Journal of Financial Economics 140 (2021) 270–291

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

What is the impact of introducing a parallel OTC market? Theory and evidence from the chinese interbank FX market☆

Craig W. Holden[a], Dong Lu[b,c,*], Volodymyr Lugovskyy[d], Daniela Puzzello[d]

[a] Kelly School of Business, Indiana University, 1309 E 10th St, Bloomington, IN 47405, United States
[b] School of Finance, Renmin University of China, Zhongguan Cun Road No.59, Haidian District, Beijing 100872, China
[c] China Financial Policy Research Center, Renmin University of China
[d] Department of Economics, Indiana University, 100 S Woodlawn Ave, Bloomington, IN 47405, United States

ARTICLE INFO

Article history:
Received 11 March 2019
Revised 23 December 2019
Accepted 20 January 2020
Available online 17 October 2020

JEL classification:
D83
G10

Keywords:
Market structure
Over-the-counter
Limit-order book
FX market

ABSTRACT

Chinese interbank foreign exchange trading was originally conducted through a centralized, anonymous limit order book (LOB). We determine the impact of the introduction of a parallel decentralized over-the-counter (OTC) market. We find that: (1) most trading migrated to the OTC, (2) the LOB price function is upward-sloping versus the OTC price function is downward-sloping, and (3) the LOB market has a single price function versus the OTC market has multiple price functions. Next, we develop a theoretical model of parallel markets that can simultaneously explain all of these empirical findings. We test a new model prediction and find support.

1. Introduction

Three widely used trading mechanisms for continuous securities trading are: *dealer* markets where dealers post prices, *limit order book (LOB)* markets where limit orders post prices, and *hybrid* markets where both dealers and limit orders post prices (Madhavan, 1992). Dealer markets may feature centralized bid-ask quotations and order routing (e.g., pre-1997 NASDAQ, pre-1997 London Stock Exchange). Alternatively, *over-the-counter (OTC)* markets are dealer markets that are decentralized, where customers search for dealers and then engage in bilateral negotiations with them. Traditionally, OTC dealers were queried by phone and by voice. In recent years, new electronic platforms have emerged (e.g., MarketAxess), which allow customers to query a large number of OTC dealers and then electronically complete the trade (Hendershott and Madhavan, 2015).

In practice, we observe that LOB and hybrid mechanisms dominate trading in stocks and options, whereas the OTC mechanism dominates trading in foreign exchanges, bonds, spot commodities, and nonstandard derivatives (see, Duffie et al., 2005; 2007; Duffie, 2012). Why is this so? One hypothesis about trading mechanisms is that the LOB or hybrid mechanism should predominate, because

☆ We thank the CFETS for providing the data. We thank Jennifer Conrad, Darrell Duffie, Nandini Gupta, Andriy Shkilko, Noah Stoffman, Chuck Trzcinka, Liyan Yang, Shengxing Zhang, Wenyu Zhu and seminar participants at the SFS Cavalcade North America and Indiana University for useful comments and an anonymous referee for very detailed and constructive suggestions. We are solely responsible for any errors. This research is supported by the National Natural Science Foundation of China (grant no. 71903191).

* Corresponding author at: School of Finance, Renmin University of China, Zhongguan Cun Road No.59, Haidian District, Beijing 100872, China.
E-mail addresses: cholden@indiana.edu (C.W. Holden), donglu@ruc.edu.cn (D. Lu), vlugovsk@iu.edu (V. Lugovskyy), dpuzzell@indiana.edu (D. Puzzello).

https://doi.org/10.1016/j.jfineco.2020.10.004
0304-405X/

theoretically the transaction costs of an LOB or hybrid should be much less than a dealer market, and much evidence supports this (see Glosten, 1994; Jain, 2005b; Abudy and Wohl, 2017). A second hypothesis is that the OTC should predominate when institutional traders are more prevalent, because they can use their larger average trade size to bargain for a better price in an OTC mechanism and have enough market power to get what they want (see Biais and Green, 2019). It is difficult to test these two hypotheses because we rarely see the introduction of a parallel trading mechanism to an existing market.[1]

We examine such a rare event. Specifically, we examine the Chinese interbank foreign exchange (FX) market. In this market, banks trade Chinese currency, called Chinese Yuan (CNY) or Renminbi (RMB),[2] for the US dollar (USD) and other major currencies. Prior to 2006, 100% of the trading in this market was done on an anonymous, centralized LOB. At the start of 2006, a parallel OTC market was introduced (see Fig. 1). The OTC was introduced to help accommodate rapidly growing volume (i.e., CNY FX volume had more than tripled between 1994 and 2005) and was designed to make the market structure more similar to international FX markets.[3] In a decentralized OTC market, both customers and dealers know who they are trading with (i.e., it is not anonymous). Fig. 1 illustrates that after the parallel OTC introduction, each bank could trade in either market and switch back and forth at will. Both parallel markets have continued to coexist all the way to the present. We have intraday trade data from before and after the introduction (August 2005 to December 2006) that allow us to determine the impact of this event.

Empirically, we examine three key questions about this natural experiment. First, after the parallel introduction, which trading mechanism predominates? Second, do the parallel LOB and OTC markets have upward- or downward-sloping price functions (i.e., prices relative to trade size)? One hypothesis is that both markets will have downward-sloping functions because FX is one of the asset classes that tends to have downward-sloping price functions (Edwards et al., 2007). An alternative hypothesis is that the LOB will be upward-sloping and the OTC will be downward-sloping, because price functions are primarily determined by the trading mechanism and these are the most common patterns for these trading mechanisms.

Our third key question is: Do the LOB and OTC markets have a single price function or multiple price functions (i.e., do different trading clienteles face the same or different price functions)? One hypothesis is that *anonymous* trading systems (e.g., LOB) must have a single price function, because there is no basis for distinguishing different orders, but that *non–anonymous* systems (e.g., OTC) lead to multiple price functions, because more powerful clienteles can bargain for better prices than less powerful clienteles. An alternative hypothesis is that competition between dealers is so strong that they compete to a single price function.

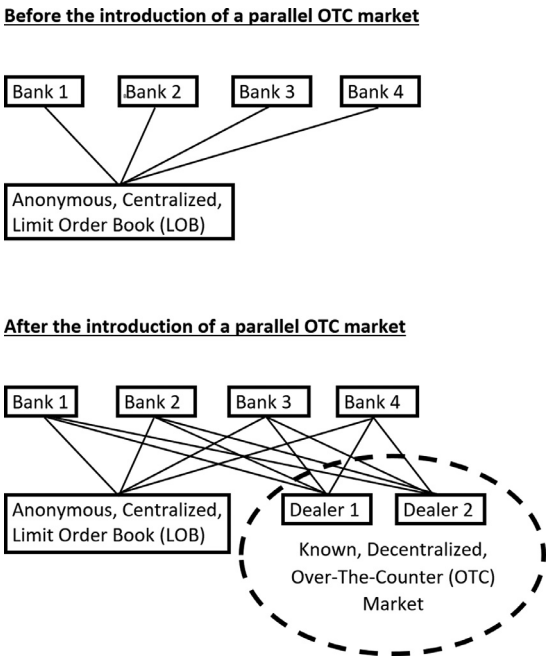

Fig. 1. Trading mechanisms in the Chinese Interbank FX market. *Note:* The labels "Bank 1", "Bank 2", etc. are meant to be inclusive of all market participants. The large majority of market participants are banks, but a small minority are nonbanks.

We find that the vast majority of trading migrates from the LOB to the OTC in a very short amount of time (less than six months). To determine how and why, we examine transaction costs in both venues. We find that OTC transaction costs are lower than LOB transaction costs for medium and large trades, but the reverse is true for small trades. This seems like the key reason for the migration of medium and large trades to the OTC. We also find that the transaction costs in LOB increased upon the introduction of the parallel OTC, which supports the idea that liquidity suppliers moved to the OTC, as well as liquidity demanders. These results support the hypothesis that the OTC would dominate due to powerful institutional traders getting their way and reject the hypothesis that the LOB would dominate due to its inherently superior efficiency.

Turning to the slope of the price functions, we find that the LOB is upward-sloping and the OTC is downward-sloping. These findings reject the hypotheses that the asset class is the primary determinant of the price function and support the hypothesis that the trading mechanism is the primary determinant of the price function slope (i.e., LOBs are upward-sloping and OTC markets are downward-sloping).

Turning to the number of price functions, we find that the LOB has a single price function, while the OTC has

[1] Jain (2005a) documents the global shift from floor trading to electronic trading in stock markets around the world. In most cases, an electronic system is introduced in parallel to the existing trading floor, and then eventually the trading floor is closed. However, this is a different dimension of trading than what we address in this paper.

[2] In this paper, we will use CNY and RMB interchangeably to denote the Chinese currency.

[3] The following section provides a longer, more elaborate explanation for why the OTC was introduced. For the description and surveys of the international FX market, see King et al. (2013) and the Triennial Survey of FX and OTC derivatives trading by the Bank for International Settlements (BIS).

multiple price functions—better prices for large banks, intermediate prices for medium banks, and worse prices for small banks.[4] If bargaining power increases in the bank size, then this finding supports the notion that bargaining power serves as the basis for different price functions. These results support the hypotheses that the anonymous versus non-anonymous character of the trading system determines the number of price functions and reject the idea that competition among OTC dealers is sufficient to force a single price function. These results are consistent with the Duffie et al. (2005) model, which is also supported by previous empirical findings.[5]

Next, we turn to our key theoretical question. Given our empirical results, how can we theoretically explain the simultaneous existence of an upward-sloping price function in the LOB market and multiple downward-sloping price functions in the OTC market? It is a difficult theoretical challenge to formulate the basis for having two parallel markets exhibit such different properties when simultaneously trading the same asset (i.e., trading the same currency pair). We develop a theoretical model of parallel LOB and OTC markets. The LOB market is based on adverse selection in a market with anonymous orders (Kyle, 1985; Glosten and Milgrom, 1985; Seppi, 1997). The OTC part is based on the search model of Vayanos and Wang (2011) with known trading counterparties who bargain over price. We include multiple classes of agents who can endogenously choose between the two venues. Importantly, we are able to show sufficient conditions for a theoretical model to exhibit the key empirically-relevant properties (i.e., simultaneous existence of an upward-sloping price function in the LOB market and multiple downward-sloping price functions in the OTC).

Here is the structure of the model. First, in the OTC you know who you are trading with. Thus, multiperiod reputational considerations imply that informed traders will not be able to repeatedly burn the same OTC dealer, because the dealer will refuse to trade with them. Therefore, the only people who can trade in the OTC are uninformed traders. Second, in an LOB market, all trading is anonymous. Thus, informed traders *can* exploit their information in an LOB market that contains a mix of informed and uninformed traders. The liquidity suppliers in an LOB market can afford to lose money to informed traders as long as they gain money from the uninformed traders and at least break even. The adverse selection in an LOB market yields an upward-sloping price function, because informed trade size is increasing in the extremity of their information. This is analogous to the upward-sloping price function in the Kyle (1985) model for a dealer market.

Third, if all that was going on in the OTC market was uninformed traders, then you would expect a flat price function. We add a fixed cost per trade and thus obtain a downward-sloping price function. Fourth, in the OTC, the liquidity demander and the dealer bargain with each other over the price. Assuming that different trader clienteles have different degrees of bargaining power, we obtain multiple, downward-sloping price functions. We then put the LOB and OTC markets together in parallel and split the uninformed traders into one nondiscretionary class and multiple discretionary classes, where the latter classes can choose to trade in either market. Finally, we calibrate a numerical example of the theoretical model and obtain price functions for the LOB and OTC markets that are qualitatively similar to the empirical price functions for both markets.

We solve for the optimal trade size, at which point a given discretionary trader class switches from trading on the LOB to the OTC. We show that it is optimal for trader classes with more bargaining power to switch to the OTC at smaller trade sizes. Finally, we empirically test this new prediction. We find that in the first six months after the introduction of parallel markets (Jan-June 2006), the switching point for large, medium, and small banks are $3 million, $10 million, and $20 million, respectively. In the second six months after the introduction of parallel markets (July-Dec 2006), the switching points for large, medium, and small banks becomes $0.5 million, $5 million, and $10 million, respectively. So in both time periods, banks with more bargaining power switch to the OTC at smaller trade sizes, which confirms our model's new prediction.

Our paper is related to a rapidly growing literature in OTC markets. Theoretical papers, including Duffie et al. (2005, 2007), Lagos and Rocheteau (2009), Vayanos and Wang (2011), Duffie (2012), Atkeson et al. (2015), Geromichalos and Herrenbrueck (2016), Mattesini and Nosal (2016), Trejos and Wright (2016), and Lagos and Zhang (2020), develop search-based models in which investors search among decentralized OTC dealers who intermediate trades (see also Weill, 2020, for a comprehensive literature review of search models of OTC markets). Empirical papers, including Mende et al. (2004), Harris and Piwowar (2006), Goldstein et al. (2007), Bessembinder and Maxwell (2008), Hendershott and Madhavan (2015), and Biais and Green (2019), examine the cost of trading in various OTC markets, including corporate bonds, municipal bonds, and foreign exchange. None of these papers examine parallel LOB and OTC trading mechanisms in a customer-dealer market.

Our paper is also related to a literature on the FX market. King et al. (2013) highlights that the FX market has a two-tier structure in which customers trade with dealers in the first tier and dealers trade with other dealers (i.e., "interdealer market") in the second tier. Our paper focuses exclusively on the first tier, customer-dealer market. Ding (2009) and Osler et al. (2011) provide evidences about the shape of the price function in the FX customer-dealer market based on a single OTC market (not parallel LOB and OTC trading mechanisms in the customer-dealer

[4] Since the large majority of the participants in China's interbank FX market are banks, we simplify the exposition by using the word *bank* for all participants. We distinguish *small, medium,* and *large banks* according to the institutions' importance and influence. Specifically, the China Foreign Exchange Trading System (CFETS), the organizer of the Chinese interbank FX market, categorizes banks into three groups, which it calls *core, fundamental,* and *ordinary* banks. We use their classification system but simplify the exposition by relabeling core banks as large banks, fundamental banks as medium banks, and ordinary banks as small banks. See more explanations and institutional details in Section 2.

[5] For example, Schultz (2001) examines the OTC market for corporate bonds and finds that the bid-ask spread is smaller for larger institutional investors.

market). Both find a downward-sloping price function in the OTC market, and both studies have relatively small samples. Ding (2009) studies one dealer in one currency pair for seven days in 2004 with a sample size of 970 trades. Osler et al. (2011) studies one dealer at one bank in one currency pair for four months in 2001. Their customer-dealer sample is 1,690 trades, and their interdealer sample is 1,919 trades. Our sample comprehensively spans the entire Chinese customer-dealer FX market over 17 months with a sample size of 87,407 trades, which is 24 times larger.

Lee and Wang (2018) is closest to our theoretical model. They develop a model in which investors can choose to trade either on a centralized dealer exchange or a decentralized OTC market. In their model, OTC dealers imperfectly cream skim the more-likely-to-be-uninformed traders and thus are able to offer a narrower bid-ask spread in the OTC market. Their model is limited to unit-size trades by price-takers and thus is silent about the slope of the price function in either market and silent about different bargaining power price functions.

Biais and Green (2019) is the closest to our empirical analysis. They examine trading in US corporate bonds and find that the large majority of trading volume shifted from the centralized NYSE LOB to the decentralized OTC dealers in the 1940s. They argue that the 1940s transition happened due to the rise of institutional traders in corporate bonds who felt they could get better prices in the OTC dealer market (i.e., the growth of institutional traders tipped the "center of gravity" to the OTC).

The rest of the paper is as follows. Section 2 provides an overview of the relevant institutional background and data sources. Section 3 develops the hypotheses. Section 4 provides the empirical results. Section 5 develops a theory that can generate the transaction cost patterns that were found empirically and provides an empirical test of a new testable prediction from the model. Section 6 concludes.

2. Institutions and data

FX markets play a critical role in supporting international trade, investment, and traders' liquidity needs. Traditionally, currencies are traded bilaterally in decentralized OTC markets, where a given dealer trades with a customer-buyer one moment and with a customer-seller another moment. The major customers are corporations engaged in international trade, leveraged and unleveraged asset managers, local and regional banks, and central banks. Since the emergence of electronic trading networks in the late 1990s, most of the interdealer trading is now carried out via limit order markets run by the electronic brokers, EBS and Thomson Reuters (see King et al., 2012; Evans and Rime, 2019, for a detailed documentation of FX market microstruture). FX markets are by far the largest financial markets in terms of trade volume.[6]

In China's interbank FX market, US dollars and other foreign currencies are traded against the CNY to help financial institutions and other market participants adjust their foreign currency positions to meet liquidity demand, clear financial transactions, to meet regulatory requirements or to speculate on FX rate fluctuations.[7] The main participants are a variety of financial institutions, including big nationwide commercial banks, local subsidiaries of foreign banks, medium/small commercial banks, rural credit cooperatives, and corporations engaged in international trade. While the Chinese FX market shares many common characteristics with international FX markets, it is also characterized by some unique features. For instance, the Chinese FX market was not originally organized as a decentralized OTC market, as was the case for most international FX markets. Instead, China established a nationwide centralized LOB for FX trading. This system was established in 1994 and is operated under the China Foreign Exchange Trading System (CFETS). Like most electronic LOB markets, every order is anonymously exposed to every other order with a centralized order-crossing algorithm. The CFETS runs on a membership system. Members issue orders through a trading terminal, and all orders are matched via the electronic trade matching system based on price priority first and then time priority. The CFETS is also responsible for the clearing of funds in foreign and domestic currencies. This function is called the Central Counterparty Clearing House for the Chinese FX market.

At the beginning of 2006, the parallel OTC market was introduced. In broad terms, it was introduced to help accommodate rapidly growing volume and to shift the market structure in the direction of international FX markets, which were held up as a desirable market design. More specifically, an early official assessment by the central bank of China, People's Bank of China (PBOC), of the possibility of introducing an OTC noted that "OTC approach is a basic practice in the international FX market and account for most of the trading volume in spot transactions globally."[8]

To implement the reform in a proactive, controllable, and gradual way, PBOC parallely introduced OTC trade mechanism as of January 4, 2006 as a way to encourage full use of the comparative advantages of the OTC and the LOB. The official public announcement from the PBOC also stated that the OTC market was introduced "to enhance the market liquidity and promote the price discovery role of the FX markets."[9] Importantly, the parallel structure mitigated the risk in introducing the OTC because its use was strictly optional. In other words, the LOB market remained open, so traders could continue to use it if they wished. In the OTC market, there are two types of market participants: customers (commercial banks and other

[6] According to the BIS's Triennial Central Bank Survey of FX and OTC derivative markets in 2019, the global trading volume in FX markets averaged $6.6 trillion per day in April 2019.

[7] The CNY was traded against the US Dollar (USD/CNY), Japanese Yen (JPY/CNY), Euro (EUR/CNY), British Pound (GBP/CNY), and Hong Kong Dollar (HKD/CNY) in the spot FX market, with USD/CNY accounting for a dominant share of more than 90%.

[8] See "Full text: PBOC on interbank spot foreign exchange market" on January 8, 2006 for details. For recent studies on China's FX market and FX policies, see Lei et al. (2020) and Lu et al. (2020).

[9] PBOC, 2006. Announcement on the introduction of OTC trading to China's interbank FX spot market.

institutions) and dealers. A dealer offers two-sided quotes, and then bilateral trades are agreed and executed by transferring currencies through the CFETS. Customer-dealer trades account for around 60% of the total trade volume, while the remaining 40% consist of interdealer trades.[10]

This rare natural experiment allows us to explore many questions about trading mechanisms. Our data are composed of the transactions of the Chinese interbank FX market, as recorded by the CFETS from August 2005 to December 2006. Thus, we have five months of data prior to the introduction of the parallel OTC market and twelve months of data after its introduction. We focus on the USD/CNY spot trades, which is by far the dominant currency pair traded in the Chinese interbank FX market. The data are comprehensive, and each record contains the party that initiates the trade, indicating whether this transaction is buyer-initiated, seller-initiated, or is an inter-dealer trade. They also contain information on the time of transaction, trade size, currency pair, and spot exchange rate. While we do not observe each institution's name, a unique ID is assigned to each market participant. The dataset provides the CFETS's classification into three groups: (1) *core* banks that are systematically important market participants with large asset scale, large influence on the RMB market and excellent internal risk management; (2) *fundamental* banks that have some influence in the RMB market and good internal risk management; and (3) *ordinary* banks that are all remaining market participants.

A subtle difference between the two parallel markets is how credit risk is handled. On the LOB, the CFETS guarantees all trades so that the trading parties bear zero credit risk. It serves the same role as a clearinghouse in a futures market. By contrast, in the OTC, the CFETS processes the transaction but does not guarantee the trade. Instead, OTC trading parties bear the credit risk. Greatly mitigating this is the fact that the credit risk in the spot FX market is very small (e.g., much smaller than the FX derivatives market). First, spot credit risk is only borne from the trade date to the settlement date, whereas FX derivative credit risk is borne over a much longer period, from the trade date to the maturity date. In China's FX spot market, the settlements are in the form of T+0, T+1 or T+2, with most trades settled in T+2, which means two business days after the transaction. Therefore, the credit (and settlement) risk is minimal.[11]

Second, spot FX is an unlevered security, whereas FX derivatives are a levered bet for which the resulting percentage change in value can be much larger. In fact, due to the low credit risk in FX spot trading, the dealers only need to allocate a tiny portion of total credit limits to FX spots while reserving a substantial part for the FX derivatives.[12] Last but not least, credit risk is so low that there was no contribution of margins by market participants to the CFETS if they trade in the FX spot market, unlike the usual practice that members contribute margins to the central clearing house.[13] Thus, there are multiple pieces of evidence that the credit risk in China's spot FX market is minimal, and so its influence on traders' choice on the two different trading venues is negligible.

3. Hypotheses

In this paper, we examine three important questions related to the LOB and the OTC markets:

(i) Which trading mechanism will predominate, the LOB or the OTC market?
(ii) Do these markets have upward- or downward-sloping price functions of trade size?
(iii) Do these markets have a single price function or multiple price functions of trade size?

There is no consensus on the answers to these questions in the existing literature. Next, we use existing theoretical and empirical findings to motivate potential answers and state them as hypotheses.

We start with the first question of whether LOB or OTC will predominate. Theoretically, Glosten (1994) argues that the LOB was "inevitable" and would predominate globally in all asset classes. He reasons that the LOB allows any possible price-quantity schedule and therefore can always undercut the more rigid pricing structure of a dealer market. Empirically, Jain (2005b) examines the leading stock exchange in 51 countries around the world and finds that pure dealer transaction costs are dramatically larger than the LOB or hybrid transaction costs. These findings imply that dealer markets should be competed out of existence. More recently, Abudy and Wohl (2017) carry the same argument over to the corporate bond market. They find that, for corporate bonds, the LOB transaction costs are much lower than the OTC dealer transaction costs. All of these findings about the superiority of the LOB provide the motivation for our first hypothesis.

Hypothesis 1. *The LOB market should predominate.*

However, the evidence of Biais and Green (2019) goes in exactly the opposite direction. They provide evidence of a huge shift in US corporate bond trading during the 1940s

[10] Initially, 13 banks were approved as dealers, and twelve of them still exist today. Although investors can also trade directly with other investors, this type of trades is very rare, accounting for less than 0.1%. This paper will focus on the customer-dealer trades that account for a dominant share in the international FX market (BIS's Triennial central bank survey of foreign exchange and OTC derivatives markets in 2019).

[11] The interbank FX market participants, most of which are banks, are trustworthy institutions. They are approved by the State Administration of Foreign Exchange (SAFE) using very strict criteria, and they have minimal counterparty risk. We survey all of the dealers in the interbank FX spot markets, and most dealers say that the default probability during the settlement process is less than 1%.

[12] Credit risks and bilateral credit limits are of concern for the FX derivatives, including FX forwards, FX swaps, FX options, and other FX derivatives, rather than FX spots. As emphasized in Duffie et al. (2005), "to trade OTC derivatives with a bank, one needs among other things, an account and a credit clearance." See also BIS's triennial central bank survey of foreign exchange and OTC derivatives markets in 2019 for a more detailed description of FX markets. From our survey to the dealers in the China's interbank FX market, FX spot transactions only take 2%-5% of the total bilateral credit lines, and market participants generally do not worry about credit risk associated with FX spot trading.

[13] See the regulatory document from the CFETS is under the name "The Announcement about the rules for trading and settlement in the interbank CNY FX spot market. (No.365 of 2005)"

from trading predominately on the centralized NYSE LOB market to trading predominately on the decentralized OTC dealer market. Their key explanation is that institutional traders felt that they could trade at a better price in the OTC dealer market and the growth institutional ownership of corporate bonds in the 1940s gave them the muscle to tip the balance of market power toward the OTC. Their evidence and explanation leads to an alternative hypothesis regarding our first question.

Hypothesis 2. *The OTC market should predominate.*

Next, we discuss hypotheses related to our second question that asks whether the price function is upward or downward sloping. An extensive literature shows that stock markets have upward-sloping price functions (see Amihud, 2002; Goyenko et al., 2009)). On the other hand, Edwards et al. (2007) documents a steep downward-sloping price function in US corporate bond trading. Therefore, one hypothesis is that the slope of the price function depends on the type of asset traded (e.g., stocks versus bonds).

Hypothesis 3. *The slope of the price function is determined by the asset class. Since both the LOB and OTC markets are trading the same asset class, namely FX, then both are expected to go in the same direction (i.e., either both upward-sloping or both downward-sloping).*

Alternatively, the slope of the pricing functions may depend on the trading mechanism. On the one hand, an LOB market, which allows you to "walk up the book" (i.e., bigger trade sizes reach price points that are further away from the bid-ask midpoint), should have an upward-sloping price function. Conversely, the evidence of Edwards et al. (2007) suggests that OTC markets typically have a downward-sloping price function.

Hypothesis 4. *The slope of the price function is determined by the trading mechanism. Specifically, the LOB market should have an upward-sloping price function and the OTC market should have a downward-sloping price function.*

Finally, we discuss hypotheses related to our third question regarding the existence of one or multiple price functions. Regarding the LOB, it is pretty clear cut that there must be a single price function. This is because trading is ex-ante anonymous, so there is no basis upon which to create multiple price functions.[14] By contrast, the OTC is more uncertain. Traders know who they are trading with in the OTC, and this allows the possibility that differences in bargaining power across different trader classes may yield multiple price functions.

Hypothesis 5. *There is a single price function in the LOB and multiple price functions in the OTC.*

Alternatively, there may be fierce competition across OTC dealers that effectively enforces the "law of one price." That is, dealers may be unwilling to make price concessions to more powerful customers, and less powerful customers may be unwilling to accept worse prices than others get.

Hypothesis 6. *Both the LOB and the OTC have a single price function.*

[14] In the LOB, the counterparties settle with the CFETS, since the CFETS is the central clearing counterparty, and the identities of the executing orders are not revealed to each another.

4. Empirical results

4.1. Which venue predominates, the LOB or the OTC market?

To answer this question, Table 1 reports the number of trades in the OTC and LOB markets, the OTC share of trade counts, and the OTC share of trade volume for spot USD/CNY trading in the Chinese interbank FX market before and after the OTC introduction. We find that the vast majority of trades and dollar volume migrated to the OTC market over a six-month transition from January 2006 to June 2006. By July 2006, 89.5% of the trades and 99.0% of the dollar volume took place on the OTC. For the first 11 months of 2006, the dollar volume market share of the OTC is higher than the trade market share. This indicates that a greater proportion of large trades migrated to the OTC market than small trades. The *t*-test on the daily data shows that the monthly difference in these two trade venues are statistically significant in every month of 2006. Fig. 2 shows the OTC market share by trade count and by dollar volume from January 2006 to December 2006, and it confirms the heavy migration of trading to the OTC. Therefore, we reject Hypothesis 1 that the LOB predominates and support Hypothesis 2 that the OTC predominates.

Next we provide more details of how the migration from LOB to OTC occurred across banks and trade sizes. Table 2 breaks out the trading into seven trade size buckets: less than \$0.5 million, \$0.5-\$1 million, \$1-\$3 million, \$3-\$5 million, \$5-\$10 million, \$10-\$20 million, and more than \$20 million. Panels A and B analyze the distribution of the original "disaggregated" trades. We consider the possibility that large parent orders may be broken up into multiple smaller child orders (especially in the LOB market), and so for robustness, we combine all of the trades from one bank on the same side (buyer-initiated versus seller-initiated) within ten minutes into an "aggregate trade." Panels C and D analyze the distribution of aggregated trades. Throughout the remainder for the paper, we analyze aggregated trades only, but our empirical results are robust to using the original disaggregated trades. Panels A and C are before the parallel OTC introduction, and panels B and D are after the six-month transition following the parallel OTC introduction.

We find that larger size trades almost entirely migrated to OTC, while smaller size trades (<\$0.5 million) were close to being equally split between the two venues. This result holds for both aggregated and disaggregated trades. Fig. 3 illustrates the migration pattern: by the end of 2006, more than 90% of transactions of the largest category of trade size (>\$3 million) while only 65% of transactions of the smallest category of trade size (<\$0.5 million) migrated to the OTC market.

Table 1
Number of trades and OTC shares over time.
This table shows the number of trades in the Limit Order Book (LOB) and Over-The-Counter (OTC) markets, the OTC share of trades, and OTC share of dollar volume for Chinese Yuan trades against the US dollar in the Chinese Interbank Foreign Exchange market. All computations are based on China Foreign Exchange Trading System (CFETS) intraday transaction data from August 2005 to December 2006. The *t*-statistic for a given month is the total for that month.

	LOB	OTC	OTC%		*T*-test of Venue Dif	
	Trades	Trades	Trade %	Volume %	Trades	$ Volume
Aug 2005	6,060		0.0%	0.0%		
Sep 2005	6,924		0.0%	0.0%		
Oct 2005	5,407		0.0%	0.0%		
Nov 2005	6,086		0.0%	0.0%		
Dec 2005	6,301		0.0%	0.0%		
Jan 2006	2,542	1,451	36.3%	74.0%	11.27	−5.90
Feb 2006	1,624	2,088	56.2%	88.3%	−2.71	−10.03
Mar 2006	1,610	3,996	71.3%	96.8%	−16.32	−19.54
Apr 2006	1,416	4,378	75.6%	94.3%	−13.59	−14.20
May 2006	1,054	3,971	79.0%	96.9%	−20.46	−22.68
Jun 2006	918	5,710	86.1%	98.5%	−16.79	−14.51
Jul 2006	907	7,709	89.5%	99.0%	−29.64	−31.01
Aug 2006	1,045	10,325	90.8%	99.1%	−28.16	−23.27
Sep 2006	1,002	11,725	92.1%	99.2%	−30.57	−23.94
Oct 2006	1,103	10,015	90.1%	97.9%	−41.16	−36.89
Nov 2006	1,006	12,395	92.5%	93.5%	−41.73	−14.05
Dec 2006	870	13,644	94.0%	89.6%	−29.45	−9.69
All of 2005	30,778		0.0%	0.0%		
All of 2006	15,097	87,407	85.3%	95.1%		
Full sample	45,875	87,407				

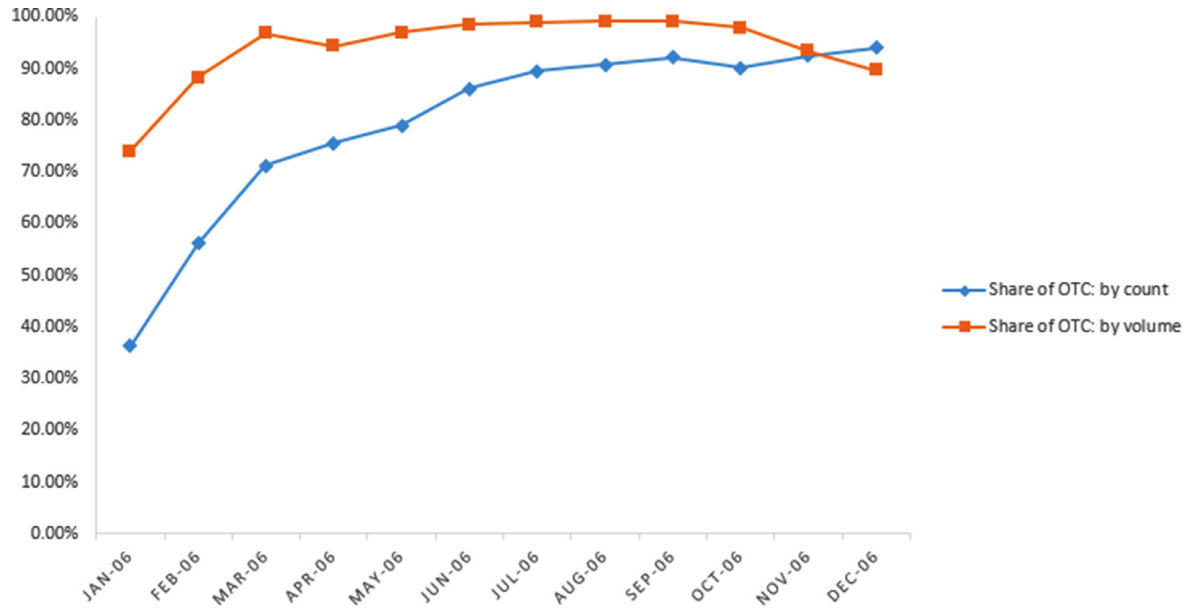

Fig. 2. The share of the OTC trades by extensive and intensive margins, 2006.

Next, we divide the banks into three groups using the CFETS's classification: large banks are core banks, medium banks are fundamental banks, and small banks are ordinary banks. Fig. 4 shows that, between January 2006 and December 2006, the market share of the large banks in the OTC increased from 40% to 60%, the market share of the medium banks decreased from 60% to 38%, and the market share of the small banks held steady at approximately 2%.

4.2. What determines the slope sign of the price function: asset class or trading mechanism?

Do the LOB and OTC markets have upward-sloping or downward-sloping price functions? To answer this question, we compute a variant of the one-way relative effective spread, which is a standard measure of transaction costs. The standard version uses the midpoint of the quoted bid and the quoted ask as the benchmark of "true

Table 2
Number of trades and volume share by trade size.
This table shows the number of trades and trade counts share (dollar volume share) of trades by trade size in the Limit Order Book (LOB) and Over-The-Counter (OTC) markets for Chinese Yuan trades against the US dollar in the Chinese Interbank Foreign Exchange market. All computations are based on China Foreign Exchange Trading System (CFETS) intraday transaction data from August 2005 to December 2006.

	LOB	OTC	OTC%		*T*-test of Venue Dif	
	Trades	Trades	Trade %	Volume %	Trades	$ Volume
Panel A. Disaggregated trades before parallel OTC introduction (Aug 2005–Dec 2005)						
< 0.5 mil	9,772		0.0%	0.0%		
0.5–1 mil	4,694		0.0%	0.0%		
1–3 mil	5,551		0.0%	0.0%		
3–5 mil	3,160		0.0%	0.0%		
5–10 mil	3,775		0.0%	0.0%		
10–20 mil	2,392		0.0%	0.0%		
> 20 mil	1,434		0.0%	0.0%		
Panel B. Disaggregated trades after parallel OTC transition (Jul 2006–Dec 2006)						
< 0.5 mil	8,669	7,676	47.0%	53.7%	2.85	−3.82
0.5–1 mil	2,985	6,501	68.5%	69.5%	−15.57	−15.68
1–3 mil	2,350	8,090	77.5%	79.1%	−25.95	−26.26
3–5 mil	637	13,100	95.4%	95.8%	−13.16	−13.00
5–10 mil	345	47,094	99.3%	99.4%	−15.31	−15.24
10–20 mil	71	2,643	97.4%	97.8%	−4.97	−4.90
> 20 mil	40	2,303	98.3%	82.8%	−3.63	−2.77
Panel C. Aggregated trades before parallel OTC introduction (Aug 2005–Dec 2005)						
< 0.5 mil	8,705		0.0%	0.0%		
0.5–1 mil	4,272		0.0%	0.0%		
1–3 mil	5,330		0.0%	0.0%		
3–5 mil	3,062		0.0%	0.0%		
5–10 mil	3,678		0.0%	0.0%		
10–20 mil	2,411		0.0%	0.0%		
> 20 mil	1,494		0.0%	0.0%		
Panel D. Aggregated trades after parallel OTC transition (Jul 2006–Dec 2006)						
< 0.5 mil	6,867	7,676	52.8%	59.6%	−2.73	−9.63
0.5–1 mil	2,635	6,501	71.2%	72.2%	−18.06	−19.31
1–3 mil	2,335	8,090	77.6%	79.3%	−26.33	−28.88
3–5 mil	666	13,100	95.2%	95.7%	−13.81	−13.80
5–10 mil	387	47,094	99.2%	99.4%	−16.61	−16.48
10–20 mil	91	2,643	96.7%	97.3%	−5.56	−5.74
> 20 mil	44	2,303	98.1%	82.5%	−3.41	−2.77

value." However, we do not have quoted bid and quoted ask data, so we cannot use this benchmark. But we do know which trades are interdealer trades and which side of each interdealer trade is demanding liquidity, allowing us to construct a unique, new benchmark of true value as the interdealer bid-ask midpoint. Specifically, it is the average of the most recent interdealer traded ask (i.e., the price of the most recent interdealer liquidity-demanding buy) and the most recent interdealer traded bid (i.e., the price of the most recent interdealer liquidity-demanding sell). There is typically only a short amount of time from a customer-dealer trade to the most recent interdealer trades, so this benchmark is fresh.

The variant of one-way relative effective spread that we use is

$$Relative\ Effective\ Spread_k = D_k(ln(P_k) - ln(M_k)),$$

where D_k is an indicator variable that equals $+1$ if the kth customer-dealer trade is buyer-initiated and -1 if the kth customer-dealer trade is seller-initiated, P_k is the price of the kth trade, $M_k = (B_k + A_k)/2$ is the interdealer bid-ask midpoint before the kth customer-dealer trade, B_k is the most recent interdealer traded bid before the kth customer-dealer trade, and A_k is the most recent interdealer traded ask before the kth customer-dealer trade. We compute the average one-way relative effective spread for a given day on a volume-weighted basis.[15]

We start by reporting the basic time-series pattern of one-way relative effective spread in both markets. Table 3 reports the one-way relative effective spread in basis points (bps) by month.[16] At the beginning of 2006, the one-way relative effective spreads are lower in LOB than in OTC. Then, as the migration to the OTC progresses, the spreads become nearly equal, and eventually the spreads are lower in the OTC. Interpreting the *t*-test in the last column, the one-way relative effective spreads are significantly lower in LOB in the early months, then become insignificantly different in the middle of 2006, and eventually become significantly lower in OTC by the end of 2006.

[15] Commissions and fees are included in prices in both the OTC and LOB. Therefore the one-way effective spread captures total transaction costs, including commissions and fees.

[16] For very small numbers, it is convenient to report them in basis points, where one basis point is 0.0001. Said differently, every reported number is multiplied by 10,000.

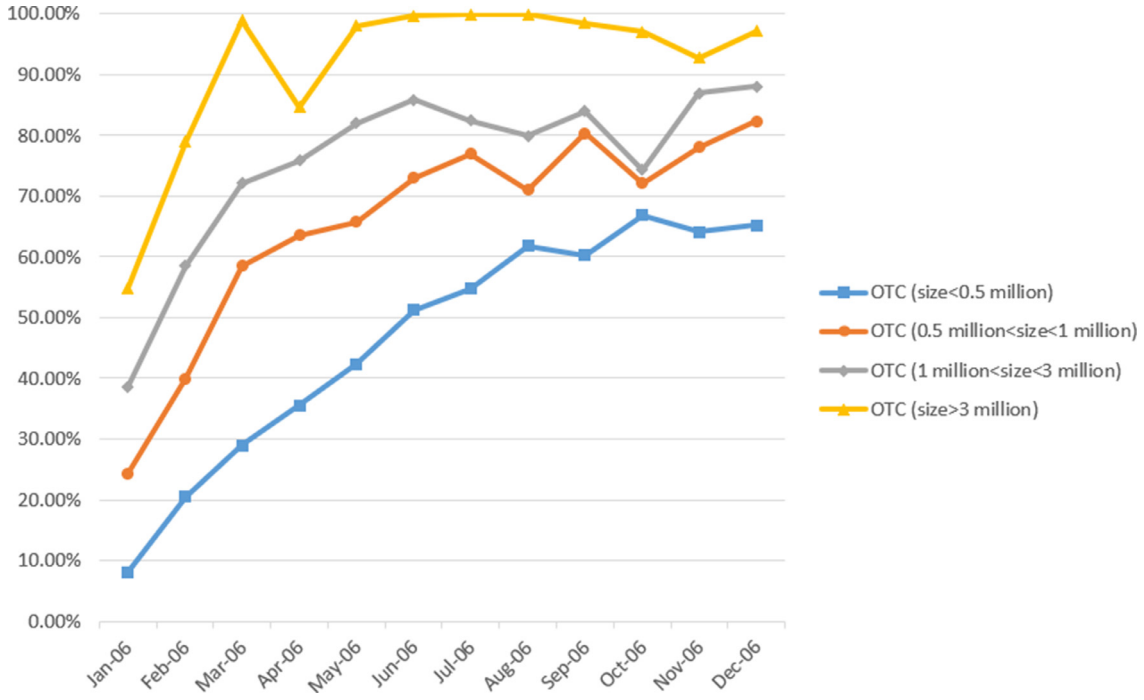

Fig. 3. The share of different trade sizes in OTC, 2006.

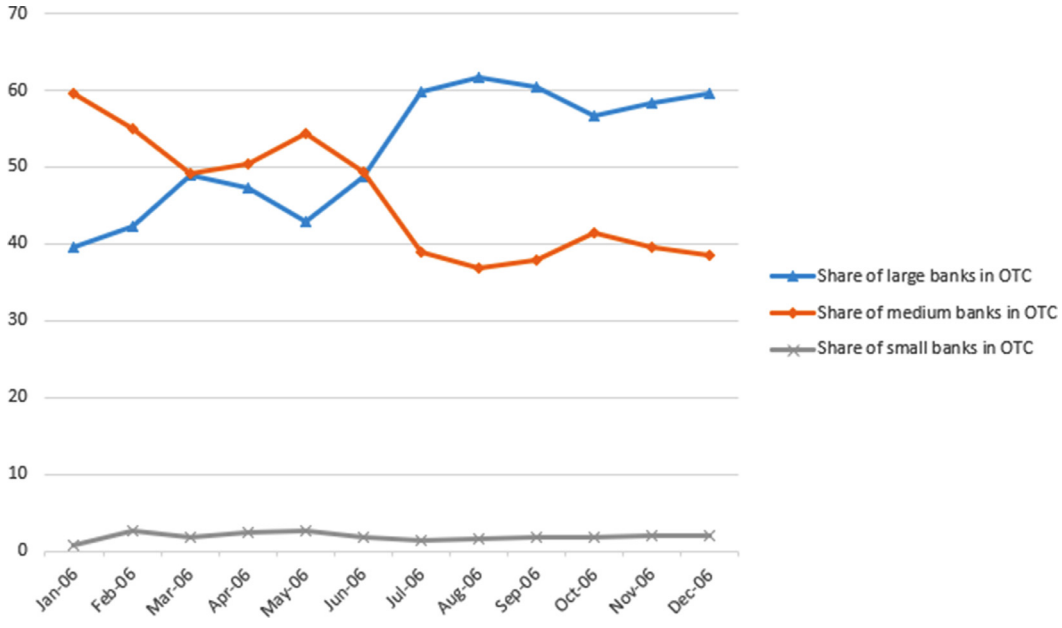

Fig. 4. Bank group market share in the OTC, 2006.

Table 4 shows one-way relative effective spreads in basis points by venue and by trade size during the parallel OTC transition (January-June 2006, Panel A) and after the parallel OTC transition (July-December 2006, Panel B). The one-way relative effective spread increases in trade size in LOB, while there is an opposite trend in OTC. This provides evidence in support of Hypothesis 4 that the slope is determined by the trading mechanism and against Hypothesis 3 that the slope is determined by the asset class.

4.3. Are there single or multiple price functions within LOB and OTC?

To answer this question, we look at the one-way relative effective spread across bank and trade sizes. Again,

Table 3
One-way relative effective spread over time.
This table shows the one-way relative effective spread in basis points (bps) over time in the Limit Order Book (LOB) and Over-The-Counter (OTC) market for Chinese Yuan trades against the US dollar in the Chinese Interbank Foreign Exchange market. All computations are based on China Foreign Exchange Trading System (CFETS) intraday transaction data from August 2005 to December 2006. The *t*-statistic for a given month is the total for that month.

	LOB (bps)	OTC (bps)	T-test
Aug 2005	2.24		
Sep 2005	1.87		
Oct 2005	1.29		
Nov 2005	0.97		
Dec 2005	0.80		
Jan 2006	2.94	3.55	−1.25
Feb 2006	1.15	4.09	−6.24
Mar 2006	2.11	3.60	−3.78
Apr 2006	3.67	5.10	−2.04
May 2006	4.22	6.25	−2.66
Jun 2006	3.50	4.54	−1.86
Jul 2006	3.73	3.59	0.22
Aug 2006	3.79	4.17	−0.63
Sep 2006	4.19	4.82	−0.71
Oct 2006	3.57	3.52	0.11
Nov 2006	4.92	2.80	4.71
Dec 2006	5.16	3.53	3.39

Table 4
One-way relative effective spread by trade size.
This table shows the one-way relative effective spread in basis points (bps) by trade size in the Limit Order Book (LOB) and Over-The-Counter (OTC) markets for Chinese Yuan trades against the US dollar in the Chinese Interbank Foreign Exchange market. All computations are based on China Foreign Exchange Trading System (CFETS) intraday transaction data from January 2006 to December 2006.

	LOB (bps)	OTC (bps)	*T*-test
Panel A. During the parallel OTC transition (Jan 2006–Jun 2006)			
< 0.5 mil	1.79	6.51	−8.31
0.5–1 mil	2.13	5.87	−4.67
1–3 mil	2.82	4.71	−3.12
3–5 mil	3.87	4.23	−1.25
5–10 mil	5.98	3.36	2.81
10–20 mil	7.62	2.42	3.58
> 20 mil	9.20	1.65	7.38
Panel B. After the parallel OTC transition (Jul 2006–Dec 2006)			
< 0.5 mil	3.57	6.23	−5.41
0.5–1 mil	4.25	5.64	−1.38
1–3 mil	5.19	4.18	1.60
3–5 mil	6.16	3.67	3.09
5–10 mil	7.11	2.56	5.98
10–20 mil	8.35	1.54	6.42
> 20 mil	9.57	0.68	8.75

we divide the banks into three groups using the CFETS's classification: large banks are core banks, the medium banks are fundamental banks, and the small banks are ordinary banks.

Fig. 5 plots the one-way relative effective spread in basis points for both markets by trade size and by bank size clienteles during the transition period of January 2006 to June 2006. This figure anticipates the finding of the Section 4.4 that the differences in spreads across bank clienteles are not significantly different in the LOB, and therefore it plots a single aggregate price function for the LOB. In Fig. 5 we find that the following is true of both markets: (i) large banks have lower one-way relative effective spreads than medium-sized banks for all trade sizes, and (ii) medium-sized banks have lower one-way relative effective spreads than small banks for all trade sizes. And it confirms that the LOB has an upward-pricing function and the OTC has a downward-sloping price function for each bank size. Fig. 5 also makes it vividly clear how large the price differences are in the OTC. In particular, large banks have dramatically lower one-way relative effective spreads than the medium and small banks. This evidence is suggestive that there are multiple price functions in the OTC. Fig. 6 shows that the same patterns hold up after the transition.

To summarize, after the introduction of OTC in 2006, trades in China's FX market migrated from LOB to OTC. Within six months, most of the large size trades migrated to OTC while a substantial portion of small size trades remained in LOB. Small size trades and small banks tend to trade in LOB, while large size trades and big banks are more likely to choose OTC. While we controlled for some differences in investors and trade types in the analysis above, a more rigorous multivariate approach is needed due to the potential endogeneity of customers' choices of trading venues.

4.4. Multivariate analysis controlling for selection bias

In this section we use trade-level data to conduct a more rigorous econometric test of the determinants of the one-way relative effective spread in OTC and LOB. We are concerned that the trades that go to the two markets are not random but instead are strongly selected. Specifically, we show that larger trades and trades from larger banks tend to go to the OTC, while smaller trades and trades from small banks tend to go to the LOB (see Figs. 3 and 4). This selection bias would be present in the single equation regression and would make the estimates biased. Therefore, we use the Heckman two-stage regression to address the potential endogeneity issue associated with choice of trading venues.[17]

The first stage of the Heckman model is given by the following Probit, where the dependent variable Y is a binary choice dummy equal to 1 if trade is executed in OTC and 0 if trade is executed in LOB:

$$Y = \Phi(Bank,\ TradeSize,\ Transition,\ Controls, \epsilon), \qquad (1)$$

where *Bank* is the set of dummy variables for large and medium banks (small banks are used as the default group) as defined in Section 4.3; *TradeSize* is coded as 0, 1, 2, 3, 4, 5, 6 for trades below 0.5 million, 0.5–1 million, 1–3 million, 3–5 million, 5–10 million, 10–20 million, and over 20 million, respectively; *Transition* is 1 for the trades within

[17] See Heckman (1979) for the details of this approach and Hendershott and Madhavan (2015) for an application of this method in US corporate bond market.

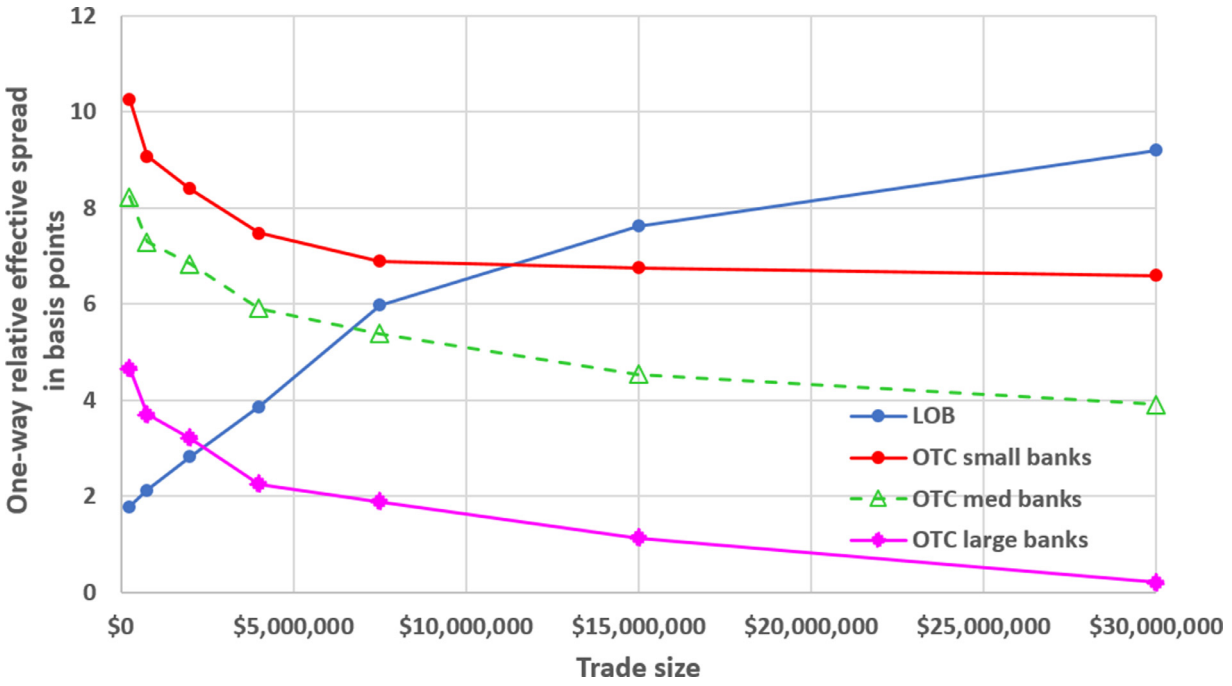

Fig. 5. Empirical one-way relative effective spread in basis points by venue, bank size, and trade size: during the adjustment (Jan 2006–Jun 2006).

Fig. 6. Empirical one-way relative effective spread in basis points by venue, bank size, and trade size: after the adjustment (Jul 2006–Dec 2006).

the first six months after the introduction of OTC and 0 otherwise.[18]

Controls include variables to control for market conditions: *Return* denotes the log change of the USD/CNY exchange rate in the current trading hour, and *Volatility* denotes the realized volatility of the exchange rate in the current trading hour. Furthermore, the dummy variable *buy* is one if the trade is buyer-initiated and zero if the trade is seller-initiated. The slope of the price function for customers buying in the OTC market needs not be the same as the slope of the price function for customers selling in

[18] This variable controls for possible transition effects (e.g., it might take time for investors to adapt to the OTC market).

Table 5
Venue selection: first stage probit model.
This table provides probit models for the binary choice between OTC and LOB. The dependent variable is one if OTC is selected. Independent variables include dummy variables for bank's size category and trade size category and for the first 6 months after the introduction of the parallel OTC market. We also include controls for market conditions and calendar effects. *, **, and *** mean the 10%, 5%, 1% level of significance, respectively.

	Heckman two-stage analysis First-stage Probit OTC versus LOB
Buy	9.19***
	(0.27)
Bank characteristics	
Medium	0.59**
	(0.30)
Large	0.03
	(0.31)
Market conditions	
Return	0.21*
	(0.13)
Volatility	−303.91***
	(33.50)
Calendar effect	
Monday	0.09**
	(0.04)
Friday	0.01
	(0.05)
End-of-month	0.14
	(0.13)
Trade size category	
$0.5–1 million	0.39***
	(0.09)
$1-3 million	0.74***
	(0.09)
$3-5 million	1.93***
	(0.14)
$5-10 million	3.10***
	(0.18)
$10-20 million	2.42***
	(0.23)
Above $20 million	2.93***
	(0.29)
Transition dummy	−0.86***
	(0.25)
Number of observations	101,905

the OTC market, and the same is true of the LOB market as well. Given these potential differences, the *buy* parameter picks up the differential likelihood of trading in the OTC for customer buys versus customer sells. The calendar time dummies *Monday* and *Friday* are used for the beginning and the end of the week to capture the potential inventory effect associated with the weekend. The dummy *End of Month* is one for the last trading day of the month to control for the fact that banks may need to rebalance their foreign currency holdings by the end of the month to satisfy regulatory requirements.

Table 5 reports the first-stage Probit model's results. Our main findings are that: (i) medium banks choose OTC over LOB with higher probability than small banks; and (ii) larger trades are more likely to be executed in the OTC market. Both results are consistent with our previous findings that OTC attracts larger trades and banks. The transition dummy has a negative coefficient, which indicates that OTC is less likely to be chosen during the transition than after the transition. This suggests that, during the transition, market participants were still learning about the new OTC trade institution, and thus they were less likely to use it.

We also find that the coefficient of *buy* is statistically significant and positive, meaning that customer buys have a higher chance of choosing the OTC than customer sells. This is likely because it is easier to buy a large amount of US dollars from the dealers in the OTC market than buy a large amount of the more constrained Chinese currency (which is equivalent to selling US dollars) in the OTC market.[19] For other control variables of the market condition, the positive and significant coefficient of *Return* indicates that trades are more likely to be executed in the OTC market through dealers if the USD appreciates against the CNY. We hypothesize that when foreign currency (USD) appreciates against domestic currency (CNY), many investors would like to buy US dollars (sell Chinese currency), and so they tend to choose OTC where the trading needs could be more easily met through the dealers. Additionally, OTC trades are more likely on Monday and less likely under volatile conditions.

In the second stage, we use the estimates from the first-stage Probit to correct for the potential selection bias. The dependent variable is the one-way relative effective spread of a trade in the OTC and LOB. The independent variables are given by trade size *Size* (the log-transform of the absolute value of trade size), its quadratic term $Size^2$, large and medium bank size dummies, transition dummy, controls, and sensitivity adjustment terms *Inv Mill OTC* and *Inv Mill LOB*.

Table 6 reports the result of the second-stage regression. We find that in the OTC, the *Large* and *Medium* bank dummies are significantly negative, meaning that there are multiple price functions in the OTC market. Assuming that larger banks have more bargaining power, it makes intuitive sense that more powerful banks are able to bargain for lower transaction costs. By contrast, in the LOB, the *Large* and *Medium* bank dummies are statistically insignificant, meaning that there is a single price function in the LOB market.[20] This provides evidence in support of Hypothesis 5 that there is a single price function in the LOB and multiple price functions in the OTC, because the former is anonymous and in the latter you know who you are trading with, and so power customers can bargain for a better price. This evidence is against Hypothesis 6 that there is a single price function in both markets.

In the OTC market, the coefficients on *Size* and $Size^2$ indicate a negative relationship between transaction costs

[19] Hendershott and Madhavan (2015) find similar results (i.e., they report investors' buying orders for corporate bonds are more likely to be executed in the OTC dealer markets).

[20] As a robustness check, we divide the banks into high, medium, and low groups based on trading activity. We obtain the same qualitative results from the same Heckman two-stage regressions. Namely, there are three downward-sloping price functions for the OTC market and a single upward-sloping price function for the LOB market.

Table 6
Multivariate and selection bias analysis.
This table shows the determinants of relative effective spread by venue. The dependent variable is the relative effective spread in basis points in each venue. Independent variables include dummy variables for bank's size category and trade size category and for the first 6 months after the introduction of the parallel OTC market. *, **, and *** mean the 10%, 5%, 1% level of significance, respectively.

	Heckman two-stage analysis	
	Second stage	
	OTC Effective Spread	LOB Effective Spread
Buy	−0.48***	−0.88
	(0.16)	(0.80)
Bank characteristics		
Medium	−2.47***	−0.44
	(0.34)	(0.55)
Large	−3.98***	−0.79
	(0.37)	(0.51)
Trade size characteristics		
Size	1.60***	−3.51**
	(0.38)	(1.47)
$Size^2$	−0.09***	0.14**
	(0.01)	(0.06)
Market condition		
Return	26.29	0.24
	(24.62)	(0.52)
Volatility	−107.56***	154.46***
	(30.95)	(51.63)
Transition dummy	0.36	−1.84***
	(0.39)	(0.40)
Inverse Mills ratio for OTC	0.53***	
	(0.13)	
Inverse Mills ratio for LOB		−0.16
		(0.21)
Number of observations	87,407	14,498

and trade sizes. Fig. 7 plots the predicted transaction cost in the OTC market against the range of trade sizes for small, medium, and large banks, during and after the transition. The transaction cost decreases at a faster rate as trade size gets larger. Contrary to OTC, in LOB, the relationship between transaction costs and trade size is positive. Fig. 8 plots the predicted transaction cost in LOB market against the range of trade sizes for small, medium and large banks, during and after the transition. In the LOB, liquidity became significantly worse after the transition.

An institutional detail in the trade venues selection are the bilateral credit lines in the OTC FX markets, which are important for whether prices are transactable by a given potential counterparty. In other words, to guarantee all of the investors can freely choose to trade either in LOB or OTC, they should be able to trade in OTC FX market with dealers. Our data do not contain the detailed bilateral credit lines information; however, we collect information about bilateral credit lines from the dealers. From our survey of dealers, they indicate their bilateral credit lines cover nearly all of the market participants (over 95%), and the credit limits are large enough to support intraday FX spot trading.[21] In the trading data, we find that only 16 market participants (5.7% of the total market participants) only trade in LOB and never traded in the OTC. However, we could not distinguish the two alternative reasons that they did not trade in the OTC market: either they are not willing to trade in OTC, or they are not able to trade in OTC due to the lack of bilateral credit lines with the dealers. In a robustness check we drop the 16 market participants who never trade in the OTC market, and we find that our empirical results remain robust.[22]

5. Theory

In this section we develop a theoretical model of parallel LOB and OTC markets as follows. First, we develop a model of a centralized LOB exchange based on an adverse selection model. Second, we develop a model of a decentralized OTC market based on a search model. Third, we develop a model of parallel markets that combines the centralized LOB market and the decentralized OTC market and allows one class of traders to endogenously choose which market to trade in. Fourth, we provide a numerical illustration that compares the centralized LOB only (before the OTC is introduced) and parallel markets (after the OTC is introduced). Finally, we develop an additional empirical prediction about the impact of introducing a parallel OTC market and test it.

5.1. The centralized LOB market

5.1.1. The setup

Assume there are three classes of traders. First, there are many risk-neutral, uninformed liquidity suppliers. Their goal in supplying liquidity is to earn a positive expected profit. At time t, each liquidity supplier has the opportunity to submit either one limit buy or one limit sell order to the LOB. They continue submitting limit orders until nobody wishes to submit any additional limit order. These limit orders set the bid and ask prices available on the LOB.

Second, there is one risk-neutral, informed liquidity demander. At time $t+1$, the informed trader receives a private signal about the time $t+2$ value of the foreign exchange. With probability ρ, the informed trader arrives at time $t+1$ and submits a market order of size x to optimally exploit this private information (i.e., the informed trader chooses x to maximize his/her expected profit).

Third, there are uninformed liquidity demanders. These traders have exogenous liquidity needs that motivate them to trade. With probability $1-\rho$, the trader who arrives on time $t+1$ is uninformed, and this trader submits a market order of size u that is drawn from a continuous uniform distribution over the interval $[-M, M]$. Both x and u are expressed in units of domestic currency, where a positive amount means a market buy order and a negative amount means a market sell order.

The model of centralized LOB is similar to the adverse selection models of Kyle (1985), Glosten and Milgrom (1985), and Easley and O'Hara (1987). Asymmetric information is present in FX markets as discussed by

[21] From our survey with dealers, they indicate the chance that the dealers run out of bilateral credit with a counterparty in the intraday FX spot trading is minimal, usually less than 1%.

[22] The results are available upon request.

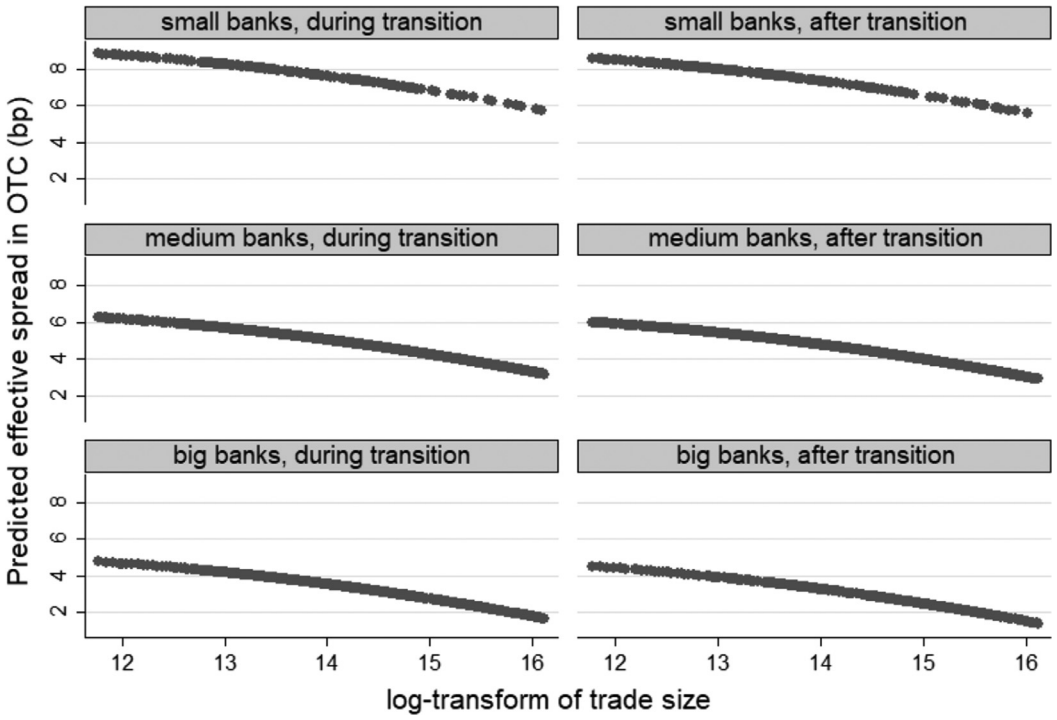

Fig. 7. One-way relative effective spread predicted by the Heckman two-stage for the OTC market: by bank size and trade size, during and after the transition. *Note:* The bottom 1 percentile and top 1 percentiles of trade size values are eliminated from the range of trade sizes. Log-transform of trade size ranging from 11.92 to 16.12 is equivalent to the trade size moving from $0.15 million to $10 million.

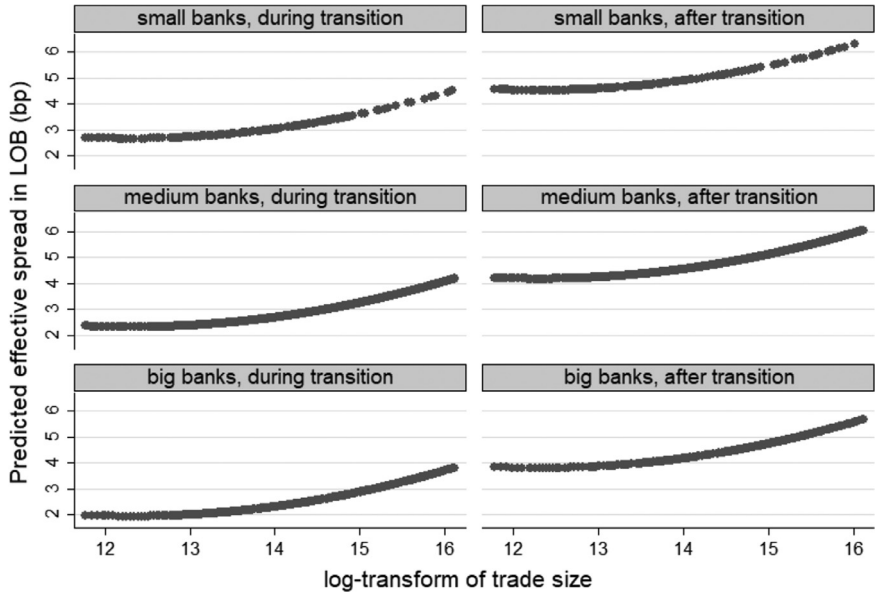

Fig. 8. One-way relative effective spread predicted by the Heckman two-stage for the LOB market: by bank size and trade size, during and after the transition. *Note:* The bottom 1 percentile and top 1 percentile of trade size values are eliminated from the range of trade sizes. Log-transform of trade size ranging from 11.92 to 16.12 is equivalent to the trade size moving from $0.15 million to $10 million.

Lyons (2006), Evans and Lyons (2008); King et al. (2013), and Michaelides et al. (2019). On an LOB, all orders that are submitted are anonymous (i.e., nobody knows the identity of who submitted each order). Importantly, nobody knows if any particular market order was submitted by an informed or uninformed trader.

A single foreign currency is traded for domestic currency. Let v be the time $t+2$ value of the foreign cur-

rency in units of domestic currency. Assume that v is uniformly distributed on the K equally-spaced discrete values $[v_1, v_2, v_3, \ldots, v_K]$. Let π_k be the probability that $v = v_k$, where $k = 1, 2, 3, \ldots, K$. The uniform distribution implies that all of the probabilities are $\pi_k = \frac{1}{K}$ and the unconditional mean is $E[v] = \frac{v_1+v_K}{2}$.

The bid and ask sides are separable and symmetric in this model. For convenience, we will focus the exposition on the ask side (i.e., where liquidity-demanding market buys trade against liquidity-supplying limit sells).[23] Assume that available prices on the ask side are on equally-spaced ticks $[p_1, p_2, p_3, \ldots, p_N, p_{N+1}]$, where p_1 is the first price above $E[v]$, p_N is the last price below v_K, and p_{N+1} is the first price that is greater than or equal to v_K.

Initially, the limit order book is empty. At time t, a liquidity supplier submits a limit sell order for the optimal quantity Q_1 at the first ask price p_1 above $E[v]$. Then another liquidity supplier submits a limit sell for the optimal quantity Q_2 at the second ask price p_2, and so on. This continues until the sum of the price times quantity of all of the limit sells reaches the maximum uninformed market buy quantity M and stops there, since there are no further opportunities to profit at the expense of an uninformed trader.[24]

Then at time $t+1$, either an informed or an uninformed trader arrives. That trader submits a market order. In the case of a market buy, the exchange will match it with limit sell order(s) based on the standard rule of price priority. That is, first the market buy will be executed against the limit sell at the best price p_1. If the market buy is not completely executed, then the remaining quantity will "walk up the book" and execute against the limit sell at the second best price p_2. And so on, until the market buy is fully executed.[25]The exchange is also assumed to follow the standard rule of time priority, in which the first limit order at a given price is executed before the second limit order at that price. As will be shown below, the first limit sell at a given price is optimally sized so as to obtain 100% of the marginal profit available at that price. Therefore, it is never optimal to submit a second limit sell at the same price. Then at time $t+2$, the foreign currency value v is realized and endogenous agent profits or losses are realized.

5.1.2. The informed trader's problem

We begin the analysis by considering the informed trader's optimal trading problem at time $t+1$. Previously at time t, liquidity suppliers filled up the ask side of the limit order book by submitting a limit sell for Q_1 at p_1, a limit sell for Q_2 at p_2, and so on, up to as much as a limit sell for Q_{N+1} at p_{N+1}. Then at time $t+1$, a trader arrives who is either as informed or uninformed.

Suppose that the trader who arrives is informed. Let s be the informed trader's time $t+1$ signal, which is equal to the time $t+2$ foreign currency value ($s = v$). What is the informed trader's optimal time $t+1$ order to exploit this private information?

It turns out that the optimal informed trader strategy is very simple. If the informed trader's private signal is greater than the lowest ask price ($s > p_1$), then the informed trader should buy the full quantity available from limit sells with prices below s and profit on the price difference. In other words, the optimal market buy order is equal to the sum of all limit sell quantities with prices below v as given by

$$x = \sum_{n=1}^{N+1} Q_n I_n, \tag{2}$$

where I_n is an indicator variable that is 1 when $p_n < v$ and is 0 otherwise.[26]

The informed trader's profit is the sum of the price difference times the quantity as given by

$$\sum_{n=1}^{N+1} (v - p_n) Q_n I_n. \tag{3}$$

5.1.3. The liquidity supplier's problem

Given the informed trader's optimal strategy at time $t+1$, we roll back to time t and analyze the liquidity supplier's problem. Without loss of generality, other liquidity suppliers have already submitted a limit sell for Q_1 at p_1, a limit sell for Q_2 at p_2, and so on, up to a limit sell for Q_{n-1} at p_{n-1}. Now consider a liquidity supplier who is considering submitting a limit sell of size Q_n at price p_n, where $n \in \{1, 2, \ldots, N+1\}$.[27]

The liquidity supplier needs to consider two possible scenarios for the limit sell's execution: (1) an informed trader arrives and imposes a loss on the liquidity supplier, and (2) an uninformed trader arrives and provides a profit to the liquidity supplier. In the first scenario, an informed trader arrives with probability ρ and follows the optimal informed trading strategy described above. Specifically, if the informed trader's signal is greater than the ask price ($s > p_n$), then the informed trader will submit a market buy that is large enough to execute in full the limit sell at the price p_n. In this case, the liquidity supplier will suffer a loss that is equal to the value difference $(v - p_n)$ times the full quantity Q_n.

In the second scenario, an uninformed trader arrives with probability $1 - \rho$ and submits a market order of size u, which is drawn from the uniform distribution over the interval $[-M, M]$. There are three possible outcomes for the limit sell. First, if u is negative, then the uninformed trader submits a market sell, or if u is positive, but less than the sum of the first $n-1$ limit sells, then the uninformed trader submits a market buy that finishes executing before

[23] The bid side (i.e., where liquidity-demanding market sells trade against liquidity-supplying limit buys) is incorporated via a symmetric analysis.

[24] Symmetrically, other liquidity suppliers submit optimally-sized limit buys at the available bid prices until they reach the maximum uninformed market sell quantity $-M$.

[25] Symmetrically, the exchange will match a market sell order against the limit buy at the best bid price, then the second best bid price, and so on until it is fully executed.

[26] Symmetrically, if the informed trader's private signal is below the highest bid price, then the informed trader should submit a market sell order equal to the sum of all limit buy quantities with prices above v.

[27] In other words, when $n = 1$, this is the first limit sell and nothing has been previously submitted.

it reaches the limit sell. In either case, none of the *nth* limit sell executes. Second, if u is positive and in an intermediate range, then the market buy is large enough to reach the limit sell but does not fully execute it. In this case, the average amount of the limit sell that executes is half of Q_n due to the uniform distribution of u. Third, if u is positive and greater than the sum of the first n limit sells, then the limit sell executes completely. The average quantity of the limit sell that executes in these three cases can be summarized as follows:

$$\begin{cases} 0 & \text{when } u < \sum_{i=1}^{n-1} Q_i \\ \frac{Q_n}{2} & \text{when } \sum_{i=1}^{n-1} Q_i < u < \sum_{i=1}^{n} Q_i \\ Q_n & \text{when } u > \sum_{i=1}^{n} Q_i. \end{cases} \tag{4}$$

Combining everything above, the liquidity supplier's profit on a limit order of size Q_n at price p_n is

$$\rho \sum_{k=1}^{K} \pi_k (v_k - p_n)(-Q_n) I_k + (1-\rho)\left(\frac{Q_n}{2M}\right)(E[v] - p_n)\left(\frac{-Q_n}{2}\right) + (1-\rho)\left(\frac{M - \sum_{i=1}^{n-1} Q_i - Q_n}{2M}\right)(E[v] - p_n)(-Q_n), \tag{5}$$

where I_k is an indicator variable that is 1 when $v_k > p_n$ and is 0 otherwise. The first term is the loss of the limit sell to the informed trader.[28] The second and third terms are the gain of the limit sell from the uninformed trader in the two scenarios of partial and complete execution, respectively.[29]

The liquidity supplier's objective function is quadratic in Q_n. Taking the derivative with respect to Q_n, we get the first order condition with the maximized value of Q_n (which we write as Q_n^*). The solution to the liquidity supplier's problem in the centralized LOB is summarized in the following proposition.

Proposition 1. Centralized LOB. *For a liquidity supplier, the optimal size of thelimit sell at price p_n is*

$$Q_n^* = \frac{\rho 2M \sum_{k=1}^{K} \pi_k (v_k - p_n) I_k}{(1-\rho)(E[v] - p_n)} + M - \sum_{i=1}^{n-1} Q_i \quad \text{for } n = 1, 2, \ldots, N \text{ and} \tag{6}$$

$$Q_{N+1}^* = \frac{M - \sum_{i=1}^{N} p_i Q_i}{p_{N+1}}. \tag{7}$$

Proof of Proposition 1. Taking the derivative of Eq. (5) yields the first order condition, which is linear in Q_n^*. Solving for Q_n^* yields Eq. (6). The second derivative of Eq. (5) with respect to Q_n is $(1-\rho)\left(\frac{E[v]-p_n}{2M}\right)$, which is negative because $p_n > E[v]$. This verifies the second order condition (i.e., that the profit function is concave). If the currency amount sum of the first N limit sells is less than M, then it is optimal to submit the $N+1$ limit sell at p_{N+1} for the full remaining amount, because there is no informed trading against this price and thus it is strictly profitable. □

Limit sells are added to the limit order book by applying Eq. (6) recursively for $Q_1^*, Q_2^*, \ldots, Q_N^*$. If the currency amount sum of all limit sells ($\sum_{i=1}^{N} p_i Q_i^*$) is equal to M, then the ask side of the limit order book is full and the expression for Q_{N+1}^* is equal to zero.

Alternatively, if the currency amount sum of all limit sells is less than M, then it is optimal for another liquidity supplier to submit one more limit sell at the price p_{N+1} for the remaining amount to reach M. By construction, the price $p_{N+1} \geq v_K$. Since there is no realization of v that is greater than p_{N+1}, it is impossible for an informed trader to profit at the expense of a limit sell at p_{N+1}. Therefore, the average profit of a limit sell at p_{N+1} is strictly positive due to trading against uninformed traders only. Thus, it is optimal to submit a limit sell at p_{N+1} for the full remaining quantity. There is no reason for the currency amount sum to go beyond M, since that is the largest uninformed market buy.

A related point is that liquidity suppliers never submit a second limit sell at any available price. To see why, recall that the liquidity supplier's profit function Eq. (5) is quadratic in Q_n. For $Q_n = 0$, the profit is zero. For larger values of Q_n, the profit strictly increases, because the marginal profit is strictly positive. When Q_n reaches Q_n^*, then the profit reaches the peak amount, because the marginal profit (i.e., the left-hand side of the first order condition) equals zero. Beyond this point, the marginal profit goes negative (i.e., you lose more to the informed trader than you gain from the uninformed trader). Thus, submitting a second limit sell would lead to an expected loss and so will not be done.

This completes the development of a centralized LOB exchange model based on adverse selection.

5.2. The decentralized OTC market

Next, we develop a model of a decentralized OTC market based on a search model similar to Vayanos and Wang (2011). The decentralized OTC market differs from the centralized LOB market in three important ways. First, bid and ask prices are *not* posted in a centralized location. Therefore, liquidity demanders must search for counterparties to trade with by contacting liquidity suppliers one at a time.

Second, once a liquidity demander finds a liquidity supplier ("dealer"), the two parties bargain over what price to trade at. We will consider the case in which heterogeneous liquidity demanders have different degrees of bargaining power based on their degree of market power. Thus, different liquidity demanders will obtain different trade prices for the same size trade.

Third, the two parties know who they are trading with (i.e., trading is *not* anonymous), so reputations matter. In a

[28] Specifically, it takes the following form: (the probability that the informed trader arrives) times (the probability of different values of v_k) times (the informed trader's profit margin) times (the limit order size) times (the indicator that $v_k > p_n$).

[29] Specifically, they take the following form: (the probability that the uninformed trader arrives) times (the probability that the uninformed market order results in partial or full execution, respectively) times (the uninformed trader's average loss margin) times (the average size executed in the partial or full execution scenarios, respectively).

multiperiod setting, an informed trader who burned a liquidity supplier would be blacklisted by the liquidity supplier. If the same informed trader persisted in burning other liquidity suppliers, he would soon run out of counterparties to trade with. Thus, an OTC market with even one informed trader would soon break down. The straightforward conclusion is that the long-term equilibrium in an OTC market can only be based on uninformed traders. We do not formally model the multiperiod reputational process, but we incorporate that key idea by basing our model on the assumption that only uninformed traders trade in the OTC market.[30] Thus, only *uninformed* liquidity demanders and liquidity suppliers trade on the OTC market.

As mentioned above, we allow the liquidity demanders to be heterogeneous. We assume that there are J classes of such traders. The classes are indexed by $j = 1, 2, \ldots, J$. Let ϕ_j be the liquidity supplier bargaining power when negotiating with a type j liquidity demander. For example, in the numerical illustration below, we will assume that there are three classes ($J = 3$). We designate $j = 1$ as the "big banks" (i.e., the liquidity demander class with the most bargaining power), $j = 2$ as the "medium banks" with medium bargaining power, and $j = 3$ as the "small banks" with the least bargaining power. If big banks have the most bargaining power, then the liquidity suppliers negotiating with the big banks would have least bargaining power, so ϕ_1 would be the smallest. By the same logic, we get the ordering $\phi_1 < \phi_2 < \phi_3$.

We assume that there are limits to the prices that any of the liquidity demanders are willing to trade at. Let $E[v] + C$ be the maximum reservation price of the liquidity demanders, and let $E[v] - C$ be the minimum reservation price, where C is a constant.

We also assume that a liquidity supplier faces a fixed order processing cost per trade. Let f be this fixed cost per trade, which is independent of the trade size. The risk-neutral liquidity suppliers are unwilling to trade at a loss, so they must be able to recover the fixed cost on each trade. Thus, the lowest ask (highest bid) price that liquidity suppliers are willing to quote for a buy (sell) quantity U is $p(U) = E[v] + If/|U|$, where $I = +1$ for a buy and $I = -1$ for a sell. The highest ask (lowest bid) price that liquidity demanders are willing to pay is the maximum (minimum) reservation price $E[v] + IC$. These two prices provide the upper and lower bound for an ask (bid) price. We assume that the two parties bargain and arrive at an intermediate price such that the liquidity supplier gets ϕ_j proportion of the gain (i.e., the difference between the upper and lower bound) and the liquidity demander of type j gets $1 - \phi_j$ proportion of the gain.

The equilibrium decentralized OTC prices are summarized in the following proposition.

Proposition 2. Decentralized OTC. *For a liquidity demander of type j, the equilibrium decentralized OTC prices are*

$$p(U) = E[v] + I\frac{f}{|U|} + I\phi_j\left(C - \frac{f}{|U|}\right). \tag{8}$$

[30] Harris (2003) provides anecdotal evidence that the decentralized, non-anonymous search market for large stock block trades is based on uninformed traders.

Proof of Proposition 2. The second term allows for the recovery of the fixed order processing cost. The third term is the gain to the liquidity supplier based on the relative bargaining power of the liquidity supplier. □

Notice that the decentralized OTC prices do *not* depend on the relative proportions of the two classes of agents but only on their relative bargaining power ϕ_j. This completes the development of a decentralized OTC model based on search.

5.3. Parallel markets

Now we develop a model of parallel markets by combining the centralized LOB model and the decentralized OTC model. Of the three classes of traders, informed traders trade exclusively in the centralized LOB market and liquidity suppliers trade in both markets. For convenience, we split the uninformed liquidity demanders into two subgroups: (1) discretionary and (2) nondiscretionary. The nondiscretionary uninformed liquidity demanders trade exclusively in the centralized LOB market. This is a simple, widely used device to make sure that there are always at least *some* uninformed traders in the centralized LOB market so that it does not collapse due to having 100% informed traders.

The discretionary liquidity demanders are heterogenous and get to choose whether to trade on the centralized LOB exchange or the decentralized OTC market. They are divided into J classes that have different degrees of bargaining power. As before, ϕ_j is the liquidity supplier bargaining power when negotiating with a type j discretionary liquidity demander, where $j = 1, 2, \ldots, J$.

Importantly, discretionary liquidity demanders can choose which of the two venues they want to trade in as a function of (1) their trader type j and (2) their realized market order size u. In other words, we could see different choices being made by small banks versus medium banks versus large banks, and each type of bank could make different choices for small orders versus medium orders versus large orders. Thus, for each combination of trader type and order size, the discretionary liquidity demanders consider the price they could trade at on the centralized LOB exchange and in the decentralized OTC market and trade where they can get the best price.

The critical switching quantity $S^*_{j,n}$ for a liquidity demander of type j whose market order finishes executing at the price p_n if sent to the LOB in a parallel markets setting is summarized in the following proposition.

Proposition 3. Optimal choice of trading venue in parallel markets. *For a liquidity demander of type j whose market order finishes executing at the price p_n if sent to the LOB, the critical switching quantity is*

$$S^*_{j,n} = \frac{-b_{j,n} + \sqrt{b^2_{j,n} - 4a_{j,n}c_{j,n}}}{2a_{j,n}} \tag{9}$$

$$\text{where} \quad a_{j,n} = p_n^2 - E[v]p_n - I\phi_j C p_n \tag{10}$$

$$b_{j,n} = 2d_n p_n - E[v](d_n + p_n e_n)$$

$$- I\phi_j(Cd_n + Cp_ne_n - f) - If \tag{11}$$

$$c_{j,n} = (d_n)^2 - E[v]d_ne_n - I\phi_j e_n(Cd_n - f) - Ifd_n \tag{12}$$

$$d_n = \sum_{i=1}^{n-1} p_iQ_i \tag{13}$$

$$e_n = \sum_{i=1}^{n-1} Q_i. \tag{14}$$

For a liquidity demander of type j, the optimal trading venue is the LOB when the absolute market order size is smaller than the switching quantity ($|U| < \sum_{i=1}^{n-1} Q_i + S^*_{j,n}$) and the OTC market when the absolute market order size is larger than the switching quantity ($|U| > \sum_{i=1}^{n-1} Q_i + S^*_{j,n}$).

Proof of Proposition 3. Set the right-hand side of Eq. (8) equal to the weighted average price on the LOB than finishes executing at the price p_n, which is $\left(\sum_{i=1}^{n-1} p_iQ_i + p_nS_{j,n}\right)/\left(\sum_{i=1}^{n-1} Q_i + S_{j,n}\right) = \left(d_{j,n} + p_nS_{j,n}\right)/\left(e_{j,n} + S_{j,n}\right)$. Then set $|U| = d_{j,n} + p_nS_{j,n}$, and you have a quadratic equation in $S_{j,n}$. Denote the quadratic formula coefficients $a_{j,n}$, $b_{j,n}$, and $c_{j,n}$ and the optimal quadratic solution $S^*_{j,n}$. □

Given that the discretionary liquidity demander of type j whose order finishes executing at the price p_n is switching to the dealer market for trade sizes larger that the critical switching quantity $S^*_{j,n}$, we need to revise the liquidity supplier's analysis to account for this. This is a simple modification of the prior analysis. Let $\hat{Q}_n$ be the revised limit sell quantity at p_n. Given that a liquidity demander of type j switches on a given price point p_n, the limit sell order faces a lower probability of an informed trader ρ_{j-1} for the initial quantity up to the critical switching quantity $S^*_{j,n}$ and a higher probability of an informed trader ρ_j for the remaining quantity $\hat{Q}_n - S^*_{j,n}$. The liquidity suppliers profit simply has two terms in place of each prior term—one for below the switching quantity and one for above the switching quantity. Thus, the revised liquidity supplier's profit is

$$\begin{aligned}
&\rho_{j-1}\sum_{k=1}^{K}\pi_k(v_k - p_n)(-S_{j,n})I_k \\
&+ \rho_j\sum_{k=1}^{K}\pi_k(v_k - p_n)(-\hat{Q}_n + S_{j,n})I_k \\
&+ (1-\rho_{j-1})\left(\frac{S_{j,n}}{2M}\right)(E[v] - p_n)\left(\frac{-S_{j,n}}{2}\right) \\
&+ (1-\rho_j)\left(\frac{\hat{Q}_n - S_{j,n}}{2M}\right)(E[v] - p_n)\left(\frac{-\hat{Q}_n + S_{j,n}}{2}\right) \\
&+ (1-\rho_{j-1})\left(\frac{M - \sum_{i=1}^{n-1} Q_i - S_{j,n}}{2M}\right)(E[v] - p_n)(-S_{j,n}) \\
&+ (1-\rho_j)\left(\frac{M - \sum_{i=1}^{n-1} Q_i - \hat{Q}_n + S_{j,n}}{2M}\right)(E[v] - p_n)(-\hat{Q}_n + S_{j,n}).
\end{aligned} \tag{15}$$

The solution to the modified liquidity supplier's problem under parallel markets is summarized in the following proposition.

Proposition 4. Optimal limit sell under parallel markets. *When the critical switching quantity for discretionary liquidity demander of type j on the price p_n is less than the optimal size of the limit sell on a centralized LOB ($S^*_{j,n} < Q^*_n$), then the revised optimal size of the limit sell at price p_n is*

$$\hat{Q}^*_n = \frac{\rho_j 2M\sum_{k=1}^{K}\pi_k(v_k - p_n)I_k}{(1-\rho_j)(E[v] - p_n)} + M - \sum_{i=1}^{n-1} Q_i \quad \text{for } n = 1, 2, \ldots, N \text{ and} \tag{16}$$

$$\hat{Q}^*_{N+1} = \frac{M - \sum_{i=1}^{N} p_iQ_i}{p_{N+1}}. \tag{17}$$

The set of critical switching quantities $S^*_1, S^*_2, \ldots, S^*_J$ and revised optimal limit order sizes $\hat{Q}^*_1, \hat{Q}^*_2, \ldots, \hat{Q}^*_{N+1}$ fully specify the parallel markets equilibrium.

Proof of Proposition 4. Taking the derivative of Eq. (15) yields the first order condition, which is linear in the maximized value of $\hat{Q}_n$ (which we write as $\hat{Q}^*_n$). Solving for $\hat{Q}^*_n$ yields Eq. (16). The second derivative of Eq. (15) with respect to Q_n is $(1-\rho_j)\left(\frac{E[v]-p_n}{2M}\right)$, which is negative because $p_n > E[v]$. This verifies the second order condition (i.e., that the revised profit function is concave). If the currency amount sum of the first N limit sells is less than M, then it is optimal to submit the $N+1$ limit sell at p_{N+1} for the full remaining amount, because there is no informed trading against this price and thus it is strictly profitable. □

Interestingly, all of the terms with $S^*_{j,n}$ drop away when taking the derivative or cancel out. So the revised optimal size of the limit sell on price p_n only differs from the original optimal size due to the variable ρ_j, which is the probability of informed trading past the switching point. The probability of informed trading is strictly higher past the switching point ($\rho_j > \rho_{j-1}$), because when the discretionary liquidity traders of type j switch to the OTC, there are fewer uninformed traders left in the LOB market to provide camouflage for the informed traders.

Given the parallel markets equilibrium, it is straightforward to derive a new testable prediction from the model, which is summarized in the following proposition.

Proposition 5. The impact of bargaining power on the critical switching quantities. *The critical switching quantity $S^*_{j,n}$ is strictly increasing in the bargaining power of the liquidity supplier who trades with the discretionary trader of type j (ϕ_j).*

Proof of Proposition 5. Take the derivative of the price $p(U)$ in Eq. (8) with respect to the bargaining power of the discretionary trader of type j (ϕ_j) and you obtain $I(C - \frac{f}{|U|})$, which is strictly greater than zero. Therefore, as ϕ_j increases, the OTC price curve will strictly shift up-

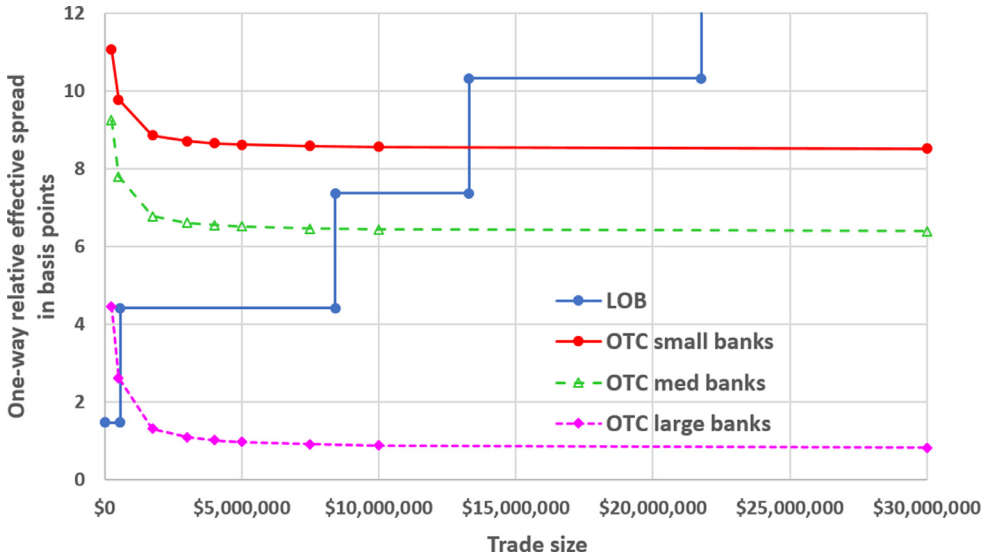

Fig. 9. A theoretical model of parallel LOB and OTC markets.

wards. Therefore, it will intersect the average price curve of the LOB at a strictly greater quantity. □

Intuitively, large banks are discretionary liquidity demander with the most bargaining power, and thus the liquidity suppliers that trade with them have the least bargaining power; therefore under Proposition 5, the large banks will switch to the OTC at the smallest quantity. Conversely, small banks are the discretionary liquidity demander with the least bargaining power, and thus the liquidity suppliers that trade with them have the most bargaining power; therefore Proposition 5, the small banks will switch to the OTC at the largest quantity.

5.4. Numerical illustration

This section provides a numerical illustration of the parallel markets model. The allowed ranges of the model parameters are: (1) all parameters must be strictly positive, (2) ρ, ϕ_1, ϕ_2, and ϕ_3 must be between zero and one, and (3) $v_K > v_1$. From these allowed ranges, we select the following parameter values: v_1 = ¥6.77, v_K = ¥6.79, M = \$50,000,000, K = 80, f = ¥320, C = ¥0.009, ρ = 10%, δ = ¥0.001, ϕ_1 = 0.03, ϕ_2 = 0.24, and ϕ_3 = 0.32. These parameters are calibrated so as to make the theoretical graph (Fig. 9) look roughly similar to the two empirical graphs (Figs. 5 and 6). The reason for doing this is to illustrate the potential realism of the theoretical model.

Fig. 9 illustrates the resulting price functions of the parallel markets model. The x-axis is the trade size, and the y-axis is the one-way relative effective spread in basis points. The blue, solid step function is the LOB price function. It is weakly upward sloping (i.e., larger quantities pay the same or higher spread). The three downward-sloping curves are the OTC price functions. The red solid curve with circle points is for small banks. The green large-dashed curve with triangle points is for medium banks. The purple small-dashed curve with diamond points is for large banks. We see that banks with greater bargaining power pay a smaller spread for all trade sizes. Thus, the figure shows that the theoretical model has all of the qualitative characteristics we found empirically as highlighted in Figs. 5 and 6.

The qualitative characteristics of the theoretical model are due to the structure of the model and are completely insensitive to the particular parameter values selected from the allowed range. For example, having a fixed cost in the OTC market generates a downward-sloping price function. This is because as the trade size increases, the fixed cost becomes a smaller percentage of the trade size. Modifying the fixed cost f and the reservation cost C shifts the intercept and slope of the OTC price curve, but it is always downward sloping for any f and C values. Similarly, the adverse selection feature of the LOB makes the price function upward sloping, as in the Kyle (1985) model. This is because liquidity suppliers rationally anticipate that the optimal informed trader strategy is increasing in trade size. Modifying the probability of an informed trader ρ or the price tick size δ shifts the LOB price function around, but the economics of adverse selection always lead to an upward-sloping price function. Having different degrees of bargaining power for different trading clienteles leads to multiple OTC price functions. Modifying the three bargaining power parameters, ϕ_1, ϕ_2, and ϕ_3, affects the degree of separation between the three OTC price functions, but there are always multiple price functions so long as the bargain power parameters are different. Finally, the parameters v_1, v_K, and M are arbitrary scales for the upper and lower bounds of the v and u distributions, and the parameter K is an arbitrary scale for the number of discrete points in the v distribution—none of which matter for qualitative characteristics of the model.

The figure also illustrates when it is optimal to switch from the LOB to the OTC. For the large banks, the LOB step function and the large bank OTC curve cross at a small trade size. For medium banks, the LOB step function and

Table 7
The relative effective spread difference (LOB - OTC) by trade size and by bank size clienteles.

This table shows the relative effective spread difference (Limit Order Book - Over The Counter Market) in basis points (bps) for Chinese Yuan trades against the US dollar in the Chinese Interbank Foreign Exchange market. The spread difference is broken out by trade size and by bank size clienteles. All computations are based on China Foreign Exchange Trading System (CFETS) intraday transaction data from January 2006 to December 2006. *, **, and *** mean the 10%, 5%, 1% level of significance, respectively, based on the *t*-test of the difference between LOB and OTC.

	Venue difference (LOB minus OTC)		
	Small banks (bps)	Medium banks (bps)	Large banks (bps)
Panel A. During the parallel OTC transition (Jan 2006–Jun 2006)			
< 0.5 mil	−7.32***	−5.87***	−2.70***
0.5–1 mil	−6.22***	−4.87***	−1.66***
1–3 mil	−5.14***	−3.99***	−0.76***
3–5 mil	−3.56***	−2.51***	0.81**
5–10 mil	−2.23***	−1.34**	1.76***
10–20 mil	−1.28**	0.18**	3.29***
> 20 mil	1.60**	3.03**	5.17***
Panel B. After the parallel OTC transition (Jul 2006–Dec 2006)			
< 0.5 mil	−5.02***	−3.79***	−0.42***
0.5–1 mil	−3.96***	−2.73***	0.40**
1–3 mil	−3.28***	−1.77***	1.46***
3–5 mil	−1.57***	−0.41**	2.90***
5–10 mil	−0.22**	0.88***	4.10***
10–20 mil	0.85***	2.12***	5.54***
> 20 mil	3.74***	5.10***	7.24***

the medium bank OTC curve cross at a medium trade size. For small banks, the LOB step function and the small bank OTC curve cross at a large trade size. In summary, banks with more bargaining power switch to the OTC at smaller sizes.

5.5. Empirical test of the new prediction

Finally, we test the new prediction. Table 7 shows the relative effective spread difference (LOB - OTC) by trade size and bank size clienteles. If the difference is positive (negative), it means the relative effective spread in LOB is higher (lower) than in OTC. *, **, and *** mean that the difference is significantly different from zero in a *t*-test at the 10%, 5%, and 1% levels, respectively. We start with Panel A during the transition. Focusing on large banks, for trades smaller than $3 million the spread difference is significantly negative, meaning it is significantly cheaper to trade on the LOB. For trades larger than $3 million, the spread difference is significantly positive, meaning that it is significantly cheaper to trade on the OTC. So, $3 million is the switching quantity for large banks. For medium banks, the spread difference switches from significantly negative to significantly positive at $10 million. For small banks, the spread difference switches from significantly negative to significantly positive at $20 million. Thus we see support for the prediction that banks with more bargaining power switch from the LOB to the OTC at smaller sizes.

We repeat this analysis using the data in Panel B after the transition. Starting with the large banks, the spread difference switches from significantly negative to significantly positive (meaning it switches from the LOB being significantly cheaper to the OTC being significantly cheaper) at $0.5 million. For medium banks, the spread switches from significantly negative to significantly positive at $5 million. For small banks, the spread difference switches from significantly negative to significantly positive at $10 million. Again, we see support for the prediction that banks with more bargaining power switch from the LOB to the OTC at smaller sizes.[31]

In summary, the evidence supports the additional prediction that comes out of the parallel markets model, which in turn supports the model.

5.6. Discussion of differentially and partially informed traders in the OTC

So far we have made the argument that an informed trader who burned a liquidity supplier would be blacklisted, and so informed traders are not consistent with equilibrium in the non-anonymous OTC market. Now we consider what would happen in an alternative model in which a fraction of the OTC trader types was differentially and partially informed.

For simplicity, start from the case in which there is only one liquidity demander class ($J = 1$) with a given degree of bargaining power. Keep the fixed cost per trade, and add multiple informed trader types who are both differentially and partially informed. The result would be a U-shaped price function. To see this, start from a fixed cost per trade and no informed, which yields a downward-sloping price function. Now sprinkle in a tiny amount of risk-neutral informed trader types with noisy signals. All informed trader types maximize their profit by trading the maximum amount because this yields both the highest profit margin and the largest quantity. But this would cause liquidity suppliers to lose money. To obtain an equilibrium, start raising the rightmost part of the price function to obtain a U-shaped function. This has two effects. First, it causes the different informed trader types to spread out over the rising part of the U-shape (i.e., informed trader types with moderate signals would trade smaller amounts to avoid the price impact, and informed trader types with extreme signals would optimally trade larger amounts because they still have a meaningful profit margin, even at higher prices). Second, it allows liquidity suppliers to break even, because the rising part of the U-shape is above their cost (i.e., they obtain profits from uninformed trades in the rising part, which offsets their losses to the informed). As you increase the fraction of informed traders, the rising part of the U-shape increases higher. In the limit, if you zeroed-out the fixed cost, then the price function would be strictly upward-sloped. The only trouble with a U-shaped or upward-sloped price function is that empirical evidence from the OTC market rejects this prediction. Thus, we do not find evidence that informed traders are a meaningful factor in the OTC market.

[31] As a robustness check, we divide the banks into high, medium, and low groups based on trading activity. We obtain the same support for the prediction that banks with more bargaining power switch from the LOB to the OTC at smaller sizes. This support holds true both during the transition and after the transition.

Now return to the three liquidity demander classes ($J = 3$) with different degrees of bargaining power, as in the original model. Keep the fixed cost per trade, and keep the newly added multiple informed trader types who are both differentially and partially informed. The result is three U-shaped price functions: the highest U for small banks, the middle U for medium banks, and the lowest U for large banks. We assume that the upward-sloping LOB price function intersects these three U-shaped functions in the leftmost, downward-sloping part of the respective U-shaped price functions as seems likely from simple geometry. More importantly, this is consistent with the empirical patterns observed in Figs. 5and 6, which present the empirical relative effective spread in both OTC and LOB, in the first six-month period and in the second six-month period, respectively. Then this alternative model would yield the same result as Proposition 5. That is, the large banks with the most "southwestern" price function would intersect the rising LOB price function at the leftmost point (i.e., switch to the OTC at the smallest quantity). The medium banks are the middle would switch at a middle quantity. The small banks with the most "northeastern" price function intersect the rising LOB price function at the rightmost point (i.e., switch at the largest quantity). Thus, the ordering of the critical switching quantities is robust to introducing some differentially and partially informed traders into the OTC market.

6. Conclusion

Originally, the Chinese Interbank FX market was a centralized anonymous limit order book. In 2006, a parallel OTC market was introduced. We use this rare event to answer the following questions. Which trading mechanism would predominate, the LOB or the OTC market? Do these markets have upward- or downward-sloping price functions of trade size? Do these markets have a single price function or multiple price functions of trade size? We find that: (1) the vast majority of trading migrated to the OTC market over a six-month transition; (2) the LOB price function is upward-sloping (i.e., higher transaction costs for larger trades), whereas the OTC price function is downward-sloping (i.e., lower transaction costs for larger trades), and (3) the LOB market has a single price function (i.e., everyone gets the same price function), whereas the OTC market has multiple price functions (i.e., larger banks get better prices). We also develop a theoretical model of parallel markets that can simultaneously explain a single upward-sloping price function on the LOB and multiple downward-sloping price functions on the OTC market. The model generates an additional empirical prediction that the critical trade size at which you switch from the LOB to the OTC is negatively related to your bargaining power. We test this prediction and find support for it.

References

Abudy, M., Wohl, A., 2017. Corporate bond trading on a limit order book exchange. Rev. Finance 22, 1413–1440.

Amihud, Y., 2002. Illiquidity and stock returns: cross section and time-series effects. J. Financ. Mark. 5, 31–56.

Atkeson, A., Eisfeldt, A., Weill, P., 2015. Entry and exit in OTC derivatives markets. Econometrica 83, 2231–2292.

Bessembinder, H., Maxwell, W., 2008. Markets transparency and the corporate bond market. J. Econ. Perspect. 22, 217–234.

Biais, B., Green, R., 2019. The microstructure of the bond market in the 20th century. Rev. Econ. Dyn. 33, 250–271.

Ding, L., 2009. Bid–ask spread and order size in the foreign exchange market: an empirical investigation. Int. J. Finance Econ. 14 (1), 98–105.

Duffie, D., 2012. Dark Markets: Asset Pricing and Information Transmission in Over-the-Counter Markets. Princeton, Princeton University Press.

Duffie, D., Gârleanu, N., Pedersen, L., 2005. Over-the-counter markets. Econometrica 73, 1815–1847.

Duffie, D., Gârleanu, N., Pedersen, L., 2007. Valuation in over-the-counter markets. Rev. Financ. Stud. 20, 1865–1900.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. J. Financ. Econ. 19, 69–90.

Edwards, A., Harris, L., Piwowar, M., 2007. Corporate bond market transaction costs and transparency. J. Finance 62, 1421–1451.

Evans, M.D., Lyons, R.K., 2008. How is macro news transmitted to exchange rates? J. Financ. Econ. 88, 26–50.

Evans, M.D., Rime, D., 2019. Microstructure of foreign exchange markets. Oxford Research Encyclopedia of Economics and Finance doi:10.1093/acrefore/9780190625979.013.31.

Geromichalos, A., Herrenbrueck, L., 2016. Monetary policy, asset prices and liquidity in over-the-counter markets. J. Money Credit Bank. 48, 35–79.

Glosten, L., 1994. Is the electronic open limit order book inevitable? J. Finance 49, 1127–1161.

Glosten, L., Milgrom, P., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. J. Financ. Econ. 14, 71–100.

Goldstein, M., Hotchkiss, E., Sirri, E., 2007. Transparency and liquidity: A controlled experiment on corporate bonds. Rev. Financ. Stud. 20, 235–273.

Goyenko, R., Holden, C., Trzcinka, C., 2009. Do liquidity measures measure liquidity? J. Financ. Econ. 92, 153–181.

Harris, L., 2003. Trading and Exchanges: Market Microstructure for Practitioners. Oxford University Press.

Harris, L., Piwowar, M.S., 2006. Secondary trading costs in the municipal bond market. J. Finance 61, 1361–1397.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Hendershott, T., Madhavan, A., 2015. Click or call? Auction versus search in the over-the-counter market. J. Finance 70, 419–447.

Jain, P., 2005. Financial market design and the equity premium: Electronic versus floor trading. J. Finance 40, 2955–2985.

Jain, P., 2005. Institutional Design and Liquidity at Stock Exchanges Around the World. University of Memphis. Unpublished working paper

King, M., Osler, C., Rime, D., 2012. Foreign exchange market structure, players, and evolution. In: James, J., Marsh, I.W, Sarno, L. (Eds.), Handbook of Exchange Rates. Hoboke, Wiley, pp. 1–44.

King, M., Osler, C., Rime, D., 2013. The market microstructure approach to foreign exchange: looking back and looking forward. J. Int. Money. Finance 38, 95–119.

Kyle, A., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1335.

Lagos, R., Rocheteau, G., 2009. Liquidity in asset markets with search frictions. Econometrica 77, 403–426.

Lagos, R., Zhang, S., 2020. Turnover liquidity and the transmission of monetary policy. Am. Econ. Rev. 110, 1635–1672.

Lee, T., Wang, C., 2018. Why trade over-the-counter? When investors want price fiscrimination. Unpublished working paper. University of Toronto.

Lei, X., Lu, D., Kasa, K., 2020. "Wait and see" or "Fear of Floating"? Macroecon. Dyn. 1–52. doi:10.1017/S1365100520000383.

Lu, D., Xia, T., Zhou, H., 2020. Foreign exchange intervention and monetary policy rules under a managed floating regime: evidence from China. Unpublished working paper. Renmin University.

Lyons, R.K., 2006. The Microstructure Approach to Exchange Rates. Cambridge, MIT Press.

Madhavan, A., 1992. Trading mechanisms in securities markets. J. Finance 47, 607–641.

Mattesini, F., Nosal, E., 2016. Liquidity and asset prices in a monetary model with OTC asset markets. J. Econ. Theory 164, 187–217.

Mende, A., Menkhoff, L., Osler, C., 2004. Asymmetric information and the cross-section of currency spreads. Unpublished working paper. Brandeis University.
Michaelides, A., Milidonis, A., Nishiotis, G.P., 2019. Private information in currency markets. J. Financ. Econ. 131, 643–665.
Osler, C., Mende, A., Menkhoff, L., 2011. Price discovery in currency markets. J. Int. Money Finance 30, 1696–1718.
Schultz, P., 2001. Corporate bond trading costs: a peek behind the curtain. J. Finance 41, 677–698.
Seppi, D., 1997. Liquidity provision with limit orders and strategic specialist. Rev. Financ. Stud. 10, 103–150.
Trejos, A., Wright, R., 2016. Search-based models of money and finance: an integrated approach. J. Econ. Theory 164, 10–31.
Vayanos, D., Wang, J., 2011. Theories of liquidity. Found. Trends Finance 6, 221–317.
Weill, P., 2020. The Search theory of OTC markets. Unpublished working paper. National Bureau of Economic Research.