Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Does the Prohibition of Trade-Through Hurt Liquidity Demanders?

Ningyuan Chen, Pin Gao, Steven Kou

To cite this article:
Ningyuan Chen, Pin Gao, Steven Kou (2023) Does the Prohibition of Trade-Through Hurt Liquidity Demanders?. Operations Research

Published online in Articles in Advance 26 Apr 2023

. https://doi.org/10.1287/opre.2023.2454

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–14
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Contextual Areas

Does the Prohibition of Trade-Through Hurt Liquidity Demanders?

Ningyuan Chen,[a] Pin Gao,[b,c,*] Steven Kou[d]

[a] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [b] School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China; [c] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518172, China; [d] Department of Finance, Questrom School of Business, Boston University, Boston, Massachusetts 02215

*Corresponding author

Contact: ningyuan.chen@utoronto.ca, https://orcid.org/0000-0002-3948-1011 (NC); gaopin@cuhk.edu.cn, https://orcid.org/0000-0001-7434-7902 (PG); kou@bu.edu, https://orcid.org/0000-0003-4457-8384 (SK)

Received: January 16, 2019
Revised: August 13, 2020; October 29, 2021
Accepted: March 21, 2023
Published Online in Articles in Advance: April 26, 2023

Area of Review: Financial Engineering

https://doi.org/10.1287/opre.2023.2454

Abstract. The order protect rule (OPR) in the United States generally prohibits any trade-through, that is, a market order that is not executed at the best possible price among fast (electronic and automated) trading venues. By deriving upper and lower bounds for the difference in the execution costs in a dynamic model, we find that, although trade-through allows for flexible trading strategies and may benefit the liquidity demander, the benefit is insignificant in most cases, especially for small trades and stocks with fast resilience. Therefore, considering other benefits of the OPR studied in the literature, this study supports the regulation and suggests that the current separate regulations for fast and slow venues may be extended to differentiate stocks with fast and slow resilience speeds.

Funding: P. Gao received financial support from the National Natural Science Foundation of China [Grants 72201234 and 72192805] and the Research Grant Council of Hong Kong (the Collaborative Research Fund) [Grant C6032-21G].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.2454.

Keywords: market microstructure • order execution • dynamic programming • model-free • equilibrium

1. Introduction

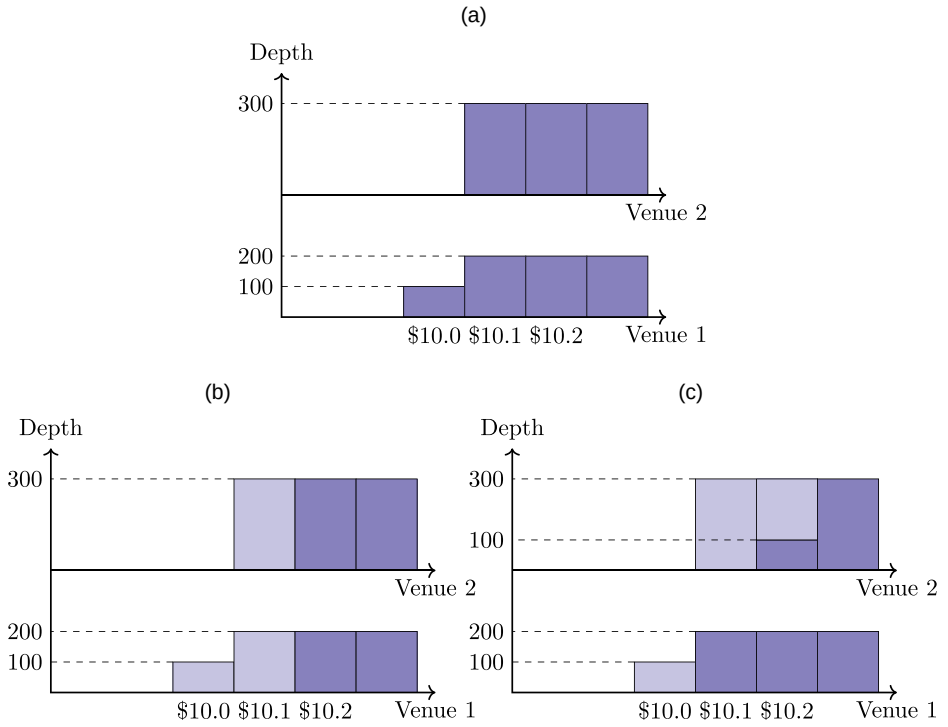
A stock can be traded at multiple venues, including primary and regional exchanges, crossing networks, and electronic communication networks, with each venue maintaining its own order books. To encourage competition among various trading venues and promote efficiency, the U.S. Securities and Exchange Commission established the regulation national market system (Reg NMS) in 2005 (Securities and Exchange Commission 2005). One particular rule of Reg NMS, the order protection rule (OPR), is related to the prohibition of trade-throughs.

A trade-through is defined to be a market order (after proper routing) that is not executed at the best possible price among all trading venues. Note that splitting orders optimally between different venues (smart routing) is not considered a trade-through by the OPR rule. Consider two venues illustrated in panel (a) of Figure 1: there are 100 limit sell orders placed at \$10.0 in venue 1 and other limit orders are displayed similarly. Suppose an investor intends to buy 600 shares through market buy orders. Panel (b) of Figure 1 is not considered a trade-through, and the cost for the investor is $\$10.0 \times 100 + \$10.1 \times 200 + \$10.2 \times 300 = \$6,050$. On the other hand, panel (c) is considered a trade-through because the fraction of the market order executed at \$10.2 in venue 2 could have been guaranteed a better price if routed to venue 1 instead. The cost in this case is $\$10.0 \times 100 + \$10.1 \times 300 + \$10.2 \times 200 = \$6,070$.

Trade-through may occur because of various reasons: (i) Brokers may lack the necessary technology (e.g., smart routers) to submit the orders to the venue with the best price immediately. (ii) Brokers may act against the best interests of their clients, for example, internalizing their inventory to avoid monitoring costs and fees, resulting in trade-throughs. (iii) Investors, especially institutional traders, may execute their market orders at inferior prices intentionally for faster execution or less market impact; this is the focus of the current paper.

Before the implementation of Reg NMS, orders could be executed at inferior prices to the best quote among all venues. It is estimated that 1.5% of all trades in the U.S. equity market were trade-throughs in 2001 (Bessembinder 2003) and even more for exchange-listed equity options (Battalio et al. 2004). To protect investors, trade-throughs are prohibited in the regulation OPR with the exception that trading through a better price on a slow (manual and floor-based exchanges) market is allowed. The response time in fast markets is usually less than one second, whereas that in slow markets can be as large

Figure 1. (Color online) The Ask Sides (All Limit Sell Orders) of Two Order Books



Notes. The lighter area indicates the orders matched to the market order. (a) Before trade. (b) Not a trade-through. (c) A trade-through.

as 15 seconds (Securities and Exchange Commission 2005). This exception is introduced in OPR because investors usually value execution speed as much as price.

The OPR has aroused controversy among academics and practitioners as trade-throughs are allowed in other countries, such as Australia. Supporters believe that the protection against trade-throughs ensures a better price for market orders and protects investors from conflict of interest. For example, the execution cost for the investor in panel (c) in Figure 1 is higher than (b). However, opponents argue that, although involuntary (or "bad") trade-throughs are prevented (i.e., reasons (i) and (ii)), intentional (or "good") trade-throughs (e.g., reason (iii)) are also banned, which may impose unnecessary constraints on investors' trading strategies when managing their portfolios. For example, although panel (c) in Figure 1 has a higher execution price than (b), the scenario in panel (b) may end up having a larger long-term market impact. It could happen because of two reasons: first, a large trade at venue one, which has relatively small market depth, tends to have a larger impact; second, scenario (b) leads to a higher national best bid and offer than (c). Thus, in the long run, panel (b) may make future order executions more costly, and investors are willing to trade short-term benefits for long-term gains; that is, trade-through may be desirable.

Our study focuses on how the prohibition of trade-throughs affects a liquidity demander who may have intentionally traded through. More precisely, we consider an investor, who demands a certain amount of an asset by a given time and submits market orders in multiple periods to a limit order market with multiple venues with the objective to minimize the expected total execution cost. The decision making of the liquidity demander involves how to split the whole order into smaller market orders over time and between different venues.

1.1. Counterfactual Analysis

It is difficult to compare the effects of allowing versus prohibiting trade-through as, at any given time, only one case occurs. In other words, one cannot observe the effects of trade-through and the prohibition of trade-through simultaneously. Therefore, we conduct a counterfactual analysis comparing the optimal trading costs of the investor with and without the prohibition of trade-through. In the traditional counterfactual analysis (see, e.g., Imbens and Rubin 2015), at any given time, each individual can only belong to either the treatment group or the control group but not both. Because of this difficulty, in a counterfactual analysis, (i) the dynamics for the treatment and control groups may have totally different dynamics, and (ii) some external assumptions (such as the famous ignorability condition proposed by Rubin 1976), which cannot be verified internally, must be imposed to link the dynamics of the treatment and control groups; these external assumptions must be reasonable and play a role such as axioms.

The same difficulties are also encountered here: (1) We have to assume two different model-free (quite general) equilibrium dynamics with and without trade-through; see Section 2.1. (2) When deriving our upper bound for the benefit of allowing trade-through in Theorem 1, we impose Assumption 2 to link the equilibrium dynamics with and without trade-through, which states that the prohibition of trade-through increases liquidity provision (i.e., stimulating the placement of limit orders). This assumption is consistent with the empirical evidence in Foucault and Menkveld (2008) and is associated with stronger resilience and a deeper market. It is important to point out that we do not assume anything about the magnitude of the increase in liquidity provision associated with the prohibition of trade-through. By conducting the counterfactual analysis, we hope to give a fair (apples-to-apples) comparison of the impacts of allowing versus prohibiting trade-through on liquidity demanders under quite robust equilibrium dynamics.

1.2. Our Contribution and Policy Implications

We conduct a counterfactual analysis that compares the total execution costs when trade-through is allowed or prohibited. The main theoretical result of the paper is a set of upper and lower bounds, indicating that the benefit of trade-through or, conversely, the extra cost to liquidity demanders when trade-through is prohibited is not large, especially for small trades and liquid stocks. The quantification of the benefit may yield significant policy implications. More precisely, once the trading frequency is fixed, the gap between the optimal costs with and without the prohibition of trade-through is mainly determined by three factors. If (i) the liquidity demand of the investor is low (as for most retail investors), (ii) the stock has large market depth, or (iii) the limit orders are filling inside the quote rapidly (strong resilience), then prohibiting trade-through does not impose a significant restriction. Our simple bound captures only the most important factors, and its effectiveness is demonstrated in a case study: the trade-through benefit is negligible for BHP Billiton (BHP), a stock traded on the Australian Securities Exchange (ASX).

Our findings have a policy implication. One of the major arguments against the OPR is the failure to consider the diversity of investors, especially those who value price impact over execution cost, because they may hold large positions of assets. We model the decision making of this group of investors and conclude that, in many cases, they are not greatly handicapped by the implementation of the OPR in terms of their total execution costs.[1] Therefore, the cost of the regulation may be outweighed by its benefit, which is already well-documented in Foucault and Menkveld (2008) and O'Hara and Ye (2011) as the prevention of trade-through helps to improve efficiency and attract liquidity provision in a market with multiple trading venues.

Our result suggests an extension to the current exception rules in OPR to differentiate stocks with fast and slow resilience speeds. In particular, we find that the resilience speed is crucial in determining whether trade-through is needed. For stocks with fast resilience, the trade-through benefit for liquidity demanders is negligible, and prohibiting trade-through does not constrain their trading strategies; for illiquid stocks with slow resilience, trade-through might be beneficial.

Methodologically, to the best of our knowledge, this paper is the first to present a general dynamic model to investigate the benefit of trade-through with interdependent trading venues. The model incorporates the following features: (i) order splitting over time and between venues, (ii) stochastic liquidity provision with arbitrary distribution and dependence structure, and (iii) a counterfactual analysis that compares the outcomes of two scenarios with and without trade-through. The generality of the model makes the conclusions and policy implications more robust. Despite the generality and complexity of the model, we are able to obtain simple upper and lower bounds for the benefit of intentional trade-through, whose effectiveness is illustrated via a case study in Section 6.

1.3. Literature Review

Blume (2007) and O'Hara (2004) express their concerns about the OPR: because the investors are highly diversified, some of them may not focus on minimizing the execution cost, but rather worry more about the market impact; based on the presumption that all investors prefer better execution price of the current trade, the OPR may, in fact, impose an unnecessary restriction on trading strategies. We complement their results by giving examples and market conditions such that trading through can indeed reduce price impact and achieve a lower total execution cost over time.

Positive viewpoints regarding the OPR are given in Foucault and Menkveld (2008), O'Hara and Ye (2011), and Hendershott and Jones (2005). Based on an equilibrium model of an incumbent and an entrant market, Foucault and Menkveld (2008) find empirically that the prohibition of trade-throughs plays an important role in enhancing liquidity provision. O'Hara and Ye (2011) find that market fragmentation, that is, the availability of multiple trading venues, improves execution quality and efficiency; they attribute the improvement to the implementation of the OPR and argue that whether the investors can benefit from market fragmentation is unclear without the protection. Hendershott and Jones (2005) empirically compare the market quality of three active exchange-traded funds (ETFs) before and after the relaxation of the OPR on ETFs. According to their findings, the market quality, measured by the effective spread and others, is essentially unchanged.

Complementary to these results, we show that the "downside" of prohibiting trade-through for a liquidity demand is insignificant for small trades and liquid stocks. Thus, considering the role of the OPR in improving market quality (O'Hara and Ye 2011), attracting liquidity provision (Foucault and Menkveld 2008), and preventing involuntary trade-throughs, we may conclude that the enforcement of the OPR arguably does more good than harm. Moreover, our findings complement Foucault and Menkveld (2008) on the connection between trade-through and liquidity provision with their results focusing on liquidity providers and ours on liquidity demanders.

Our dynamic trading model is related to the literature studying optimal order execution; see, for example, Predoiu et al. (2011), Obizhaeva and Wang (2013), Tsoukalas et al. (2019), and Chen et al. (2018). Here are some new features of our model. (i) The limit order profile, or its market depth, can be a function of price and is more general than block-shaped models in Obizhaeva and Wang (2013), Chen et al. (2018), and Tsoukalas et al. (2019). (ii) There are multiple trading venues whose dynamics are interdependent. To our best knowledge, optimal order execution is only considered in a single order book except this paper and Tsoukalas et al. (2019). The latter studies the dynamic portfolio execution of correlated assets rather than a single asset in a fragmented market. (iii) We assume the market depth follows a Markov process. It requires the investor to make dynamic decisions rather than following a predetermined strategy. This is similar to Chen et al. (2018), whereas in other papers, the market depth is usually a deterministic process.

2. Basic Setting of the Counterfactual Analysis

A stock can be traded in m venues with each venue maintaining its own limit order book. Consider the snapshots of the ask sides of the order books, including all limit sell orders, at time points $0 = t_0 < t_1 < \cdots < t_n = T$. The snapshots are taken before the transactions at t_i, and we denote the time point right after the transactions at t_i by the subscript $i+$. The time-t_i state of the order books is characterized by the following primitives, observed or inferred by market participants: M_i, the fundamental value of the stock; $d_{i,j}$, the spread between the fundamental value and the best ask price of venue j's order book; $\mathcal{D}_{i,j}(\cdot)$, the market depth of venue j; s_i, the transient equilibrium spread which is introduced shortly after.

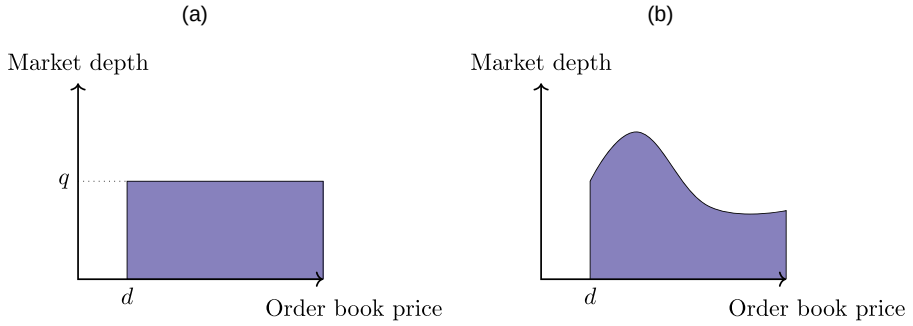
Given the fundamental value and the spreads, the best ask price of venue j is $M_i + d_{i,j}$. The market depth $\mathcal{D}_{i,j}(\cdot) : \mathbb{R}^+ \longmapsto \mathbb{R}^+$ represents the density of limit orders in venue j adjusted for the best ask price; that is, there are $\mathcal{D}_{i,j}(p)dp$ shares of limit orders inside the price interval $[M_i + d_{i,j} + p, M_i + d_{i,j} + p + dp)$. The size of limit orders and prices are continuous variables in our model. Note that we do not require the function $\mathcal{D}$ to be continuous. The fundamental value, the spreads, and the market depth uniquely determine the profiles of the ask sides of the limit order books.

We also introduce an unobservable quantity called the *transient equilibrium spread* s_i at time t_i. The transient equilibrium price $M_i + s_i$ can be interpreted as follows: if we hypothetically aggregate the limit orders of all venues at t_i, thus eliminating market fragmentation, then the best ask price of the consolidated order book becomes $M_i + s_i$. Note that $s_i \neq \min_j d_{i,j}$ because, upon aggregating, some information is diffused to the whole market and the limit order profile may change. (We do not try to model such information diffusion explicitly.) The transient equilibrium price is essential to capture the correlated liquidity provision in the venues.

We give three examples of limit order profiles:

- A block-shaped limit order book: Suppose $\mathcal{D}(p) \equiv q$ for $p \geq 0$, that is, there are q shares of limit orders within a price unit, independent of the price p. This example is illustrated in the left panel of Figure 2. Although extremely simple, this type of limit order book is used widely in the existing models because of its analytical tractability.
- A more general limit order book: As supported by empirical evidence (e.g., Biais et al. 1995), the market depth typically first increases and then decreases as the

Figure 2. (Color online) Two Examples of the Order Book Profiles



Notes. The price d represents the best ask, and the y-axis represents the density or market depth of the limit orders. (a) A block-shaped limit order book. (b) A real limit order book.

price increases from the best ask. It suggests a limit order profile as in the right panel of Figure 2.

- The example in panel (a) of Figure 1 with continuous tick sizes: The best ask price in venue 1 is $M_0 + d_{0,1} = 9.95$, and we define $\mathcal{D}_{0,1}(p)$ for venue 1 as

$$\mathcal{D}(p) = 1,000\mathbb{I}_{\{p\in[0,0.1)\}} + 2,000\mathbb{I}_{\{p\in[0.1,+\infty)\}}. \tag{1}$$

The major difference from Figure 1 is that the tick sizes are infinitesimal tick sizes as an approximation for the reality. For venue 2, it is block shaped.

2.1. Notations and Trading Mechanism

Consider an investor (liquidity demander) who trades on dates $0 = t_0 < t_1 < \cdots < t_n = T$ in order to purchase x_0 shares by time T. To narrow our analysis down to intentional trade-throughs, we consider only one liquidity demander who can submit market orders without any delay or commission fees. The cases of latency and multiple liquidity demanders is studied in Section 5. Suppose, at time t_i, a market order of size $u_{i,j}$ is submitted to venue j. Thus, $\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j} = x_0$.[2] The limit sell orders listed on venue j are consumed from price low to high, and its best ask price increases instantaneously. Upon the transaction of $u_{i,j}$, the spread $d_{i,j}$ is pushed up to $d_{i+,j}$. The impact of this transaction on other venues $k \neq j$ does not occur instantaneously at t_i.

The marginal cost of the order adjusted for the best ask (i.e., we subtract the best ask price before the transaction from the marginal cost), ϕ, depends on the order book's market depth and the order size. More precisely, if we define

$$\phi(\mathcal{D}_{i,j}, u_{i,j}) \triangleq \inf\left\{\overline{p}\,\middle|\,\int_0^{\overline{p}} \mathcal{D}_{i,j}(p)dp = u_{i,j}\right\} \tag{2}$$

to be the marginal cost of the market order, adjusted for the best ask price, then the spread immediately after the trade is given by

$$d_{i+,j} = d_{i,j} + \phi(\mathcal{D}_{i,j}, u_{i,j}). \tag{3}$$

As an example, for block-shaped order books (panel (a) of Figure 2) considered in Obizhaeva and Wang (2013) and Chen et al. (2018), $\phi(\mathcal{D}_{i,j}, u_{i,j}) = u_{i,j}/q$, where q is the constant number of limit orders within unit price.

Applying to Figure 1 with market depth in (1), for venue 1, $\phi(\mathcal{D}, u) = \mathbb{I}_{\{u\le 100\}}u/1,000 + \mathbb{I}_{\{u>100\}}(u-100)/2,000$. There is a subtle difference when modeling real-world limit order books with discrete ticks. In particular, when ticks are discrete as in Figure 1, the unadjusted marginal cost in venue 1 is a constant 10 for $u \le 100$ and 10.1 for $u \in (100, 300]$. In our model, in contrast, it can be seen as a linear approximation to the piecewise constant function: the marginal cost is $9.95 + u/1,000$ for $u \le 100$ and $10.05 + (u-100)/2,000$ for $u \in (100, 300]$. The difference becomes inconsequential as the world has seen a shift toward smaller tick sizes in most exchanges.

The transactions have a price impact on the transient equilibrium spread s_i at time t_i as well. We consider a general form:

$$s_{i+} = s_i + g\left(\mathcal{D}_{i,1}, \ldots, \mathcal{D}_{i,m}, u_{i,1}, \ldots, u_{i,m}\right). \tag{4}$$

We do not make any assumptions on the function $g(\cdot)$ here.

2.2. Model-Free Equilibrium Dynamics

Next, we describe the equilibrium dynamics of the limit order books between t_{i+}, the time right after the trade, and t_{i+1}. As limit and market orders of market participants are submitted and processed, the state of the order books evolves between t_{i+} and t_{i+1}. The dynamics may arise from the equilibrium behavior of traders in the market. Although we do not specify the equilibrium explicitly, the dynamics may cover a wide range of models because of the generality (we do not require any distributional assumption and, thus, refer to the dynamics as "model-free"). To clarify the information structure, we denote the σ-field generated by the available information before the trade by the investor at t_i by $\mathcal{F}_i$ and denote the information right after the trade at t_i by $\mathcal{F}_{i+} = \sigma\{\mathcal{F}_i, u_{i,1}, \ldots, u_{i,m}\}$. The dynamics of M, $\mathcal{D}$, d, and s are introduced subsequently.

2.2.1. The Dynamics of M_i. We assume M_i is common knowledge and a martingale adapted to $\mathcal{F}_i$, which is consistent with the efficient market hypothesis. That is, $\mathsf{E}[M_{i+1}|\mathcal{F}_i] = M_i$. This is a standard assumption in many equilibrium models (Goettler et al. 2005). Because the investor we are considering is not an insider, $\mathcal{F}_{i+}$ does not provide extra information for M_{i+1} compared with $\mathcal{F}_i$.

2.2.2. The Dynamics of $\mathcal{D}_{i,j}$. The submission and cancellation of limit orders outside the quotes give rise to the change of the market depth. We make a very mild Markovian assumption that the distribution of $\{\mathcal{D}_{i+1,j}(\cdot)\}_{j=1}^{m}$ only depends on the limit order book profiles and the investor's orders at t_i but not on the entire history, whereas no assumption is imposed on the distribution itself. That is, the distribution of $\{\mathcal{D}_{i+1,j}\}_{j=1}^{m}$ depends on $(t_i, s_i, \{d_{i,j}\}_{j=1}^{m}, \{\mathcal{D}_{i,j}\}_{j=1}^{m}, \{u_{i,j}\}_{i=1}^{m})$. Here, M_i does not affect the future market depth because it is common knowledge of the market. Because $\mathcal{D}_{i,j}(\cdot)$ itself is a function, it can represent any general state of the limit order book. Moreover, the market depth across venues, $\mathcal{D}_{i,j}(\cdot)$ for different j, is allowed to be dependent in an arbitrary manner.

Because $\mathcal{D}_{i,j}(\cdot)$ and its dynamics are not restricted to a particular form, one may, for example, calibrate a functional time series using the market data or assume

block-shaped order books ($\mathcal{D}$ is a constant function), which follows a Markov chain as in Chen et al. (2018). Such generality allows our model to reproduce various empirical phenomena of the market depth, for example, its mean exhibiting a reverse U-shape during a day (Ahn et al. 2001), the correlation of liquidity effects across venues, etc. This, in turn, makes our conclusion and policy implications more robust.

2.2.3. The Dynamics of s_i and $d_{i,j}$. Between t_i and t_{i+1}, new limit orders are placed inside the spread, and existing limit orders might be canceled, giving rise to the dynamics of s_i and $d_{i,j}$. Denote $\gamma_i \triangleq (\log(s_{i+}) - \log(s_{i+1}))/(t_{i+1} - t_i)$ and $\rho_{i,j} \triangleq (\log(d_{i+,j} - s_{i+}) - \log(d_{i+1,j} - s_{i+1}))/(t_{i+1} - t_i)$, which is equivalent to

$$\begin{aligned} s_{i+1} &= \exp(-\gamma_i(t_{i+1} - t_i))s_{i+}, \\ d_{i+1,j} - s_{i+1} &= \exp(-\rho_{i,j}(t_{i+1} - t_i))(d_{i+,j} - s_{i+}). \end{aligned} \tag{5}$$

In the literature, γ_i and $\rho_{i,j}$ are referred as the resilience rates because, when γ_i and $\rho_{i,j}$ are large, s_{i+1} and $d_{i+1,j} - s_{i+1}$ revert to zero at a higher rate. In practice, the best ask prices tend to revert to the transient equilibrium price, and the transient equilibrium price reverts to zero over time if no transaction occurs. Such resilience effect of the order books is well-documented in empirical studies (Biais et al. 1995).

We make a mild assumption that the nonnegative random variables γ_i and $\rho_{i,j}$ are Markovian; that is, their distribution only depends on $(t_i, s_i, \{d_{i,j}\}_{j=1}^m, \{\mathcal{D}_{i,j}\}_{j=1}^m, \{u_{i,j}\}_{i=1}^m)$. In the theoretical literature (see, e.g., Obizhaeva and Wang 2013, Chen et al. 2018, Tsoukalas et al. 2019), the rates γ_i and $\rho_{i,j}$ are usually assumed to be deterministic, automatically satisfying our Markovian assumption. In the empirical literature, autoregressive models are more popular (Large 2007, Lo and Hall 2015) and are also Markovian if the lag is one.

In summary, we make the following mild assumption regarding the equilibrium dynamics.

Assumption 1 (Equilibrium Dynamics). *The fundamental value M_i is common knowledge and a martingale; the distribution of the market depth $\{\mathcal{D}_{i+1,j}\}_{j=1}^m$ and the resilience rates $\gamma_i \geq 0$ and $\rho_{i,j} \geq 0$ depend on $(t_i, s_i, \{d_{i,j}\}_{j=1}^m, \{\mathcal{D}_{i,j}\}_{j=1}^m, \{u_{i,j}\}_{i=1}^m)$. Note that the change of the market depth and the resilience effect of the order book may depend on $\{u_{i,j}\}_{i=1}^m$, the trading decision of the investor.*

In the literature, most equilibrium models of market structure study a single limit order book, which usually follows a Markov process (Foucault et al. 2005, Goettler et al. 2005, Roşu 2009): the order book profiles in the next period are determined by the profiles in the current period. To the best of our knowledge, no papers study a dynamic game-theoretical model of multiple trading venues. Our model allows for a general stochastic structure and is, thus, able to capture dependence between different trading venues. For example, to reflect the positive correlation of future liquidity provision in the venues, we can let the limit order profiles in the next period $\mathcal{D}_{i+1,j}$ be positively correlated for different venue j.

With this setup, a transaction at t_i in venue j instantaneously increases the price of venue j, $d_{i,j}$, and the transient equilibrium spread s_i. By the resilience effect, the price impact diffuses to other order books in the fragmented market by (5). A possible scenario of market dynamics after panel (b) of Figure 1 is illustrated in Figure 3.

2.3. Optimal Execution Cost

The liquidity demander's objective is to minimize the expected total cost of purchasing x_0 shares over the period $[0, T]$.[3] With Notation (2), the cost resulting from the market order $u_{i,j}$ is

$$\begin{aligned} &\int_0^{\phi(\mathcal{D}_{i,j}, u_{i,j})} (p + M_i + d_{i,j})\mathcal{D}_{i,j}(p)dp \\ &\quad = u_{i,j}(M_i + d_{i,j}) + \Phi(\mathcal{D}_{i,j}, u_{i,j}), \end{aligned}$$

where $\Phi(\mathcal{D}, u) \triangleq \int_0^{\phi(\mathcal{D},u)} p\mathcal{D}(p)dp$. For example, applying to Figure 1 with continuous tick sizes, the order execution cost of venue 2 in panel (c) is $\int_0^{\phi(\mathcal{D}_{0,2}, 500)} (10.05 + p) \times \mathcal{D}_{0,2}(p)dp = 5,066.7$ as $\mathcal{D}_{0,2}(p) \equiv 3,000$ and $\phi(\mathcal{D}_{0,2}, 500) = 1/6$. This is slightly different from the scenario with discrete tick sizes as shown in panel (c) with the cost $\$10.1 \times 300 + \$10.2 \times 200 = \$5,070$.

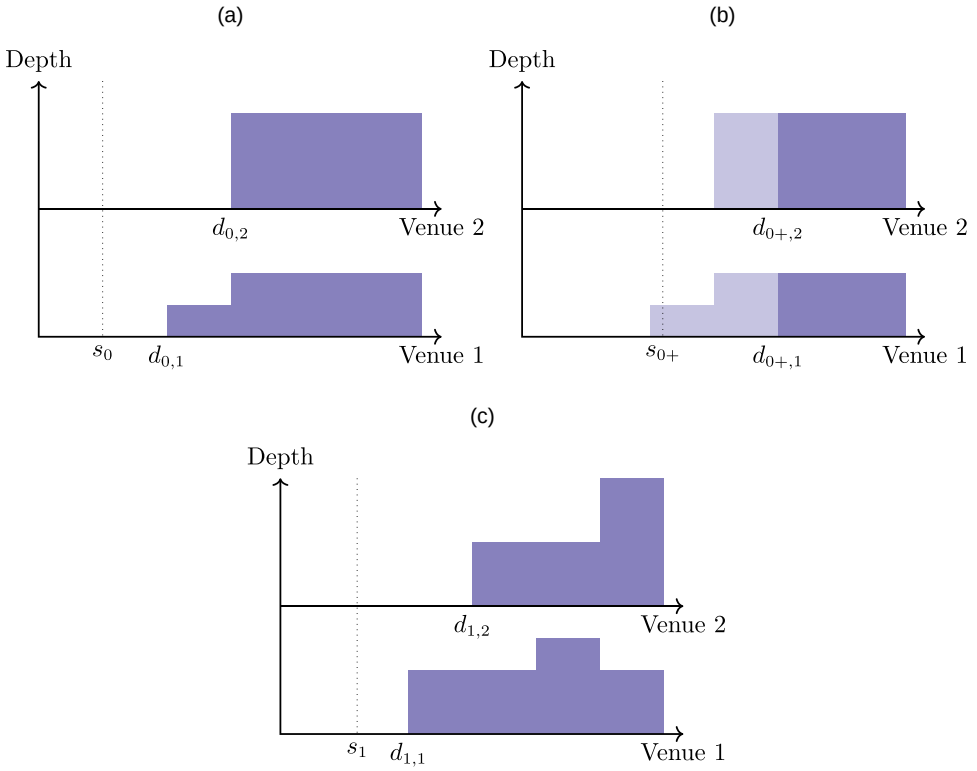
The cost has two components: The price effect $u_{i,j}(M_i + d_{i,j})$, which is linear in the order size, and the market microstructure effect, which incurs the extra cost $\Phi(\mathcal{D}_{i,j}, u_{i,j})$, including higher order terms. Note that, under market microstructure, the notion that the stock is traded at a single price, which corresponds to an infinite market depth, is no longer true. Rather, the price is determined by a local supply/demand curve, which is captured by the market depth $\mathcal{D}$ in our model. For example, if $\mathcal{D}_{i,j}$ follows the left panel of Figure 2, then the marginal cost is $\phi(\mathcal{D}, u) = u/q$ and the extra execution cost is $\Phi(\mathcal{D}, u) = u^2/2q$. This is one of the few examples in which ϕ and Φ can be analytically computed from $\mathcal{D}$.

In summary, the liquidity demander chooses $\{u_{i,j}\}_{j=1}^m \in \mathcal{F}_i$ to minimize the execution cost

$$\mathsf{E}\left[\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j}(M_i + d_{i,j}) + \Phi(\mathcal{D}_{i,j}, u_{i,j}) \,\middle|\, M_0, x_0, \{d_{0,j}\}_{j=1}^m, s_0, \{\mathcal{D}_{0,j}\}_{j=1}^m\right] \tag{6}$$

where the price dynamics are specified in (3)–(5), M_i is a martingale, and the information structure is given by Assumption 1. In the following, we may omit the initial condition of the order books when there is no confusion.

Figure 3. (Color online) An Illustration of the Trading Mechanism Based on Panel (b) of Figure 1



(c) Before trade at t_1

Notes. After the market order is placed at t_0, limit orders are consumed. Meanwhile, s_0 and $d_{0,j}$ are increased. After a period $t_1 - t_0 > 0$, new limit orders are submitted, and some existing ones are canceled. As a result, $d_{1,1}$, $d_{1,2}$, and s_1 all revert to zero. (a) Before trade at t_0. (b) After trade at t_0. (c) Before trade at t_1.

2.4. Counterfactual Analysis of Allowing/Prohibiting Trade-Throughs

Because trade-through is currently prohibited in the U.S. stock market, our analysis involves evaluating the order execution cost of the investor in a counterfactual setting in which trade-through is allowed. In the two scenarios (trade-through is allowed or prohibited), the equilibrium dynamics may differ significantly because of the impact of trade-through on liquidity provision (Foucault and Menkveld 2008), which, in turn, affects the optimal execution cost.

To encode such a policy impact on the equilibrium dynamics, we use $\mathcal{T}$ and $\mathcal{N}$ to denote the order book states or parameters under "trade-through" and "no trade-through." For example, $\rho^{\mathcal{T}}_{i,j}$ denotes the resilience rate of venue j between t_{i+} and t_{i+1} in the universe in which trade-through is allowed. More precisely, the following quantities related to the equilibrium dynamics may differ when trade-through is allowed or prohibited: the resilience effect γ_i and $\rho_{i,j}$ and the dynamics of the market depth $\mathcal{D}_{i,j}$ for $i \geq 1$. For simplicity, we sometimes use $\mathsf{E}^{\mathcal{N}}$ and $\mathsf{E}^{\mathcal{T}}$ (such as Equation (6)) to denote the two scenarios instead of adding superscript to all the quantities.

Perhaps a more direct impact of prohibiting trade-through on the liquidity demander is that the set of trading policies is restricted. A trade-through occurs when a market order is not executed at the best possible price among all venues at that time. In our model, the orders $\{u_{i,j}\}_{j=1}^m$ are considered a trade-through if there exist j and k such that $d_{i+,j} > d_{i+,k}$ and $u_{i,j} > 0$, that is, the marginal cost of the order $u_{i,j}$ could have been lowered if submitted to venue k. Before the establishment of Reg NMS, there is no constraint on $u_{i,j}$, and in particular, trade-through is allowed. The corresponding policy set is

$$\Theta^{\mathcal{T}} = \left\{ u_{i,j} \middle| \sum_{i=0}^{n} \sum_{j=1}^{m} u_{i,j} = x_0,\ u_{i,j} \geq 0,\ u_{i,j} \text{ is } \mathcal{F}_i\text{-measurable},\ i = 0,\dots,n,\ j = 1,\dots,m \right\}. \quad (7)$$

If trade-through is prohibited, then only those policies in $\Theta^{\mathcal{T}}$ that do not trade through are now admissible:

$$\Theta^{\mathcal{N}} = \{u_{i,j} \in \Theta^{\mathcal{T}} \,|\, \text{If } u_{i,j} > 0, \text{ then } d_{i+,k} \geq d_{i+,j} \text{ for any } k \neq j,\ i = 0,\dots,n\}. \quad (8)$$

Consequently, starting from the same initial market environment, we are comparing two worlds,

$$J^{\mathcal{T}} \triangleq \min_{u_{i,j}\in\Theta^{\mathcal{T}}} \mathsf{E}^{\mathcal{T}}\left[\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j}(M_i + d_{i,j}) + \Phi(\mathcal{D}_{i,j}, u_{i,j}) \,\middle|\, M_0, x_0, \{d_{0,j}\}_{j=1}^{m}, s_0, \{\mathcal{D}_{0,j}\}_{j=1}^{m}\right]$$

$$J^{\mathcal{N}} \triangleq \min_{u_{i,j}\in\Theta^{\mathcal{N}}} \mathsf{E}^{\mathcal{N}}\left[\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j}(M_i + d_{ij}) + \Phi(\mathcal{D}_{i,j}, u_{i,j}) \,\middle|\, M_0, x_0, \{d_{0,j}\}_{j=1}^{m}, s_0, \{\mathcal{D}_{0,j}\}_{j=1}^{m}\right],$$

to gauge the benefit of trade-through to the liquidity demander.

Similar to Predoiu et al. (2011) and Obizhaeva and Wang (2013), the fundamental value doesn't affect the decision making. More precisely,

$$J^{\mathcal{T}} = \min_{u_{i,j}\in\Theta^{\mathcal{T}}} \mathsf{E}^{\mathcal{T}}\left[\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j}d_{i,j} + \Phi(\mathcal{D}_{i,j}, u_{i,j})\right] + M_0x_0$$

$$J^{\mathcal{N}} = \min_{u_{i,j}\in\Theta^{\mathcal{N}}} \mathsf{E}^{\mathcal{N}}\left[\sum_{i=0}^{n}\sum_{j=1}^{m} u_{i,j}d_{i,j} + \Phi(\mathcal{D}_{i,j}, u_{i,j})\right] + M_0x_0.$$

Therefore, we set $M_i \equiv 0$ without loss of generality. Moreover, we consider equal-spaced trading times and let $t_i = i\Delta t$.

The expected total execution cost is essentially equivalent to the effective spread, a measure for execution quality commonly used in empirical studies (e.g., Hendershott and Jones 2005, O'Hara and Ye 2011). To see this, note that the effective spread is defined as the difference between the quote midpoint and the transaction price. In our model, the quote midpoint can be interpreted as M_t because we only focus on one side of the order books. For the transaction $u_{i,j}$, the effective spread is $d_{i,j} + p$ for the infinitesimal $\mathcal{D}_{i,j}(p)dp$ shares of the market order given that $p \le \phi(\mathcal{D}_{i,j}, u_{i,j})$. Therefore, the average effective spread for this transaction is $\int_0^{\phi(\mathcal{D}_{i,j},u_{i,j})} (d_{i,j}+p)\mathcal{D}_{i,j}(p)dp$, which equals to $d_{i,j}u_{i,j} + \Phi(\mathcal{D}_{i,j}, u_{i,j})$. Hence, the overall effective spread is the total execution cost minus the portion caused by the fluctuation of the fundamental value, and the latter term is zero under the assumption $M_i \equiv 0$.

3. The Magnitude of the Trade-Through Benefit

In this section, we quantify the magnitude of $J^{\mathcal{N}} - J^{\mathcal{T}}$, that is, the benefit of trade-through, by studying a dynamic model. A dynamic model is necessary to answer this question as, in Section EC.1 of the online supplement, we prove that it is never optimal to trade through when submitting a one-period market order and trade-through may be desirable in dynamic trading over periods. In other words, to understand the benefit of trade-through, a static model is not sufficient.

The analytical characterization of the difference $J^{\mathcal{N}} - J^{\mathcal{T}}$ is virtually infeasible because of the generality of our setup. The main mathematical result of this section is upper and lower bounds for the trade-through benefit in the form of a simple and explicit function of the market parameters. The result provides new insight into the debate about the OPR as it quantifies the extra cost of a liquidity demander and the change of effective spread after the regulation is enforced. The bounds are also computed in an empirical setting in Section 4 using parameters calibrated from real data.

3.1. Main Assumption

Just as with the standard counterfactual analysis, we need to impose some external assumptions (such as the ignorability condition in the counterfactual analysis). If we allow the dynamics of $\mathcal{N}$ and $\mathcal{T}$ to be arbitrarily different, then one cannot expect to obtain a meaningful bound for $J^{\mathcal{N}} - J^{\mathcal{T}}$. Therefore, we impose the following assumption.

Assumption 2. *We assume that the prohibition of trade-through increases liquidity provision, which is associated with stronger resilience and a deeper market. More precisely, given the same market condition at t_i, that is, $(s_i^{\mathcal{T}}, \{d_{i,j}^{\mathcal{T}}\}_{j=1}^m, \{\mathcal{D}_{i,j}^{\mathcal{T}}\}_{j=1}^m) = (s_i^{\mathcal{N}}, \{d_{i,j}^{\mathcal{N}}\}_{j=1}^m, \{\mathcal{D}_{i,j}^{\mathcal{N}}\}_{j=1}^m)$, and that the aggregate order sizes in that period are identical, $\sum_{j=1}^m u_{i,j}^{\mathcal{T}} = \sum_{j=1}^m u_{i,j}^{\mathcal{N}}$, then the following hold: (i) The spreads in the next period when trade-through is allowed stochastically dominate those when trade-through is prohibited in the sense of first order stochastic dominance: $\mathsf{E}[f(s_{i+1}^{\mathcal{T}}, d_{i+1,1}^{\mathcal{T}}, \ldots, d_{i+1,m}^{\mathcal{T}})] \ge \mathsf{E}[f(s_{i+1}^{\mathcal{N}}, d_{i+1,1}^{\mathcal{N}}, \ldots, d_{i+1,m}^{\mathcal{N}})]$ for any increasing function $f(\cdot)$.[4] Note that the superscripts $\mathcal{T}$ and $\mathcal{N}$ denote the dynamics under trade-through and no trade-through, respectively. (ii) The market depth in the next period when trade-through is prohibited stochastically dominates that when trade-through is allowed in the sense of first order stochastic dominance. That is, $\mathsf{E}[f(\mathcal{D}_{i+1,1}^{\mathcal{T}}, \ldots, \mathcal{D}_{i+1,m}^{\mathcal{T}})] \le \mathsf{E}[f(\mathcal{D}_{i+1,1}^{\mathcal{N}}, \ldots, \mathcal{D}_{i+1,m}^{\mathcal{N}})]$ for any increasing functional $f(\cdot)$.[5]*

Assumption 2 merely states a mild relationship between the equilibrium dynamics: when trade-through is prohibited, the resilience effect tends to be stronger and the market tends to be deeper in the sense of first order stochastic dominance. It is consistent with the empirical evidence (Foucault and Menkveld 2008): prohibiting trade-through increases liquidity provision, which is associated with stronger resilience and a deeper market.

3.2. The Upper Bound

We first list additional technical assumptions and their interpretations.

Assumption 3. *For the dynamics with trade-through, in addition to Assumption 1, we assume that, conditional on* $(t_i, \{\mathcal{D}^{\mathcal{T}}_{i,j}\}_{j=1}^m, \{u_{i,j}\}_{j=1}^m)$, $(\{\mathcal{D}^{\mathcal{T}}_{i+1,j}\}_{j=1}^m, \gamma^{\mathcal{T}}_i, \{\rho^{\mathcal{T}}_{i,j}\}_{j=1}^m)$ *depends only on the new information, that is, independent of the old information* $(s_i, \{d_{i,j}\}_{j=1}^m)$.

We do not impose the assumption on the dynamics when trade-through is prohibited, which has a complicated constrained set for trading strategies. For the trade-through dynamics, the assumption can be interpreted as follows: conditional on the trade information and the current market depth, it is the new information associated with the market orders $\{u_{i,j}\}_{j=1}^m$ that drives the dynamics of the market depth and the price resilience, not the old information in the prices that has already been digested. Under Assumption 3, the resilience effect and the market depth in the next period can still depend on the old information (i.e., $(s_i, \{d_{i,j}\}_{j=1}^m)$) through $(t_i, \{\mathcal{D}^{\mathcal{T}}_{i,j}\}_{j=1}^m, \{u_{i,j}\}_{j=1}^m)$. Assumption 3 is slightly stronger than Assumption 1 because of the conditional independence and is easily satisfied in the optimal order execution literature (Obizhaeva and Wang 2013, Tsoukalas et al. 2019) in which the resilience rates are usually deterministic.

Assumption 4. *The function* $g(\mathcal{D}_1, \ldots, \mathcal{D}_m, u_1, \ldots, u_m)$ *in* (4) *satisfies* $\frac{\partial g}{\partial u_j} \in \left[0, \frac{1}{\min_p\{\mathcal{D}_j(p)\}}\right]$.

Recall that $g(\mathcal{D}_1, \ldots, \mathcal{D}_m, u_1, \ldots, u_m)$ is the price impact on the transient equilibrium spread s_i at time t_i if the market depth is $\mathcal{D}_{i,j} = \mathcal{D}_j$ and the market orders are $u_{i,j} = u_j$. The interpretation of Assumption 4 is as follows: the individual orders sent to venue j tend to increase the transient equilibrium price as well, and thus, $\frac{\partial g}{\partial u_j} \geq 0$. The price impact of venue j, denoted as $\phi(\mathcal{D}_j, u_j)$, satisfies $\frac{\partial \phi}{\partial u_j} \leq \frac{1}{\min_p\{\mathcal{D}_j(p)\}}$ according to Definition (2). Assumption 4 states that the price impact on the hypothetical consolidated limit order book is less than that on individual venues.

Assumption 5. *There exists a constant* $C_1^{\mathcal{T}} \leq 1$ *such that, for any given* $(s_i^{\mathcal{T}}, \{d_{i,j}^{\mathcal{T}}\}_{j=1}^m, \{\mathcal{D}_{i,j}^{\mathcal{T}}\}_{j=1}^m, \{u_{i,j}\}_{i=1}^m)$, *we have* $\exp(-\rho_{i,j}^{\mathcal{T}}\Delta t) \leq C_1^{\mathcal{T}}$ *and* $\exp(-\gamma_i^{\mathcal{T}}\Delta t) \leq C_1^{\mathcal{T}}$ *almost surely for all i and j.*

Assumption 5 automatically holds if there are resilience effects across all trading venues. In fact, the constant $C_1^{\mathcal{T}}$ measures the minimal resilience effect of all venues as well as the transient equilibrium price over the horizon. When the resilience effect is strong, $C_1^{\mathcal{T}}$ is expected to be much less than one as in the empirical study in Section 4. In practice, $C_1^{\mathcal{T}}$ can be estimated using the half-life of liquidity shocks and the trading frequency (see Section 4 for more details).

Assumption 6. *There exist positive constants* $\overline{D}^{\mathcal{T}}$ *and* $\underline{D}^{\mathcal{T}}$ *such that* $\overline{D}^{\mathcal{T}} \geq \mathcal{D}_{i,j}^{\mathcal{T}}(\cdot) \geq \underline{D}^{\mathcal{T}}$ *almost surely.*

Assumption 6 is mild as it requires that the market depth is uniformly bounded. Although $\overline{D}^{\mathcal{T}}$ does not appear in the formal results, the proofs depend on this upper bound.

Assumption 7. *For* $\mathcal{I} \in \{\mathcal{N}, \mathcal{T}\}$, *starting from the market condition* $(s_i^{\mathcal{I}}, \{d_{i,j}^{\mathcal{I}}\}_{j=1}^m, \{\mathcal{D}_{i,j}^{\mathcal{I}}\}_{j=1}^m)$ *at* t_i, *the optimal execution cost in the remaining horizon is increasing in* $(s_i^{\mathcal{I}}, \{d_{i,j}^{\mathcal{I}}\}_{j=1}^m)$ *and decreasing in the market depth* $\{\mathcal{D}_{i,j}^{\mathcal{I}}\}_{j=1}^m$ *in the sense of Endnotes* 4 *and* 5.

Assumption 7 is also mild. It essentially states that a deeper market and smaller bid–ask spread reduces the execution cost. Next, we provide an upper bound for the benefit of trade-through.

Theorem 1. *Suppose Assumptions 2–7 hold. Then, the benefit of trade-through is bounded by* $J^{\mathcal{N}} - J^{\mathcal{T}} \leq \frac{C_1^{\mathcal{T}} x_0^2}{2\underline{D}^{\mathcal{T}}}$.

There are two difficulties in deriving such a bound. First, the equilibrium dynamics of both scenarios are very general and unstructured. To compare the execution costs under the two dynamics, we have to rely on the limited information, such as $C_1^{\mathcal{T}}$ and x_0. Second, the constraint corresponding to the prohibition of trade-through is nonconvex. Finding the optimal solution of a nonconvex objective is usually theoretically intractable. To prove the result, we first strengthen the bound and use backward induction in i. The optimal solution under trade-through is relaxed to a suboptimal solution so that the trading processes under the two dynamics can be connected. To complete the inductive step, we prove a Lipschitz-like condition for the cost with respect to the starting market environment. The details are shown in the online appendix.

Although the optimal execution problem is high-dimensional and intractable, the upper bound is surprisingly simple as the upper bound doesn't involve the best ask prices or transient equilibrium price. It only depends on the total liquidity demand x_0, the minimal market depth $\underline{D}^{\mathcal{T}}$, and the resilience effect $C_1^{\mathcal{T}}$. Despite its simplicity, the upper bound turns out to be very effective in our empirical study: it reasonably estimates the order of magnitude of the trade-through benefit that can be compared with the total execution cost, and in most cases, the benefit is negligible.

Another feature of the upper bound is that it only depends on the equilibrium dynamics of "$\mathcal{T}$," that is, when trade-through is allowed. This makes the empirical evaluation of the bound simple because we cannot evaluate parameters in a counterfactual setting in practice.

The upper bound doesn't depend on the number of trading dates n, and it holds universally for all $n \geq 0$. In fact, with the same technique, we may derive a tighter n-dependent upper bound, which coincides with Online

Theorem EC.1 when $n = 0$. In other words, Theorem 1 is attained when $n = +\infty$. For simplicity, we only present the universal upper bound.

It is not surprising to see that the bound is quadratic rather than linear in the liquidity demand x. Indeed, recall that linearity corresponds to the classic price effect, whereas trade-through, closely related to market microstructure, produces higher order terms as a result of increasing marginal cost. In particular, block-shaped limit order books give rise to the second order approximation in which $\mathcal{D} \equiv \underline{D}$ and $\phi(\mathcal{D}, u) = u^2/2\underline{D}$.

The upper bound becomes meaningless for limit order books with some price gaps as $\underline{D}^{\mathcal{T}} = 0$. The gaps in prices might happen for microcap stocks but rarely for small-, middle-, and large-cap stocks. For microcap stocks, it is more common to trade a large position using hidden orders, dark pools, or iceberg orders. Hence, it is outside the scope of this paper.

3.3. A Matching Lower Bound

The quantity provided in Theorem 1 is an upper bound for the trade-through benefit. Could the trade-through benefit be significantly lower than the upper bound? In this section, we attempt to answer this question by providing a lower bound that can almost match the upper bound in Theorem 1 up to a constant factor.

Theorem 2. *For any $C_1^{\mathcal{T}} \in (0,1)$, $\underline{D}^{\mathcal{T}} > 0$ and $x_0 > 0$, there exists an instance satisfying Assumptions 2–7 such that* $J^{\mathcal{N}} - J^{\mathcal{T}} \geq \frac{(C_1^{\mathcal{T}})^2 x_0^2}{24\underline{D}^{\mathcal{T}}}$.

Pairing Theorem 2 with Theorem 1, we can see that the upper and lower bounds have the same form, which depends on $x_0^2/\underline{D}^{\mathcal{T}}$, except for the constants and the order of $C_1^{\mathcal{T}}$ (quadratic instead of linear in the lower bound).

Note that the lower bound doesn't contradict Online Theorem EC.1, which states that the benefit can be zero when $n = 0$. This is because n is arbitrary in the lower bound, and one can choose an adversarial case for the lower bound such that $n \neq 0$ in the construction.

3.4. Implications

According to the upper and lower bounds, there are three factors that affect the potential benefit of trade-through: the liquidity demand, the minimal market depth, and the resilience effect. In particular, if the liquidity demand is small ($x_0 \to 0$), the market is deep ($\underline{D}^{\mathcal{T}} \to \infty$), or the resilience of the order books is high ($C_1^{\mathcal{T}} \to 0$), then trade-through is not beneficial ($J^{\mathcal{N}} - J^{\mathcal{T}} \to 0$). Because the total execution cost can be translated to the effective spread, a measure of market quality that concerns practitioners and policy makers, these factors may provide practical and regulatory implications: if the factors are significant and, thus, entering the asymptotic regime, then there is virtually no benefit of trading through, and the execution quality, in terms of the total execution cost, does not deteriorate after the OPR was implemented.

1. Liquidity demand: For the initial liquidity demand x_0, the benefit of trade-through is at most $J^{\mathcal{N}} - J^{\mathcal{T}} = O(x_0^2)$, which scales quadratically in the liquidity demand. Note that the total execution costs $J^{\mathcal{T}} + M_0 x_0$ and $J^{\mathcal{N}} + M_0 x_0$ are linear in x_0. For small trades, the linear term dominates the higher order ones, and thus, the percentage benefit of trade-through is of the same order as x_0. Therefore, small retail investors are mostly concerned about the best ask prices rather than the market microstructure. The OPR enforced by Reg NMS, which governs the microstructure regime, does not have much influence on retail investors.

2. Market depth: For stocks with large market depth, all transactions are executed approximately at the best quote. The price impact is negligible so that whether to trade through doesn't make a big difference. In general, market microstructure becomes less important compared with the price effect as the market depth increases. Therefore, if the stock has abundant liquidity, the regulation doesn't significantly restrict investors; more precisely, the percentage benefit of trade-trough scales with the reciprocal of the market depth.

3. Resilience effect: With large $\rho_{i,j}$ and γ_i, the order books are able to rapidly recover from the market impact of past transactions. This is a common feature of liquid stocks as limit orders constantly flow in to fuel the mean reversion of the spread. For such stocks, current transactions hardly have any impact on future ones. Therefore, a dynamic order execution can be approximately decomposed into a series of one-period executions. As shown in Online Theorem EC.1, trade-through is not beneficial in this case as the gap $J^{\mathcal{N}} - J^{\mathcal{T}}$ diminishes as $C_1^{\mathcal{T}} \to 0$.

The salient role played by the resilience effect suggests that the exception rules regarding fast or slow markets should be extended to differentiate stocks with slow and fast resilience. Currently, fast markets in terms of execution speed cannot be traded through. Our results have two implications: (i) The definition of fast markets may also include liquid stocks with fast resilience; prohibiting trade-through does not incur high costs for the liquidity demanders trading those stocks. (ii) The exception might be granted for stocks with slow resilience even if they are traded in fast markets.

Another factor that indirectly affects the benefit of trade-through is trading frequency. For a fixed T, if the interval Δt is large, then the resilience effect is stronger because there are more limit orders filling in between consecutive trades of the investor. As a result, $C_1^{\mathcal{T}}$ is smaller. This implies that a high-frequency trader has a stronger motivation to trade through because trade-through is more profitable for the trader.

Our result also complements Foucault and Menkveld (2008) on the relationship between trade-through and

liquidity provision. According to Foucault and Menkveld (2008), protection against trade-throughs attracts liquidity providers, which, in turn, makes trade-through less beneficial for liquidity demanders, according to our finding. Such a virtuous cycle makes the prevention of trade-throughs favorable for both groups of market participants.

In summary, for small trades and stocks with abundant liquidity provision, large market depths, and rapid replenishment of liquidity, the benefit of trade-through and the change of effective spread is unsubstantial. Hence, the implementation of the OPR doesn't impose a significant restriction.

4. A Case Study on ASX Data

Now, we present a case study to calibrate the parameters and estimate the trade-through benefit. The data we use contain all order events for BHP in the ASX,[6] which consists of two periods, from September 1, 2009, to December 1, 2009, and from June 4, 2019, to August 19, 2019. We refer to the two data sets as samples 1 and 2, respectively. Besides ASX, BHP Billiton is also traded in Chi-X in Australia. The market integrity rule of the Australian Securities and Investments Commission ensures that the investors' instructions on how to execute an order must be followed by the brokers, including trade-throughs (Australian Securities and Investments Commission 2015). However, we are unable to obtain data from Chi-X. Hence, we have to rely on the parameters estimated from the ASX data to approximate the characteristics of Chi-X. The normal trading time of each date spans from 10:00 a.m. to 4:00 p.m. Following Lo and Hall (2015), we focus on the order events between 10:15 a.m. and 3:45 p.m. to remove effects resulting from the opening and closing of the market.

We use the variables in Table 1 from the data and the econometric approach detailed in Online Section EC.3 to obtain the summary statistics in Table 2. Based on the empirical results in Table 2 as reference points, we can derive the upper and lower bounds for the benefit of trading through for different levels of market depth and resilience rates.

More precisely, Theorem 1 implies that the absolute benefit of trade-through is bounded above by $C_1^{\mathcal{T}} x_0^2 / 2\underline{D}^{\mathcal{T}}$. To obtain the ranges for $\underline{D}^{\mathcal{T}}$ and x_0, we can use the quantiles of the time series for the market depth and average daily trading volume, respectively; the value of M_0 is determined by the average stock price. As for $C_1^{\mathcal{T}}$, in Online Section EC.3.2, we get the bounds (see (EC.21)):

$$(1/2)^{(\Delta t / t_{hl}^{min})} \le C_1^{\mathcal{T}} \le (1/2)^{(\Delta t / t_{hl}^{max})},$$

where Δt is the trading interval and t_{hl} is the half-life of price impact estimated using the approach detailed in Online Section EC.3.1.1. To convert the absolute benefit to the percentage benefit of trade-through, we need to divide the upper bound by the average trading cost, which can be approximated by $M_0 x_0$. Thus, using Theorem 1, an easy upper bound for the percentage trade-through benefit is

$$\overline{B} = \frac{C_1^{\mathcal{T}} x_0}{2\underline{D}^{\mathcal{T}} M_0} \le \frac{1}{2} \cdot \frac{1}{2^{\Delta t / t_{hl}^{max}}} \cdot \frac{x_0}{M_0 \underline{D}^{\mathcal{T}}}. \tag{9}$$

Similarly, using Theorem 2, an easy lower bound for the percentage trade-through benefit is

$$\underline{B} = \frac{(C_1^{\mathcal{T}})^2 x_0}{24\underline{D}^{\mathcal{T}} M_0} \ge \frac{1}{24} \cdot \frac{1}{2^{2\Delta t / t_{hl}^{min}}} \cdot \frac{x_0}{M_0 \underline{D}^{\mathcal{T}}}. \tag{10}$$

Here, we use t_{hl}^{max} and t_{hl}^{min} to denote the maximal and minimal half-life in the time series.

In Table 3, we calculate the bounds for the trade-through benefit for BHP according to the bounds in (9) and (10) by setting $M_0 = 38.4$ and $x_0 = 10^5$. The percentage trade-through benefit is less than 0.01% unless (i) the trading interval is less than 40 seconds, (ii) the minimum market depth is less than 300 shares per Australian dollar (AUD), and (iii) the half-life of the price impact is longer than 1.5 seconds. Therefore, the trade-through benefit is almost negligible for stocks with fast resilience (half-life less than one second) but may not be so for stocks with slow resilience.

The empirical study can provide a simple regulatory framework to make trade-through exemptions. In particular, the regulator can first estimate the range of half-life, for example, by using the econometric approach detailed in Online Section EC.3. The range of half-life, combined with other statistics, such as the average market depth and trading volume, can be used in (9) and (10) for the regulators to decide whether to exempt stocks from trade-through restrictions by checking whether the bounds for trade-through benefits exceed some predetermined threshold. For example, a more lenient rule toward investors would grant trade-through exemptions if the upper bound in (9) is bigger than 0.01%, that is,

$$\frac{1}{2} \cdot \frac{1}{2^{\Delta t / t_{hl}^{max}}} \cdot \frac{x_0}{M_0 \underline{D}^{\mathcal{T}}} \ge 0.0001.$$

On the other hand, a more restrictive rule may only grant trade-through exemptions if the lower bound in (10) is bigger than 0.01%, that is,

$$\frac{1}{24} \cdot \frac{1}{2^{2\Delta t / t_{hl}^{min}}} \cdot \frac{x_0}{M_0 \underline{D}^{\mathcal{T}}} \ge 0.0001.$$

Table 1. The Required Data for the Econometric Model

Symbol	Definition
s_t	The bid–ask spread
Δq_t	The change of midquote compared with the previous order book event
x_t^b	Buy order dummy variable
x_t^s	Sell order dummy variable
$v_t^{a,i}$	Log depth at the ith best ask price (thousands)
$v_t^{b,i}$	Log depth at the ith best bid price (thousands)

Table 2. The Estimated Range of Parameters

Parameters Unit	$\underline{D}^{\mathcal{T}}$ Shares/AUD	t_{hl} Seconds	M_0 AUD	x_0 Shares
Sample 1: September 1 to December 1, 2009	$[3\times10^2, 5\times10^3]$	[1.56, 5.00]	38.4	7×10^5
Sample 2: June 4 to August 19, 2019	$[1.5\times10^3, 10^4]$	[0.13, 2.43]	39.8	4×10^4

Notes. We use the 1%–5% quantile of the market depth time series to obtain the range for $\underline{D}^{\mathcal{T}}$. The range for the half-life t_{hl} is estimated in Online Section EC.3.1.1. The values of M_0 and x_0 are determined by the average stock price and one 10th of the average daily trading volume, respectively.

5. Model Extensions

In this section, we study two important extensions of the base model: trading latency and competition of multiple liquidity demanders.

5.1. Latency

One important factor for liquidity demanders is the delay or trading latency (Moallemi and Sağlam 2013). More precisely, after the investor observes the state of the limit order books and submits a market order, the actual execution of the order occurs after a delay at which point the limit order books may have changed. Consider a single investor trading on two dates in m venues. The market depth and resilience are general and follow the setup in Section 2. To capture the latency, we assume that the investor determines the market orders to submit at t_1 before observing the resilience between t_0 and t_1 and the market depth at t_1. Specifically, the following sequence of events happen: (i) when trade-through is prohibited, the investor decides the size of two market orders at t_0, which are the total trading volumes at t_0 and t_1, respectively. The orders are then split among the m venues on the two dates such that there is no trade-through. (ii) With trade-through, the investor decides $2m$ market order sizes at t_0, which are to be traded on the m venues and two dates, respectively. Note that all the order sizes are determined before the resilience, such as $\gamma_{0,j}$ and ρ_0, between the two dates realizes, whereas the market orders at t_1 are executed after it realizes. The next result implies that the upper bound in Theorem 1 can be applied to this setting as well.

Proposition 1. *The benefit of trade-through under latency is less than that without latency under the same parameter setting as specified earlier.*

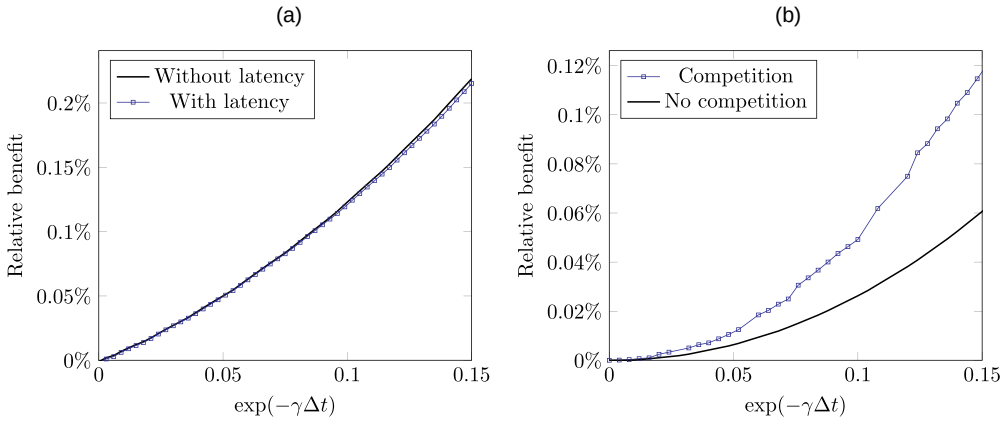
To demonstrate the magnitude of trade-through benefit, we consider the following example with two venues. The market depth is deterministic: $\mathcal{D}_{0,1} = \mathcal{D}_{1,2} \equiv q$ and $\mathcal{D}_{0,2} = \mathcal{D}_{1,1} \equiv kq$. At time 0, $s_0 = d_{0,1} = d_{0,2} = 0$. Moreover, we set $\gamma_0 = \rho_{0,1} = \rho_{1,1}$, where γ_0 is a random variable such that $\gamma_0 = (1/2)\gamma$ or γ with equal probabilities (we vary γ in the experiments). We set $k = 3$ and $q = 4{,}000$. The investor has initial liquidity demand $x_0 = 10^5$, and the initial fundamental value is $M_0 = 38.4$. To give a comparison, we also calculate the benefit of trade-through in the base model when there is no latency. The optimal trading strategies are calculated using the Bellman equation and the backward induction after discretizing the state space. Figure 4(a) illustrates the difference in the costs when trade-through is allowed or prohibited.

Table 3. The Upper and Lower Bounds ((9) and (10)) of the Percentage Benefit of Trade-Through for Different Half-Life (Seconds), Trading Interval Δt ($1/\Delta t$ Trades per Second), and Minimum Market Depth $\underline{D}^{\mathcal{T}}$ (Shares per AUD)

t_{hl}	Δt	$\underline{D}^{\mathcal{T}} = 3\times10^2$, %	$\underline{D}^{\mathcal{T}} = 10^3$, %	$\underline{D}^{\mathcal{T}} = 3\times10^3$, %	$\underline{D}^{\mathcal{T}} = 10^4$, %
			Slow resilience		
6.9	20	[0.66, 58.74]	[0.20, 17.62]	[0.07, 5.87]	[0.02, 1.76]
	40	[0.01, 7.95]	[0.00, 2.38]	[0.00, 0.79]	[0.0, 0.24]
	60	[0.00, 1.08]	[0.00, 0.32]	[0.00, 0.11]	[0.00, 0.03]
3.5	20	[0.01, 7.95]	[0.00, 2.38]	[0.00, 0.79]	[0.00, 0.24]
	40	[0.00, 0.15]	[0.00, 0.04]	[0.00, 0.01]	[0.00, 0.00]
	60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
2.3	20	[0.00, 1.08]	[0.00, 0.32]	[0.00, 0.11]	[0.00, 0.03]
	40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
1.5	20	[0.00, 0.03]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
	40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
			Fast resilience		
1.3	20 or 40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
0.9	20 or 40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
0.7	20 or 40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
0.6	20 or 40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]
0.5	20 or 40 or 60	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]	[0.00, 0.00]

Note. The upper (lower) panel includes the scenarios with slow (fast) resilience (the percentage trade-through benefit more (less) than 0.01%).

Figure 4. (Color online) The Benefit of Trade-Through in the Presence of Trading Latency or Multiple Liquidity Demanders



Notes. (a) Trading latency. (b) Multiple liquidity demanders.

The trading latency seems to have almost no impact on the benefit of trade-through for a reasonable range of parameters. Note that the choice and range of parameters in this example are consistent with the estimates in the empirical study in Section 4.

5.2. Multiple Liquidity Demanders

We extend the model to multiple liquidity demanders and investigate the effect of competition on the benefit of trade-through. Consider two investors (say a and b) and two venues ($m = 2$) on two trading dates ($n = 1$). The market depth is deterministic: $\mathcal{D}_{0,1} = \mathcal{D}_{1,2} \equiv q$ and $\mathcal{D}_{0,2} = \mathcal{D}_{1,1} \equiv kq$ for $k > 1$. At t_0, we let $s_0 = d_{0,1} = d_{0,2} = 0$. Moreover, we set $\gamma_0 = \rho_{0,1} = \rho_{1,1} \equiv \gamma$. Both investors have initial liquidity demand x_0. The initial fundamental value is M_0. We let investors trade simultaneously on each date and share the costs. Because it is hard to define trade-through when orders of both investors are mixed, we only investigate symmetric Nash equilibria. That is, both investors trade the same number of orders as an equilibrium strategy, and trade-through is defined analogously after the aggregate order.

More precisely, the symmetric Nash equilibria with and without trade-through are defined as follows: Without trade-through, on each date, the market receives the orders from the investors and automatically splits the orders into different venues such that the aggregate order is executed at the best possible price among all trading venues. Suppose an aggregate order of size U_i is submitted to market at t_i. Let $J_0(U_0)$ be the aggregate trading cost incurred at t_0 and let $J_1(U_0, U_1)$ be the total trading cost incurred at t_1. When investor $I \in \{a, b\}$ submits a market order of size u_i at t_i, the total trading cost of investor I at t_0 is $\frac{u_0^I}{u_0^a+u_0^b} J_0(u_0^a + u_0^b)$, and the cost at t_1 is $\frac{u_1^I}{u_1^a+u_1^b} J_1(u_0^a + u_0^b, u_1^a + u_1^b)$; that is, we allocate the cost pro rata among the two investors. As each investor demands x_0 units, both submit a market order of size u_0^* (respectively, $x_0 - u_0^*$) at t_0 (t_1), at which the symmetric Nash equilibrium is defined as $u_0^* = \arg\min_{u_0 \le x_0} \left\{ \frac{u_0}{u_0+u_0^*} J_0(u_0 + u_0^*) + \frac{x_0-u_0}{2x_0-u_0-u_0^*} J_1(u_0 + u_0^*, 2x_0 - u_0 - u_0^*) \right\}$.

With trade-through, each investor needs to determine how many to trade on each venue. The symmetric Nash equilibria is defined similarly: Suppose an aggregate market order of size U_{ij} is submitted to venue j at t_i. Let $J_{0j}(U_{01}, U_{02})$ be the aggregate trading cost incurred on venue j at t_0 and $J_{1j}(U_{01}, U_{02}, U_{11}, U_{12})$ the total trading cost on venue j and t_1. When investor $I \in \{a, b\}$ submits a market order of size u_{ij} to venue j at t_i, the trading cost is $\frac{u_{0j}^I}{u_{0j}^a+u_{0j}^b} J_{0j}(u_{01}^a + u_{01}^b, u_{02}^a + u_{02}^b)$, and the cost on venue j at t_1 is $\frac{u_{1j}^I}{u_{1j}^a+u_{1j}^b} J_{1j}(u_{01}^a + u_{01}^b, u_{02}^a + u_{02}^b, u_{11}^a + u_{11}^b, u_{12}^a + u_{12}^b)$. We can define the symmetric Nash equilibria similarly.

We first provide a result for the symmetric equilibrium strategy when trade-through is prohibited.

Proposition 2. *In the symmetric Nash equilibrium, when trade-through is prohibited, both investors trade* $\min\left\{\frac{x_0}{3(1-e^{-\gamma\Delta t})}, x_0\right\}$ *shares at* t_0 *and the remaining at* t_1.

When trade-through is allowed, the equilibrium strategies are not available analytically, and we have to resort to numerical solutions. In particular, we let $q = 4{,}000$, $k = 3$, $x_0 = 10^5$, $M_0 = 38.4$ similar to the calibration of the empirical study and vary the value of γ. To solve the Nash equilibrium, we discretize the state space (remaining shares) of both investors and the strategy space. The value functions can then be solved using backward induction. To give a comparison, we also calculate the benefit of trade-through when there is only one investor as analyzed in our main model. Figure 4(b) illustrates the difference in the costs when trade-through is allowed or prohibited of both investors in competition. The benefit

of trade-through seems to be comparable to that in the case of a single investor. More specifically, the trade-through benefit is less than 0.12% for a reasonable range of parameters.

Acknowledgments

The authors thank John Birge (editor-in-chief), the associate editor, and the reviewers for their constructive comments that considerably improved the paper.

Endnotes

[1] As shown in Section 2.3, the total execution cost is equivalent to the average effective half-spread, a common empirical measure for market quality.

[2] The time frame we are considering, T, is in the order of days, and the period between trades $t_{i+1} - t_i$ is of the order of seconds to hours. See Section 4 for a detailed empirical study. For the commission fees, because we consider a fixed number of shares x_0 and a fixed number of trades $n + 1$ (unless the investor chooses not to trade at some t_i), our conclusion doesn't change for a fee structure that is affine in the order size. That is, the commission fee at t_i is $a + (\sum_{j=1}^{m} u_{i,j})b$. In practice, many brokerage firms charge no commission fee at all for stock trades and make money via bid–ask spreads and advertisement. We acknowledge that the asymmetric nonlinear commission fee structures could be a factor that drives trade-throughs in practice; see Foucault et al. (2013) for a study on this topic.

[3] We do not consider the submission of limit orders from this investor because of two reasons: First, as we list in the introduction, so far, the only argument against the OPR is that some investors may have traded through intentionally. This is why we are interested in market orders primarily. It is only for market orders that intentional trade-through becomes a viable option, whereas limit orders can only be traded through passively. Second, from the modeling side, incorporating the submission of limit orders requires more information on the equilibrium dynamics of the limit order book, for example, the filling probability of limit orders at each price and how it interacts with the market beliefs. More modeling assumptions tend to make the claim less robust.

[4] A function $f(d_1, \ldots, d_{m+1})$ is called increasing if $d_j \geq d'_j$ for all j implies that $f(d_1, \ldots, d_{m+1}) \geq f(d'_1, \ldots, d'_{m+1})$.

[5] A functional $f(\cdot)$ is called increasing if $\mathcal{D}_j(p) \geq \mathcal{D}'_j(p)$ for all j and $p \geq 0$ implies that $f(\mathcal{D}_1, \ldots, \mathcal{D}_m) \geq f(\mathcal{D}'_1, \ldots, \mathcal{D}'_m)$.

[6] We acquire two types of data from ASX: one records real-time trading events, and the other records real-time five-level order book depth. We use the first to determine whether the update of the limit order book in the second is due to trading events or order cancellations.

References

Ahn HJ, Bae KH, Chan K (2001) Limit orders, depth, and volatility: Evidence from the stock exchange of Hong Kong. *J. Finance* 56(2):767–788.

Australian Securities and Investments Commission (2015) RG 223 guidance on ASIC market integrity rules for competition in exchange markets. RG 223.82. Accessed April 9, 2018, http://download.asic.gov.au/media/3225864/rg223-published-4-may-2015.pdf.

Battalio R, Hatch B, Jennings R (2004) Toward a national market system for US exchange–listed equity options. *J. Finance* 59(2): 933–962.

Bessembinder H (2003) Quote-based competition and trade execution costs in NYSE-listed stocks. *J. Financial Econom.* 70(3): 385–422.

Biais B, Hillion P, Spatt C (1995) An empirical analysis of the limit order book and the order flow in the Paris Bourse. *J. Finance* 50(5):1655–1689.

Blume ME (2007) Competition and fragmentation in the equity markets: The effect of Regulation NMS. Working paper, University of Pennsylvania, Philadelphia.

Chen N, Kou S, Wang C (2018) A partitioning algorithm for Markov decision processes with applications to market microstructure. *Management Sci.* 64(2):784–803.

Foucault T, Menkveld AJ (2008) Competition for order flow and smart order routing systems. *J. Finance* 63(1):119–158.

Foucault T, Kadan O, Kandel E (2005) Limit order book as a market for liquidity. *Rev. Financial Stud.* 18(4):1171–1217.

Foucault T, Kadan O, Kandel E (2013) Liquidity cycles and make/take fees in electronic markets. *J. Finance* 68(1):299–341.

Goettler RL, Parlour CA, Rajan U (2005) Equilibrium in a dynamic limit order market. *J. Finance* 60(5):2149–2192.

Hendershott T, Jones CM (2005) Trade-through prohibitions and market quality. *J. Financial Markets* 8(1):1–23.

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, Cambridge, UK).

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

Lo DK, Hall AD (2015) Resiliency of the limit order book. *J. Econom. Dynamics Control* 61:222–244.

Moallemi CC, Sağlam M (2013) OR forum—The cost of latency in high-frequency trading. *Oper. Res.* 61(5):1070–1086.

Obizhaeva AA, Wang J (2013) Optimal trading strategy and supply/demand dynamics. *J. Financial Markets* 16(1):1–32.

O'Hara M (2004) Searching for a new center: US securities markets in transition. *Econom. Rev.* 89(4):37–52.

O'Hara M, Ye M (2011) Is market fragmentation harming market quality? *J. Financial Econom.* 100(3):459–474.

Predoiu S, Shaikhet G, Shreve S (2011) Optimal execution in a general one-sided limit-order book. *SIAM J. Financial Math.* 2(1): 183–212.

Roşu I (2009) A dynamic model of the limit order book. *Rev. Financial Stud.* 22(11):4601–4641.

Rubin D (1976) Inference and missing data. *Biometrika* 63(3):581–592.

Securities and Exchange Commission (2005). Regulation NMS. Release No. 34-51808, Washington, DC.

Tsoukalas G, Wang J, Giesecke K (2019) Dynamic portfolio execution. *Management Sci.* 65(5):2015–2040.

Ningyuan Chen is an assistant professor in the department of management at the University of Toronto, Mississauga. He is also cross-appointed at the Rotman School of Management, University of Toronto.

Pin Gao is an assistant professor at the school of data science, the Chinese University of Hong Kong, Shenzhen. His research interests include revenue management, innovation, algorithm design, and mechanism design.

Steven Kou is a Questrom professor in management and professor of finance at Boston University. He teaches courses on FinTech and quantitative finance. He is a fellow of the Institute of Mathematical Statistics and won the Erlang Prize from INFORMS in 2002. Some of his research results have been incorporated into standard MBA textbooks and are implemented in commercial software packages and terminals, for example, in Bloomberg Terminals.