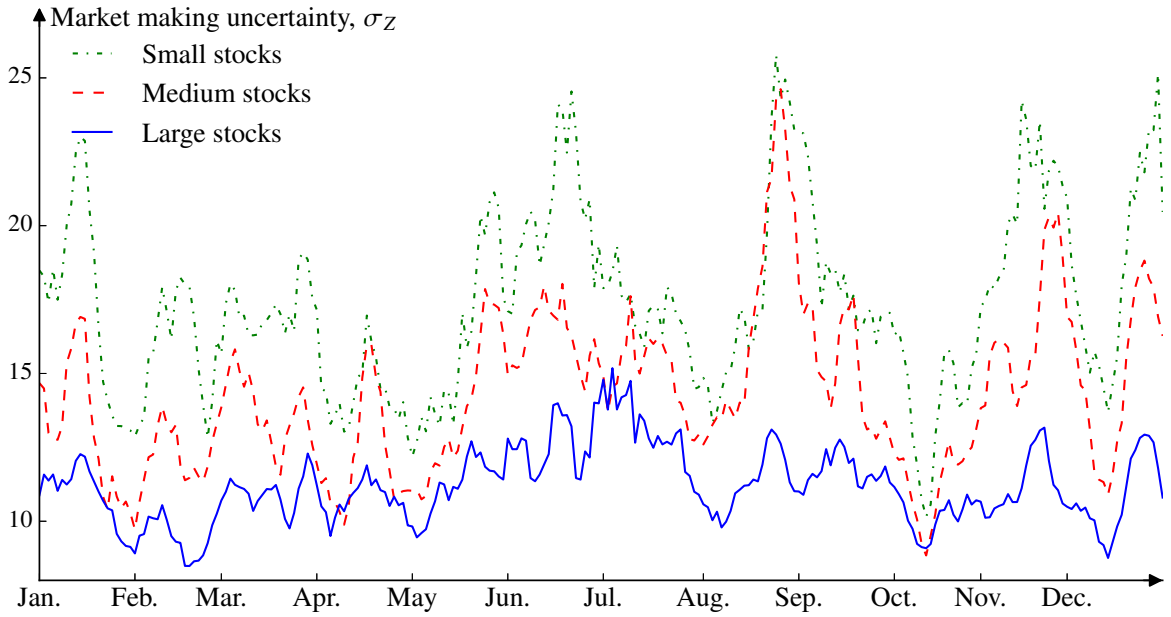


Figure 7: Time series of market making uncertainty. This figure plots the estimated the market making uncertainty—the dispersion of the net contemporaneous price impact, $\hat{\sigma}_Z$—over the one-year sample period of 2014. The series is plotted for large, medium, and small stocks respectively. The 5-day (weekly) moving averages of the σ_Z estimates are used for each size tercile throughout in 2014.

the best ask price). The bid side is defined symmetrically. The corresponding Herfindahl index is

$$h := \sum_j \left(\frac{\tau_j^{\text{Ask}} + \tau_j^{\text{Bid}}}{\sum_k \left(\tau_k^{\text{Ask}} + \tau_k^{\text{Bid}} \right)} \right)^2 .$$

Note that for the first three Herfindahl measures, interpolated time stamps in milliseconds are used to overcome the data limitation of Monthly TAQ data (Holden and Jacobsen, 2014). Rounding errors due to the interpolation are acknowledged.

To investigate whether and how uncertain market making connects with the competitiveness in the quoting activity, the following regressions are performed:

$$h(i,t) \sim \sigma_Z(i,t) + \text{controls}(i,t) + \text{stock fixed effects}(i) + \text{residual}(i,t)$$

where for stock i on day t, $h(i,t)$ is one of the four Herfindahl indices above and $\sigma_Z(i,t)$ is the

						Percentile in the full sample				
	Unit	All	Large	Medium	Small	5%	25%	50%	75%	95%
h_{Time}	%	16.46	14.87	16.15	17.94	10.60	13.15	15.16	17.39	23.40
$h_{\text{Alone time}}$	%	28.18	25.14	26.24	32.01	15.23	19.66	23.77	32.37	51.60
$h_{\text{Depth}\times\text{Time}}$	%	21.81	20.48	20.83	23.61	12.68	17.38	20.03	23.89	33.32
h_{Improves}	%	31.91	27.64	29.55	37.00	18.14	22.94	27.62	37.83	54.33
quotes/trades		11.64	10.61	10.58	13.32	4.72	7.16	9.74	13.52	21.56
ln($-volume)		82.20	216.20	48.47	10.53	3.48	11.97	37.54	98.42	364.02
net $-volume	$million	-0.30	-1.03	-0.04	-0.03	-7.69	-1.03	-0.02	0.81	6.16
spread	bps	11.59	3.70	7.16	21.38	2.03	3.63	6.40	12.47	25.95
depth	$1,000	0.03	0.07	0.03	0.02	0.01	0.01	0.01	0.02	0.16
return volatility, 1s	bps	0.95	0.66	0.80	1.30	0.41	0.58	0.77	1.08	1.80
variance ratio, 1s/10s		1.13	1.11	1.17	1.12	0.59	0.85	1.01	1.21	1.98
count		111,471	37,198	37,088	37,185					

Table 2: Summary statistics of the Herfindahl indices and other market quality measures. This table reports the summary statistics of the Herfindahl indices measuring the competitiveness of quoting activity as well as other market liquidity and volatility variables.

estimate of market making uncertainty from the structural model. The control variables include other structural estimates and a number of market quality measures. Specifically, for each stock-day, the following statistics are calculated to measure liquidity: "Quotes/trades" is the quote-to-trade ratio, defined as the number of quote activities (at the best prices) divided by the number of trades. "ln($-volume)" is the log dollar volume. "Net $-volume" is the net dollar volume. "Spread" is the time-weighted average bid-ask spread divided by the midquote price, measured in basis points. "Depth" is the time-weighted average dollar depth at the best quotes. The following two measures of market volatility is also included: "Volatility" is the realized standard deviation of returns sampled at one-second frequency. "Variance ratio" is the ratio of the one-second realized return variance over the 10-second realized return variance. Table 2 provides summary statistics for these variables.

	h_{Time}		$h_{\text{Alone time}}$		$h_{\text{Depth}\times\text{Time}}$		h_{Improves}	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Market making uncertainty								
$\sigma_Z/100$	0.29**	0.20*	0.95**	0.78**	0.37*	0.33*	-0.41	-0.40
	(2.81)	(2.10)	(3.04)	(2.81)	(2.49)	(2.26)	(-1.65)	(-1.77)
Other structural variables								
$\lambda/100$	3.82**	2.16*	12.24**	7.87**	4.07**	2.12	5.08*	0.07
	(4.03)	(2.40)	(4.32)	(2.98)	(2.65)	(1.44)	(2.19)	(0.03)
$\psi/100$	-3.78*	-2.93	-0.43	3.11	-1.00	0.35	1.43	6.34
	(-2.10)	(-1.65)	(-0.08)	(0.61)	(-0.36)	(0.12)	(0.33)	(1.49)
ϕ	-0.25**	-0.21**	-0.23*	-0.15	-0.34**	-0.25**	-0.27*	0.03
	(-5.55)	(-4.66)	(-2.26)	(-1.45)	(-4.62)	(-3.52)	(-2.56)	(0.26)
σ_η	0.64**	0.73**	-1.36**	-0.42	-0.26	0.42	3.82**	3.26**
	(4.60)	(5.14)	(-3.03)	(-0.94)	(-1.28)	(1.76)	(11.38)	(8.72)
skew[z_t]	0.00	0.00	0.02**	0.02**	0.01**	0.01**	-0.02**	-0.01**
	(1.21)	(1.00)	(3.17)	(2.91)	(2.96)	(3.01)	(-3.16)	(-2.64)
skew[η_t]	0.00	0.00	0.01	0.01	0.00	0.01	0.01*	0.02**
	(0.91)	(1.10)	(1.21)	(1.56)	(1.10)	(1.48)	(2.51)	(3.23)
Liquidity variables								
quote/trade		-0.01		-0.02		0.06**		0.29**
		(-1.14)		(-0.48)		(4.04)		(9.13)
ln($-volume)		-0.50**		-1.29**		-0.55**		-0.78**
		(-4.93)		(-3.78)		(-3.67)		(-3.21)
net $-volume		0.01*		0.03**		0.01**		0.01
		(2.27)		(2.78)		(2.63)		(0.99)
spread		0.01		0.03		-0.06**		-0.13**
		(0.66)		(0.70)		(-2.70)		(-2.79)

Continue on the next page ...

... continued from the previous page

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Liquidity variables (continued)								
$-depth		-5.42**		3.08		7.31**		-0.36
		(-5.03)		(1.46)		(4.81)		(-0.19)
Volatility variables								
volatility, 1s		-0.02		-0.20		-0.02		0.69*
		(-0.16)		(-0.64)		(-0.12)		(2.38)
variance ratio, 1s/10s		0.04		0.01		0.07		-0.05
		(1.48)		(0.07)		(1.55)		(-0.79)
stock dummies	yes	yes	yes	yes	yes	yes	yes	yes
R^2	0.004	0.012	0.003	0.006	0.001	0.007	0.021	0.054
observations	109,079	109,015	109,079	109,015	109,079	109,015	109,079	109,015

Table 3: Uncertain market making and competition in quoting activity. This table shows the result of regressing $h(i,t)$ on $\sigma_Z(i,t)$, where h is one of the four Herfindahl indices (time at best prices, time alone at best prices, depth×time at best prices, and number of quote improvements) estimated for each stock i on each trading day t. The explanatory variable is the market making uncertainty measure $\sigma_z(i,t)$ for stock i on day t (scaled by 1/100). The coefficients of other controls (see table 2) are also tabulated. For all regression specifications, a firm fixed-effect dummy is included. Two-way cluster (stock and day) robust standard errors are computed and the t-statistics are reported in brackets. The superscripts * and ** indicate 5% and 1% statistical significance, respectively, based on two-sided t-tests.

The estimation results are presented in table 3.[10] For each Herfindahl index, two separate models, with and without the market quality controls, are estimated with different explanatory variables. A significant positive coefficient of σ_Z on the Herfindahl indices is seen across all model specifications, except for quote improvements. As high Herfindahl indices suggest lack of competition, the regressions provide evidence consistent with the theory prediction that lack of market making competition is associated with more severe uncertain market making, proxied by

[10] In the regression, all structural variables are winsorized at 2% at tails, because outlier-like estimates are found in the numerical estimation. The results are robust to a range of winsorization thresholds.

the price impact dispersion σ_Z.

6.4 Market making uncertainty over the years

Conceptually, uncertain market making is not necessarily a new phenomenon pertaining to the market structure nowadays. So long as there are frictions preventing market makers from perfect competition in a short time interval, the argument for a random pricing equilibrium works. How the magnitude of market making uncertainty evolves over the years—especially in the recent decade witnessing the leap of technology—is an empirical question.

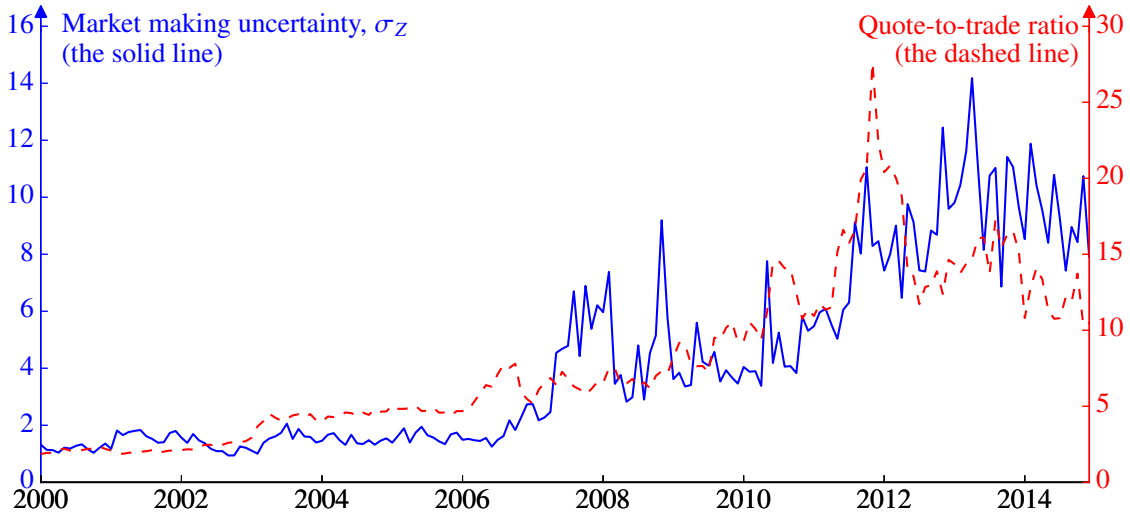
This subsection takes a first stab into this question by examining the evolution of σ_z over a fifteen-year sample period from the beginning of 2000 to the end of 2014. The sample is comprised of the 19 Dow Jones Industrial Average index components stocks that survived all these fifteen years. For each stock, the TAQ data of every trading Wednesday are collected and used to estimate the structural model (section 5) for that stock-Wednesday. The intraday TAQ data is aggregated into 4,680 5-second snapshots based on clock time[11] on which the structural model is estimated for each stock-Wednesday. This procedure yields 19 time series, each containing roughly 750 stock-Wednesday observations spanning the fifteen years. These 19 time series are then aggregated into a monthly average series which is plotted in figure 8.

Market making uncertainty is illustrated by the solid line in panel (a). A clear rising trend, most prominently after 2007, is observed. Before 2007, market making uncertainty remained at a relatively stable level of σ_Z slightly below 2.0. That is, before 2007, the net contemporaneous price impact has a dispersion of roughly 2 times of its mean. This dispersion soon swelled to roughly 6 times (of the mean) in mid 2007 and has since been rising up until roughly 10 times in 2014.

Panel (b) shows the general reduction trend in the permanent price impact (λ) over the fifteen-year period. The noticeable exception stands around the 2008-2010 financial crisis. The general

[11] Compared to subsection 6.2, the choice of 5-second snapshots here is made in order to account for the relatively low-frequency of market events in earlier years. Other sampling frequencies (e.g., 1-second, 10-second, and 30-second) have been explored and the patterns shown in figure 8 below remain qualitatively robust.

(a) Market making uncertainty



(b) Price impact

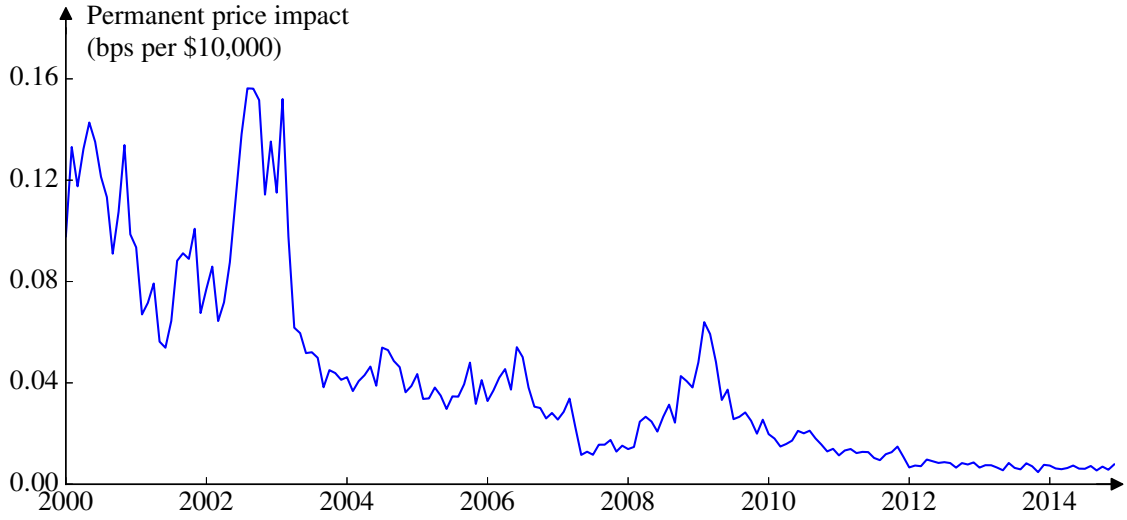


Figure 8: Fifteen years of market making uncertainty and price impact. This figure plots both the market making uncertainty σ_Z in panel (a) and the price impacts λ and $\lambda+\psi$ in panel (b) for a sample period of fifteen years, from the beginning of 2000 to the end of 2014. The dashed line in panel (a) also plots the quote-to-trade ratio (computed using only the quotes at the best prices and all trades) for the same period. The sample is comprised of the 19 Dow Jones Industrial Average index component stocks that survived all these fifteen years. The estimates are based on the structural model developed in section 5.

reduction in the price impact is consistent with the finding of improved market liquidity and reduced trading cost over the years (see, e.g., Angel, Harris, and Spatt, 2010; Hendershott, Jones, and Menkveld, 2011). Comparing panel (a) and (b) suggests that although the *long-run* price impact is reducing, in sharp contrast, its short-run dispersion—the second moment—does not seem narrowing.

This rising trend of market making uncertainty coincides in time with several other well-known phenomena, e.g. increasing algorithm and high-frequency trading, increasing quote-to-trade ratio, and market fragmentation. See, for example, the increasing quote-to-trade ratio demonstrated by the dashed line in panel (a). All of these market structure changes could have, at least to some extent, impaired market makers' ability to perfectly monitor the market in a short period of time. (To emphasize, the structural analysis in this paper is mainly descriptive and does not attempt to draw any causal conclusion.)

The turning point of 2007 is also evidenced by findings from other researches. Skjeltorp, Sojli, and Tham (2013) illustrate how the quote-to-trade ratio, a measure of algorithmic trading, evolves from 1999–2012. While a mild increase of the ratio is seen from 1999 to 2006, an abrupt rise is seen in 2007 and is most noticeable for large-cap stocks. Lyle, Naughton, and Weller (2015) find that the reduction in bid-ask spread starting from the beginning of this century stagnated around 2007. They emphasize the difference in market monitoring and algorithmic trading in general and find evidence that only the improvement in the former contributes to the reduction in the bid-ask spread. These observations are consistent with the theory of uncertain market making that too many market events (e.g. quoting activity) might negatively affect market monitoring efficiency. This in turn will worsen the pricing efficiency of order flows in the short-run, as is reflected in the rising σ_Z in panel (a) of figure 8.

7 Conclusion

This paper studies uncertain market making, a phenomenon due to the fact that 1) market makers face time constraint and that 2) they do not know how many competitors they are facing (during the constrained time). These two frictions are arguably more phenomenal in modern financial markets because of the increasingly challenge task of market making. Under uncertain market making, an analytically tractable model is developed and it is shown that there exists a random pricing equilibrium: Conditional on the order flows, market makers price the asset randomly.

By explicitly modeling the profit maximization of market makers, this paper offers several new implications for market quality based on the theory of uncertain market making. Prima facie evidence via the estimation of a structural model is provided in support of the theory of uncertain market making. Estimates show that in 2014, the order flows' (net contemporaneous) price impact has a dispersion of around 10 times of its sample mean, while in early 2000s this dispersion was only twice as large as the mean. This evidence suggests worsening short-run order flow pricing efficiency in the U.S. equity markets over the last decade.

Appendix

A Heteroskedasticity in the structural model

This appendix briefly discusses how price return heteroskedasticity is embedded in the structural model developed in section 5. Observe from equation (18) that the price return can always be written as

$$\begin{aligned}\Delta p_t &= A(L)y_t + \varsigma_t \\ \varsigma_t &= \mu_t + \frac{1-L}{1-\phi(L)}(\varepsilon_t + \lambda z_t y_t^*)\end{aligned}$$

where $A(L)$ is some generic lag polynomial. Note that the residual r_t is uncorrelated with y_t even though y_t is part of its linear construction, because μ_t, ε_t, and z_t are all assumed to be independent of y_t (assumption 3).

However, the time-t conditional variance of ς_t is *not* uncorrelated with y_t. To see this, consider a simple example of $\phi(L) = 0$ (which is consistent with, among others, Brennan and Subrahmanyam, 1996; and Sadka, 2006). Then

$$\begin{aligned}\varsigma_t^2 =&(\mu_t + \varepsilon_t + \lambda z_t y_t^*)^2 + \left(\varepsilon_{t-1} + \lambda z_{t-1} y_{t-1}^*\right)^2 \\ &- 2(\mu_t + \varepsilon_t + \lambda z_t y_t^*)\left(\varepsilon_{t-1} + \lambda z_{t-1} y_{t-1}^*\right)\end{aligned}$$

Conditional on the realized order flows, therefore,

$$\mathbb{E}\left[\varsigma_t^2 |\, y_t, y_{t-1}, \ldots\right] = \text{var}\mu + 2\text{var}\varepsilon + \lambda^2\sigma_Z^2 \cdot \left((y_t^*)^2 + \left(y_{t-1}^*\right)^2\right),$$

which depends on the realization of order flows y_t^*. Since the conditional variance of the residual to the price return is time-varying, such a structural model implies heteroskedasticity in the data.

Note that such heteroskedasticity exists only because of uncertain market making. In the case of $\sigma_Z^2 = 0$, i.e. without uncertain market making, $\mathbb{E}\left[r_t^2 |\, y_t, y_{t-1}, \ldots\right]$ is not time-varying. (This holds true irrespective of the specification of $\phi(L)$.)

Further properties of the heteroskedasticity can be derived but those are beyond the scope of this current manuscript. To emphasize, this appendix shows that uncertain market making contributes to the price return heteroskedasticity. It is acknowledged that alternative sources of heteroskedasticity exist (e.g., model misspecification, time-varying parameters, etc.). Further research is required to understand to what extent the heteroskedasticity predicted by uncertain market making can be differentiated from alternative explanations.

B Proofs

Lemma 1

Proof. Under assumption 1, the differential equation (9) simplifies to, using binomial formula,

$$(1-\theta F(z)) - (m-1)\theta \dot{F}(z) z = 0$$

and it can be verified easily that the function $F(\cdot)$ stated in the lemma satisfies the above differential equation. In solving the differential equation, a positive constant a arises and establishes the support for Z_i. □

Lemma 2

Proof. The market makers' strategy follows lemma 1. The insider's strategy is given by $x(v) = v/(2(1+\zeta)\Lambda)$, as derived in the paragraphs immediately before the proposition. It remains to solve $\zeta = \mathbb{E}Z$, where Z is defined as the minimum price impact markup among all active market makers' prices. To do so, note the c.d.f. of Z, conditional on $M \geq 1$, is:

$$\begin{aligned}
\mathbb{P}(Z < z |\, M \geq 1) &= \sum_{k=1}^{m} \binom{m}{k} \theta^k \cdot (1-\theta)^{m-k} \Big(1 - (1-F(z))^k\Big) / (1-(1-\theta)^m) \\
&= \sum_{k=0}^{m} \binom{m}{k} \theta^k \cdot (1-\theta)^{m-k} \Big(1 - (1-F(z))^k\Big) / (1-(1-\theta)^m) \\
&= (1-(1-\theta F(z))^m) / (1-(1-\theta)^m)
\end{aligned}$$

where the second equality adds a zero in the numerator, so that the third equality follows binomial formula. Substituting $F(\cdot)$ with the solution from lemma 1 yields

$$G(z) := \mathbb{P}(Z < z |\, M \geq 1) = \frac{1-(1-\theta)^m (a/z)^{\frac{m}{m-1}}}{1-(1-\theta)^m} \tag{B.1}$$

and density

$$\dot{G}(z) = \frac{\frac{m}{m-1}(1-\theta)^m a^{\frac{m}{m-1}} z^{-\frac{m}{m-1}-1}}{1-(1-\theta)^m} \tag{B.2}$$

for $(1-\theta)^{m-1} a \leq z \leq a$. Using this density, it can be evaluated that

$$\zeta = \mathbb{E}Z = \frac{(1-\theta)^{m-1}\theta m a}{1-(1-\theta)^m}.$$

Hence, the insider's aggressiveness parameter satisfies $\beta = 1/(2(1+\zeta)\lambda)$. From market makers' learning, $\lambda = \beta\sigma_V^2/(\beta^2\sigma_V^2 + \sigma_U^2)$. Together, these two conditions solve the equilibrium coefficients as stated in the lemma. □

Lemma 3

Proof. The expression of $G(z;\theta,m)$ is derived in equation (B.1). Its support is on $[(1-\theta)^{m-1}a, a]$, whose left bound decreases as either m or θ increases and the right bound does not change. Hence, when θ to θ' (or when m increases to m'), it suffices to verify whether $G(z;\cdot)$ increases or decreases for z on the original support $[(1-\theta)^{m-1}a, a]$. From expression (B.1), it is easy to verify that $\partial G(\cdot)/\partial m < 0$ and $\partial G(\cdot)/\partial\theta < 0$. Therefore, as either m or θ increases, $G(z;\cdot)$ also increases, proving the stochastic dominance. □

Lemma 4

Proof. The first step in this proof is to establish that the expected profit function (14) is monotone decreasing in the attention argument θ_j. To see this, substitute β_j with the expression of lemma 2 to get (for notation simplicity, the argument m is omitted)

$$\pi_j(\theta_j) = \frac{\left(1-\theta_j\right)^{m-1}}{\sqrt{1+2\zeta(\theta_j)}} a\sigma_{V,j}\sigma_{U,j}, \tag{B.3}$$

where ζ_j is written as a function of θ_j following lemma 2. Evaluating the derivative of ζ with respect to θ_j yields

$$\frac{\partial\zeta}{\partial\theta_j} = \frac{m-1}{1-\theta_j}\zeta(\theta_j) + \frac{\zeta(\theta_j)}{\theta_j} - \frac{\zeta(\theta_j)^2}{\theta_j a} > -\frac{m-1}{1-\theta_j}\zeta(\theta_j).$$

The last inequality follows because by definition, $\zeta(\theta_j) = \mathbb{E}Z \le a$, where Z has a support with upper bound a. Hence, evaluating the derivative of π_j with respect to θ_j gives

$$\begin{aligned}
\frac{\partial\pi_j}{\partial\theta_j} &= -\frac{(1-\theta_j)^{m-1}}{(1+2\zeta(\theta_j))^{3/2}}\left[\frac{m-1}{1-\theta_j}(1+2\zeta(\theta_j)) + \frac{\partial\zeta}{\partial\theta_j}\right] \\
&< -\frac{(1-\theta_j)^{m-1}}{(1+2\zeta(\theta_j))^{3/2}}\left[\frac{m-1}{1-\theta_j}(1+2\zeta(\theta_j)) - \frac{m-1}{1-\theta_j}\zeta(\theta_j)\right] \\
&= -\frac{(1-\theta_j)^{m-1}}{(1+2\zeta(\theta_j))^{3/2}}\frac{m-1}{1-\theta_j}\Big(1+\zeta(\theta_j)\Big) < 0.
\end{aligned} \tag{B.4}$$

Above it is shown that π_j is a monotone decreasing function in θ_j on $0 \le \theta_j \le 1$. It has

maximum of $\pi_j(0) = a\sigma_{V,j}\sigma_{U,j}$ and minimum of $\pi_j(1) = 0$. Because the monotonicity is strict, there exists an inverse function defined as

$$\pi_j^{-1}(\pi) = \begin{cases} \theta_j, & \text{for } \pi \in [0, a\sigma_{V,j}\sigma_{U,j}] \\ 0, & \text{for } \pi > a\sigma_{V,j}\sigma_{U,j} \end{cases}.$$

This inverse function is also strictly monotone decreasing on $[0, a\sigma_{V,j}\sigma_{U,j}]$.

Summing over all marketplaces, $\sum_{j=1}^{n} \pi_j^{-1}(\cdot)$ is a continuous function monotone decreasing from n to 0 on the support of $\left[0, \max_j \left\{a\sigma_{V,j}\sigma U, j\right\}\right]$. Clearly, by continuity, there exists a unique π^* on this support such that

$$\sum_{j=1}^{n} \pi_j^{-1}(\pi^*) = 1.$$

This π^* is the equilibrium expected profit for all market makers. Correspondingly, the equilibrium attention allocated to marketplace j is solved from the inverse function: $\theta_j = \pi_j^{-1}(\pi^*)$. Note that it is possible that some of the marketplaces are assigned with no attention at all, if the maximum expected profit is too low in those marketplaces (i.e. if $a\sigma_{U,j}\sigma_{V,j} < \pi^*$). □

Proposition 1

Proof. From lemma 3, it immediately follows that as either θ or m reduces, the expected markup ζ (= $\mathbb{E}Z$) increases. Importantly, ζ is uniquely determined by θ, m, and a following the expression given in lemma 2. It then remains to connect ζ to β and λ.

First, from lemma 2, $\beta = \sigma_U/(\sigma_V\sqrt{1+2\zeta})$, which is clearly decreasing in ζ. Hence, as θ or m reduces, ζ increases, and the insider's aggressiveness β reduces.

Second, from equation (6), it can be easily seen that λ is monotone increasing in β for $0 \leq \beta \leq \sigma_U/\sigma_V$. Therefore, as θ or m reduces, ζ increases, and both β and λ decrease.

Finally, from the insider's first order condition, the total trading cost $(1+\zeta)\lambda$ is simply the inverse of 2β. As a result, the total trading cost increases with market making uncertainty. □

Proposition 2

Proof. The skewness of Δp conditional on order flow is defined as

$$\frac{\mathbb{E}\left[(\Delta p - \mathbb{E}[\Delta p|\, Y\,])^3 |\, Y\right]}{(\mathrm{var}[\Delta p|\, Y\,])^{3/2}} = \frac{\mathbb{E}\left[(Z-\zeta)^3\right]}{(\mathrm{var}[1+Z])^{3/2}},$$

the sign of which only depends on the numerator. Using the density of Z derived in equation (B.2), it can be easily shown that the numerator above is indeed positive for all $m \geq 2$. □

Proposition 3

Proof. Consider two arbitrary independent random variables, Y and Λ. The kurtosis of a random variable is defined as $\kappa[Y] = \mathbb{E}(Y - \mathbb{E}Y)^4/(\mathbb{E}(Y - \mathbb{E}Y)^2)^2$. Note that kurtosis is invariant of the mean of the random variable. Without loss of generality, therefore, suppose $\mathbb{E}Y = 0$. Then

$$\kappa[\Lambda Y] = \frac{\mathbb{E}(\Lambda Y)^4}{(\mathbb{E}(\Lambda Y)^2)^2} = \frac{\mathbb{E}\Lambda^4 \mathbb{E}Y^4}{(\mathbb{E}\Lambda^2)^2(\mathbb{E}Y^2)^2} = \frac{\mathbb{E}\Lambda^4}{(\mathbb{E}\Lambda^2)^2}\kappa[Y].$$

Denote $X = \Lambda^2$. Then $\mathbb{E}\Lambda^4 = \mathbb{E}X^2 = \text{var}X + (\mathbb{E}X)^2$. Hence,

$$\kappa[\Lambda Y] = \left(\frac{\text{var}X}{(\mathbb{E}X)^2} + 1\right)\kappa[Y] \geq \kappa[Y].$$

The inequality is strict as long as $\text{var}X > 0$, i.e. when Λ is not degenerate. In the model, $\Lambda = (1 + Z)\lambda$, where Z reflects market making uncertainty. □

Proposition 4

Proof. From the proof of lemma 4, it is established that there exists an inverse function $\theta_j = \pi_j^{-1}(\pi)$ strictly decreasing on $\pi \in [0, a \cdot \sigma_{U,j} \cdot \sigma_{V,j}]$. Fix any expected profit level π. Then by implicit function theorem,

$$\frac{\partial \theta_j}{\partial \sigma_{U,j}} = -\frac{\partial \pi_j / \partial \sigma_{U,j}}{\partial \pi_j / \partial \theta_j} > 0,$$

where the inequality holds because the numerator $\partial \pi_j / \partial \sigma_{U,j} > 0$ (from equation B.3) and the denominator is negative (equation B.4).

Now consider an equilibrium where $\sum_{j=1}^{n} \pi_j^{-1}(\pi^*) = 1$ for some π^* (see the proof of lemma 4) and denote by θ_j^* the equilibrium attention in marketplace j. Suppose in marketplace h, σ_{Uh}^2 increases (and nothing else changes). Then, $\pi_h^{-1}(\pi; \sigma_{U,h}^{'}) \geq \pi_h^{-1}(\pi; \sigma_{U,h})$ for $\sigma_{U,h}^{'} \geq \sigma_{U,h}$ at any fixed level of π. Holding everything else the same, hence,

$$\pi_h^{-1}(\pi^*; \sigma_{U,h}^{'}) + \sum_{j \neq h} \pi_j^{-1}(\pi^*) > 1,$$

implying the new equilibrium must be some $\pi^{\#} > \pi^*$, as each of these inverse function is monotone decreasing. In this new equilibrium, $\theta_j^{\#} = \pi_j^{-1}(\pi^{\#}) < \pi_j^{-1}(\pi^*) = \theta_j^*$, $\forall j \neq h$, because π_j^{-1} is monotone decreasing. That is, in the new equilibrium, the attention allocated to all marketplaces

other than h is lower than in the old equilibrium. Since the sum of all attention allocated should be unity,

$$\theta_h^{\#} = 1 - \sum_{j \neq h} \pi_j^{-1}(\pi^{\#}) > 1 - \sum_{j \neq h} \pi_j^{-1}(\pi^*) = \theta_h^*.$$

Therefore, in the new equilibrium, the attention allocated to marketplace h, where $\sigma_{U,h}$ increased to $\sigma'_{U,h}$, is higher than in the old equilibrium. The same analysis holds true for shocks in $\sigma_{V,h}$ and is omitted for brevity. □

Corollary 1

Proof. From the proof of proposition 4, if either σ^2_{Uj} or σ^2_{Vj} increases, in the new equilibrium, the attention allocated to all marketplaces other than j is lower than in the old equilibrium. Proposition 1 then applies. □

References

Angel, James J., Lawrence E. Harris, and Chester S. Spatt. 2010. "Equity Trading in the 21st Century." Working paper.

Back, Kerry and Shmuel Baruch. 2013. "Strategic Liquidity Provision in Limit Order Markets." *Econometrica* 81 (1):363–392.

Back, Kerry, Kevin Crotty, and Tao Li. 2014. "Estimating the Order-Flow Component of Security Returns." Working paper.

Banerjee, Snehal and Brett Green. 2015. "Signal or noise? Uncertainty and learning about whether other traders are informed." *Journal of Financial Economics* 117 (2):398–423.

Baruch, Shmuel and Lawrence R. Glosten. 2013. "Fleeting Orders and the Competitive Equilibrium." Working paper.

Berman, Gregg E. 2014. "What Drives the Complexity and Speed of our Markets?" Speech.

Biais, Bruno, David Martimort, and Jean-Charles Rochet. 2000. "Competing Mechanisms in a Common Value Environment." *Econometrica* 68 (4):799–837.

———. 2013. "Corrigendum to 'Competing Mechanisms in a Common Value Environment'." *Econometrica* 81 (1):393–406.

Bondarenko, Oleg. 2001. "Competing market makers, liquidity provision, and bid-ask spareds." *Journal of Financial Markets* 4 (3):269–308.

Brennan, Michael J. and Avanidhar Subrahmanyam. 1996. "Market Microstructure and Assetp Pricing: On the Compensation for Illiquidity in Stock Returns." *Journal of Financial Economics* 41:441–464.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2014. "High Frequency Trading and the 2008 Short Sale Ban." Working paper.

Burdett, Kenneth and Kenneth L. Judd. 1983. "Equilibrium Price Dispersion." *Econometrica* 51 (4):955–969.

Butters, Gerard R. 1977. "Equilibrium Distributions of Sales and Advertising Prices." *The Review of Economic Studies* 44 (3):465–491.

Cao, Melanie and Shouyong Shi. 2001. "Screening, Bidding, and the Loan Market Tightness." *European Finance Review* 5 (1-2):21–61.

Cespa, Giovanni and Thierry Foucault. 2014. "Illiquidity Contagion and Liquidity Crashes." *The Review of Financial Studies* 27 (6):1615–1660.

Corwin, Shane A. and Jay F. Coughenour. 2008. "Limited Attention and the Allocation of Effort in Securities Trading." *The Journal of Finance* 63 (6):3031–3067.

Dennert, Jürgen. 1993. "Price Competition between Market Makers." *The Review of Economic Studies* 60 (3):735–751.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2015. "Benchmarks in Search Markets." Working paper.

Ellis, Katrina, Roni Michaely, and Maureen O'Hara. 2000. "The Accuracy of Trade Classification Rules: Evidence from Nasdaq." *Journal of Financical and Quantitative Analysis* 35 (4):529–551.

Glosten, Lawrence R. 1989. "Insider Trading, Liquidity, and the Role of the Monopolist Specialist." *Journal of Business* 62 (2):211–235.

Glosten, Lawrence R. and Lawrence E. Harris. 1988. "Estimating the Components of the Bid-Ask Spread." *Journal of Financial Economics* 21:123–142.

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents." *Journal of Financial Economics* 42 (1):71–100.

Hasbrouck, Joel. 1991. "The Summary Informativeness of Stock Trades: An Econometric Analy-

sis." *Review of Financial Studies* 4 (3):571–595.

———. 1993. "Assessing the Quality of a Security Market: A New Approach to Transaction-Cost Measurement." *Review of Financial Studies* 6 (1):191–212.

———. 2007. *Empirical Market Microstructure: the Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, New York.

———. 2015. "High Frequency Quoting: Short-Term Volatility in Bids and Offers." Working paper.

Hausch, Donald B. and Lode Li. 1993. "A Common value Auction Model with Endogenous Entry and Information Acquisition." *Economic Theory* 3 (2):315–334.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" *The Journal of Finance* 66 (1):1–33.

Hendershott, Terrence and Albert J. Menkveld. 2014. "Price pressures." *Journal of Financial economics* 114 (3):405–423.

Holden, Craig W. and Stacey Jacobsen. 2014. "Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions." *The Journal of Finance* 69 (4):1747–1785.

Jovanovic, Boyan and Albert J. Menkveld. 2015. "Dispersion and Skewness of Bid Prices." Working paper.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica* 53 (6):1315–1336.

———. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies* 56 (3):317–356.

Lyle, Matthew R., James P. Naughton, and Brian M. Weller. 2015. "How Does Algorithmic Trading Improve Market Quality?" Working paper.

Menkveld, Albert J. 2013. "High Frequency Trading and the *New Market* Makers." *Journal of Financial Markets* 16 (4):712–740.

Menkveld, Albert J., Siem Jan Koopman, and André Lucas. 2007. "Modelling Round-the-Clock Price Discovery for Cross-Listed Stocks using State Space Methods." *Journal of Business & Economic Statistics* 25 (2):213–225.

O'Hara, Maureen. 2015. "High Frequency Market Microstructure." *Journal of Financial Economics* 116 (2):257–270.

Rossi, Stefano and Katrin Tinn. 2014. "Man or machine? Rational trading without information

about fundamentals." Working paper.

Sadka, Ronnie. 2006. "Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk." *Journal of Financial Economics* 80 (2):309–349.

Salop, Steven and Joseph Stiglitz. 1977. "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion." *The Review of Economic Studies* 44 (3):493–510.

Skjeltorp, Johannes A., Elvira Sojli, and Wing Wah Tham. 2013. "Trading on Algos." Working paper.

Varian, Hal R. 1980. "A Model of Sales." *American Economic Review* 70 (4):651–659.

Yang, Ming. 2015. "A Market Order Model When Insider May Not Exist." Working paper.

List of Figures

List of Tables