Modeling bid and ask price dynamics with an extended Hawkes process and its empirical applications for high-frequency stock market data

Kyungsub Lee [*]   Byoung Ki Seo [†]

Abstract

This study proposes a versatile model for the dynamics of the best bid and ask prices using an extended Hawkes process. The model incorporates the zero intensities of the spread-narrowing processes at the minimum bid-ask spread, spread-dependent intensities, possible negative excitement, and nonnegative intensities. We apply the model to high-frequency best bid and ask price data from US stock markets. The empirical findings demonstrate a spread-narrowing tendency, excitations of the intensities caused by previous events, the impact of flash crashes, characteristic trends in fast trading over time, and the different features of market participants in the various exchanges.

1 Introduction

A considerable volume of research on high-frequency trading, quotes, and financial data is available. High-frequency quotes and trading, which are the main sources of high-frequency financial data, are considered to be (automated) trading generally based on a (mathematical) algorithm that can generate a large number of quotes and trades over a short time horizon. However, the precise definition varies across studies and regulatory entities. Numerous studies consider the impact of high-frequency trading on market quality in terms of liquidity, transaction costs, volatility, price discovery, social benefits, and its role in market crashes.

Studies of market microstructures and high-frequency data rely on statistical models that can precisely analyze market data such as order duration, transaction time, and price changes. One effective method is based on the Hawkes process (Hawkes, 1971a,b; Hawkes and Oakes, 1974), a type of point process for estimating the parameters related to the impacts of events. Researchers first applied Hawkes models to seismology, but presently apply them extensively throughout the natural and social sciences, including finance. For example, the earliest studies applying the Hawkes process to financial markets include Hewlett (2006), who introduces a symmetric bivariate Hawkes process to examine the arrival times of trades and price impacts; Large (2007), who examines market resilience; and Bowsher (2007), who analyzes the relationship between trading times and mid-price changes. Since then, the model has often been applied to financial data. Bacry et al. (2015), Law and Viens (2015), and Hawkes (2018) summarize recent developments in the application of the Hawkes process to finance.

[*] ksublee@yu.ac.kr, Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 38541, Korea.

[†] (corresponding author) bkseo@unist.ac.kr, School of Business Administration, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea.

The basic Hawkes model captures several interesting properties of price dynamics at the micro level (Chavez-Demoulin and McGill, 2012; Bacry and Muzy, 2014; Da Fonseca and Zaatour, 2014) such as order clustering, microstructure noise, and the impacts of past events. The Hawkes process is suitable for modeling the intensities of a series of events such as price changes in tick structure and order arrivals. In addition, other studies aim to clarify the complex microstructure of the limit order book and price dynamics using extended models (Zheng et al., 2014; Morariu-Patrichi and Pakkanen, 2018; Hainaut and Goutte, 2019; Jang et al., 2020). Our approach is also based on modifying the existing Hawkes model to allow applications to bid and ask price dynamics. Owing to its versatility, this extended Hawkes process allows us to infer useful information from high-frequency financial data and understand the market microstructure and high-frequency activities.

Specifically, we develop a rigorous and novel probability model that describes the nature of bid-ask price dynamics. The proposed model incorporates characteristics that cannot be described by the basic Hawkes model. Bid-ask spreads are positive and our model guarantees this property. If the spread widens, competition among market makers to replenish liquidity increases under typical market conditions and thus the proposed model has spread-level-dependent intensities.

In stock market microstructures, a bid or ask price change is expected to increase the likelihood of events such as new order arrivals and successive price changes occurring. The basic Hawkes model can describe these excitation properties. However, if price changes occur such that spreads decline, intensity may not increase. For example, if the spread reduces to a tick size (i.e., it reaches the minimum level), then no further events that reduce the spread occur; that is, the associated intensities that narrow the spread become zeros. Since the basic Hawkes model always has positive intensities, it needs to be modified using, for example, negative excitements. To add a negative excitement term into the model, the model design must ensure that intensity is nonnegative to satisfy mathematical rigor.

Related ideas can be found in the past literature. An extended Hawkes process that partially describes non-exciting relationships was proposed in Brémaud and Massoulié (1996). Along that line, Hansen et al. (2015) considered multivariate counting processes with inhibitory effects in a non-linear setting whose importance is also mentioned in Bacry et al. (2016).

We apply this model to high-frequency financial data and make the following interesting findings. First, we find evidence of a growing number of high-frequency traders even though not all stocks show the same monotone increasing growth pattern over time. Second, liquidity provision and removal behaviors have different responsiveness. In all the stocks in our sample, we find higher instantaneous responsiveness in liquidity provision than in removal. Third, we also observe different features from different exchanges under market-making activity. Being a highly fragmented market, stock market exchanges have different fee structures, policies, and technology, which leads their participants to have different characteristics.

We also report the dramatic changes in market conditions under special circumstances such as the Flash Crash of 2010. Although we perform most of our analyses on a daily basis, because of the nature of high-frequency financial data, sufficient data are available for a few minutes, thereby allowing us to conduct the analysis on a near real-time basis. Lee and Seo (2017a) also show this property. Using sophisticatedly designed parameters, we can thus observe how the behaviors of high-frequency traders change during the crash, recovery period, and next day.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 presents the extended Hawkes model of the dynamics of the best bid and ask prices.

The proposed model incorporates the zero intensities of the spread-narrowing processes at the minimum bid-ask spread, spread-dependent intensities, possible negative excitement, and non-negative intensities. Section 5 discusses the findings from our empirical study performed by applying the presented model to high-frequency financial data. The results demonstrate increasing trends in low-latency quotes and trading, the characteristics of stocks and exchanges, a comparison between liquidity provision and removal, and the interesting behaviors of high-frequency traders during and after the Flash Crash. Section 6 concludes.

2 Literature review

This section reviews studies analyzing high-frequency trading and high-frequency financial data. The main advantage of high-frequency trading, especially quotes, is the supply of liquidity. Many studies demonstrate that the liquidity provided by active high-frequency quotes and large numbers of outstanding orders reduces the bid-ask spread and that transaction costs decrease accordingly (Bershova and Rakhlin, 2013; Chaboud et al., 2014; Stoll, 2014). In addition, Hagströmer and Nordén (2013) and Menkveld (2013) report that many high-frequency traders are liquidity providers. However, a proportion of this liquidity is an illusion in the abundance of fleeting orders (Hasbrouck and Saar, 2009), and some researchers express skepticism about whether it is sustainable during a market crisis (Cvitanic and Kirilenko, 2010), even though high-frequency trading is not generally considered to be the primary originator of the Flash Crash (Kirilenko et al., 2017).

The rapidity of high-frequency trading is based on high-end, state-of-the-art computing, network equipment, and costly financial services such as colocation, which involves installing traders' servers as close to trade matching engines as possible. However, while this rapidity can help provide liquidity (Brogaard et al., 2015), high-frequency traders armed with new equipment may take advantage of low-frequency traders (Jarrow and Protter, 2012; van Kervel, 2015; Biais et al., 2015). Theoretical and empirical research thus suggests that the rapidity of high-frequency trading can be harmful as well as beneficial (Foucault and Moinas, 2018). The general assumption is that the speed of high-frequency trading is a positive aspect of price discovery as it allows the market to rapidly process new information and reflect it in new prices (Manahov and Hudson, 2014). However, whether information processing and quote updates that occur just a few seconds or milliseconds earlier are of practical utility remains an open question (Brogaard et al., 2014).

Because volatility is an important indicator of the market conditions, measuring the volatility of prices or returns more accurately using high-frequency financial data has long been a topic of study (Zhou, 1996; Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2003). In addition, the effect of high-frequency trading on volatility is under debate (Bollen and Whaley, 2015; Kelejian and Mukerji, 2016; Hasbrouck, 2018). In some ways, high-frequency activities such as the exploitation of arbitrage opportunities in high-frequency trading, crowding effect, flickering quotes, and high order cancellation rates are sources of higher volatility, but some studies find no significant statistical relations between high-frequency activities and volatility. Chung and Lee (2016), Menkveld (2016), and Virgilio (2019) provide excellent summaries of the arguments in the literature above.

The research interest in high-frequency trading is not limited to the above discussion; it extends to the optimal strategies of market makers and investors based on high-frequency trading

(Avellaneda and Stoikov, 2008; Guilbaud and Pham, 2013; Cartea and Jaimungal, 2013; Cartea et al., 2018; Choi et al., 2021) and limit order book modeling (Cont and de Larrard, 2013; Toke, 2015). Bouchaud et al. (2002), Hollifield et al. (2004), and Huang et al. (2015) describe the statistical properties of limit order books.

3 The model

3.1 Definition

In this section, we model the best bid and ask price dynamics of financial assets using point processes. The best bid and ask price processes consist of

A^u: the number of upward movements in the best ask process,

A^d: the number of downward movements in the best ask process,

B^u: the number of upward movements in the best bid process, and

B^d: the number of downward movements in the best bid process,

defined on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ satisfying the usual conditions. Each process is a random counting measure that counts the number of events in any Borel set in the timeline, $\mathbb{R}$. We can represent each process using a stochastic process (i.e., nondecreasing integer-valued random step functions) by denoting $A^u(t) = A^u(\omega, (0, t])$ for $\omega \in \Omega$ and similarly denoting the other counting measures. We assume the point processes are simple, i.e.,

$$\mathbb{P}\{N(\{t\}) = 0 \text{ or } 1 \text{ for all } t\} = 1$$

for $N = A^u, A^d, B^u$ or B^d. We can categorize the counting processes further as

$$\begin{aligned} A^u, B^d &: \text{ the bid-ask spread-widening processes} \\ A^d, B^u &: \text{ the bid-ask spread-narrowing processes.} \end{aligned}$$

We establish a system of counting processes with the corresponding conditional intensities:

$$\boldsymbol{N}_t = \begin{bmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \\ N_4(t) \end{bmatrix} = \begin{bmatrix} A^u(t) \\ A^d(t) \\ B^u(t) \\ B^d(t) \end{bmatrix}, \quad \boldsymbol{\lambda}_t = \begin{bmatrix} \lambda_1(t) \\ \lambda_2(t) \\ \lambda_3(t) \\ \lambda_4(t) \end{bmatrix} = \begin{bmatrix} \lambda_A^u(t) \\ \lambda_A^d(t) \\ \lambda_B^u(t) \\ \lambda_B^d(t) \end{bmatrix}.$$

We can express the conditional intensities as the expected number of events over an infinitesimal interval $\mathrm{d}t$:

$$\lambda_i(t)\mathrm{d}t \approx \mathbb{E}[N_i(t + \mathrm{d}t) - N_i(t) | \mathcal{F}_{t-}],$$

where $\mathcal{F}_{t-}$ is the σ-field generated by $\boldsymbol{N}$ at times up to but not including t. Thus, the intensities are left continuous and the counting processes are right continuous. We have two types of notations, number-based, such as N_i and λ_i, and Roman-based, A^u and λ_A^u. We use them interchangeably depending on the context.

The following model is used in our main empirical study in Section 5. Slightly different versions of the model are also examined in Section 4. The motivation and property of the model is further explained in Subsection 3.2.

Definition 1. The intensities of the up and down movements of the best bid and ask prices are assumed to be

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu}_t + \int_{-\infty}^{t} \boldsymbol{h}(t,u)\mathrm{d}\boldsymbol{N}_u, \tag{1}$$

where

$$\boldsymbol{\mu}_t = \begin{bmatrix} \mu \\ f(\ell(t-)) \\ f(\ell(t-)) \\ \mu \end{bmatrix},$$

where ℓ denotes the level of the bid-ask spread relative to the mid-price and

$$f(\ell(t-)) = \eta\ell(t-) \tag{2}$$

with a constant parameter, $\eta > 0$. The relative level is defined by

$$\ell(t) = \frac{L(t)}{p(t)},$$

where $p(t)$ is the mid-price and $L(t) \in \{0, 1, 2, \cdots\}$ is the absolute level of the bid-ask spread with $L(t) = 0$ implying the minimum level. The matrix of excitement terms is based on an exponential kernel:

$$\boldsymbol{h}(t,u) = \mathrm{e}^{-\beta(t-u)} \begin{bmatrix} \alpha_{s1} & \alpha_m & \alpha_{s2} & 0 \\ \alpha_{w1} & \alpha_{n1}(u) & \alpha_{n1}(u) & \alpha_{w2} \\ \alpha_{w2} & \alpha_{n2}(u) & \alpha_{n2}(u) & \alpha_{w1} \\ 0 & \alpha_{s2} & \alpha_m & \alpha_{s1} \end{bmatrix},$$

where $\alpha_{s1}, \alpha_{s2}, \alpha_m, \alpha_{w1}$, and α_{w2} are parameters under the condition that

$$\alpha_{s1} + \alpha_{s2} + \alpha_m < \beta \tag{3}$$

and α_{n1} and α_{n2} are stochastic terms with

$$\alpha_{n1}(u) = -\sum_{j=1}^{4} \lambda_{2j}(u) + \xi\ell(u), \quad \alpha_{n2}(u) = -\sum_{j=1}^{4} \lambda_{3j}(u) + \xi\ell(u), \tag{4}$$

for constant $\xi \geq 0$ and λ_{ij} is a component of λ_i such that

$$\lambda_{ij}(t) = \int_{-\infty}^{t} h_{ij}(t,u)\mathrm{d}N_j(u).$$

We use relative level rather than absolute level in the modeling of $\boldsymbol{\mu}$ and $\boldsymbol{h}$ for empirical research. In the US stock market, the tick size is fixed at \$0.01, while the price of a stock varies from stock to stock. However, for mathematical analysis, the absolute level model is more tractable, which will be discussed in Subsection 4.3.

In the definition, we use the left continuous version of $\ell(t-)$ to define $\boldsymbol{\mu}$ to ensure that the intensity process is left continuous with a right limit. The base intensities of the spread-narrowing processes, A^d and B^u, increase as the spread increases. When $\ell(t-) = 0$ (i.e., the spread is at its minimum), the base intensities of A^d and B^u are equal to zero.

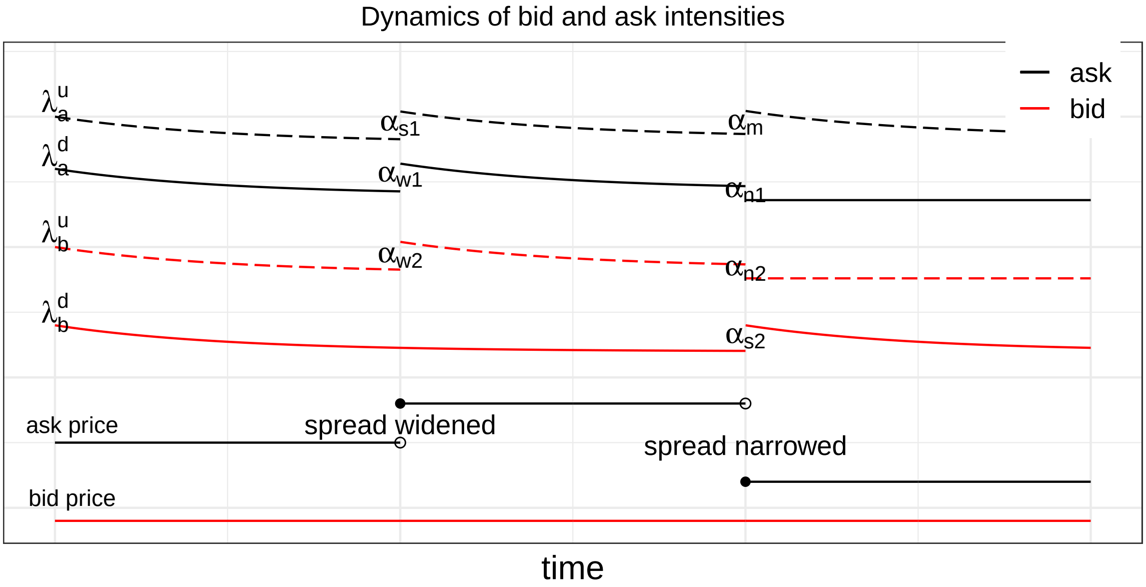


Figure 1: Illustration of the intensity processes in the proposed model

The stochastic terms α_{n1} and α_{n2} control the amounts of changes in the spread-narrowing intensities right after the spread narrows. We introduce these terms as stochastic to ensure mathematical rigor which will be discussed in the next subsection. We cannot rule out the possibility that ξ has different values for α_{n1} and α_{n2} in practice, but we assume that ξ takes the same value in both for model parsimony.

The coefficients in the top right and left bottom corners of the matrix $\boldsymbol{h}$ are intentionally set equal to zero for model parsimony, since the interaction between upward movement in the best ask and downward movement in the best bid can negligible especially in ultra-high-frequency. The empirical evidence will be discussed in Subsection 5.3.1.

An example of intensity changes by the events under the model in Definition 1 is shown in Figure 1. An hypothetical situation where the bid-ask spread widens first and then decreases to the minimum level is illustrated. Note that the spread-narrowing intensities λ_b^u and λ_a^d become zeros right after the spread hits the minimum.

3.2 Discussion of the model

Previous studies often assume that the mid-price process follows a self and mutually excited Hawkes process, which has empirical support (Bacry et al., 2013; Da Fonseca and Zaatour, 2014; Lee and Seo, 2017b). Self-excitement means that upward or downward movements of mid-prices are excited by past events in their own processes and are often considered to be due to the order splitting. Mutual excitement means that downward movement excites upward movement and vice versa. Because the mid-price is the mean of the best bid and ask prices, it is natural to assume that we can use a Hawkes-type model to represent the bid and ask price processes in this context. The intensities in the Hawkes model consist of the base intensities and excitations by past events. The typical intensity process in the Hawkes model takes the form of Eq. (1)

with constant base intensity $\boldsymbol{\mu}$ and constant excitation kernel $\boldsymbol{h}$.

Our goal of modeling is to find the latent information in the level-one limit order book by modeling the base intensities and excitation matrix appropriately. To define a model as in Definition 1, the followings are considered. The model is proposed to incorporate the following properties in addition to Hawkes-type features in the mid-price process.

Zero intensity with minimum spread

When the spread between the best bid and ask prices reaches the minimum tick size, the ask price does not fall and the bid price does not rise before the spread widens. Mathematically, the intensities of the spread-narrowing processes, $A^d(t)$ and $B^u(t)$, should be zero at the minimum level of the bid-ask spread until a new event occurs. Thus, $\lambda_A^d(t)$ and $\lambda_B^u(t)$ do not follow the intensities of the classical Hawkes model in which the intensities can be close to zero, but are always greater than zero. Hence, it is necessary to tweak the model to set both the background intensities and the excitations equal to zero when the spread reaches the minimum.

Spread-dependent intensity

When the bid-ask spread is wide, we assume that it tends to narrow depending on its level. This property may be due to market makers who quote limit orders with aggressive prices between the widened bid and ask orders, providing market liquidity. As the spread increases, the potential profit for market makers increases, and market making hence becomes more active. We can observe this tendency in two ways. One is the increase in the base intensities, regardless of previous events, and the other comes from the excitations due to past events. Our model includes both aspects. The base intensity rates of the spread-narrowing processes, A^d and B^u, increase as the spread widens. In addition, the overall intensities of the processes are excited immediately after the spread-widening events by the Hawkes-type kernel.

Possible negative excitement when the spread narrows

If spread-narrowing events occur, we can assume that these activities can slow the rate of the additional arrivals of aggressive quotes. In other words, the intensities of the spread-narrowing processes, $A^d(t)$ and $B^u(t)$, may diminish, or even vanish, when the spread becomes minimal. We can express decreases in intensities due to a series of activities as a negative term in the Hawkes-type kernel.

Avoiding negative intensity

As mentioned above, negative excitements (i.e., diminishing intensity) can occur, but we should control this amount to ensure that intensity is nonnegative. Thus, the amount of excitement in the spread-narrowing processes after a spread-narrowing event depends on the intensities of the narrowing processes.

Simultaneous movement of bid and ask prices

Theoretically, bid and ask prices can move simultaneously in rare circumstances. For example, assume 100 volumes of the best ask price in a limit order book. When a marketable limit order to buy 120 shares at the best ask price arrives, it consumes the existing best ask orders and the previous second-best ask order becomes the best. In addition, the remaining market order of 20 volumes converts to the bid order, which becomes the best bid. This feature is difficult to interpret as a mathematical model, since we assume simple processes in general, where only one jump occurs in an infinitesimal time interval. In practice, we assume a small time interval between two events. We can model this using the network, computational, and transaction delays in a trading system, and hence from raw data, or by data preprocessing, as we explain in Section 5.2.

The parameters in the excitation kernel, $\alpha_{s1}, \alpha_{s2}, \alpha_m, \alpha_{w1}$, and α_{w2}, increase the intensities after certain events. Increased intensities induce successive arrivals of corresponding events. Our empirical study shows that the parameter estimates are large owing to the abundant high-frequency trading in the highly liquid US equity markets. Next, we discuss the role of each constant parameter.

- α_{w1}: This parameter is related to liquidity provision. If α_{w1} is large and the other parameters are small, then the bid and ask prices are likely to follow the series of movements in Figure 3a when a spread-widening event occurs. From the market-making perspective, this parameter is associated with the replenishment speed for limit orders. When a large market order sweeps the existing best limit orders or all the best orders are canceled, the bid-ask spread widens. If α_{w1} is sufficiently large, then liquidity providers tend to replenish the depleted limit orders immediately. These activities may occur through (automated) market makers or rebate-seekers in maker-taker exchanges to supply liquidity.

- α_{w2}: This parameter is also related to liquidity provision and is similar to α_{w1}. The relevant bid and ask prices behaviors with large α_{w2} is presented in Figure 3b when a spread-widening event occurs. With large α_{w2}, the widened bid-ask gap tends to narrow immediately, but by a different type of limit order from a depleted limit order. If these events occur consecutively in the same direction, then significant mid-price changes may eventually result.

- α_{s1}: This parameter is related to removing liquidity. If α_{s1} is large, then we can observe increments in the ask price or decrements in the bid price by successive limit order depletions, as in Figure 3c. If the order depletion in the same direction happens in a very short time, it may be related to anticipatory order cancellation, which means that when a large market order arrives, the market maker considers it to be an informed trader's action and immediately cancels its limit orders to prevent adverse selection (Fox et al., 2019). If the market is sufficiently liquid, as the US equity market then successive depletions and the corresponding widened spread are likely to be replenished immediately with large α_ws.

- α_{s2}: This parameter is also related to liquidity removal and related movements are described in Figure 3d when a spread-narrowing event occurs. As in the case of α_{w2}, if these events occur consecutively, then price changes may eventually result.

- α_m: This parameter is also related to removing liquidity. When α_m is large, if a better limit order than the current best price arrives, then the new limit order is executed or canceled immediately, and the price rebounds to its original status, as in Figure 3e. This result may occur because market participants may consider the new order to be good and high-frequency traders execute the market order immediately. Alternatively, a trader may issue a limit order aggressively and cancel the order instantly to avoid adverse selection, especially in an uncertain market. Order cancellations are abundant in stock markets. Scholtus and van Dijk (2015) report that 60% of all orders are canceled within one second in the S&P 500 exchange-traded funds traded on the Nasdaq from 2009 to 2011. Hasbrouck and Saar (2009) refer to limit orders canceled immediately as fleeting

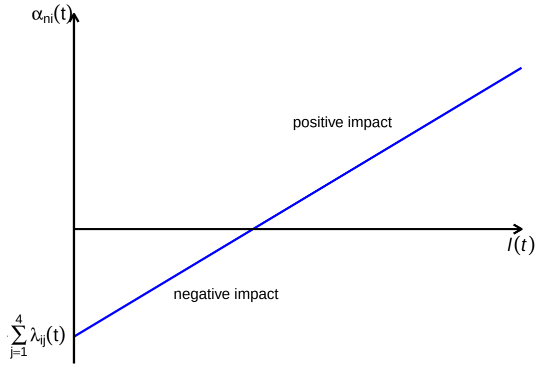


Figure 2: $\alpha_{ni}(t)$ versus the relative spread in a linear model

orders. Repeated order submissions and cancellations are also known as flickering quotes, a concept introduced more than two decades ago (Harris, 1999).

The terms α_n can be negative in contrast to the classical Hawkes model in which t he α parameters are nonnegative. Although $\alpha_n(t)$s may be negative, however, they are bounded by $-\sum_{j=1}^{4} \lambda_{ij}(t)$ below, and hence λs are always nonnegative. More precisely, for $i = 3$ or 4, and assume that there is a spread narrowing event at time u, then

$$\lambda_i(u+) = \lambda_i(u) + \alpha_{nm}(u) \geq \lambda_i(u) - \sum_{j=1}^{4} \lambda_{ij}(u) = \eta\ell(u) \geq 0$$

where $m = 1$ for $i = 2$ and $m = 2$ for $i = 3$. In addition, under the symmetric setting, i.e., ξ is the same in α_{n1} and α_{n2}, we have $\lambda_2 = \lambda_3$ and $\alpha_{n1} = \alpha_{n2}$.

Figure 2 shows the impact of the spread-narrowing intensities when the spread narrows. The figure shows that the impact can be positive or negative depending on the level. If the level is high, even if the spread narrows, the speed at which it narrows can still increase. If the level is sufficiently low, then the spread-narrowing event reduces the rate at which the spread decreases. The y-intercept of the line is $-\sum_{j=1}^{4} \lambda_{ij}(t)$, a $\mathcal{F}_{t-}$-measurable random variable, and the slope ξ is a constant parameter. This stochastic y-intercept ensures that the negative excitement does not make the intensities less than zero.

We can rewrite the intensity processes as

$$
\begin{aligned}
\lambda_A^u(t) &= \mu + \int_{-\infty}^t \alpha_{s1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^u(u) + \int_{-\infty}^t \alpha_m\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^d(u) + \int_{-\infty}^t \alpha_{s2}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^u(u)\\
\lambda_A^d(t) &= f(\ell(t-)) + \int_{-\infty}^t \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^u(u) + \int_{-\infty}^t \alpha_{w2}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^d(u)\\
&\qquad + \int_{-\infty}^t \alpha_{n1}(u)\mathrm{e}^{-\beta(t-u)}(\mathrm{d}N_A^d(u) + \mathrm{d}N_B^u(u))\\
\lambda_B^u(t) &= f(\ell(t-)) + \int_{-\infty}^t \alpha_{w2}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^u(u) + \int_{-\infty}^t \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^d(u)\\
&\qquad + \int_{-\infty}^t \alpha_{n2}(u)\mathrm{e}^{-\beta(t-u)}(\mathrm{d}N_A^d(u) + \mathrm{d}N_B^u(u))\\
\lambda_B^d(t) &= \mu + \int_{-\infty}^t \alpha_{s2}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^d(u) + \int_{-\infty}^t \alpha_m\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^u(u) + \int_{-\infty}^t \alpha_{s1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^d(u).
\end{aligned}
$$

In the classical Hawkes model, the parameter αs are restricted by β for nonexplosive solutions. Intuitively, a large α causes intensity to increase rapidly. For example, if α is greater than β, which describes the decrease in intensity over time, then the increment in intensity caused by α is much larger than the decrease caused by β, and intensity explodes. By contrast, in our model, α_w, which relates to the spread-narrowing processes, can have a larger number than β. If λ_A^d or λ_B^u rapidly increases owing to a large α_w, then the spread goes to zero and λ_A^d or λ_B^u vanishes according to the model definition. This feature makes it possible to estimate α_w without restriction which represents market resiliency.

Note that to fully characterize the mid price and spread processes, we need additional information about the size of movement in the price processes. The mid-price process is

$$
p(t) = p(0) + \int_0^t \delta_A^u(u)\mathrm{d}N_A^u(u) + \int_0^t \delta_B^u(u)\mathrm{d}N_B^u(u) - \int_0^t \delta_A^d(u)\mathrm{d}N_A^d(u) - \int_0^t \delta_B^d(u)\mathrm{d}N_B^d(u)
$$

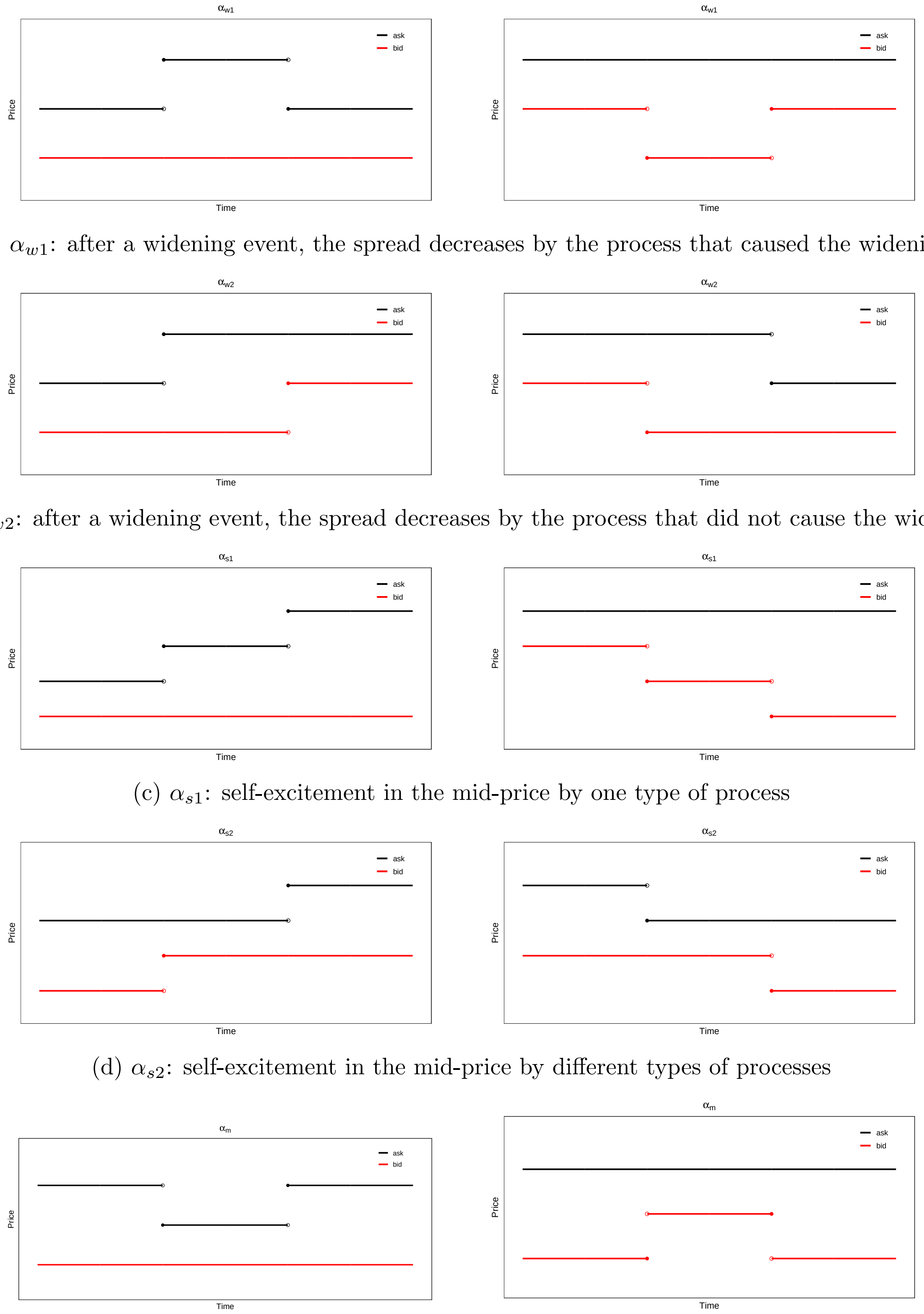
where $\delta_A^u, \delta_B^u, \delta_A^d$ and δ_B^d denote the corresponding jump sizes. Similarly, the spread process is

$$
L(t) = L(0) + \int_0^t \delta_A^u(u)\mathrm{d}N_A^u(u) + \int_0^t \delta_B^d(u)\mathrm{d}N_B^d(u) - \int_0^t \delta_A^d(u)\mathrm{d}N_A^d(u) - \int_0^t \delta_B^u(u)\mathrm{d}N_B^u(u).
$$

However, the inclusion of the jump size to the model make introduce marked point processes, which is beyond the scope of our paper. We assume that the jump sizes at time t are exogenously given, i.e., $\delta(t)$s are $\mathcal{F}_t$-measurable but not measurable to the internal history of $\boldsymbol{N}$. In our model, only if the jump sizes are all constant unity, then the mid-price and spread processes are endogenously determined.

4 Mathematical analysis

In this section, we investigate the mathematical properties of the model. However, since the proposed model in Definition 1 is too complicated, slightly different versions are investigated instead in Subsections 4.2 and 4.3. This does not provide a complete proof of the stability of the proposed model, but will indirectly provide insight.



(a) α_{w1}: after a widening event, the spread decreases by the process that caused the widening

(b) α_{w2}: after a widening event, the spread decreases by the process that did not cause the widening

(c) α_{s1}: self-excitement in the mid-price by one type of process

(d) α_{s2}: self-excitement in the mid-price by different types of processes

(e) α_m: mutual excitement in the mid-price by one type of process

Figure 3: Possible bid and ask price movements when the corresponding α is large and the other αs are close to zero

4.1 Reset property

The stochastic terms α_ns reset the excitement terms in the spread-narrowing intensities, λ_2 and λ_3, to $\xi\ell$ right after a spread-narrowing event. Let $\{\tau_k\}$ be a sequence of the spread-narrowing event times. For $i = 2$ and 3,

$$\lambda_i(\tau_k+) = f(\tau_k) + \xi\ell(\tau_k)$$

where the excitement term is reset to $\xi\ell(\tau_k)$ which only depends on the current level of spread. In addition, for $\tau_k < t < \tau_{k+1}$,

$$\lambda_i(t) = f(\ell(t-)) + \xi\ell(\tau_k)\mathrm{e}^{-\beta(t-\tau_k)} + \int_{\tau_k}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_1(u) + \int_{\tau_k}^{t} \alpha_{w2}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_4(u). \tag{5}$$

The above equation implies that whenever a spread-narrowing event happens, the excitement term reset to $\xi\ell$ with decreasing rate β and up to the next spread-narrowing event, typical constant Hawkes mutually excitement terms, α_ws, are added with decreasing rate β on the occurrences of N_1 and N_4. This is because the excited parts of the spread-narrowing intensities before time τ_k are canceled out by

$$-\sum_{j=1}^{4} \lambda_{ij}(\tau_k), \text{ for } i = 2 \text{ and } 3.$$

Especially, when the spread hits the minimum by a spread-narrowing event, the reset term $\xi\ell$ becomes zero and hence the excitements of intensities of the spread-narrowing processes also become zero.

4.2 Mid-price process under constant model

In this subsection, we examine the properties of the mid-price process from a slightly different version of our model. In the linear model of f and α_n in Definition 1, there is a rather complicated aspect of the analysis, so let's simplify the model a little bit. Because the mid-price is the average of the bid and ask prices, we can derive the mid-price process from the bid and ask price processes. Let N^u and N^d be the counting processes for the upward and downward movements of the mid-price process; then,

$$N^u = N^u_A + N^u_B, \quad N^d = N^d_A + N^d_B$$

and

$$\lambda^u = \lambda^u_A + \lambda^u_B, \quad \lambda^d = \lambda^d_A + \lambda^d_B.$$

Thus, in the symmetric setting (i.e., $\alpha_n := \alpha_{n1} = \alpha_{n2}$),

$$\begin{aligned}
\lambda^u(t) = \mu + f(\ell(t-)) &+ \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u_A(u) + \int_{-\infty}^{t} (\alpha_{s2} + \alpha_n(u))\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u_B(u) \\
&+ \int_{-\infty}^{t} (\alpha_m + \alpha_n(u))\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d_A(u) + \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d_B(u) \\
\lambda^d(t) = \mu + f(\ell(t-)) &+ \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u_A(u) + \int_{-\infty}^{t} (\alpha_m + \alpha_n(u))\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u_B(u) \\
&+ \int_{-\infty}^{t} (\alpha_{s2} + \alpha_n(u))\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d_A(u) + \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d_B(u).
\end{aligned}$$

As mentioned before, we adopt a simpler model in this subsection, a constant model for f and α_n; that is,

$$f(\ell(t-)) = \mathbf{1_{\{\ell(t-)>0\}}}\eta$$

and

$$\alpha_{n1}(t) = -\mathbf{1_{\{\ell(t)=0\}}}\sum_{\mathbf{j=1}}^{\mathbf{4}}\lambda_{\mathbf{2j}}(\mathbf{t}) + \mathbf{1_{\{\ell(t)>0\}}}\xi, \quad \alpha_{\mathbf{n2}}(\mathbf{t}) = -\mathbf{1_{\{\ell(t)=0\}}}\sum_{\mathbf{j=1}}^{\mathbf{4}}\lambda_{\mathbf{3j}}(\mathbf{t}) + \mathbf{1_{\{\ell(t)>0\}}}\xi,$$

for some constants $\eta > 0$ and $\xi > 0$. This model is simpler than the linear model, but it consistent with the general concept in Subsection 3.1. This model guarantees zero intensities of the spread-narrowing processes at the minimum spread as in the linear model. The base intensities for the spread-narrowing processes become zero right after the spread hit the minimum due to the fact that $f(0) = 0$, and the excitation parts are canceled out by

$$-\sum_{j=1}^{4}\lambda_{2j}(t).$$

By canceling the exact amount of the existing excitation parts, it guarantees the nonnegative intensities. Once the spread becomes larger than zero, the model acts like a basic Hawkes model with constant parameters η and ξ.

Under the special circumstances, the mid-price process of the constant model is similar to that of the classical Hawkes model. We assume that the spread is sufficiently large; that is, $\mathbb{P}(\ell(t) = 0)$ is close to zero under the constant model. Then,

$$\begin{aligned}
\lambda^u(t) \approx \mu + \eta &+ \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^u(u) + \int_{-\infty}^{t} (\alpha_{s2} + \xi)\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^u(u)\\
&+ \int_{-\infty}^{t} (\alpha_m + \xi)\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^d(u) + \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^d(u)\\
\lambda^d(t) \approx \mu + \eta &+ \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^u(u) + \int_{-\infty}^{t} (\alpha_m + \xi)\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^u(u)\\
&+ \int_{-\infty}^{t} (\alpha_{s2} + \xi)\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_A^d(u) + \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_B^d(u).
\end{aligned}$$

For further simplification, suppose that $\alpha_{s1} + \alpha_{w2} = \alpha_{s2} + \xi$ and $\alpha_{w1} = \alpha_m + \xi$, which is used for demonstration purposes. Then,

$$\begin{aligned}
\lambda^u(t) &\approx \mu + \eta + \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u(u) + \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d(u)\\
\lambda^d(t) &\approx \mu + \eta + \int_{-\infty}^{t} \alpha_{w1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^u(u) + \int_{-\infty}^{t} (\alpha_{s1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N^d(u),
\end{aligned}$$

which is a typical symmetric self- and mutually excited Hawkes process for the mid-price dynamics. To describe the full dynamics of the mid-price process, we require a jump size distribution.

4.3 Spread process under absolute level model

This subsection investigate the stability in mean of the spread process under a slightly different version. The spread is the difference between the bid and ask prices; thus, the bid and ask price processes can induce counting processes for the number of upward or downward movements of the spread process. Let N_s^u and N_s^d be the counting processes for the upward and downward movements of the spread process, respectively. Then,

$$N_s^u = N_A^u + N_B^d, \quad N_s^d = N_A^d + N_B^u$$

and

$$\lambda_s^u = \lambda_A^u + \lambda_B^d, \quad \lambda_s^d = \lambda_A^d + \lambda_B^u.$$

Thus, in the symmetric setting,

$$\lambda_s^u(t) = 2\mu + \int_{-\infty}^t \alpha_{s1} \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^u(u) + \int_{-\infty}^t (\alpha_{s2} + \alpha_m) \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^d(u) \tag{6}$$

$$\lambda_s^d(t) = 2f(L(t-)) + \int_{-\infty}^t (\alpha_{w1} + \alpha_{w2}) \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^u(u) + \int_{-\infty}^t \alpha_n(u) \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^d(u). \tag{7}$$

The intensity of the upward movements of the spread process is the classical Hawkes model and that of the downward movements is the Hawkes model with stochastic excitement. As mentioned previously, to describe the full dynamics of the spread process, we require a jump size distribution of up and down price movements, which is beyond the scope of this study. In addition, for mathematical analysis, it is more tractable when modeling with absolute level of spread (especially in the base intensity) in Eq. (7).

Assume that the jump size is one so that the spread process is represented by the difference between N_s^u and N_s^d. In addition, let us start with a simple case such that

$$\lambda_s^u(t) = 2\mu + \int_{-\infty}^t \alpha_{s1} \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^u(u) + \int_{-\infty}^t (\alpha_{s2} + \alpha_m) \mathrm{e}^{-\beta(t-u)} \mathrm{d}N_s^d(u) \tag{8}$$

$$\lambda_s^d(t) = 2\eta L(t-) \tag{9}$$

i.e., without the excitement terms in $\lambda_s^d(t)$. Since the differential form of $\lambda_s^u(t)$ can be written as

$$\mathrm{d}\lambda_s^u(t) = \beta(2\mu - \lambda_s^u(t))\mathrm{d}t + \alpha_{s1}\mathrm{d}N_s^u(t) + (\alpha_{s2} + \alpha_m)\mathrm{d}N_s^d(u),$$

we have

$$\mathbb{E}[\lambda_s^u(t)|\mathcal{F}_0] = \lambda_s^u(0) + \int_0^t 2\{\beta\mu + (\alpha_{s1} - \beta)\mathbb{E}[\lambda_s^u(u)] + (\alpha_{s2} + \alpha_m)\mathbb{E}[\lambda_s^d(u)]\}\mathrm{d}u.$$

In addition,

$$\lambda_s^d(t) = 2\eta L(t-) = 2\eta(L(0) + N_s^u(t-) - N_s^d(t-))$$

and

$$\mathbb{E}[\lambda_s^d(t)|\mathcal{F}_0] = 2\eta L(0) + \int_0^t 2\eta(\mathbb{E}[\lambda_s^u(u)] - \mathbb{E}[\lambda_s^d(u)])\mathrm{d}u.$$

Thus, we construct a system of differential equation,

$$\begin{aligned}
\frac{\mathrm{d}\mathbb{E}[\lambda_s^u(t)|\mathcal{F}_0]}{\mathrm{d}t} &= 2\beta\mu + (\alpha_{s1} - \beta)\mathbb{E}[\lambda_s^u(t)] + (\alpha_{s2} + \alpha_m)\mathbb{E}[\lambda_s^d(t)] \\
\frac{\mathrm{d}\mathbb{E}[\lambda_s^d(t)|\mathcal{F}_0]}{\mathrm{d}t} &= 2\eta(\mathbb{E}[\lambda_s^u(t)] - \mathbb{E}[\lambda_s^d(t)]).
\end{aligned}$$

The matrix of the system of differential equation is

$$A = \begin{bmatrix} \alpha_{s1} - \beta & \alpha_{s2} + \alpha_m \\ 2\eta & -2\eta \end{bmatrix}.$$

It is known that the stability conditions of the system is that the trace of A is negative and the determinant is positive, i.e.,

$$\mathrm{tr}(A) = \alpha_{s1} - \beta - 2\eta < 0, \quad \det(A) = 2\eta(\beta - \alpha_{s1} - \alpha_{s2} - \alpha_m) > 0.$$

Since $\eta > 0$, if Eq. (3) holds, then this stability condition satisfies. In addition, under the steady-state of the differential equation system,

$$\begin{bmatrix} \mathbb{E}[\lambda_s^u(t)] \\ \mathbb{E}[\lambda_s^d(t)] \end{bmatrix} = -A^{-1} \begin{bmatrix} 2\beta\mu \\ 0 \end{bmatrix}$$

and this yields that the steady-state average intensity is

$$\mathbb{E}[\lambda_s^u(t)] = \mathbb{E}[\lambda_s^d(t)] = \frac{2\beta\mu}{\beta - \alpha_{s1} - \alpha_{s_2} - \alpha_m}$$

and

$$\mathbb{E}[L(t)] = \frac{\mathbb{E}[\lambda_s^d(t)]}{2\eta} = \frac{\beta\mu}{\eta(\beta - \alpha_{s1} - \alpha_{s_2} - \alpha_m)}.$$

Therefore, under the Eqs. (8) and (9), the expected intensities and the absolute level of the spread are finite.

Adding exciting terms to λ_s^d in Eq. (9) does not change much of the discussion above. As in the previous, let $\{\tau_k\}$ be a sequence of the spread-narrowing event times. For $\tau_k < t < \tau_{k+1}$,

$$\lambda_s^u(t) = 2\mu + (\lambda_s^u(\tau_k) - 2\mu)\mathrm{e}^{-\beta(t-\tau_k)} + \int_{\tau_k}^t \alpha_{s1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_s^u(u)$$

and $\lambda_s^u(t)$ only has a self-exciting feature, and hence if Eq. (3) holds, there are only finite number of events by N_s^u during the period $[\tau_k, \tau_{k+1}]$ almost surely and $\lambda_s^u(t)$ will not explode to infinity. In addition, for $\tau_k < t < \tau_{k+1}$, by recalling the reset property for $\lambda_s^d = \lambda_2 + \lambda_3$ in Eq (5),

$$\lambda_s^d(t) = 2\eta L(t-) + \xi L(\tau_k)\mathrm{e}^{-\beta(t-\tau_k)} + \int_{\tau_k}^t (\alpha_{w1} + \alpha_{w2})\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_s^u(u)$$

is also finite almost surely, because there are only finite events by N_s^u (even when $\alpha_{w1}+\alpha_{w2} > \beta$).

Since the processes are stable for $\tau_k < t < \tau_{k+1}$, it is sufficient to check the stability at each time of $\{\tau_k\}$. By recalling the reset property, the intensities right after a spread-narrowing event can be represented by

$$\lambda_s^u(\tau_k+) = 2\mu + \int_{-\infty}^{\tau_k} \alpha_{s1}\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_s^u(u) + \int_{-\infty}^{\tau_k} (\alpha_{s2} + \alpha_m)\mathrm{e}^{-\beta(t-u)}\mathrm{d}N_s^d(u) \tag{10}$$

$$\lambda_s^d(\tau_k+) = 2\eta L(\tau_k). \tag{11}$$

These are discrete-time versions of Eqs. (8) and (9) and as long as Eq. (3) holds, we can expect the stability at times of $\{\tau_k\}$.

5 Empirical study

Our model allows us to examine various aspects of the first level order book in high-frequency. This section includes the following topics.

- Since the intensities are functions of the bid-ask spread level, the model captures the narrowing tendency of the spread when the spread is large, see Subsection 5.3.2.
- The parameter αs measure the responsiveness rates of the market to the previous events, see also Subsection 5.3.2.
- For each stock, by comparing liquidity providing parameters, α_ws, and removing parameters, α_m and α_ss, we can examine the market-making speed in various stocks, see Subsection 5.3.3.
- We also found some stocks received particular attentions from ultra-high-frequency liquidity providers, see Subsection 5.3.4.
- The estimation results also show that the propensities of participants differ for each stock exchange, see Subsection 5.4.
- The estimation result of η in an almost continuous-time provides another perspective during the Flash Crash, see Subsection 5.5.

Before the main empirical analysis results, we explain the method, data and model selection.

5.1 Likelihood estimation and simulation study

We estimate the parameters of the model using the maximum log-likelihood method. The method to calculate the log-likelihood of our model is similar to that for calculating the log-likelihood of the classical Hawkes process. For discussions on maximum likelihood estimations for the stationary point process, see Ogata (1978) and Ozaki (1979). The log-likelihood of the model up to time T is

$$\log \mathcal{L}(\theta, T) = \sum_{i=1}^{4} \left(\int_0^T \log \lambda_i(t) \mathrm{d}N_i(t) - \int_0^T \lambda_i(t) \mathrm{d}t \right),$$

where the intensity functions should be considered to be left continuous and θ denotes the set of model parameters. We calculate the location of the maximum log-likelihood using a numerical optimizer in R; for more details, see Nash (2014).

We conduct simulations to check whether the numerical optimizer works well in our model. We generate 500 simulated sample paths using presumed parameter values and then estimate the parameters for each sample path. For the simulation of the Hawkes process, we can apply the thinning algorithm of Ogata (1981) or the exact method of Dassios et al. (2013). Thus, we obtain 500 estimation results for each parameter. Each sample path has 10,000 observations of bid or ask price changes. The presumed parameter values in the simulation are the estimates from the real stock data in the later empirical study. Likewise, the distributions of the jump sizes in the price processes are based on empirical distributions. Table 1 summarizes the results.

Table 1: Simulation study with 500 sample paths

	μ	η	α_{s1}	α_{s2}	α_m	α_{w1}	α_{w2}	β	ξ
True	0.080	0.100	4.000	26.00	5.000	11.00	7.000	50.00	2.700
Mean	0.080	0.100	3.993	26.02	5.003	11.02	6.973	50.03	2.705
Std.	0.003	0.002	0.360	0.658	0.277	0.477	0.404	0.796	0.085
True	0.170	0.140	200.0	250.0	150.0	300.0	330.0	1200	50.00
Mean	0.170	0.140	199.6	252.3	151.3	301.5	332.0	1207	50.04
Std.	0.003	0.003	8.369	8.362	7.327	9.371	9.210	7.570	2.441

The row labeled "mean" is the average of the 500 estimates calculated from the 500 sample paths and the row labeled "std." is the standard deviation of the estimates. The results show that the numerical estimator is unbiased with sufficiently small standard deviations.

Since the log-likelihood function is not convex, it may be questionable whether the numerical optimizer can find the maximum well. The following simulation study demonstrates that if true β is large and the initial value of β is also large, then sometimes the appropriate convergence for αs, ξ, and β might not occur. The difficulty can be mitigated with large sample sizes and small initial points of optimization.

Figure 4 is for simulation result testing convergence rate with true $\beta = 400$ in the left and true $\beta = 1,600$ in the right. For other parameters, we use $\mu = 0.1, \eta = 0, \xi = \beta/10$ and all αs are a quarter of β, for simplicity. With various fixed initial β_0, represented in the x-axis, the rate of successful convergences is examined with 500 randomly generated sample paths. The convergence is defined as success when the root mean squared error between true parameter values and estimates is less than 0.2 The initial values of μ_0 and η_0 are randomly chosen between 0 and 10, and αs and ξ are also randomly chosen between 0 and β_0.

The left panel of the figure shows that when β is relatively small such as 400, in both sample sizes 5,000 and 10,000, the optimization converges well even with relatively large β_0. For relatively large β as in the right panel, we recommend relatively small initial points for β and αs. Numerical optimization with a rather small initial value of β leads to a successful result in most cases. With sample size 10,000 and true $\beta = 1,600$, with reasonable choices of the initial values of β such as around from 10 to 1,000, the optimizations converge well.

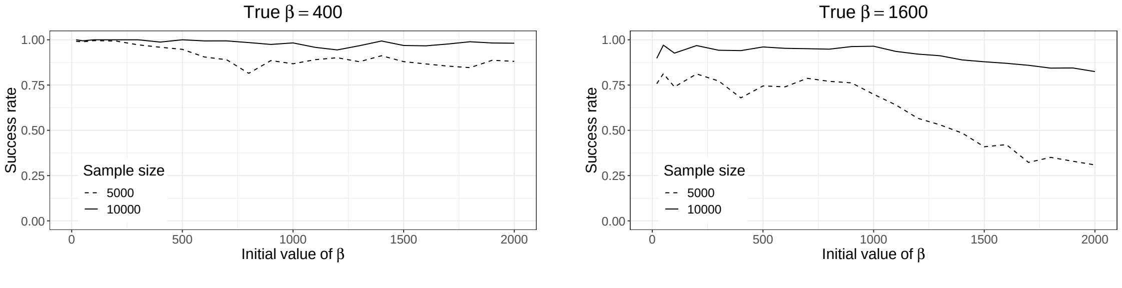


Figure 4: The success rate of estimation to find the global maximum with various initial points (x-axis) with true $\beta = 400$ (left) and $\beta = 1,600$ (right)

Table 2: US stock market exchanges

Abbreviation	Exchange
NYSE	New York Stock Exchange
NSDQ	Nasdaq Stock Market
Arca	NYSE Arca
BZX	Cboe BZX exchange, formerly Bats Global Markets
BYX	Cboe BYX exchange, formerly Bats-Y Global Markets
EDGA	Cboe EDGA exchange, formerly Direct Edge Exchange A
EDGX	Cboe EDGX exchange, formerly Direct Edge Exchange X
BSTN	Nasdaq OMX BX, formerly Boston Stock Exchange
PSX	Nasdaq PSX
NSX	National Cincinnati Stock Exchange
IEX	Investors Exchange

5.2 Data

The high-frequency data for the analysis consist of the arrival time stamps and prices of the best bid and ask quotes from various stock exchanges in the US managed by the Consolidated Tape Association or Unlisted Trading Privileges (UTP). We obtain the limit order information for NYSE-listed stocks from the Consolidated Tape System (CTS) with time resolutions in milliseconds up to July 2015, microseconds up to around September 2018, and nanoseconds for the rest. The series of time stamps of the Nasdaq-listed securities' limit order events managed by the UTP Plan are recorded in milliseconds up to July 2015, microseconds up to around October 2016, and nanoseconds for the rest. The sample for the empirical study ranges from March 2009 to December 2019. We use data from 10:00 to 15:30 to reduce the intraday seasonal effects of market opening and closing.

We focus on two types of best bid and ask price processes: the national best bid and offer (NBBO) price processes and the best bid and offer (BBO) price processes per stock exchange, as the US stock market is highly fragmented. See Table 2 for more details. More precisely, the NBBO price data consist of the official NBBO data from the NYSE and the quotes completed by Wharton Research Data Services (WRDS). The quotes supplemented by the WRDS are those that contain both the national best bid and ask prices that the NYSE does not add[1]. Because only the displayed limit orders generate the quote data, we base our analysis solely on these values. In addition, since our model depends on the spread, we use samples with a sufficient spread (i.e., with prices sufficiently large, around \$100 or higher, relative to the tick size \$0.01).

Table 3 shows the percentage of NBBO contributions by exchange for the IBM and AMZN stocks. The NYSE, Arca, NSDQ, and BZX mainly determine the NBBO for the NYSE-listed stock IBM. The Arca, NSDQ, BZX, and, recently, PSX determine the NBBO for the Nasdaq-listed stock AMZN. A higher percentage indicates more substantial liquidity provided by the corresponding exchange.

Because it is not possible to apply some of the original data directly to the Hawkes estimation, we apply data preprocessing. For example, some records of different quotes have the same

[1] https://wrds-www.wharton.upenn.edu/pages/support/data-overview/wrds-overview-taq/

Table 3: Percentage of BBOs from each exchange in the NBBO

					IBM					
Year	NYSE	Arca	NSDQ	BZX	BYX	EDGA	EDGX	BSTN	PSX	NSX
2009	20.1	20.3	27.6	22.9	0.0	0.0	0.0	1.8	0.0	2.8
2010	45.8	14.6	26.6	6.6	0.0	0.8	1.0	1.3	0.0	1.0
2011	20.2	11.5	20.7	8.0	10.4	7.1	3.3	11.4	2.0	5.3
2012	16.2	12.1	23.2	7.8	8.8	2.1	5.3	10.7	0.4	13.0
2013	22.5	13.0	18.8	11.1	5.8	3.8	7.2	7.6	2.3	6.8
2014	16.0	10.5	25.8	8.1	11.7	7.0	6.6	6.4	4.3	2.9
2015	25.3	10.3	25.6	7.7	9.7	5.8	5.9	6.0	3.6	0.0
2016	50.1	8.8	23.4	4.8	1.9	0.6	4.8	2.1	1.1	0.0
2017	35.5	12.0	23.7	8.2	2.8	1.0	8.8	3.0	1.7	0.0
2018	27.5	6.5	18.9	16.3	2.8	1.1	6.5	3.5	2.8	2.1
					AMZN					
2009	0.0	17.5	43.7	27.0	0.0	0.0	0.0	3.6	0.0	1.9
2010	0.0	18.3	52.3	11.1	0.0	0.8	1.7	1.2	0.0	10.8
2011	0.0	17.7	23.3	6.6	8.2	7.3	3.2	11.9	2.0	18.1
2012	0.0	15.4	21.1	5.7	8.9	2.9	4.2	31.0	1.5	9.0
2013	0.0	9.6	21.2	5.4	7.0	6.2	6.8	21.6	8.8	12.7
2014	0.0	10.4	31.1	5.8	11.6	10.0	9.8	10.6	7.8	0.7
2015	0.0	14.2	33.2	8.0	9.8	6.3	8.2	13.9	6.5	0.0
2016	0.0	15.3	31.3	8.7	3.6	1.1	7.1	1.4	8.7	0.0
2017	0.0	7.3	24.9	6.1	3.2	0.4	4.0	2.4	33.8	0.0
2018	0.0	3.2	21.0	7.5	0.8	0.2	2.6	0.4	22.4	19.1

time stamps. Because the probability of two or more events occurring at the same time is almost surely zero in the Hawkes-type model, these time records require modification. We assume that the same time quotes are nearly simultaneously released limit orders and we relocate the times of these records to equidistant arrivals within the time resolution.

In some cases, the best bid and ask prices change at the same time. Similar to the previous case, we separate the orders into two successive events that we assume occur sequentially within the time resolution. Whether we assume that the bid or ask price changes first depends on the spread. For example, if we assume that the bid quote occurs first and the bid price becomes greater than or equal to the ask price as a result, then the assumption is false. In this case, it is reasonable to assume that the ask quote occurs first. If it does not matter which quote occurs first, then we randomly select the order of the quotes.

Before March 2009, the percentage of data with these issues is as high as 20%. From March 2009 (more precisely, February 25, 2009), the quality of the data improves dramatically. For example, from March 2009, around 2% of the data have the above issues, and in 2018, less than 0.1% of the data do. Because the proportion of abnormal data is small, we assume that the data preprocessing described above does not have a significant impact on our statistical analysis.

In addition, we need to modify the NBBO data before the estimation. The national best bid and ask prices can be equal or sometimes reversed for reasons such as network delays. The following is an example of IBM's NBBO on January 3, 2018. At time 10:03:09.057956, the national best bid price was \$159.33, disclosed by exchange X (Nasdaq PSX). After about a second, the new national best ask price was $159.33, disclosed by exchange P (NYSE Arca).

Time	Bid	Ask	Exchange
10:03:09.057956	159.33		X
	⋮		
10:03:10.107464		159.33	P
10:03:10.107586	159.32		X

To conserve space, we do not show the best ask prices for all the exchanges, but $159.33 was the prevailing ask price across all the exchanges at this moment. However, because the best bid price was still $159.33, this best ask price was not logically correct. This result may have been due to network or computational delays, and we consider the following scenario in this case.

The following table lists the transaction details near the time of the event above. Using these details, we can presume the following scenario, although it is not possible to know the actual process exactly.

Time	Price	Exchange
10:03:09.060686	159.33	P
10:03:09.150237	159.33	K
10:03:10.107383	159.33	K
10:03:10.107446	159.33	K
10:03:10.107650	159.33	X

A large sell order arrived at exchange P at time 10:03:09.060686. It consumed all existing bid limit orders at exchange P, but was still not filled completely. This order was routed to other exchanges, and the trading system noticed that the volume was still insufficient across all the exchanges. For the existing matched quantity, the transactions proceeded in parallel on exchanges K and X, and the insufficient quantity was posted as an ask limit order on exchange P at time 10:03:10.107464, as in the previous table. The transaction processes took more time and were eventually finalized in exchange X at time 10:03:10.107586. The best bid price was updated when all the limit bid orders were consumed at $159.33. Transaction reporting ended slightly later and finally finished at 10:03.10.107650 (combined with the previous transactions in exchange K). For physical reasons, this type of delay seems to be inevitable. For example, Ding et al. (2014) report that the discrepancy between the NBBOs and synthetic NBBOs calculated using data from the exchanges is around one to two milliseconds.

Events that violate the assumptions of our model constitute about 1–2% of all events. Because the data set is so large, it is impossible to identify the cause of these anomalies in each case. We assume that these data do not significantly change the statistical properties and proceed with the estimation after removing these records.

5.3 Estimation results with the NBBO

5.3.1 Model selection

This subsection covers the results of various model selection issues. First, we discuss the estimation results from the basic Hawkes model for the bid and ask price processes, even though the model is unsuitable because it does not guarantee the positivity and reverting property of the spread process. Consider the basic Hawkes process with constant parameters:

$$\text{Basic Hawkes model: } \boldsymbol{\mu}_t = \begin{bmatrix} \mu \\ \mu \\ \mu \\ \mu \end{bmatrix}, \quad \boldsymbol{h}(t,u) = \mathrm{e}^{-\beta(t-u)} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{bmatrix}.$$

We can obtain some intuitions by fitting the data to the basic Hawkes model.

Figure 5a plots the daily estimated α_{21} and α_{34}, while Figure 5b plots α_{24} and α_{31}. These terms represent the excitation terms of the upward movements of the ask price and downward movements of the bid price caused by events that widen the spread (i.e., A^u and B^d). To improve the clarity of the graph, we omit the standard errors, but note that these values are significantly greater than zero. The estimates of α_{24} and α_{31} are similar, as are those of α_{21} and α_{34}. The parameters α_{21} and α_{34} are related to α_{w1} in our model, whereas α_{24} and α_{31} are related to α_{w2}.

Figures 5c and 5d present the estimates of $\alpha_{11}, \alpha_{44}, \alpha_{12}$, and α_{43}. The parameters α_{11} and α_{44} are typical self-exciting terms in the Hawkes model and their values are similar. The parameters α_{12} and α_{43} are typical mutually exciting terms in the Hawkes model and, again, they have similar values. Figure 5e illustrates the other mutually exciting terms α_{13} and α_{42}, showing that these values are also positive and similar.

Figure 5f shows the estimates of α_{14} and α_{41}. These values are much lower than those of the other parameters. Thus, if the ask price is up, then the tendency of the bid price to decline is relatively less significant. We can apply the same argument to the case in which the bid price is down. Because these values are relatively small, we assume that the corresponding parameters in our proposed model are zeros for model parsimony. We discuss more details in the next model comparison part.

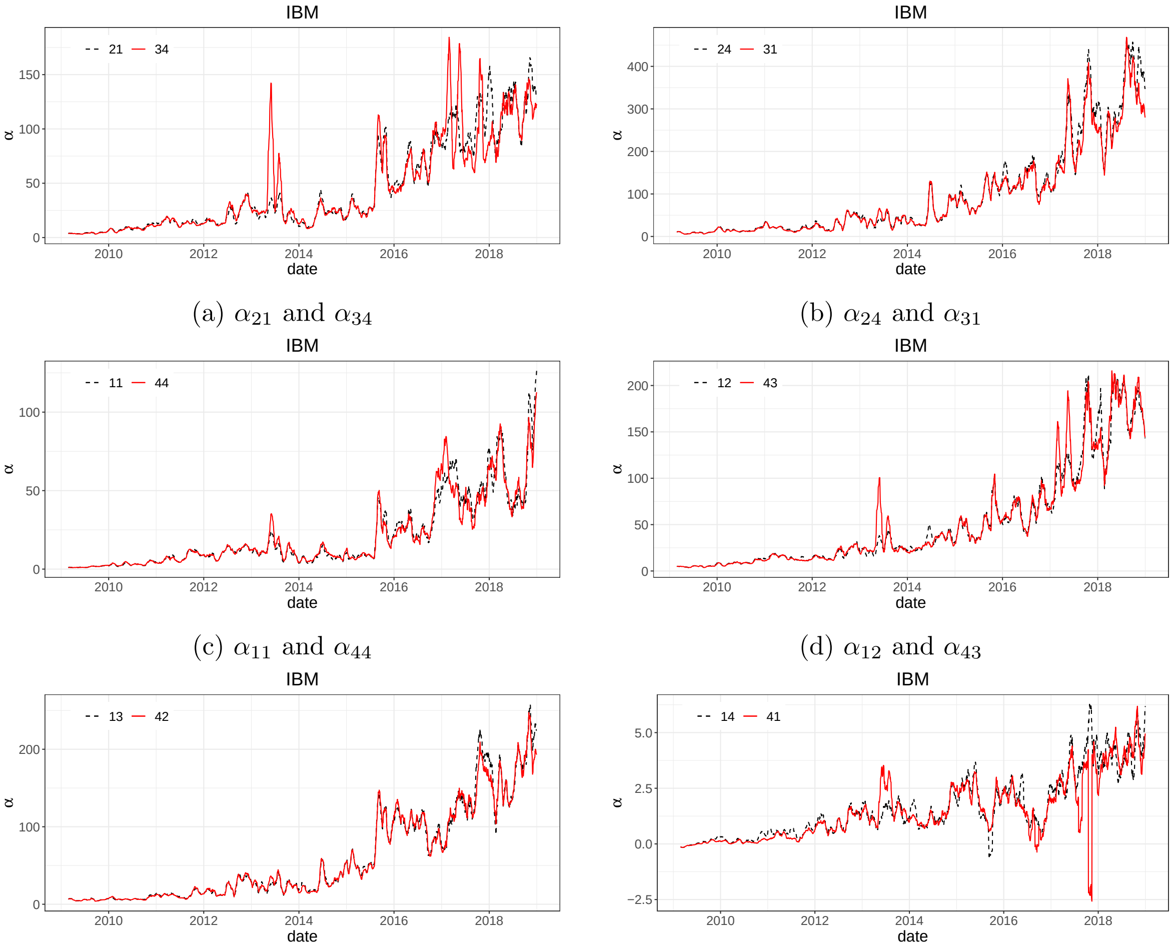
To highlight the difference between the basic Hawkes model and proposed model, Figure 6 shows the sample paths for the bid and ask processes. For the simulation, the parameter values are assumed to be one of the estimates obtained in the empirical analysis. As shown in Figure 6a, the basic Hawkes model has negative or unrealistically wide spreads. The proposed model, on the contrary, incorporates the basic nature of the bid-ask spread.

Figure 7 shows the Akaike information criterion (AIC) and Bayesian information criterion (BIC) using IBM's NBBO data for the first quarter of 2017. The proposed model (solid black line) is preferred, as it has lower values than the basic Hawkes model (dashed blue line) in this period.

Second, we examine the model assuming that the coefficients of the top right and bottom left of $\mathbf{h}$ are not zeros.

$$\text{Extended model I:} \quad \boldsymbol{h}(t,u) = \mathrm{e}^{-\beta(t-u)} \begin{bmatrix} \alpha_{s1} & \alpha_{m} & \alpha_{s2} & \alpha_{14} \\ \alpha_{w_1} & \alpha_{n1}(u) & \alpha_{n1}(u) & \alpha_{w_2} \\ \alpha_{w_2} & \alpha_{n2}(u) & \alpha_{n2}(u) & \alpha_{w_1} \\ \alpha_{41} & \alpha_{s2} & \alpha_{m} & \alpha_{s1} \end{bmatrix}$$

and the other parts are as in Definition 1. The assumption in our model that the corner coefficients of $\mathbf{h}$ are zeros does not imply that these coefficients are always zeros. Based on the empirical results analyzed daily, on some of the days, the null hypothesis that the corner coefficients are zeros is not rejected while on other days this hypothesis is rejected. However,



(a) α_{21} and α_{34}

(b) α_{24} and α_{31}

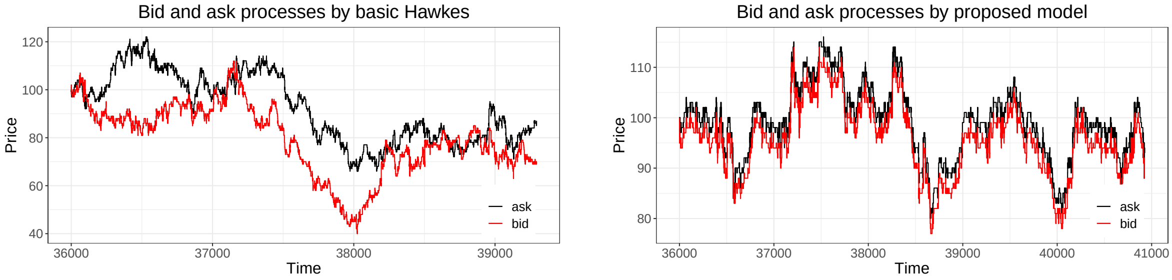
(c) α_{11} and α_{44}

(d) α_{12} and α_{43}

(e) α_{13} and α_{42}

(f) α_{14} and α_{41}

Figure 5: Daily estimates for the NBBO of IBM under the basic Hawkes model from 2009 to 2018

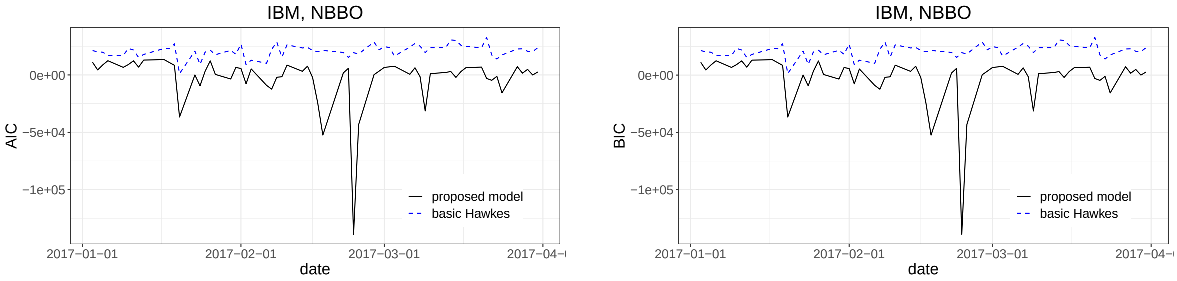


(a) basic Hawkes

(b) proposed model

Figure 6: Comparison of the bid and ask processes between the basic Hawkes model and proposed model

even on the day that this hypothesis is rejected, the coefficients have smaller values than the

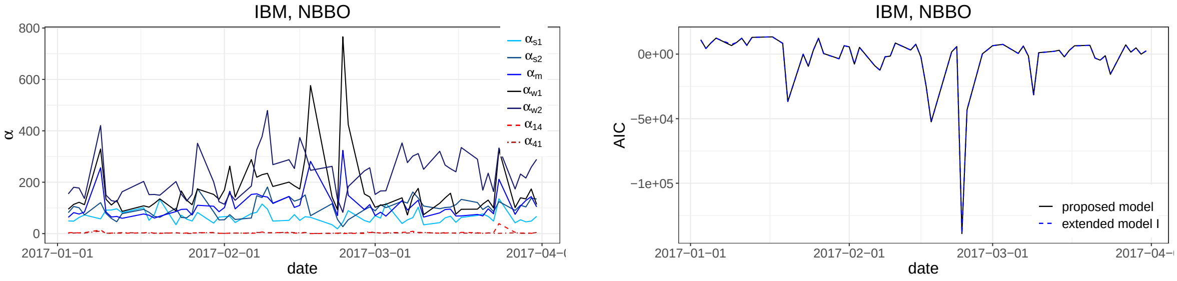


(a) AIC

(b) BIC

Figure 7: Comparison of the AIC and BIC between the proposed model and basic Hawkes model

other coefficients in the matrix $\boldsymbol{h}$, as shown in Figure 8a. The estimates of α_{14} and $\alpha 41$ are shown as the red lines in the figure. These parameters may only be significant on certain days when unexpected news causes a liquidity shock with sudden spread widening (Jiang et al., 2011; Christensen et al., 2014) and can reflect the bilateral spread-widening event. However, the inclusion of these two parameters does not significantly improve the model from the AIC point of view, as shown in Figure 8b on typical days.



(a) αs

(b) AIC

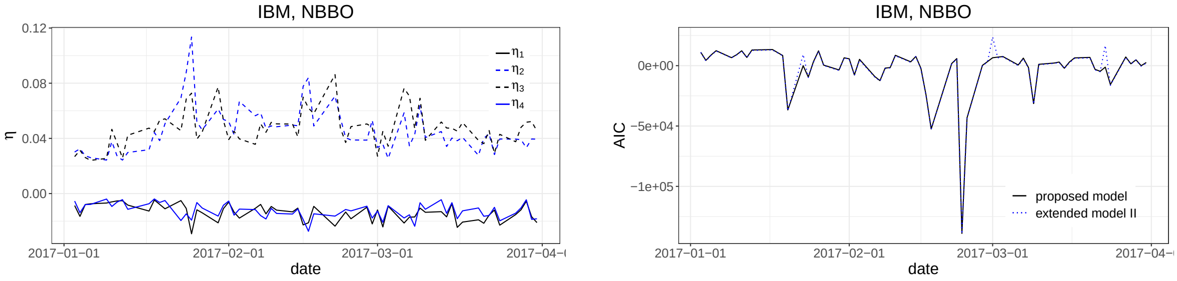
Figure 8: Estimates of the αs under extended model I (left) and comparison of the AIC between the proposed model and extended model I (right)

Third, we examine the model with different μs and ηs in μ:

$$\text{Extended model II:} \qquad \boldsymbol{\mu}_t = \begin{bmatrix} \mu_1 + \eta_1 \ell(t-) \\ \eta_2 \ell(t-) \\ \eta_3 \ell(t-) \\ \mu_4 + \eta_4 \ell(t-) \end{bmatrix}$$

and the other parts are as in Definition1. One may think that when the spreads increase and market makers provide liquidity more actively, the intensities of bid-ask narrowing A_d and B_u increase and the intensities of bid-ask widening A_u and B_d decrease. Thus, one can assume that the base intensities of A_d and B_u are also functions of the spread. Indeed, this negative relationship between the spread and intensities of the spread-widening processes is observed to some extent, as shown in Figure 9a, but is close to zero compared with the other coefficients. It also does not significantly improve the model from the perspective of the AIC, as shown in

Figure 9b. We do not show a graph of μ_1 and μ_4 because of a lack of space, but their estimates are found to be similar.



(a) αs

(b) AIC

Figure 9: Estimates of the αs under extended model II (left) and comparison of the AIC between the proposed model and extended model II (right)

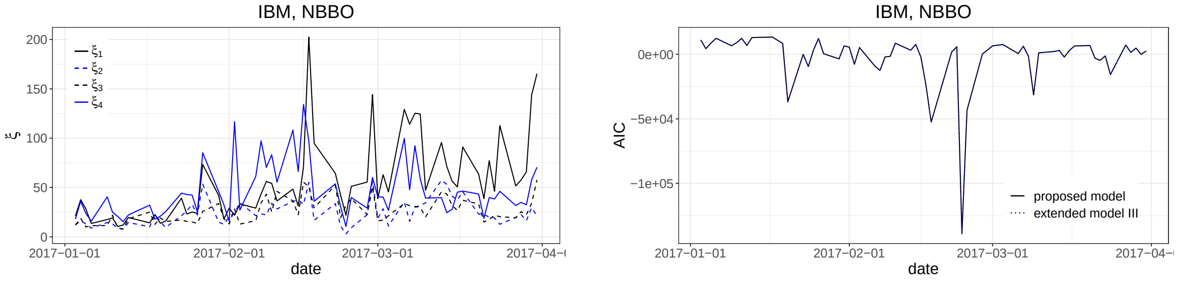
Fourth, we examine the model with different ξs. To see if it can be justified to assume that the α_ns in the middle of the matrix $\boldsymbol{h}$ are equal, we extend the model so that the α_ns have different ξs:

$$\text{Extended model III:} \quad \boldsymbol{h}(t,u) = \mathrm{e}^{-\beta(t-u)} \begin{bmatrix} \alpha_{s1} & \alpha_m & \alpha_{s2} & 0 \\ \alpha_{w_1} & \alpha_{n1}(u) & \alpha_{n2}(u) & \alpha_{w_2} \\ \alpha_{w_2} & \alpha_{n3}(u) & \alpha_{n4}(u) & \alpha_{w_1} \\ 0 & \alpha_{s2} & \alpha_m & \alpha_{s1} \end{bmatrix}$$

where

$$\alpha_{ni}(t) = -\sum_{j=1}^{4} \lambda_{kj}(t) + \xi_i \ell(t)$$

and the other parts are as in Definition 3.1. As in Figure 10a, $\xi_1 \approx \xi_4$ and $\xi_2 \approx \xi_3$ and hence the model extension can serve as an alternative. However, no significant benefit can be obtained from the perspective of the AIC, as shown in Figure 10b.



(a) ξs

(b) AIC

Figure 10: Estimates of the ξs under extended model III (left) and comparison of the AIC between the proposed model and extended model III (right)

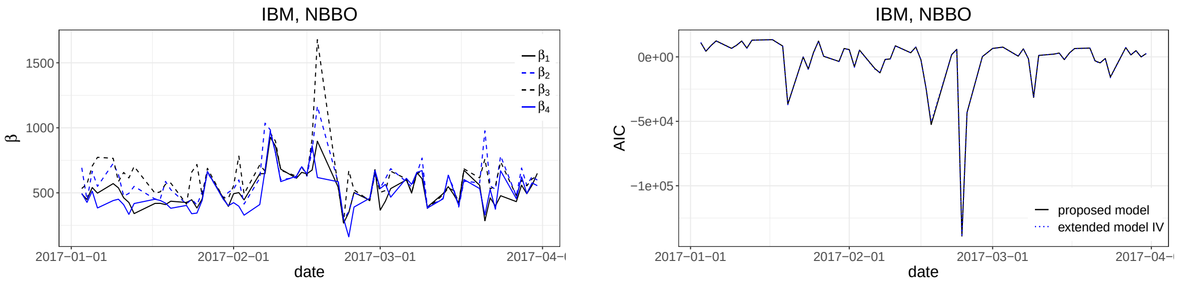
Fifth, one may think that the decaying parameter β does not have to be the same and we compare the proposed model with an extended model with a column-wise different β:

$$\text{Extended model IV:}\quad \boldsymbol{h}(t,u) = \mathbf{B}(t-u) \circ \begin{bmatrix} \alpha_{s1} & \alpha_m & \alpha_{s2} & 0 \\ \alpha_{w_1} & \alpha_{n1}(u) & \alpha_{n2}(u) & \alpha_{w_2} \\ \alpha_{w_2} & \alpha_{n3}(u) & \alpha_{n4}(u) & \alpha_{w_1} \\ 0 & \alpha_{s2} & \alpha_m & \alpha_{s1} \end{bmatrix}$$

where $\circ$ denotes the element-wise multiplication (the Hadamard product),

$$\mathbf{B}(t-u) = \begin{bmatrix} \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_4(t-u)} \\ \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_4(t-u)} \\ \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_4(t-u)} \\ \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_4(t-u)} \end{bmatrix}$$

and the other parts are as in Definition 3.1. As shown in Figure 11, the column-wise βs seem to have the same value and there is no improvement in terms of the AIC.



(a) βs

(b) AIC

Figure 11: Estimates of the βs under extended model IV (left) and comparison of the AIC between the proposed model and extended model IV

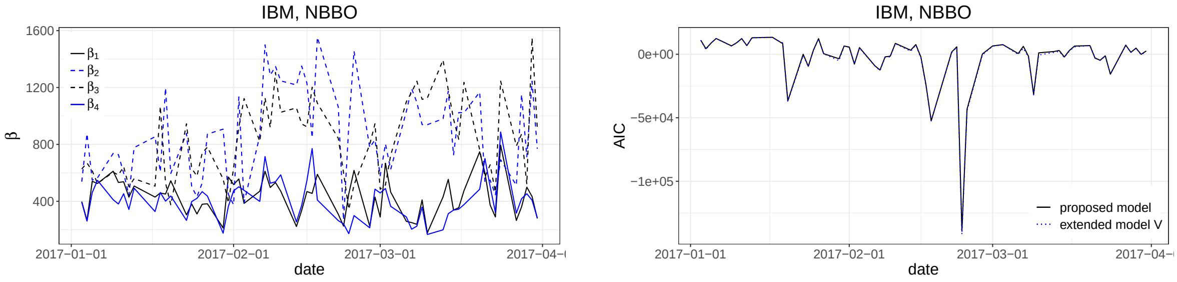
Sixth, we compare the proposed model with an extended model with a row-wise different β:

$$\text{Extended model V:}\quad \boldsymbol{h}(t,u) = \mathbf{B}(t-u) \circ \begin{bmatrix} \alpha_{s1} & \alpha_m & \alpha_{s2} & 0 \\ \alpha_{w_1} & \alpha_{n1}(u) & \alpha_{n2}(u) & \alpha_{w_2} \\ \alpha_{w_2} & \alpha_{n3}(u) & \alpha_{n4}(u) & \alpha_{w_1} \\ 0 & \alpha_{s2} & \alpha_m & \alpha_{s1} \end{bmatrix}$$

where

$$\mathbf{B}(t-u) = \begin{bmatrix} \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_1(t-u)} & \mathrm{e}^{-\beta_1(t-u)} \\ \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_2(t-u)} & \mathrm{e}^{-\beta_2(t-u)} \\ \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_3(t-u)} & \mathrm{e}^{-\beta_3(t-u)} \\ \mathrm{e}^{-\beta_4(t-u)} & \mathrm{e}^{-\beta_4(t-u)} & \mathrm{e}^{-\beta_4(t-u)} & \mathrm{e}^{-\beta_4(t-u)} \end{bmatrix}$$

and the other parts are as in Definition 3.1. Extended model V also shows slightly improved performance in terms of the AIC, even though $\beta_1 \approx \beta_4$ and $\beta_2 \approx \beta_3$, which is consistent with the results reported by Lee and Seo (2017b). This slight improvement is insufficient to use the extended model at the expense of increased model complexity. Therefore, the empirical studies presented hereafter are conducted based on the proposed model in Definition 1.



(a) βs

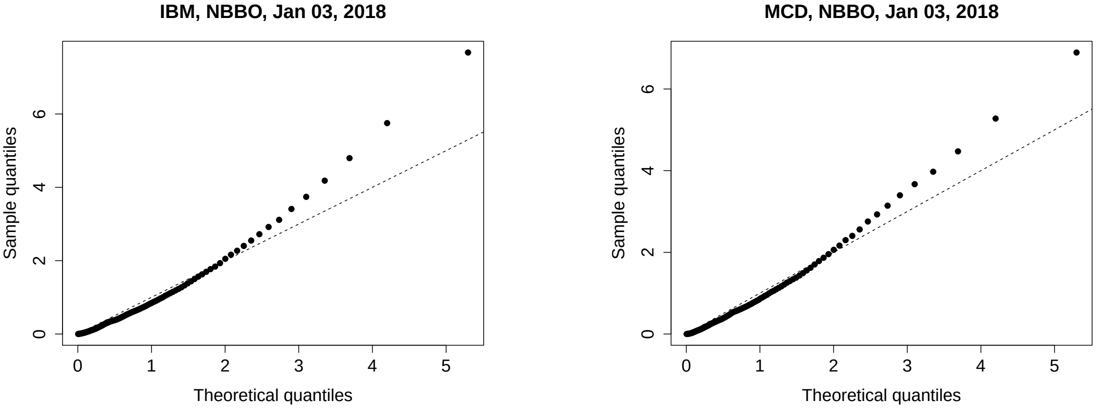
(b) AIC

Figure 12: Estimates of the βs under extended model V (left) and comparison of the AIC between the proposed model and extended model V

In Figure 13, quantile-quantile plots of the residuals versus the unit exponential distribution are presented. The residuals are defined by

$$\bigcup_{1\leq i\leq 4}\left\{\int_{t_{i,j}}^{t_{i,j+1}}\hat{\lambda}_i(u)\mathrm{d}u\right\}$$

where $\hat{\lambda}_i$ are fitted intensities of our model and $t_{i,j}$ are event times. The plot shows that the residuals follow a slightly fatter tail distribution. In order to obtain a better fitting result, it is necessary to introduce a multi-kernel with different βs, which leads more complex model. This is beyond our scope. Instead we focus on describing the activities at the ultra-high-frequency level, so the extension to multiple kernels is left as a future work.



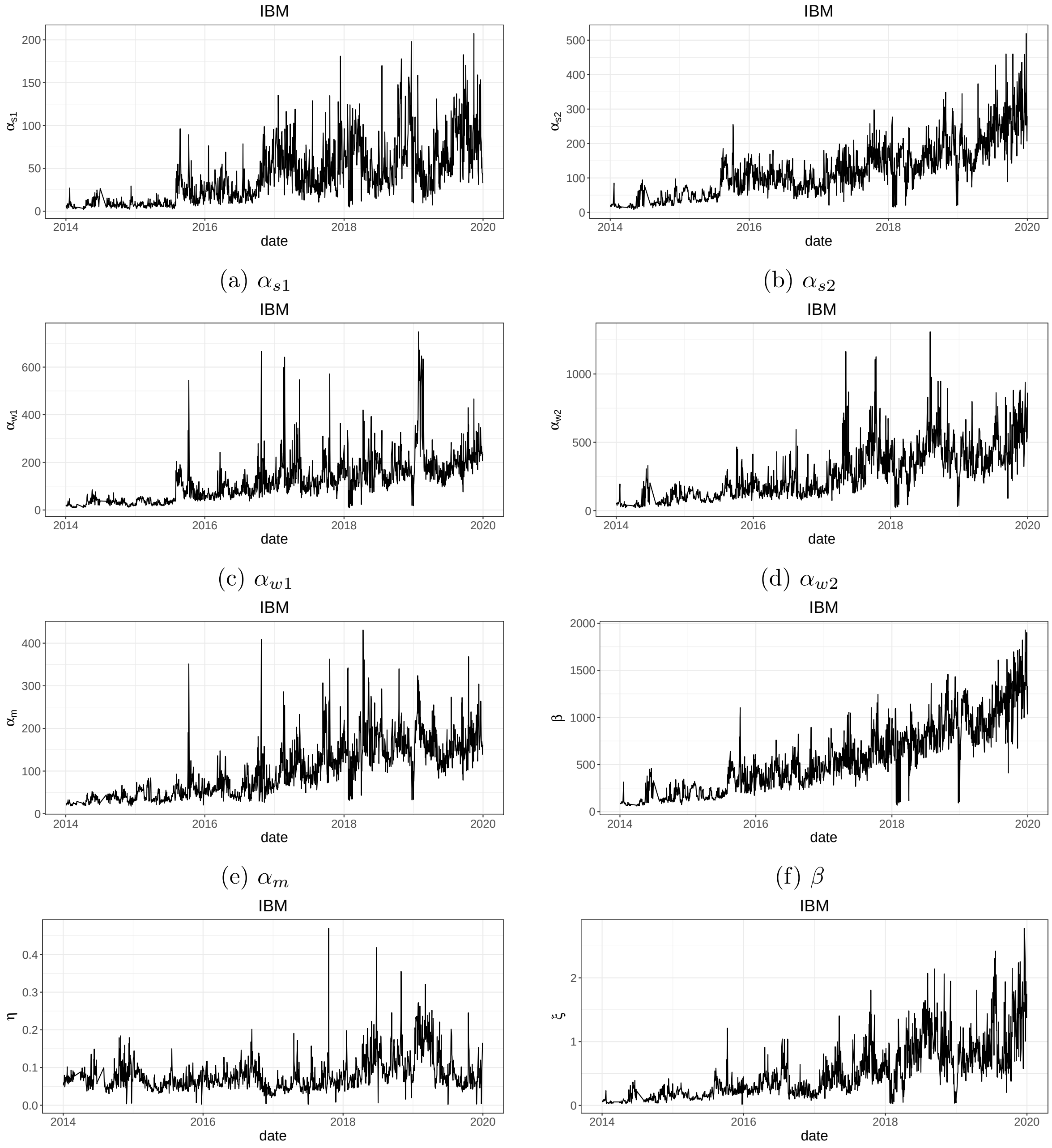
(a) IBM

(b) MCD

Figure 13: Q-Q plot of residual process of our model

5.3.2 Basic results

This section reports the basic results obtained by the NBBO. Some of these results may seem obvious, but they are worth checking, while other results are new. Figure 14 illustrates the dynamics of the estimates from the NBBO of the IBM stock under the model defined in Definition 3.1 from 2014 to 2019. We compute all the estimates on a daily basis, and hence have around 250 estimates for each parameter in each year. The time unit for the intensities is one second. For the detailed results of the estimations with standard errors, see Appendix A.



(a) α_{s1}

(b) α_{s2}

(c) α_{w1}

(d) α_{w2}

(e) α_m

(f) β

(g) η

(h) ξ

Figure 14: Dynamics of the daily estimated parameters for the NBBO of IBM from 2014 to 2019

First, the results show that the narrowing tendency of the bid-ask spread depends on the spread. The estimates of liquidity parameter η are significantly positive throughout the period, as Figure 14g shows. Recall that we use $\eta\ell(t-)$ in the base intensity of the spread-narrowing processes (i.e., downward movements of ask prices and upward movements of bid prices). A positive η implies that with a larger bid-ask spread, it is more likely that the spread will narrow as the base intensities of the narrowing movements increase. This finding is in line with our intuition because the larger the spread is, the more likely it is that market makers will earn profits and will submit aggressive limit orders. Note that η exists only in the base intensity, but not in the kernel; hence, is not related to the previous events, but only to the current spread. In addition, this parameter changes little over time and has a stable value in general.

Not every day shows a strong spread-narrowing tendency. If liquidity providers are concerned about adverse selection (i.e., they expect the realized spread to be negative), then they will not try to reduce the spread. For example, on January 5, 2017, IBM's liquidity parameter η is 0.025, which is less than half of the average ηs estimated in 2017, and we can assume that market makers were particularly careful about adverse selection on this day. In addition, when market disruption occurs, liquidity parameter η can come close to zero during the day, which we discuss in Subsection 5.5.

Second, the figure shows a large excitement parameter α, which implies an abundance of automated high-frequency quotes and trading. In our model, some limit order arrivals depend on the spread described by η, whereas the rest relate to the excitations caused by past events, as the parameter αs also affect the arrivals or removals of limit orders, which changes the best bid or ask prices. The estimates of the αs range in magnitude from tens to hundreds in general, which means that in the event of a limit order arrival or removal that changes the best bid or ask price, the intensities increase by tens to hundreds, and hence the probability of subsequent events dramatically increases. Because the decaying parameter β is also large, the increased probability also decreases sharply over time. For example, if an event increases intensity to 500, then approximately five successive events can also occur in the next 0.01 seconds. Intensity also tends to converge to the background rate μ owing to the large decaying parameter β, provided that no additional events occur.

This instantaneous responsiveness is high; as human behavior is unlikely to be the cause, it may be closely related to high-frequency automated algorithmic trading. Many trading venues ensure that transactions are fast, especially in quote revisions. Chan (2017) demonstrates that better financial services and technologies such as colocation systems in which the trading servers are in the same facilities as the execution venues can reduce the latency to microseconds.

Third, the excitement parameters αs show an increasing trend (similar to the decaying parameter) in general. One reason for this increasing trend is the more accurate time resolution. As mentioned before, from August 2015 to around September 2018, the data are recorded in microseconds compared with the previous frequency of milliseconds for the stocks reported by the CTS. This change is noticeable on the graph, as we observe peaks around August 2015 for the excitement parameters. However, even when excluding the effect of the improved resolution, the overall estimates increase, which means lower latencies for order submissions and the data feed in high-frequency trading over time. In addition, when the time resolution improves to nanoseconds (around September 2018), we find no notable abrupt changes. This result implies that the resolution of microseconds is sufficient for this analysis.

Because the daily estimates of the excitement parameter αs are too variable over time, Figure 15 presents each α estimated from IBM's NBBO as a moving average. We compute the moving average using 20 successive days from 2009 to 2019. The increasing trends for all the parameters are clear. Overall, α_{w2} is the largest, α_{s1} is the smallest, and the rest are similar. The large α_ws mean that liquidity-providing actions occur with fast reactions in response to spread-widening events, indicating that market making is active.

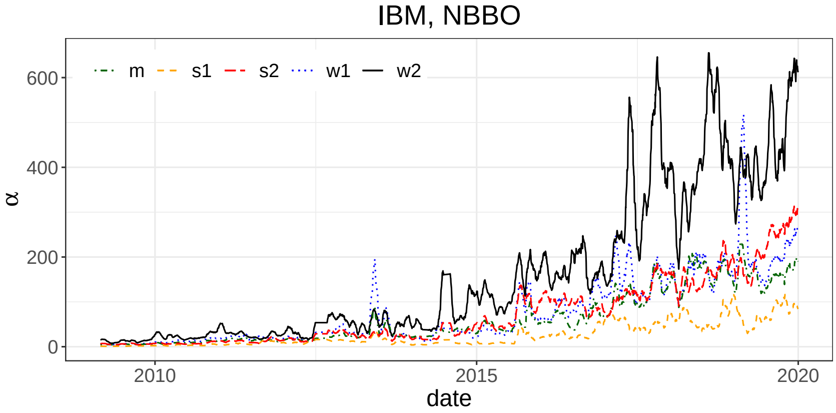


Figure 15: Estimates of the αs for IBM's NBBO, smoothed by the moving average of 20 successive days

We define

$$\bar{\alpha} = \frac{\alpha_{s1} + \alpha_{s2} + \alpha_m + \alpha_{w1} + \alpha_{w2}}{5}$$

and consider this value to be a measure of the overall instantaneous responsiveness of each stock. Of course, we cannot assert that successive events occur solely based on some decision making or algorithm depending on preceding events, but this instantaneous responsiveness can measure how fast a series of successive potentially related events occurs. Figure 16 plots the moving average of $\bar{\alpha}$ from 2009 to 2019 for the NBBO of Accenture, Berkshire Hathaway Inc. Class B, and IBM. Instantaneous responsiveness generally increases over time owing to technological innovations such as enhanced and cheaper hardware, faster communication speeds, new software, and better financial services. Including these stocks, most stocks show a typical increasing pattern, but not all equities show monotone increases over time. We discuss other examples in the next subsection.

5.3.3 Liquidity provision versus depletion (removal)

Figure 17 shows the estimation results of the ratio between the parameters related to liquidity provision and those related to liquidity depletion for various stocks. The parameter for the former is the mean of α_{w1} and α_{w2}, whereas that for the latter is the mean of α_{s1}, α_{s2}, and α_m. The figure shows the average values of the parameter estimates in 2016 and 2017. In all the examined stocks, the average estimates of the liquidity provision parameters are larger than the average liquidity depletion parameters. This result implies that actions to provide liquidity generally occur faster than do actions that remove liquidity in response to previous events.

Among those, Amazon in 2017 has the smallest ratio of liquidity provision, at slightly greater than 0.5, and Visa has a value close to 1 in 2016. Market markers' speed competition in providing liquidity can be well understood, since the fastest provider of limit orders preempts the advantageous position in the limit order queue. For brevity, we illustrate only the estimates

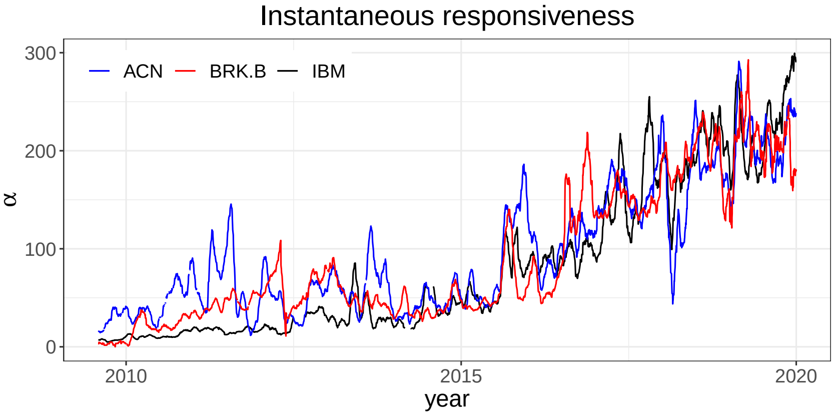
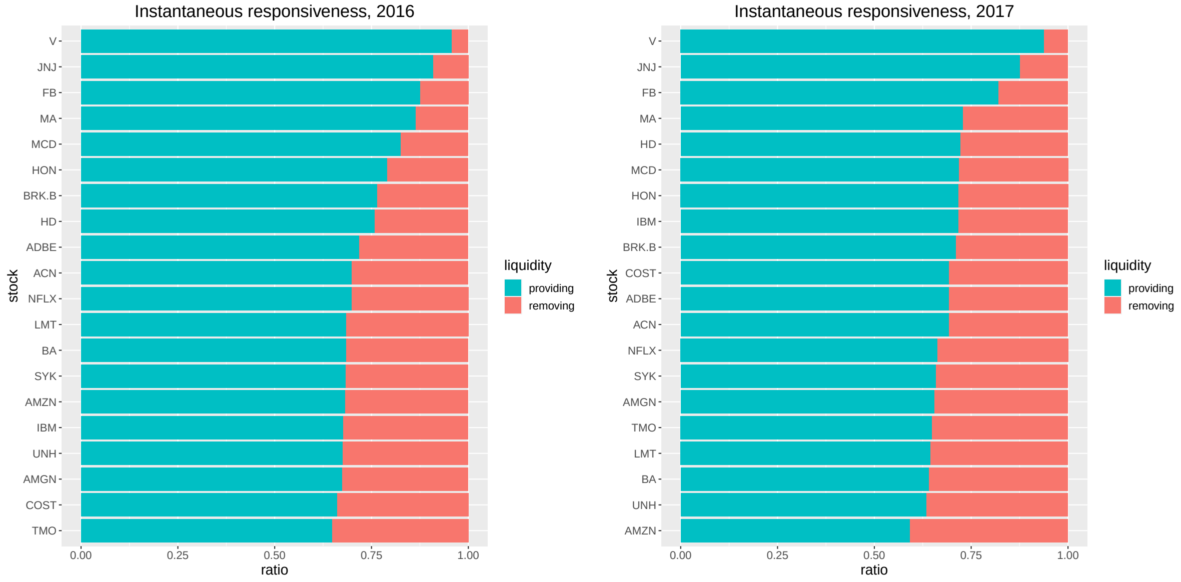


Figure 16: Moving average of $\bar{\alpha}$ for the NBBOs of Accenture, Berkshire Hathaway Inc. Class B, and IBM



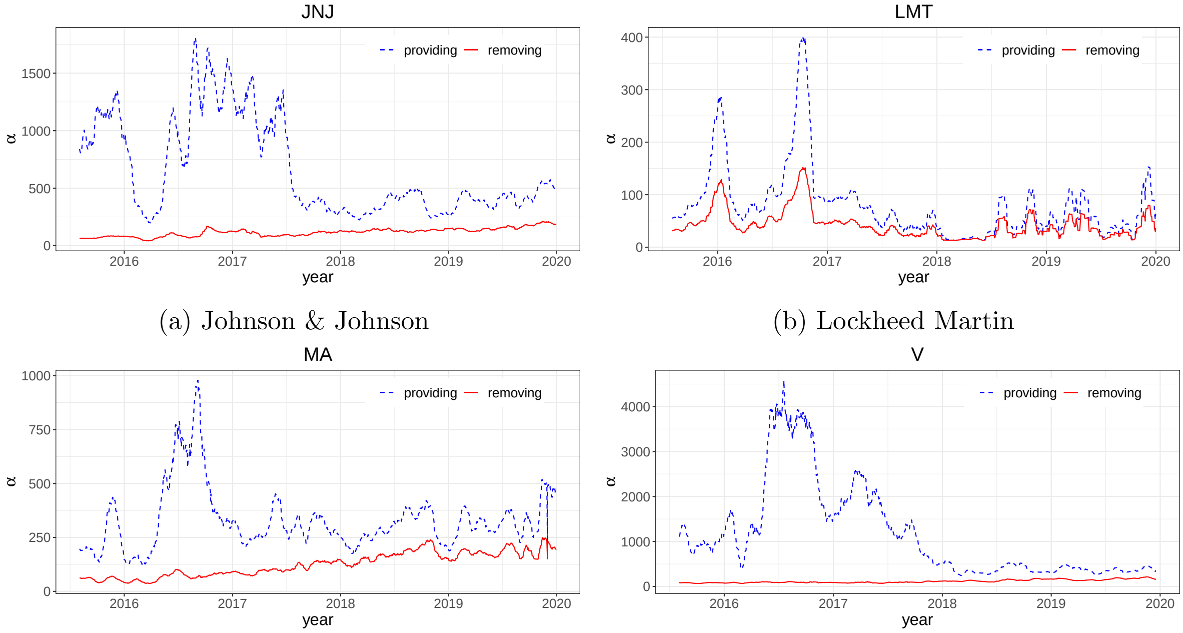
(a) 2016

(b) 2017

Figure 17: Responsiveness ratio measured using the liquidity provision and depletion parameters for various stocks, 2016 and 2017

for 2016 and 2017, but the ratio changes over time. We can also clearly see that the ratios of liquidity provision are generally larger in 2016.

Some stocks show an interesting pattern that peaks around 2016 and diminishes in the liquidity provision parameter in Figure 18. This figure shows the smoothed dynamics of the liquidity provision (blue dashed line) and depletion α parameters (red solid line). For example, the largest liquidity provision parameter for Visa is over 4,000, which is large compared with the α values in Figure 16, suggesting that ultra-high-frequency market makers paid much attention to these stocks during this period. However, market makers might conclude that such ultra-fast reaction times do not necessarily improve profits. Indeed, to reduce costs, the liquidity provision parameters converge to much lower values after 2018, which are similar to the levels of α in Figure 16. We discuss the reaction speed for liquidity provision, which is particularly high in 2016, in Section 5.4.2 in more detail.



(a) Johnson & Johnson

(b) Lockheed Martin

(c) Mastercard

(d) Visa

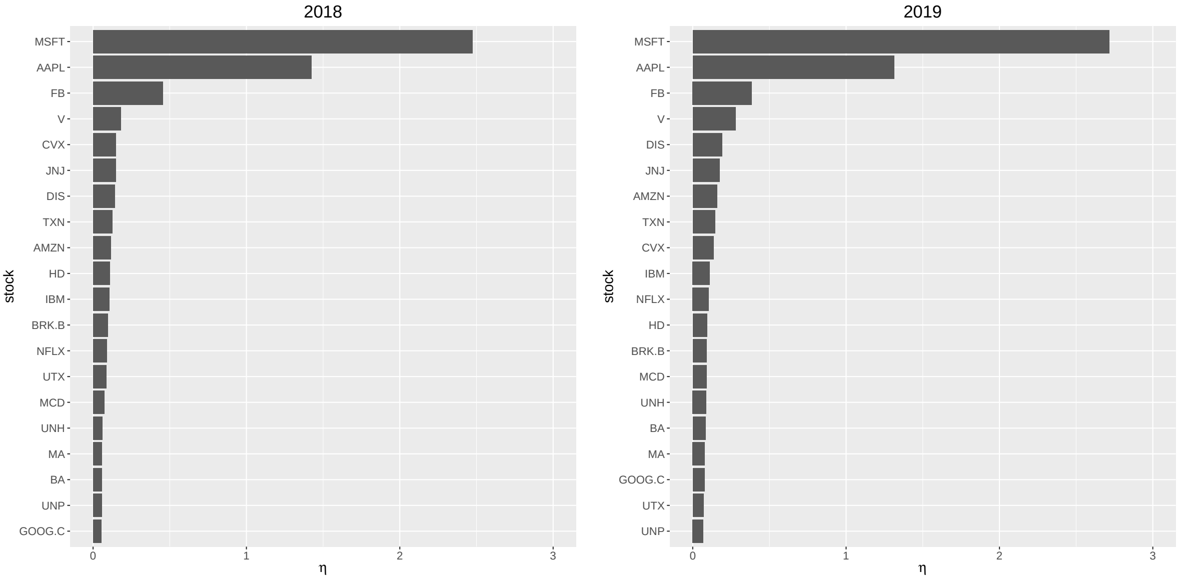
Figure 18: The 2016 peak and diminishing pattern in the liquidity provision parameter

As we show, for most stocks, $\bar{\alpha}$, or the liquidity-providing α, which represents the average reaction time, reaches levels of around tens to hundreds by 2019. In some cases, it increases steadily to that level in Figure 16; on the contrary, in other cases, it reaches this level after several years of volatile liquidity provision speeds, as in Figure 18. The behavior of high-frequency traders on most stocks appears to eventually reach similar levels over time in terms of speed and risk attitude. However, not all the stocks show the same pattern. In 2018 and 2019, a few stocks have fierce market-making competition in ultra-high-frequency trading with strong confidence in adverse selection.

5.3.4 Outliers

Some of the stocks we analyze seem to have received particular attention from ultra-high-frequency liquidity providers in terms of the liquidity parameter η. These are IT giants such as Microsoft, Apple, and Facebook. The larger the η is, the stronger market makers tend to narrow the spread through aggressive limit orders when spreads widen. This happens when market makers are confident and less concerned about adverse selection. Figure 19 shows bar graphs of the ηs estimated in 2018 and 2019 on average for 20 stocks. The η of Microsoft is large, followed by those of Apple and Facebook. For the other stocks, the ηs are at a similar level, implying a similar assessment of adverse selection risk on average.

The competition for these stocks, particularly for Microsoft, to occupy the top priority of the limit order books is intense, as Figure 20 shows. This figure presents the yearly averages



(a) 2018

(b) 2019

Figure 19: The average η for various stocks, 2018 and 2019

of the liquidity-providing αs, namely, α_{w1} and α_{w2}, with green bars and the yearly averages of the liquidity-depleting αs with red bars. Microsoft, Apple, and Facebook has the top liquidity provision speed in 2018, but responsiveness decreases from 2018 to 2019. Especially for Apple and Facebook, by 2019, the relative gap between the liquidity provision speed, measured by the average α_ws, and the other stocks narrows. However, Microsoft is still the only outlier in 2019, at least in our sample, and seems to attract more attention from high-frequency liquidity providers than the other stocks do. Although we do not know the reason for Microsoft's popularity among high-frequency traders, this result means that not all the stocks receive the same interest from high-frequency traders.

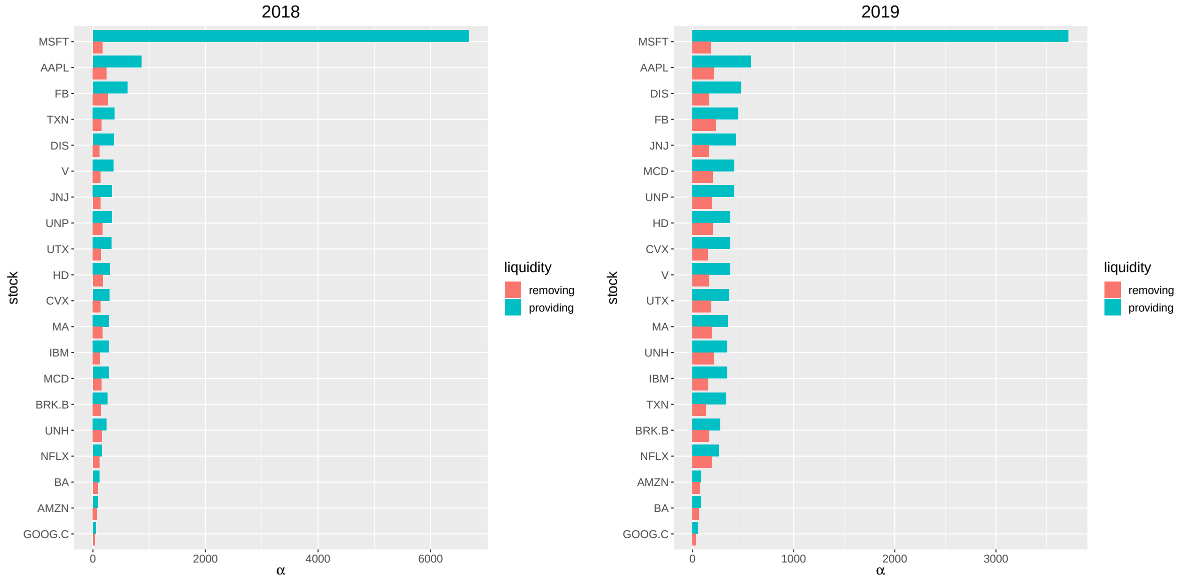
5.4 Estimation results by exchange

5.4.1 NYSE versus BZX

Using our model, we can capture the varying characteristics of the different exchanges by analyzing the quote data for each trading venue.

The left-hand side of Figure 21 plots the dynamics of the moving average $\bar{\alpha}$ for the BBO of IBM in several of the stock exchanges. In 2010 and 2011, the responsiveness of the Nasdaq, shown by the red dotted line, is the fastest, whereas the responsiveness of the Nasdaq, NYSE (black solid line), and BZX (blue dashed line) is similar. Although omitted, the Arca has similar $\bar{\alpha}$ values to the NYSE, which is not surprising as it operates as a subsidiary of the NYSE Group. Indeed, the Arca, formerly the Archipelago, merged with the NYSE in 2006. The Archipelago had electronic stock trading technology, which the NYSE did not at the time. However, from 2009, the starting point of our data analysis, we see no substantial performance difference in the speeds of the NYSE and Arca.

The Nasdaq's electronic stock trading technology originated from one of the first electronic



(a) 2018

(b) 2019

Figure 20: The average liquidity-providing and -depleting αs for various stocks, 2018 and 2019

communication networks, Island, which had a good reputation for fast electronic trading in the early 2000s (Patterson, 2012). By adopting this technology, the Nasdaq shows relatively large $\bar{\alpha}$ values in the early 2010s. We also include a graph of the Cboe BZX Exchange, which claims to be designed for cheap and fast transactions after consecutive mergers between traditional stock exchanges and electronic trading systems. The Cboe BZX Exchange had a 6% market share in 2017 and is considered to be among top-tier exchanges such as the NYSE, NYSE Arca, and Nasdaq (Wah et al., 2018).

The right-hand side of Figure 21 shows the time-varying shape of IBM's overall responsiveness $\bar{\alpha}$ for the NYSE and Cboe BZX Exchange. We do not smooth the estimates in this figure. The black solid line represents the $\bar{\alpha}$ estimated daily by the best bid and ask prices quoted in the NYSE and the red dashed line represents the $\bar{\alpha}$ estimated by the prices in the BZX. When comparing the two exchanges, the NYSE has relatively little variation in $\bar{\alpha}$, but $\bar{\alpha}$ changes more drastically over time for the BZX. The maximum $\bar{\alpha}$ of the BZX is almost twice that of the NYSE, implying that the BZX can operate under very low latency compared with the NYSE and can attract high-frequency traders in this respect. This result is in line with the BZX's original claim of high-speed technology.

We can also see the different behaviors between NYSE and BZX traders in volatile markets in Figure 22 based on data for IBM from 2017 to 2018. The figure presents the scatterplots of α_{w1} and α_{w2}, the liquidity provision parameters, versus daily log-returns. The other excitement parameters have similar patterns. In the NYSE (left), when the absolute value of the log-return is large (i.e., on days when the market is volatile), liquidity provision parameters are relatively small, and the absolute return and excitement parameters have a negative correlation. Thus, high-frequency traders in the NYSE, especially liquidity providers, are less active in volatile conditions, perhaps concerned about adverse selection. Conversely, we can observe the opposite situation in the BZX. High-frequency quotes, indicated by the high liquidity provision

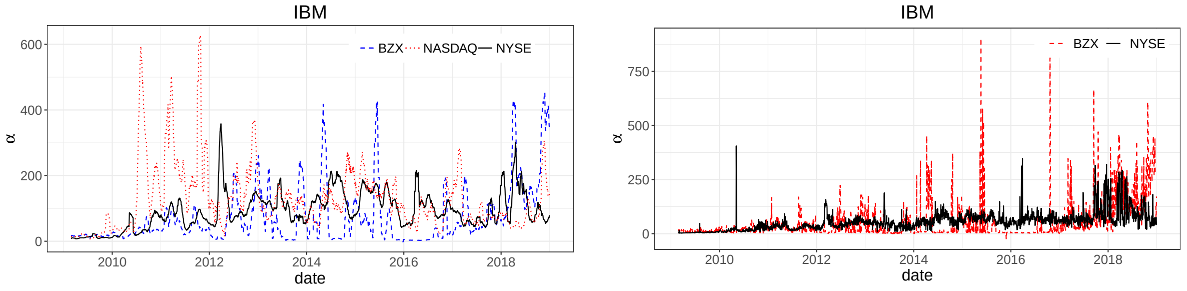


Figure 21: Smoothed $\bar{\alpha}$ for IBM from the NYSE, NSDQ, and BZX (left) and nonsmoothed $\bar{\alpha}$ for IBM from the NYSE and BZX (right)

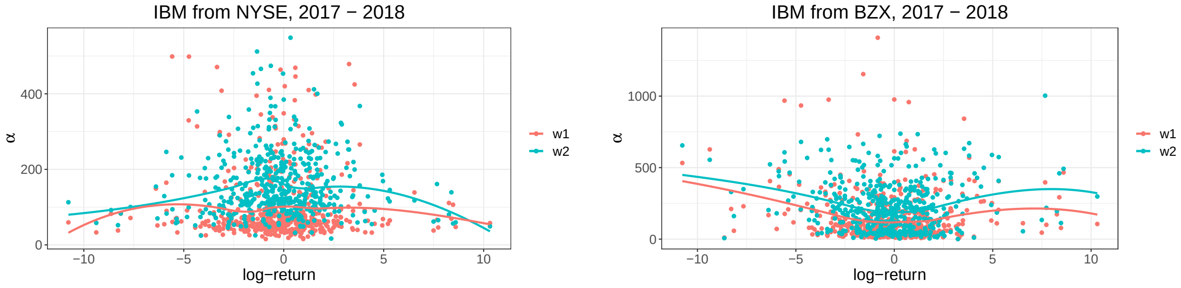


Figure 22: Scatterplot of the αs versus the daily log-return, IBM, 2017–2018, from the NYSE (left) and BZX (right)

parameter α_w values, remain active and become even more active in a volatile market in the BZX. The absolute return and liquidity provision parameters are positively correlated. Thus, high-frequency traders in the BZX tend to take aggressive positions, even in volatile markets.

5.4.2 The BZX versus BYX and the EDGA versus EDGX

In this subsection, we compare instantaneous responsiveness depending on the fee structure. The right-hand side of Figure 21 shows that the averaged excitement parameters, $\bar{\alpha}$, of the BZX approaching 2016 is close to zero; thus, the left-hand side of Figure 23 plots the $\bar{\alpha}$s of the BZX and BYX. The reason for the low averaged excitement parameter values in this period is unknown, but we suppose that the high-frequency traders, especially liquidity providers, active in the BZX before 2016 moved to the BYX at this time. The averaged excitement parameters of the BYX rises sharply during the period when the ones of the BZX is small. Conversely, after 2017, the averaged excitement parameters of the BZX rises steadily, whereas the ones of the BYX remains close to zero. The BZX provides some rebates when traders provide liquidity through a limit order and takes fees when transacting market orders. Conversely, the BYX uses an inverted fee schedule, providing rebates to traders removing liquidity through market orders and taking fees for quoting limit orders.

We can see a similar relationship for the EDGA and EDGX. The EDGX's averaged excitement parameters increases from the second half of 2018, whereas the EDGA's parameter is close to zero in later years, as shown on the right-hand side of Figure 23. In addition, the

EDGA's averaged excitement parameters rises sharply in 2016. The EDGA provides rebates for removing liquidity, whereas the EDGX provides rebates for adding liquidity.

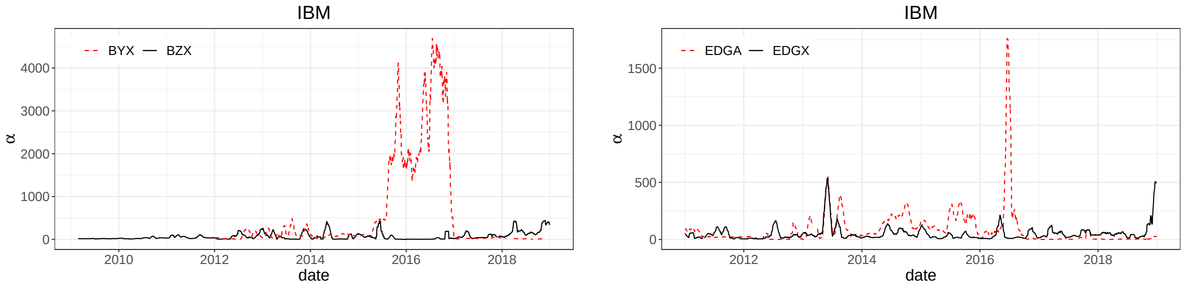


Figure 23: Smoothed dynamics of the $\bar{\alpha}$s of the BZX and BYX (left) and the EDGA and EDGX (right)

The BYX adopted an inverted fee schedule, and the EDGA also had an inverted fee structure until June 1, 2017. These exchanges have very high excitement parameter values for about one year or a few months around 2016. Thus, we discuss whether we can assume that limit order quotes and trading were more active in both exchanges during this period. Indeed, not all the excitement parameters of the BYX and EDGA are large. The estimated α_{w1} dominates all the other values on the left-hand side of Figure 24 for the BYX; the case is similar for the EDGA on the right-hand side of the figure. Recall that α_{w1} is related to the replenishment rate when the best limit orders are removed, which we can consider to be the basic job of market makers. We assume that market making that provides liquidity was active, with low latency in the BYX (and EDGA) around 2016. Thus, we can assume that market participants actively tried to profit through market making during this period. While it may seem strange to see markets on exchanges with inverted fee schedules that charge a fee to provide liquidity, inverted structures attract market orders, and hence the limit orders in these exchanges are likely to be concluded first. However, these activities were discouraged after 2017, or at least market making slowed. Although we do not show the results in detail, the α_{w1} for IBM in the BYX is still profound in 2018, with an average around 15; however, it is much smaller than the value in 2016 is. Perhaps the BYX's market makers determined that market making at a higher latency than before would be sufficiently competitive given that low latency is costly. Conversely, the α_{w1} of the EDGA is small in 2018 (average around 0.7), meaning that high-frequency market making should not have occurred in that year.

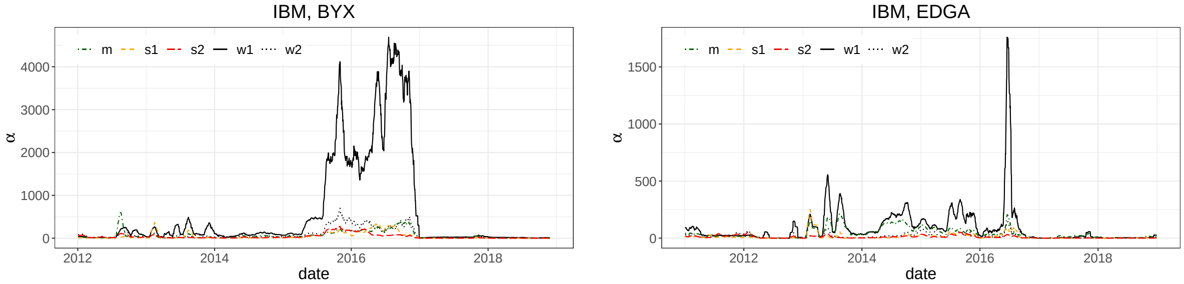


Figure 24: Smoothed dynamics of the αs of IBM from the BYX (left) and EDGA (right)

Around 2016, the best bid and ask price processes of the BYX (and EDGA) differ from the typical dynamics of liquid exchanges. The left-hand side of Figure25 shows the dynamics of the best bid and ask prices in the NYSE and the right-hand side shows the best bid and ask prices in the BYX. Although we do not show the NBBO's movements, the NYSE's movements closely resemble them. Conversely, the bid-ask spread in the BYX changes rapidly in jump size and the overall movement is different from the NBBO dynamics. We observe similar movements in the EDGA. On the right-hand side of the figure, when a transaction occurs and the existing best quote is removed, we momentarily observe a rather large spread. However, responsible market makers seem to supplement quotes quickly and the spread narrows. Except for these quotes, it appears that limit orders lack sufficient depth. The best bid and ask prices are far from the mid-price when the market maker does not quote the limit orders sufficiently near the mid-price. This shallow depth is consistent with the fact that the BYX has a maker-taker fee schedule.

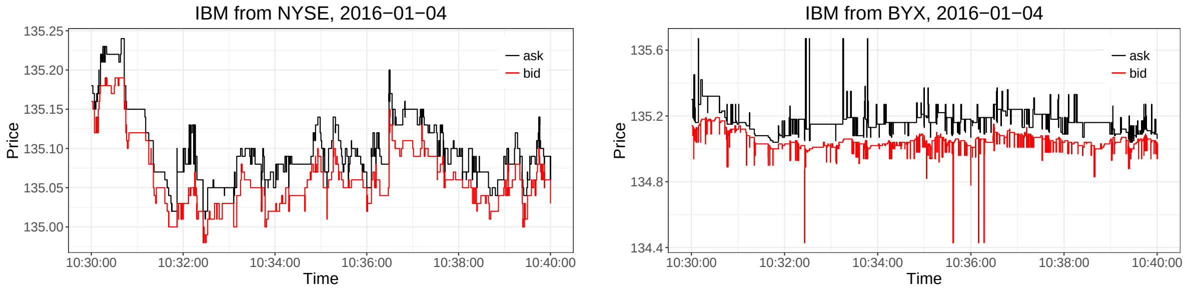


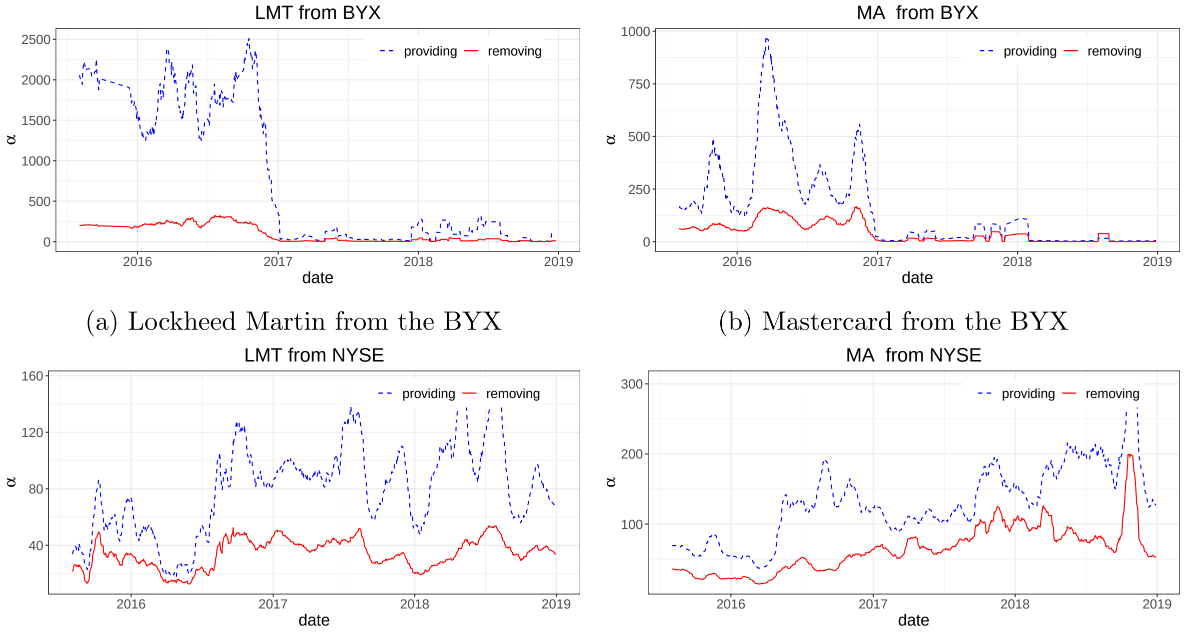
Figure 25: Comparison of IBM's tick dynamics in the NYSE (left) and BYX (right)

Figure 18 shows the large liquidity-providing αs (the mean of α_{w1} and α_{w2}) around 2016. We strongly suspect that the main source of this phenomenon is an exchange such as the BYX. For the Lockheed Martin and Mastercard stocks, Figures 26a and 26b plot the time-varying liquidity-providing and -depleting excitement parameter, respectively, using the bid and ask prices published by the BYX. The liquidity-providing αs in both stocks from the BYX remain high until the second half of 2016, but suddenly fall close to zero after 2017. On the contrary, Figures 26c and 26d show the results estimated using the data from the NYSE (the main exchange for these stocks), indicating that the αs remain relatively stable.

These two results suggest that the peak of the liquidity-providing excitement parameters around 2016 observed from the NBBO in Figure 18 did not exist in all the exchanges, but particularly occurred in exchanges such as the BYX. This result is consistent with the previous discussion. High-frequency traders thought that they would profit from ultra-high-frequency trading on specific stocks such as Lockheed Martin and Mastercard on specific exchanges such as the BYX. The exact reason is unknown; however, after 2017, it seems that high-frequency traders stopped making these types of trades.

5.4.3 The IEX

Recently, the IEX introduced the speed bump technique, which intentionally slows order processing to 350 microseconds to protect low-frequency traders (Bishop, 2017). Exchanges can use this technique to slow ultra-fast market orders to prevent orders from snipping stale quotes that do not yet reflect the state of the national bid and offer prices. The IEX can also use it to



(a) Lockheed Martin from the BYX

(b) Mastercard from the BYX

(c) Lockheed Martin from the NYSE

(d) Mastercard from the NYSE

Figure 26: Smoothed dynamics of the liquidity-providing and -depleting αs from 2015 to 2018

help protect market orders executed with the rest of the orders to complete on other exchanges from phantom liquidity by delaying the execution and transmission of the results. In addition, with the crumbling quote detection algorithm and discretionary pegged orders, in which orders aim to access liquidity more aggressively in a stable market and become passive when quotes crumble, traders in the IEX are less concerned about speed.

Unfortunately, we cannot obtain information on the invisible pegged orders of the IEX. Instead, we perform the estimation using visible limit orders, which represents about 20% of all the orders of the IEX (Aldrich and Friedman, 2019). The responsiveness we measure using the excitement parameter αs in this exchange is lower than that in the other exchanges, as Figure 27 shows, for both liquidity depletion and liquidity provision. For a better visualization, we use the moving averages of the excitement parameter αs' estimates. We can interpret these results in two ways. The IEX's liquidity providers are relatively unconcerned about improving responsiveness because they expect the speed bump technique to protect their stale quotes. Alternatively, the IEX's meticulous traders already use invisible pegged orders, meaning that we cannot observe their exact behaviors and can only present the behaviors of those traders relatively insensitive to increasing the speed of quote processing.

5.5 The Flash Crash

On May 6, 2010, the Flash Crash occurred during which stock prices fell sharply due to large sell orders and then recovered quickly. Our model can investigate the details of the behavior of market participants on this day. The left-hand side of Figure 28 plots the IBM stock price (black solid line) and dynamics of the liquidity parameter η (blue dashed line). Recall that η represents the tendency to narrow when the bid-ask spread widens.

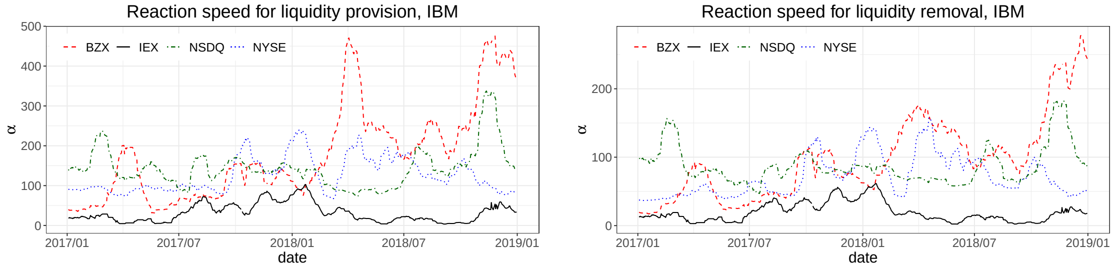


Figure 27: Estimates of IBM's αs from the IEX, smoothed by a moving average of 20 successive days

Unlike the previous cases, to examine the intraday changes, we perform the estimations every minute and base them on three-minute overlapping intervals using IBM's NBBO data on bid and ask prices on May 6, 2010. The Flash Crash occurred around 14:32. As the figure shows, the high η value indicates that liquidity providers continued to provide liquidity confidently for a few minutes immediately after the start of the price collapse. However, η then drops to almost zero, implying that liquidity providers became seriously concerned about adverse selection due to the unexpectedly large price drop (i.e., they did not try to narrow the spread). The value of η begins to rise almost at the same time as the price recovery and then returns to its original level with synchronized price fluctuations.

The right-hand side of the figure depicts the liquidity-providing α (blue dashed line) and -depleting α (red solid line) with price changes (black solid line). Interestingly, both the excitement parameters start to rise at the point the price reaches its lowest value. In addition, the two parameters are almost the same near to and immediately after the time of the lowest price. In a stable market, the liquidity-providing parameter α is greater than the liquidity-depleting α, as shown in the previous figures. Immediately after the Flash Crash, these parameter values rise; that is, market participants are more likely to respond faster to market changes in both liquidity provision and liquidity depletion, which continued and strengthened until the next day.

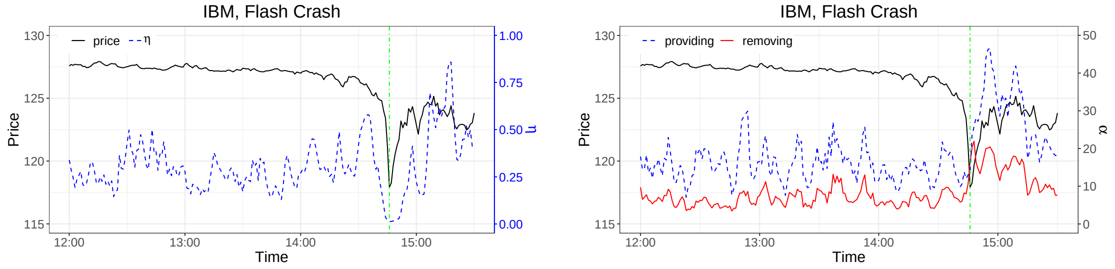


Figure 28: Stock price dynamics during the Flash Crash and intraday changes in η and α

Returning to Figure 21, we can see that the average of excitement parameters, $\bar{\alpha}$, of the NYSE reaches its maximum value in May 2010. Figure 29 provides an enlarged picture of this part of the graph together with the $\bar{\alpha}$ of AMZN. Interestingly, the peaked day with the largest excitement parameters (i.e., May 7, 2010) is the day after the Flash Crash. On this day, the excitement parameter values estimated for the NYSE and Arca are much higher than those on the other days; that is, it seems to be the busiest day of 2010 for the high-frequency traders on these exchanges. Additionally, when the market closes on that day without any particular event, their responsiveness returns to their original state after May 7, 2010. However, the BZX data indicate no such busy activities on May 7, 2010.

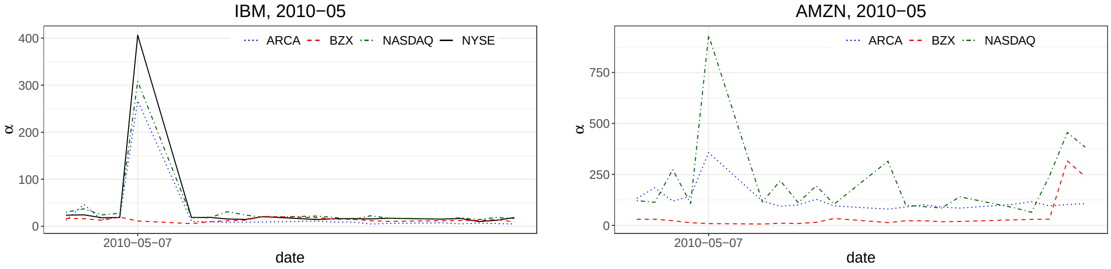


Figure 29: Dynamics of $\bar{\alpha}$ around the Flash Crash for IBM (left) and AMZN (right). The peak occurs the day after the Flash Crash.

The reason behind the above estimation results becomes clearer when we observe the best bid and ask price processes on this day. The left-hand side of Figure 30 shows the best bid and ask price processes reported by the NYSE the day after the Flash Crash, May 7, 2010. The price process moves much more irregularly in the NYSE than in the BZX, as the right-hand side of the figure shows. The NYSE best bid and ask processes on this day are more irregular than on the other days, although we do not show them to save space. The liquidity providers on the NYSE were extra careful to post limit orders on this day based on the previous Flash Crash experience, and they tried to avoid adverse selection by repeating the process of submitting and then canceling limit orders. The two largest estimates of excitement parameters on this day are $\alpha_{w1} = 883$ and $\alpha_m = 554$, representing the fast repetitions of liquidity supply and removal. Note that α_{w1} is related to supplementing quotes and α_m is related to submitting and immediately canceling (or executing) limit orders.

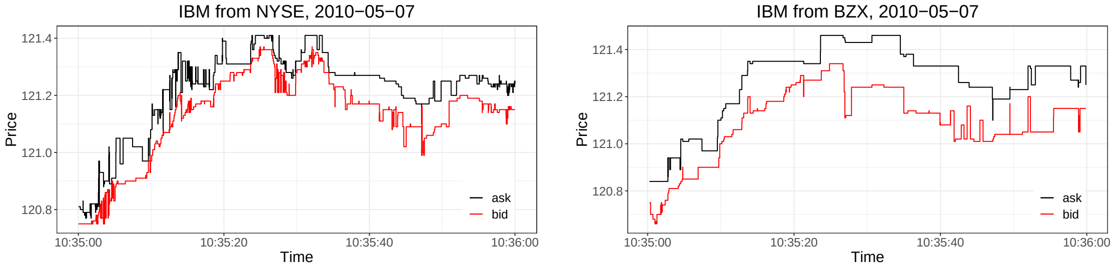


Figure 30: IBM stock tick structure the day after the Flash Crash on the NYSE (left) and BZX (right)

6 Conclusion

We propose an extended Hawkes model to describe the best bid and ask price processes. The model aims to rigorously reflect the nature of the price processes and capture the various aspects of high-frequency stock price data observed in micro- or milliseconds.

The empirical analysis shows that larger spreads tend to reduce spreads in general, but that this tendency decreases greatly when adverse selection is a concern. In addition, the movements of the bid and ask price processes not only depend on the spread, but also are excited by recent events that cause price changes. With the abundance of automated high-frequency trading, the estimated responsiveness and corresponding decay rate have become high in recent years. For most stocks, we observe increasing responsiveness over the past decade and find that some stocks show unique behavior. Specifically, responsiveness to liquidity provision increased dramatically around 2016 and then declined. Moreover, some stocks, namely Microsoft, Apple, and Facebook, have become of particular interest to high-frequency traders in recent years. These stocks, particularly Microsoft, show high responsiveness in terms of liquidity provision and a high tendency to replenish when spreads widen.

Exchanges have different responsiveness patterns. For example, the responsiveness of the NYSE is more stable than that of the Cboe BZX, which changes more drastically. The proposed model also allows us to observe the different characteristics of high-frequency traders across exchanges. We find that the speed of liquidity provision peaked around 2016, stemming from the speed of stock exchanges such as the BYX. However, the new IEX has low responsiveness relative to other exchanges.

Focusing on intraday changes in the parameters during the Flash Crash in 2010, we observe a change in confidence related to adverse selection among market makers along with a sharp decline and recovery in a stock price. We also find that responsiveness peaked on the day after the Flash Crash in response to the previous market anomaly, as traders repeatedly submitted and canceled limit orders.

The application of the proposed model is not limited to those presented here. Owing to its nature, which can incorporate a variety of properties, our proposed model, or variations of it, is able to reveal some of the information hidden in high-frequency financial data. Hence, the presented findings should provide a better understanding of the market and aid decision making and policymaking.

A Estimation result: Example

Table 4 presents the estimates and standard errors in parentheses from our model for IBM for January 2018. We calculate the standard error numerically. To save space, we report only some of the results. However, the other cases show similar standard errors.

References

E. M. Aldrich and D. Friedman. Order protection through delayed messaging. *Available at SSRN 2999059*, 2019.

T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71:579–625, 2003.

M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008.

Table 4: Estimates and standard errors (in parentheses) from our model for IBM, January 2018

Date	μ	η	α_{s1}	α_{s2}	α_m	α_{w1}	α_{w2}	β	ξ
2018-01-02	0.067	0.0693	96.32	193.2	219.5	271.6	459.6	991.1	61.44
	(0.0013)	(0.0014)	(0.3570)	(0.3306)	(0.3929)	(0.2521)	(0.2169)	(0.0668)	(0.1012)
2018-01-03	0.1004	0.1184	69.34	192.6	173.5	210.1	381.2	707.0	71.82
	(0.0016)	(0.0018)	(0.0642)	(0.0605)	(0.0818)	(0.1111)	(0.0990)	(0.0444)	(0.1146)
2018-01-04	0.0787	0.1165	66.79	158.3	193.5	201.4	479.0	718.3	89.27
	(0.0014)	(0.0022)	(0.1439)	(0.1829)	(0.2113)	(0.1049)	(0.1261)	(0.0410)	(0.1658)
2018-01-05	0.0736	0.0758	63.81	185.0	142.2	126.2	371.6	739.4	48.41
	(0.0014)	(0.0014)	(0.3090)	(0.7946)	(0.5339)	(0.3335)	(0.0165)	(0.0719)	(0.1475)
2018-01-08	0.0757	0.0886	69.77	150.3	158.9	157.6	395.6	699.0	47.31
	(0.0014)	(0.0016)	(0.1229)	(0.1692)	(0.3506)	(0.4053)	(0.1474)	(0.0543)	(0.2686)
2018-01-09	0.0768	0.0911	50.70	178.1	169.1	174.7	514.5	881.1	71.06
	(0.0014)	(0.0019)	(0.1341)	(0.1719)	(0.2671)	(1.1020)	(1.1630)	(0.0869)	(0.1191)
2018-01-10	0.0447	0.0369	71.22	201.6	119.3	135.6	281.6	831.4	43.54
	(0.0011)	(0.0008)	(0.4306)	(0.4451)	(0.6818)	(0.9216)	(0.4001)	(0.1445)	(0.0683)
2018-01-11	0.0513	0.0635	50.83	190.2	150.2	160.2	344.6	683.5	60.97
	(0.0012)	(0.0013)	(0.2936)	(0.2478)	(0.1626)	(0.3738)	(0.0587)	(0.1526)	(0.4125)
2018-01-12	0.0625	0.0669	64.45	139.4	137.8	166.9	362.0	595.0	54.68
	(0.0013)	(0.0014)	(0.2930)	(0.1787)	(0.1425)	(0.3041)	(0.1897)	(0.0838)	(0.1723)
2018-01-16	0.1099	0.1124	51.86	113.7	150.6	142.8	265.1	576.0	42.94
	(0.0017)	(0.0017)	(0.3374)	(0.2102)	(0.2304)	(0.2500)	(0.6593)	(0.1097)	(0.4373)
2018-01-17	0.1051	0.0931	55.83	110.5	135.9	141.3	256.4	541.7	40.56
	(0.0017)	(0.0014)	(0.0950)	(0.0968)	(0.1000)	(0.1193)	(0.1199)	(0.0337)	(0.2273)
2018-01-18	0.1211	0.135	97.30	166.6	159.9	186.7	386.7	750.5	55.04
	(0.0018)	(0.0019)	(0.0918)	(0.1867)	(0.0113)	(0.0626)	(0.0604)	(0.1009)	(0.0402)
2018-01-19	0.1791	0.1977	124.8	251.1	331.5	334.2	466.9	1091	97.98
	(0.0021)	(0.0022)	(0.1484)	(0.1087)	(0.0670)	(0.1072)	(0.0495)	(0.0112)	(0.0487)
2018-01-22	0.0835	0.1042	80.68	276.9	342.1	280.2	529.5	1098	94.90
	(0.0015)	(0.0016)	(0.0517)	(0.0723)	(0.0333)	(0.0955)	(0.0500)	(0.0508)	(0.0647)
2018-01-23	0.0781	0.0645	72.98	173.7	139.6	141.8	264.7	760.0	37.03
	(0.0014)	(0.0011)	(0.0704)	(0.0602)	(0.1765)	(0.1796)	(0.2232)	(0.0419)	(0.1726)
2018-01-24	0.0736	0.0621	12.04	20.31	36.15	19.56	37.39	99.01	4.355
	(0.0014)	(0.0008)	(0.2121)	(0.0651)	(0.0849)	(0.1409)	(0.1630)	(0.2076)	(0.0836)
2018-01-25	0.0544	0.0565	5.861	15.96	33.48	12.35	27.12	79.13	3.944
	(0.0012)	(0.0009)	(0.2745)	(1.1480)	(1.4180)	(0.4565)	(0.5676)	(0.1170)	(0.6511)
2018-01-26	0.0534	0.0407	35.69	117.1	104.9	81.64	247.1	550.5	39.55
	(0.0012)	(0.0009)	(0.4864)	(0.3846)	(0.4970)	(1.1400)	(0.7538)	(0.1820)	(0.2034)
2018-01-29	0.046	0.0519	4.861	15.46	31.30	8.930	20.29	68.35	2.682
	(0.0011)	(0.0008)	(0.4999)	(1.772)	(1.883)	(0.2002)	(0.5845)	(0.1414)	(0.2195)
2018-01-30	0.0567	0.0575	8.437	22.42	34.29	14.94	35.35	90.93	5.069
	(0.0012)	(0.0009)	(0.2398)	(0.4787)	(0.6678)	(0.4286)	(0.1209)	(0.0723)	(0.1531)
2018-01-31	0.0839	0.0453	70.94	146.2	98.07	94.21	253.4	649.5	32.98
	(0.0015)	(0.0008)	(0.2563)	(0.2819)	(0.3680)	(0.1556)	(0.2291)	(0.0819)	(0.2457)

E. Bacry and J.-F. Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14:1147–1166, 2014.

E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13:65–77, 2013.

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005, 2015.

E. Bacry, T. Jaisson, and J. Muzy. Estimation of slowly decreasing Hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, 16:1179–1201, 2016.

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

N. Bershova and D. Rakhlin. High-frequency trading and long-term investors: a view from the buy-side. *Journal of Investment Strategies*, 2:25–69, 2013.

B. Biais, T. Foucault, and S. Moinas. Equilibrium fast trading. *Journal of Financial Economics*, 116:292 – 313, 2015. ISSN 0304-405X.

A. Bishop. The evolution of the crumbling quote signal. *Available at SSRN 2956535*, 2017.

N. P. Bollen and R. E. Whaley. Futures market volatility: What has changed? *Journal of Futures Markets*, 35:426–454, 2015.

J.-P. Bouchaud, M. Mézard, M. Potters, et al. Statistical properties of stock order books: empirical results and models. *Quantitative finance*, 2(4):251–256, 2002.

C.-G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141:876–912, 2007.

J. Brogaard, T. Hendershott, and R. Riordan. High-Frequency Trading and Price Discovery. *The Review of Financial Studies*, 27(8):2267–2306, 2014.

J. Brogaard, B. Hagströmer, L. Nordén, and R. Riordan. Trading fast and slow: Colocation and liquidity. *The Review of Financial Studies*, 28(12):3407–3443, 2015.

P. Brémaud and L. Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24:1563 – 1588, 1996.

A. Cartea and Jaimungal. Modelling asset prices for algorithmic and high-frequency trading. *Applied Mathematical Finance*, 20:512–547, 2013.

A. Cartea, S. Jaimungal, and J. Ricci. Algorithmic trading, stochastic control, and mutually exciting processes. *SIAM Review*, 60:673–703, 2018.

A. P. Chaboud, B. Chiquoine, E. Hjalmarsson, and C. Vega. Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69:2045–2084, 2014.

E. P. Chan. *Machine trading: Deploying computer algorithms to conquer the markets.* John Wiley & Sons, Hoboken, New Jersey, 2017.

V. Chavez-Demoulin and J. McGill. High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance*, 36:3415 – 3426, 2012.

S. E. Choi, H. J. Jang, K. Lee, and H. Zheng. Optimal market-making strategies under synchronised order arrivals with deep neural networks. *Journal of Economic Dynamics and Control*, 125:104098, 2021.

K. Christensen, R. C. Oomen, and M. Podolskij. Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics*, 114(3):576–599, 2014.

K. H. Chung and A. J. Lee. High-frequency trading: Review of the literature and regulatory initiatives around the world. *Asia-Pacific Journal of Financial Studies*, 45(1):7–33, 2016.

R. Cont and A. de Larrard. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics*, 4:1–25, 2013.

J. Cvitanic and A. A. Kirilenko. High frequency traders and asset prices. Available at SSRN: https://ssrn.com/abstract=1569067, 2010.

J. Da Fonseca and R. Zaatour. Clustering and mean reversion in a Hawkes microstructure model. *Journal of Futures Markets*, 35:813–838, 2014. ISSN 1096-9934.

A. Dassios, H. Zhao, et al. Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18, 2013.

S. Ding, J. Hanna, and T. Hendershott. How slow is the NBBO? a comparison with direct exchange feeds. *Financial Review*, 49(2):313–332, 2014.

T. Foucault and S. Moinas. Is trading fast dangerous? In W. Mattli, editor, *Global Algorithmic Capital Markets: High Frequency Trading, Dark Pools, and Regulatory Challenges*, chapter 2, pages 9–27. Oxford University Press, 2018.

M. B. Fox, L. Glosten, and G. Rauterberg. *The New Stock Market: Law, Economics, and Policy.* Columbia University Press, 2019.

F. Guilbaud and H. Pham. Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94, 2013.

B. Hagströmer and L. Nordén. The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770, 2013.

D. Hainaut and S. Goutte. A switching microstructure model for stock prices. *Mathematics and Financial Economics*, 13(3):459–490, 2019.

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21:83 – 143, 2015.

L. Harris. Trading in pennies: a survey of the issues. Unpublished working paper, University of Southern California, 1999.

J. Hasbrouck. High-frequency quoting: Short-term volatility in bids and offers. *Journal of Financial and Quantitative Analysis*, 53:613–641, 2018.

J. Hasbrouck and G. Saar. Technology and liquidity provision: The blurring of traditional definitions. *Journal of financial Markets*, 12(2):143–172, 2009.

A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33:438–443, 1971a.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90, 1971b.

A. G. Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

P. Hewlett. Clustering of order arrivals, price impact and trade path optimisation. In *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 2006.

B. Hollifield, R. A. Miller, and P. Sandås. Empirical analysis of limit order markets. *The Review of Economic Studies*, 71(4):1027–1063, 2004.

W. Huang, C.-A. Lehalle, and M. Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, 110:107–122, 2015.

H. J. Jang, K. Lee, and K. Lee. Systemic risk in market microstructure of crude oil and gasoline futures prices: A Hawkes flocking model approach. *Journal of Futures Markets*, 40:247–275, 2020.

R. A. Jarrow and P. Protter. A dysfunctional role of high frequency trading in electronic markets. *International Journal of Theoretical and Applied Finance*, 15:1250022, 2012.

G. J. Jiang, I. Lo, and A. Verdelhan. Information shocks, liquidity shocks, jumps, and price discovery: Evidence from the us treasury market. *Journal of Financial and Quantitative Analysis*, 46(2):527–551, 2011.

H. H. Kelejian and P. Mukerji. Does high frequency algorithmic trading matter for non-at investors? *Research in International Business and Finance*, 37:78 – 92, 2016.

A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzin. The Flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

J. Large. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10:1–25, 2007.

B. Law and F. Viens. Hawkes processes and their applications to high-frequency data modeling. In I. Florescu, M. C. Mariani, H. E. Stanley, and F. G. Viens, editors, *Handbook of High-Frequency Trading and Modeling in Finance*, chapter 6, pages 183–219. John Wiley & Sons, 2015.

K. Lee and B. K. Seo. Marked Hawkes process modeling of price dynamics and volatility estimation. *Journal of Empirical Finance*, 40:174–200, 2017a.

K. Lee and B. K. Seo. Modeling microstructure price dynamics with symmetric Hawkes and diffusion model using ultra-high-frequency stock data. *Journal of Economic Dynamics and Control*, 79:154–183, 2017b.

V. Manahov and R. Hudson. The implications of high-frequency trading on market efficiency and price discovery. *Applied Economics Letters*, 21(16):1148–1151, 2014.

A. J. Menkveld. High frequency trading and the new market makers. *Journal of Financial Markets*, 16(4):712 – 740, 2013.

A. J. Menkveld. The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8:1–24, 2016.

M. Morariu-Patrichi and M. S. Pakkanen. State-dependent Hawkes processes and their application to limit order book modelling. *arXiv preprint arXiv:1809.08060*, 2018.

J. C. Nash. *Nonlinear parameter optimization using R tools*. John Wiley & Sons, 2014.

Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–261, 1978.

Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27:23–31, 1981.

T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31:145–155, 1979.

S. Patterson. *Dark pools: The rise of AI trading machines and the looming threat to Wall Street*. Random House, London, 2012.

M. Scholtus and D. van Dijk. High-frequency activity on NASDAQ. In G. N. Gregoriou, editor, *Handbook of high frequency trading*, chapter 1, pages 3–23. Academic Press, 2015.

H. R. Stoll. High speed equities trading: 1993–2012. *Asia-Pacific Journal of Financial Studies*, 43:767–797, 2014.

I. M. Toke. The order book as a queueing system: average depth and influence of the size of limit orders. *Quantitative Finance*, 15:795–808, 2015.

V. van Kervel. Competition for Order Flow with Fast and Slow Traders. *The Review of Financial Studies*, 28:2094–2127, 2015.

G. P. M. Virgilio. High-frequency trading: a literature review. *Financial Markets and Portfolio Management*, 33(2):183–208, 2019.

E. Wah, S. Feldman, F. Chung, A. Bishop, and D. Aisen. A comparison of execution quality across US stock exchanges. In W. Mattli, editor, *Global Algorithmic Capital Markets: High Frequency Trading, Dark Pools, and Regulatory Challenges*, chapter 5, pages 91–146. Oxford University Press, 2018.

B. Zheng, F. Roueff, and F. Abergel. Modelling bid and ask prices using constrained Hawkes processes: Ergodicity and scaling limit. *SIAM Journal on Financial Mathematics*, 5:99–136, 2014.

B. Zhou. High-frequency data and volatility in foreign-exchange rates. *Journal of Business & Economic Statistics*, 14:45–52, 1996.