Oligopolistic market making and inventory heterogeneity*

Yury Olshanskiy

May 10, 2022

Abstract

This paper explores market making under imperfect competition using a novel dataset on individual-level intraday market making. Fixed Effects estimation at aggregate and account levels confirm the significant impact of inventory distribution on market makers' activity and overall liquidity. I propose a dynamic duopoly market making model where inventory distribution shapes agents' strategic behavior and observed liquidity provision. Tight capital constraints lead to a "resting" market maker behavior, while relaxed constraints correlate with a wider bid-ask spread in response to inventory imbalances. Analyzing a grim-trigger non-Markov equilibrium, I find that collusive behavior among market makers raises liquidity prices but reduces their variability.

*Please, check the webiste www.ols-y.com for the up-to-date version of the paper.

[1] Yury Olshanskiy: MIT Sloan School of Managements, email: ols@mit.edu

This paper is a working paper that was started as a master thesis at MIT Sloan. It is actively developing with the companion work in progress "Option Market Making" joint with Hui Chen. I am grateful to my advisors for the paper Hui Chen and Leonid Kogan. I thank Allison Cole, Andrea Hamaui, Joanne Im, Pierre Jaffard, Joel Flynn, Robert C. Merton, Jonathan Parker, Indira Puri, Anton Petukhov, Roman Sigalov, Lawrence Schmidt, Tim de Silva, Justin Rand Scott, Albert Shin, Markus Schwedeler, Adrien Verdelhan, Jiang Wang, Haoxiang Zhu, Jiaheng Yu, and seminar participants at MIT for their thoughtful comments and discussions.

1 Introduction

Almost in all existing security markets, a fundamental role in observed market liquidity is played by special agents, called market makers. These agents continuously operate in the exchanges suggesting being a buyer or a seller for an appropriate price. In many markets, they are responsible for the most of observed trading volume, e.g., Carrion (2013), Menkveld (2013), and they are the agents who provide the best available buy and sell prices most of the time. The singularity of market makers in the trading environment naturally brings us to studying their risks and the way they operate as the main drivers of liquidity.

The ground for understanding incentives and concerns which the market makers face was laid by seminal papers of the late 70s and 80s when academic researchers suggested ways to model dealers that match incoming buy and sell orders. Stoll (1978), Amihud and Mendelson (1980) and others focused on the importance of inventory risk, while Glosten and Milgrom (1985) and Kyle (1985) emphasized information asymmetry. The next generation of papers, e.g. Gromb and Vayanos (2002), Comerton-Forde et al. (2010), focused on financial constraints suggesting that market makers provide less liquidity under tight financial constraints.

This paper continues studying liquidity as a byproduct of market makers' activity. Unlike many preceding papers that operate within environments characterized by market makers functioning as either monopolists or perfectly competitive entities, my focus here is on the scenario of imperfect competition among market makers. This is a particularly crucial consideration in today's modern markets, where increasing technological intricacies and entry barriers make market making accessible to only a select few economic agents. In such a setting, the heterogeneity among market makers assumes a more prominent role in defining market liquidity. This paper introduces a mechanism, which is relatively unconventional in the expanding literature on heterogeneity: it explores how heterogeneity influences the dynamics of oligopolistic behavior. Specifically, the paper concentrates on inventory imbalances, the most easily observable form of heterogeneity, and examines its implications for the competitiveness of market makers.

The paper commences by presenting an example of a substantial market featuring a limited number of active market makers. I conduct an analysis of the Chinese SSE 50 ETF Option market spanning from 2015 to 2017, representing a distinctive option market in China[1]. The uniqueness of the data stems from the comprehensive observation of the intraday behavior of market makers across a diverse set of assets. This enables drawing conclusions not only on their inventory dynamics but also on individual proactiveness in providing liquidity for each asset. The study allows for an examination of market makers' behavior

[1]The Chinese SSE 50 ETF Option market stands out as the exclusive option market on indices in China throughout the specified period.

concurrently, describing instances when they become more or less aggressive on both the bid and ask sides of the spread relative to each other.

The initial observations contradict numerous beliefs regarding market makers' behavior. It becomes evident that in many assets, market makers do not engage in highly aggressive competition with each other. The status of the leading liquidity provider undergoes frequent changes, and the same market makers exhibit the highest trading activity on both sides of the spread. Notably, market makers accumulate substantial positive and negative inventory positions that cannot be swiftly canceled.

The findings prompt consideration of modeling an oligopolistic market-making environment. I propose a model featuring multiple market makers with inventory limits, extending a continuous-time monopolistic model outlined in Avellaneda and Stoikov (2008). This model is grounded in the concept of inventory risk as a fundamental cost of providing liquidity. Market makers operate within a centralized market, offering buy and sell prices for exogenously arriving orders. They share identical preferences and hold homogeneous beliefs about the asset's fundamental value. Competition unfolds as they contend for the best bid and ask prices, managing their cash balances and inventories in a risky asset. The market makers act strategically, considering the ramifications of their actions on the future decisions of their competitors. Both equilibrium types, Markov (without collusion) and non-Markov collusive equilibria, are discussed.

I characterize the equilibrium through a family of functions following a linear system of ordinary differential equations and define the optimal quotes. While the model lacks a closed-form solution, one can employ numerical methods and approximations to explore the equilibrium. In the numerical analysis, I calibrate the model using Chinese option data. Employing Black-Scholes pricing for options and Poisson regression, I estimate the exogenous demand for the asset as a function of deviations of the bid and ask prices from the fundamental option values.

The core insight of the model is that the distribution of inventories significantly influences the strategic behavior of agents. The posted quotes of a market maker are heavily contingent on the inventories of competitors. In addition to the risks faced by market makers, inventories give rise to diverse conditions conducive to monopolistic behavior on both sides of the bid-ask spread. This dynamic can generate additional incentives to pursue specific types of inventory distributions.

In the first part, I demonstrate that market makers' "resting" behavior can arise as an equilibrium when facing tight inventory constraints. a scenario emerges where, during certain periods, only one agent is the most active on both sides of the bid-ask spread, while its competitor refrains from active participation. This result is derived without assuming any collusion-type behavior but arises naturally as an equilibrium outcome. The inventory constraints serve as natural barriers to competition, disrupting the incentives of market makers. Agents aim to push their competitors to their limits and keep them there to act as monopolists on one side of the spread. This can lead them to compete exceptionally aggressively on one side of the spread while readily yielding on the other. To the best of my knowledge, this paper is the first to model market makers as the best liquidity providers on both sides of the spread without assuming preference differences.

Moving forward, the paper explores a baseline scenario where inventory constraints are loosened, yet agents still fail to cooperate. Despite lacking incentives for maintaining a collusive equilibrium, agents engage in a more relaxed form of competition, deriving positive value from their market participation. The envisioned Bertrand competition, wherein agents set prices reflecting the inventory risk associated with buying or selling an asset, remains unattained even when agents are identical. The rationale is that each agent anticipates that if their rivals dominate the competition on one side of the spread "today", they will be less aggressive on that side "tomorrow", enabling the agent to command a higher price. This mechanism undermines the incentives for market makers to compete with each other, resulting in elevated liquidity prices.

The broader spread diminishes the arrival of external orders, prolonging the time spent in all states and further intensifying the incentive to temporarily "lay down arms." This self-amplification mechanism significantly widens the spread across a broad range of inventory distributions. Another noteworthy finding is that the utility of agents in states with substantial negative inventory, when their competitor holds a large positive inventory, is roughly equivalent to a scenario where they both offset their inventories to zero. This sets the stage for potential collusion.

This leads us to the exploration of cooperative equilibria. The paper delves into a grim-trigger cooperative equilibrium, showing the conditions for its sustainability. Market makers engage in trading on different sides of the spread, augmenting the absolute value of their inventories until one of them deviates from the cooperative strategy. This equilibrium exhibits distinct liquidity properties: the bid-ask spread undergoes significant variations with changes in inventory imbalances. This underscores the significance of the game aspect in the non-cooperative equilibrium, emphasizing its high sensitivity to inventory distributions.

The rest of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the dataset and notation used later in the paper. Section 4 summarizes the market liquidity and activity of the market makers. Section 5 introduces the dynamic model of oligopolistic market making. Section 6 discusses calibration of the model based on the observed data. Section 7 characterizes the non-cooperative equilibrium and analyzes its two types. Firstly, it examines the case when capital constraints are tight. Secondly, it studies cases with relaxed capital constraints providing both a numerical solution and discussing the asymptotic approximation of the equilibrium. Section 8 investigates the case of a cooperative equilibrium. Section 9 studies the aggregate market-level liquidity and individual-level decision-making on being active. Section 10 concludes.

2 Literature Review

The models concerning a monopolist market maker facing inventory risk were initially developed in Stoll (1978), Ho and Stoll (1981), and Amihud and Mendelson (1980). This monopolistic setup explores the tradeoff between demand size and inventory holding costs, shaping optimal reservation prices set by market makers. Prices fluctuate in tandem with market maker inventories due to the changing cost of holding additional asset units. The authors of Avellaneda and Stoikov (2008) proposed a dynamic model solvable up to a high-order approximation. Additionally, the study Guéant et al. (2013) analyzed a similar model under financial constraints, forming the theoretical basis for the present work. The primary contribution here involves transitioning to a scenario with multiple market makers and introducing a competitive "game" among them.

Recent inventory risk models developed in Cartea et al. (2014) and Ait-Sahalia and Saglam (2017) primarily focus on examining the behavior of monopolistic High-Frequency Traders (HFTs). Although the approach from Ait-Sahalia and Saglam (2017) involves HFTs trading alongside competitive low-frequency traders, my model bears minimal resemblance to this line of literature.

In models featuring perfectly competitive market makers dealing with information frictions, which originated in Kyle (1985) and Glosten and Milgrom (1985), market makers breaking even accelerates informational costs and market liquidity. However, understanding the variation in liquidity across inventory distribution proves challenging unless specific liquidity providers assume a zero-mass. The present paper bypasses information frictions to maintain an equilibrium where agents are aware of each other's inventories. Here, market makers generate profits, and inventory distribution assumes a central role.

Market makers under imperfect competition have been analyzed in a few papers. The complexity of constructing a tractable yet solvable model and limited data resources on individual market making make testing challenging. For instance, Ho and Stoll (1983) analyzed a two-period model with imperfect competition, concluding that market makers would charge a second price. While this finding is reflected in the equilibrium of this paper, it manifests differently in the dynamic model. The present paper introduces equilibrium prices affected, but not wholly determined, by competitors' inventories due to repeated interactions. Additionally, it introduces market makers who "rest" and those providing liquidity on both sides of the bid-ask spread, an aspect not previously present in the literature to the best of my knowledge.

Oligopolistic financial markets have also been studied in recent papers such as Vayanos (1999a), Du and Zhu (2017), and Kyle et al. (2017). However, these papers model agents as institutional traders, not market makers. There is no distinction between bid and ask prices in these papers, and they lack a strategic view of rivals' future market making actions.

The significance of financial constraints for market makers has been explored in existing literature, as seen in works like Gromb and Vayanos (2002), Kyle and Xiong (2001), and Brunnermeier and Pedersen (2009). Market makers reluctant to bind constraints reduce provided liquidity nearing the limit. The present paper introduces capital constraints, but their role differs. Capital constraints distort agents' strategic behavior, incentivizing them to wait until competitors reach the constraints. Simultaneously, if market makers possess similar inventories close to the boundaries, it does not necessarily diminish competition and decrease liquidity.

The growing empirical literature on market making scrutinizes modern HFT market makers and aspects of their behavior, evident in works like Menkveld (2016), Kirilenko et al. (2017a), Budish et al. (2015a), and Carrion (2013). My paper utilizes a unique dataset closely observing individual-level market maker behavior, offering new insights into modern market operations. Few papers delve into individual-level market making analysis.

The closest empirical paper, Comerton-Forde et al. (2010), investigates inventory roles at an account level in liquidity provision using daily data. They illustrate that market makers widen spreads when accumulating large positions, linking it to capital constraints. In contrast, my paper focuses differently. Firstly, it conducts inter-day analysis and closely tracks variations in market makers' activity. Secondly, large inventory holdings do not necessarily lead market makers to reduce liquidity provision. Agents consider not only their own inventories but also those of other market makers, emphasizing competition's importance. Moreover, this paper observes a broad set of similar assets and individual-level liquidity across them, enabling a dif-in-dif type identification and a more accurate handling of potential inventory endogeneity.

Finally, the authors of Reiss and Werner (1998) examine inter-dealer trading in the London Stock Exchange in the mid-90s. They identify inventory cycles among market makers, showcasing how dealers share risk by selling inventories to each other during large inventory shocks. Although my paper does not document inter-market maker trading, the provided model can explain why inter-dealer trading may not always be favorable. The inventory difference might discipline agents in sharing the market and assist in extracting monopolistic profits from different sides of the spread.

3 Data and Notation

3.1 Data

The trading activity data is sourced from the Chinese SSE 50 ETF Option market, spanning from February 9, 2015, to August 31, 2017.[2] The raw data encompasses all trade and quote information from anonymized major market participants. For each quoted order, details include the contract name, order size, trade direction, and a timestamp down to the millisecond when the order was submitted. Cancellation timestamps are observed for canceled contracts, and for executed orders, only the submission time and confirmation of consumption are noted[3].

The traded contracts comprise call and put options with varying strikes and expiration dates on the SSE 50 ETF, a stock index of the Shanghai Stock Exchange. These represent the first and only standardized options traded on the Chinese market, totaling 1024 different contracts. On average, 81 contracts are traded per day. Each call/put contract has a corresponding put/call contract with the same strike and maturity date. A contract unit consists of 10,000 options on a unit of ETF, and participants can only submit orders with a whole number of contract units.

Contract expiration dates fall on the fourth Wednesday of each expiration month only[4]. Consequently, contracts with arbitrary time to maturity are not observed, but rather a few groups of contracts with the same maturity. The introduction of new contracts followed specific rules, requiring the existence of traded contracts with the current month, next month, and the following two consecutive quarters' maturity, with a sufficiently dense set of strikes around the current value of the ETF stock[5]. New contracts were introduced if the ETF price changed significantly, and there were insufficient strikes in the corresponding direction.

Figure 1 provides a summary of information on the traded contracts, showcasing active trading for all types with nuances in posting and trading frequency and volume between calls and puts, and significant differences among contracts in moneyness and maturity.

Intraday trading occurs during the opening auction from 9:15 to 9:25, and continuous trading sessions from 9:30 to 11:30 and 13:00 to 15:00. The liquidity analysis focuses on the latter sessions. Prices are denominated in renminbi with a tick size of one basis point.

The observed market participants are accounts with the largest share of trading volume[6]. These entities

[2]Professor Hui Chen kindly provided the data. This version of the paper relies solely on quote data, with further empirical analysis using a more extensive dataset covering details about executed orders.

[3]In cases of partial consumption, cancellation time for the residual is observed

[4]If the expiration date coincides with a holiday, it is correspondingly postponed

[5]Exact rules can be found on the exchange's website; however, they are not pivotal for the paper's analysis

[6]Exact numbers will be clarified.

include major institutions and proprietary accounts, with several engaging in high-frequency trading, posting and canceling contracts with millisecond frequency. Some consistently act as market makers, providing bid and ask quotes.

I estimate the Greeks of traded contracts based on the Black-Scholes model, extracting implied values for the stock price, its volatility, and the discount factor from the observed data at ten-minute frequency intervals. Detailed information on the estimation is provided in section 6, where the calibration of the model is discussed.

Lastly, the dataset is extensive, comprising over a billion observations in raw format, expanding further when constructing limit order books and other necessary add-ons. Due to the substantial time required for data processing, many variables are aggregated to larger time intervals. An enhanced cost analysis could further benefit from more precisely identified strategies.

3.2 Notation

Let $\mathcal{T}$ denote the set of observation times. While the market is continuously observed over time, the data requires aggregation into time intervals for processing. The specific interval, whether one or ten minutes, or the spot time at the interval's conclusion, for $t \in \mathcal{T}$ will be clarified based on the context. $\mathcal{T}^1$ encompasses a defined trading date with 240-minute intervals per day. Similarly, $\mathcal{T}^{10}$ comprises 24 ten-minute intervals per day.

Denote the set of available maturity dates at time $t \in \mathcal{T}$ by $\mathbb{T}_t$. Denote the set of available strikes with date $T \in \mathbb{T}_t$ by $\mathcal{K}_t(T)$. Hence, the total set of traded contracts at time t can be defined as $\mathcal{C}_t = \left\{\{C_t(K,T), Pu_t(K<T)\}_{K\in\mathcal{K}_t(T)}\right\}_{T\in\mathbb{T}_t}$, where $C_t(K,T)/Pu_t(K,T)$ is a call/put contract with strike K and maturity date T respectively.

Suppose that for every moment s during time interval t there is the best ask $Ask^*_{k,s}$ and best bid $Bid^*_{k,s}$ prices offered in the market, $Ask^*_{k,s} > Bid^*_{k,s}$, $k \in \mathcal{C}_t$. Define price of the asset at t as the average mid-quote price,

$$P_{k,t} = \frac{1}{dt} \cdot \int_{s=t-1}^{t} \frac{Ask^*_{k,s} + Bid^*_{k,s}}{2} ds,$$

where dt is the length of interval t, which equals to t. Ignore contract k at time t if ask or bid price is not defined for some time over the period.[7]

[7] we ignore a few cases using the criteria

3.2.1 Baskets of contracts

The option contracts vary along three dimensions: moneyness, time to maturity, and the type of option. Based on these dimensions, at each time $t \in \mathcal{T}$, they are uniquely categorized into one of the baskets in the basket sets $\mathcal{B}_{1,t}$ and $\mathcal{B}_{2,t}$, denoted as $\mathcal{C}_t = \sqcup_{B \in \mathcal{B}_{1,t}} B = \sqcup_{B \in \mathcal{B}_{2,t}} B$. The definitions of the basket sets will be provided below. Contracts differing only in type, i.e., being call or put, are occasionally grouped into one basket, as a put and a call with the same maturity and strike can be readily replicated by one another with access to the underlying stock and risk-free asset[8]. This grouping results in less refined sets of baskets, denoted as $\mathcal{B}^u_{1,t}$ and $\mathcal{B}^u_2$.

I define moneyness of an option based on its delta, following definition suggested in Johnson et al. (2016):

$$\begin{cases} \text{If Delta of the contract is in } [.375, .625] & \text{then it is at the money (At-)contract} \\ \text{If Delta of the contract is smaller than } .375 & \text{then it is out the money (Out-)contract} \\ \text{If Delta of the contract is larger than } .625 & \text{then it is out the money (In-)contract} \end{cases}$$

Time to maturity is measured as the number of days for the contract before expiration and changes over the life of a contract. I use two classification of contracts based on time-to-maturity. Firstly, I follow Johnson et al. (2016),

$$\begin{cases} \text{if Maturity of the contract is larger than 40 days, it is a } \textbf{long term (LT-)contract} \\ \text{if Maturity of the contract is smaller than 40 days, it is a } \textbf{short term (ST-)contract.} \end{cases}$$

This classification forms twelve baskets of contracts: $\mathcal{B}_{1,t} = \{At, Out, In\} \times \{LT, ST\} \times \{Call, Put\}$. Given the frequency of our data we can afford even finer partition based on time-to-maturity. I define other 8 groups based on maturity, specified in Table 1 . Based on these groups I form baskets $\mathcal{B}_{2,t} = \{At, Out, In\} \times \{T1 : T8\} \times \{Call, Put\}$.

Group	T1	T2	T3	T4	T5	T6	T7	T8
Time−to−Maturity, t	$1 \le t \le 9$	$10 \le t$ $t \le 19$	$20 \le t$ $t \le 29$	$30 \le t$ $t \le 39$	$40 \le t$ $t \le 49$	$50 \le t$ $t \le 59$	$60 \le t$ $t \le 100$	$t > 100$

Table 1: Contract groups based on maturity

[8]Though short-selling constraints for the ETF can create significant frictions

Last, for any call/put-option contract with strike K and maturity T, we observe a put/call-option contract with the same strike and maturity. At the same time, both contracts are either at-the-money, or have opposite moneyness: one is Out while another is In. Aggregate In- and Out- contracts into (NotAt-) not-at-the-money contracts. Then we can classify each of the contracts into the same basket from $\mathcal{B}_1^u = \{At, NotAt\} \times \{ST, LT\}$ or $\mathcal{B}_2^u = \{At, NotAt\} \times \{T1 : T8\}$, so that put and calls with same same strike and maturity are always in the same basket.

3.2.2 Inventories and Liquidity Provision of MMs

Denote the set of agents by I and subset of market makers by MM, $MM \subset I$. For every agent $i \in I$ and $t \in \mathcal{T}$ define net buy during interval t, i.e., difference between contracts bought and sold by an agent, of a contract $k \in \mathcal{C}_t$ by $NB_{i,k,t}$. Since I do not observe time of executed orders, I have to assume that orders were immediately executed, i.e., I count executed contracts with submission time in t. Thus, there can be mis-measurement if contracts were executed later. Net inventories at t are defined as sum of net buys from the previous times including the last time interval t, $Q_{i,k,t} = \sum_{t' \in \mathcal{T}, t' \leq t} NB_{i,k,t'}$[9]. Denote the total inventory of market makers by $Q_{k,t}^{mm} = \sum_{i \in MM} Q_{i,k,t}$.

A big long position in a Call/Put can be perfectly hedged by big short positional in a Put/Call with the same strike and maturity date combined with ETF stock and bond. That makes us thinking about an aggregate inventory of an agent in the couple of the contracts. Define $Q_{i,(K,T),t}^u$ as the difference in inventory of call and put with the strike $K \in \mathcal{K}_t$ and maturity date $T \in \mathbb{T}_t$, $Q_{i,(K,T),t}^u = Q_{i,C(K,T),t} - Q_{i,Pu(K,T),t}$.

Based on cancelled contracts for which I observe both submission and cancelation time-stamps, at each spot time s for every contract $k \in \mathcal{C}_s$, I can construct a limit order book posted by agent, which consists of set of pairs, $\{Bid_{i,k,s}^n, V_{i,k,s}^{a,n}\}_{n \in N_s^b}$ and $\{Ask_{i,k,s}^n, V_{i,k,s}^{b,n}\}_{n \in N_{i,k,s}^b}$, buy/sell price on which agent has posted orders together with total posted volume on the prices. $N_{i,k,s}^a / N_{i,k,s}^b$ is the total number of different prices assuming that $Bid_{i,k,s}^n / Ask_{i,k,s}^n$ are ranged from the best to the worst price.[10]. Define $Bid_{i,k,s}^* / Ask_{i,k,s}^*$ as the best suggested prices by an agent. This limit order book is incomplete because I do not observe the traded assets, but it is not a very bad approximation because agents update it more frequently than they have executed orders often with the same prices as executed orders. When agent is active in a contract it is often when $N_{i,k,s}^z$, $z = a, b$, is larger than one with minimal difference in the prices. That means, hopefully, this sometimes incomplete limit order book captures well liquidity provision by an agent.

I use a bunch of liquidity measures built on the limit order book to evaluate market makers' activeness

[9] I also count purchased and sold contracts in the opening auction.

[10] i.e., from largest/smallest to the smallest/largest prices for Buy/Sell order. $Bid_{i,k,s}^n / Ask_{i,k,s}^n$ is n-th best bid/ask levels posted by agent.

in a contract. Define provided ask/bid prices and spread by an agent at time interval t taking average of the spot prices

$$Ask^n_{i,k,t} = \frac{1}{dt^{n,a}_{i,k,t}} \cdot \int_{s=t-1}^{t} Ask^n_{i,k,s} ds \qquad Bid^n_{i,k,t} = \frac{1}{dt^{n,b}_{i,k,t}} \cdot \int_{s=t-1}^{t} Bid^n_{i,k,s} ds$$

$$Spread_{i,k,t} = \frac{1}{dt^*_{i,k,t}} \cdot \int_{s=t-1}^{t} (Ask^*_{i,k,s} - Bid^*_{i,k,s}) ds,$$

where $dt^{n,z}_{i,k,t}$, $z = a,b$, is the total amount of time agent was providing this type of liquidity. Usually if a market maker is active in a contract $dt^*_{i,k,t}$ eqauls to the whole length of interval, i.e., to one or ten minutes. Next, suppose we observe the aggregate limit order book, which will be discussed in next section. Define average volume provided by an agent on the best price by

$$V^{a,*}_{i,k,t} = \frac{1}{dt} \cdot \int_{s=t-1}^{t} V^{a,1}_{i,k,s} \cdot \mathbb{I}\{(Ask^*_{i,k,s} = Ask^*_{k,s})\} ds,$$

where $Ask^*_{k,s}$ is the best ask price at time s and dt is the length of time interval t. Similarly, define volume on best bid $V^{b,*}_{i,k,t}$, and volumes on nth-best prices:

$$V^{a,n}_{i,k,t} = \frac{1}{dt} \cdot \int_{s=t-1}^{t} \sum_{n' \in N^a_{i,k,s}} V^{a,n'}_{i,k,s} \cdot \mathbb{I}\{(Ask^{n'}_{i,k,s} = Ask^n_{k,s})\} ds,$$

where $Ask^n_{k,s}$ is n-th best price in the aggregate limit order book. Finally define time spent on n-th best price simply as

$$TB^{a,n}_{i,k,t} = \frac{1}{dt} \cdot \int_{s=t-1}^{t} \sum_{n' \in N^a_{i,k,s}} 1 \cdot \mathbb{I}\{(Ask^{n'}_{i,k,s} = Ask^n_{k,s})\} ds.$$

3.2.3 Liquidity Measures in the option market

I do not have direct access to the aggregate limit order book posted in the market. Instead, I construct an approximation based on all visible orders, which primarily consist of orders by major market participants. Typically, this approximation includes at least three layers of different prices, with more than ten layers for actively traded assets, especially short-term in-the-money contracts. While the data covers orders from major market makers, it is important to note that this aggregate limit order book approximation may introduce bias as it does not account for executed orders.[11]

By aggregating all orders posted in the market, including those that were cancelled, similar to the previous section, I construct sets denoted as $\{Bid^n_{k,s}, V^{a,n}_{k,s}\}_{n \in N^b_s}$ and $\{Ask^n_{k,s}, V^{b,n}_{k,s}\}_{n \in N^b_s}$, where the best available prices are represented by $Bid^*_{k,s}/Ask^*_{k,s}$. Taking the average over a time interval t, I define the prices and

[11] In the new version of the paper, that is not distributed yet, I do observe the executed orders reducing the bias.

spread at time t:

$$Ask_{k,t}^{n} = \frac{1}{dt_{k,t}^{n,a}} \cdot \int_{s=t-1}^{t} Ask_{k,s}^{n} ds \qquad Bid_{k,t}^{n} = \frac{1}{dt_{k,t}^{n,b}} \cdot \int_{s=t-1}^{t} Bid_{k,s}^{n} ds$$

$$Spread_{k,t} = \frac{1}{dt_{k,t}^{*}} \cdot \int_{s=t-1}^{t} (Ask_{k,s}^{*} - Bid_{k,s}^{*}) ds,$$

Since we are specifically interested in the behavior of market makers, we define in the same way $Ask_{k,t}^{mm,n}$, $Bid_{k,t}^{mm,n}$, and $Spread_{k,t}^{mm,n}$ based on orders posted by market makers only. Almost all the time, the best bid and ask in the aggregate limit order book coincide with the best bid and ask provided by market makers: $Ask_{k,t}^{mm,*} = Ask_{k,t}^{*}$ and $Bid_{k,t}^{mm,*} = Bid_{k,t}^{*}$. Lastly, we are interested in the arrival of orders from other market participant. For those orders only, denote $VB^{mm}k,t$ and $VS^{mm}k,t$ as the volume of buy and sell orders, respectively, that market makers match during t.

4 Empirical observations

4.1 Descriptives

In this section, I add descriptive facts about liquidity measures and agents, which are needed later in the paper. Let us start with the aggregate liquidity measures. Figure 3 shows the distribution of the bid-ask spread in the data. The highest mass is concentrated between 1 and 50 bps. Figure 23 shows that spread varied over the years with the highest values in 2015, which can be both because of increasing competition between market makers and because of decreased uncertainty. Distribution of spread also massively depends on the type of contracts, as shown in Figure 24 . Figure 25 shows that spread does not vary a lot intraday except for the opening and closing times.

Liquidity is also described by the volume of the best prices and distance between the best prices. Figure 26 shows that the distance between the first and the second-best prices is substantial and equals to half of the spread on average. Similarly, the distance between the second best and third best prices.

The arrival of executed orders is different by a group of contracts. Figure 4 summarizes information on the distribution of executed orders' sizes. Demand on contracts increases when they approach the maturity date. At-the-money contracts are the most frequently traded. Out-the-money contracts, as the cheapest contracts, are traded by the largest blocks.

4.2 Competitiveness of Market Makers

To measure competitiveness of market makers, one can look at two aspects of market making. The first is the trading volume that the market makers provide and the trading opportunities which they suggest. i.e., the spread and the volume on the quotes which the market makers ensure.

The observed trading volume suggests that there are a few agents which are active in a contract over a day. Figure 5 shows that from two to four market makers are responsible for at least half of daily trading volume[12] in a specific contract. During a ten minutes interval these are usually at most just two agents and very rare more than three. (Figure 27)

There are usually leading market makers who provide the best price on a specific contract over a long period. Figure 6b illustrates this phenomenon. Over a 10-minutes interval, there is almost always an agent who spends at least 200 seconds (one-third of the time) on the best quote. Roughly in 50% cases, there is an agent who spends at least 500 seconds on the best quote. The result does not significantly depend on the maturity or type of contract.

If market makers were competing with each other aggressively, it should be observed that many market makers either changed each other on the best prices frequently or posted the best price at the same time. Figure 6a shows that at least 60% of the time, a unique agent is suggesting the best price for a substantial time. Roughly in 25% cases, there two such agents. That is very unlikely that there no or many such agents. This result does not depend on the type of contract. Therefore it should not be the case that agents aggressively change each other on the best quotes. Figure 8 shows that the average number of agents on the best, second best, and third best quotes of a contract is between one and two. The average number of agents on the best three prices can be even less than three agents. In many cases, the same agents provide both the first and the second-best bids. Therefore, though most of the market makers actively trade every contract at some point, they usually do not split the best quotes with each other.

At the same time, there can be other forms of competition. For example, agents can provide different volume on the quotes. If the posted volume on the best quote is small then many large orders will hit through the best quote. Figure 9 shows that it is not the case in the market. Average volume on the best quotes is at least twice larger than average executed order. That means that hiting through the best price is not very likely.

Finally, that is a typical case when the same agents are the most active both on the best bid and best ask. Figure 7 illustrates this phenomenon plotting the expected time to suggest the best ask/bid price conditional

[12] formal definition of the reported variable is the minimum number of agents during a day needed to explain half of observed trading volume in the contract

on spending time on the best bid/ask quote. At the plot, I intentionally drop the points with the largest mass when an agent spends zero time or all ten minutes on the conditional quote because the graph has a discontinuity at the points. We can see that the average time being best on the opposite quote smoothly increasing function. That could be explained for the small values of the X-axis as reversal action. i.e., a market maker trades in one direction in the first half of the ten minutes, accumulates inventories, and cancels the position in the remaining time. Nevertheless, this logic cannot describe the positive dependence in the right tail of the graph. If an agent spends almost all ten minutes interval on the best bid, when why would she spend the 200 seconds on the best ask?

We must mention that being most active on both sides of spread is not a dominating strategy. It roughly happens in one out of three cases. That raises an important question, why can market makers be sometimes aggressive on both sides of spread and sometimes not?

To sum up, one can say that there is very few agents which actually aggressively compete in a contract. Many market makers are "resting" taking position on quotes which are far from the best quotes. At the same time, other market makers can be active on both sides of spread at the same time.

4.3 Inventories of Market makers

The observed market makers do not just play the role of matching orders on different sides. They also accumulate quite substantial positions. Figures 11a and 28 shows that in all type of contracts there is at least one agent holding or shorting more than 1000 units of contracts[13]. For most of the contracts, excluding very actively traded short-term contracts[14], this is comparable with total daily trading volume in the contract. That does not necessarily mean that market makers hold very high inventory risk. A big position in one of contracts can be hedged by positions in other contracts. A perfect hedge for a call/put contract can be achieved only by a portfolio with opposite sign of put/call contract together with buying or shorting stock or bond.

Shorting a stock might be costly for market makers what creates asymmetry in hedging costs of long and short inventory positions. Nevertheless, even ignoring the cost, a substantial risk is observed in the hands of market makers because their Call-Put positions do not clear each other. Figures 11b and 11c shows that $Q^u_{i,(K,T),t}$ achieves large negative and positive values, i.e., there are market makers with huge imbalances in the specific type of risk. An alternative is to look at the Greeks of portfolios held by market makers. If all positions are hedged well enough, then one should expect neither big variation in the Greeks nor high

[13] i.e., 10^7 real options

[14] especially at the money

absolute values of them. Figures 10a–10b show that they do vary and have big absolute values.

The next question is how many market makers are actually accumulating the substantial positions and how large is the difference in their positions. Say that a trader has the minimal/maximal position is a contract if her contract inventory is smaller/larger than the inventories of all other market makers. We do not want to treat market makers differently if they have the inventory difference in just a few units of the contract. To deal with the problem, extend the set of minimal/maximal-inventory market makers by those whose inventories are different from the minimum/maximum inventory not more than an average trading volume in the contract of this type.[15]

Figure 12a shows that for most of the contracts, there exists a unique minimal market maker. The situation slightly improves when At- and In- contracts approach their expiration. Figure 31 , in the Appendix, shows a similar picture for the maximal market maker. Figure 12b shows that most of the time, there are not many agents with a substantial inventory in a contract. I define an inventory position to be substantial if its absolute value is larger than trading volume[16] over a 30 minutes interval. For the very long-term contracts, one sees four-five agents satisfying the criteria, because the trading volume is small. For the actively traded contracts[17] these are just one or two agents on the positive and negative sides of inventories.

To sum up, observed market makers accumulate significant imbalanced inventory positions getting exposure to different types of risk. They vary over time. At the same time, there are not many market makers which end up with the large inventories in a specific contract. Among them, most of the time, there is a unique maximal/minimal market maker. The next step is to link the market makers' positions to liquidity provision by them.

4.4 Predictive Power of Inventories

The goal of this section is to identify that inventories have predictive power for the liquidity provision in a contract. No casual relationship is claimed at this point.

Figure 13 shows that maximal and minimal inventories are associated with a high probability of being active on the provided bid-ask spread. As I discussed before, approximately in 50% cases, the same market maker is active on both sides of the spread. This rule continues holding for the maximal and minimal market makers, as shown in Appendix, Figure 30 .

Next, Figure 14 illustrates that trading volume is also associated with being maximal or minimal market

[15] I use baskets $\mathcal{B}_{2,t}$ to define similar types of the contracts. I take the respective year and the quarter to get the average volume over the similar contracts.

[16] the trading volume, as before, is taken for the contracts in the same bin $\mathcal{B}_2$ over the quarter

[17] Figure 1b shows that these are contracts with maturity smaller than 40 or 50 days.

maker but to a smaller magnitude. We can also see that probability to be the largest net buyer or net seller is between 20-30% for the most of the contracts.

Substituting the inventory position in a contract by the call-put inventory difference discussed before, $Q^u_{i,(K,T),t}$, does not change the probabilities significantly, look at similar Figures 33 - 34 in Appendix.

To summarize, having the largest (smallest) inventory positions makes probability to be active larger. Nevertheless, it does not guarantee to be the most active on any side of the spread. Moreover, the effect is hard to distinguish for the buy and sell activities. This finding is somewhat contradicting to the results of Demsetz (1968), where they predict decreasing activity of the market makers when they accumulate substantial inventories. Moreover, the intention of the market makers to reverse back to zero are scarcely observed.

5 Model Description

5.1 Market Makers and Orders Arrival

I assume that time is continuous, $t = [0, T]$ and the fundamental value of a traded asset follows arithmetic Brownian motion

$$dS_t = \sigma dW_t. \tag{1}$$

The asset has no dividends and pays S_T at time T. Consider two agents, market makers, $i = 1, 2$, with identical utility functions which has constant absolute risk aversion γ. Denote amount of cash agent i holds by $X_{i,t} \in \mathbb{R}$ and denote her inventory of the asset by $q_{i,t} \in \mathbb{N}$. Furthermore, denote the agents' inventory constraint by $\bar{Q} \in \mathbb{N}$, $|q_{i,t}| \leq \bar{Q}$.

Agents continuously provide liquidity posting bid/ask prices $S^a_{i,t}/S^b_{i,t}$ unless they achieve the bound. Say that $S^a_{i,t} = \infty$ if $q_{i,t} = -\bar{Q}$ and $S^b_{i,t} = -\infty$ if $q_{i,t} = \bar{Q}$. Denote the distance between the prices and fundamental value, known as *reservation price*, by $\delta^a_{i,t} := S^a_{i,t} - S$ and $\delta^b_{i,t} := S - S^b_{i,t}$. Then $\delta^z_t := \min\{\delta^z_{1,t}, \delta^z_{2,t}\}$ are the best quotes' distances.

Let $\lambda(x) := A \cdot e^{-\kappa \cdot x}$ for some positive constants A and κ. At every instant with probability $\lambda(\delta^a_t) \cdot dt$ ($\lambda(\delta^b_t) \cdot dt$) an outside trader buys (sells) a unit of the asset at the best available price. Denote the counting processes associated with intensity δ^z_t by N^z_t, i.e., N^b_t and N^a_t are the aggregate inflow and outflow of the assets from the hands of the market makers:

$$q_{1,t} + q_{2,t} = q_{1,0} + q_{2,0} + N^b_t - N^a_t.$$

Let us use matrix notation for cash and asset holdings: $X_t = [X_{1,t}, X_{2,t}]'$, $q_t = [q_{1,t}, q_{2,t}]'$. The information on who posts the best quote at instant t is summarized in vector $\Omega^z_t \in \mathbb{R}^2$, $z = a, b$:

$$\Omega^z_t = \begin{cases} [1,0]^\top & \delta^z_1 < \delta^z_2 \\ [0,1]^\top & \delta^z_1 > \delta^z_2 \\ [\frac{1}{2}, \frac{1}{2}]^\top & \delta^z_1 = \delta^z_2 \end{cases}$$

Then cash balances of agents satisfy the identity

$$dX_t = (S_t + \delta^a_t) \cdot \Omega^a_t \cdot dN^a_t - (S_t - \delta^b_t) \cdot \Omega^b_t \cdot dN^b_t, \tag{2}$$

and the inventories evolve according to the relation

$$dq_t = \Omega^b_t \cdot dN^b_t - \Omega^a_t \cdot dN^a_t \tag{3}$$

5.2 The game and MM's problem

The game between market makers is defined as follows. At every instant t each agent i decides on a pure action $(\delta^a_{i,t}, \delta^b_{i,t})$ conditional on observed state variables (q_t, X_t, S_t). Later I will allow decision to be dependent on pre-history $\mathcal{H}_s = \{(q_s, X_s, S_s, \{\delta^a_{j,t}, \delta^b_{j,t}\}_{j=1,2})\}_{s \in [0,t)}$. The equilibrium $\{\delta^{a*}_{j,t}, \delta^{b*}_{j,t}\}_{j=1,2}$ must constitute a subgame perfect Nash Equilibrium[18], i.e., a single deviation of any agent to $(\tilde{\delta}^a_{i,t}, \tilde{\delta}^b_{i,t}) \neq (\delta^{a*}_{i,t}, \delta^{b*}_{i,t})$ should not provide a utility gain conditional on her rival j continuing playing $(\delta^{a*}_{j,t}, \delta^{b*}_{j,t})$.

The problem of agent i is

$$u_i(t, X_t, q_t, S_t) = \max_{\{\delta^a_{i,s}, \delta^b_{i,s}\}^T_{t=s}} \mathbb{E}_t \Big[-\exp\{-\gamma \cdot (X_{i,T} + q_{i,T} \cdot S_T)\} \Big]$$
$$s.t. \;\; \{\delta^a_{j,s}{}^*, \delta^b_{j,s}{}^*\}^T_{t=s}, (1), (2), (3),$$

where $\delta^a_{i,s}, \delta^b_{i,s}$ and $\{\delta^a_{j,s}{}^*, \delta^b_{j,s}{}^*\}$ are adapted processes. $\{\delta^a_{j,s}{}^*, \delta^b_{j,s}{}^*\}$ are the optimal quotes of the rival j. Agents are forward-looking in two aspects. First of all, they trade to increase their cash-holdings and minimize risk of holding large inventory position. Second, they compete strategically taking into account the dependence of the future competitors' actions on their current decisions.

Inter-dealer trading is not allowed unless stated otherwise. As it becomes clear later, it is not a strong

[18] now I ignore some details of the equilibrium definition in the continuous time, the answers on related issues are discussed in unpublished notes Rosu (2006)

assumption, because incentives for inter-dealer trading are minimal.

Let me call an equilibrium where agents' actions do not depend on pre-history $\mathcal{H}_s$, i.e., a Markov equilibrium, *non-cooperative equilibrium*. It exists since processes for orders arrival, and evaluation of fundamental value are Markov processes. The model allows non-Markov equilibria too. For example, where agents agree to trade on different sides of spread or different times with punishment if any player defected in the past. Such equilibria I will call *cooperative equilibria*.

6 Calibration of the model

The model presented in the paper does not have a closed-form solution but allows numerical solution. Though the findings of the model seem to be unchanged over a broad set of parameters, I use data on the Chinese option market to calibrate the model for the baseline specification.

The identification of the model requires four parameters. I do not estimate γ, but estimate the rest three parameters A, κ, σ. The main problem one faces applying the model to the data is that all the parameters vary for a contract. Firstly, because of variation in demand on the contract and its properties over its life-span. Next, because of changes in global uncertainty. For example, local-crisis events can affect option demand parameters A and κ substantially, as well as the volatility of the contracts. Expected by market makers, fluctuations in the parameters may affect the equilibrium of the model in unpredicted directions.

To partially deal with the problem, I try to look at the demands for not a specific contract but for a specific type of contract and similar volatility times. The first can create minor problems because, in the discussed model, agents trade the same contract up to the expiration date. If agents realize that the contract will change properties, they will take it into account, trading it today. However, that should not be an issue with the observed arrival rate of orders. The model converges to a steady-state briefly, and these long-term changes do not affect equilibrium pricing. I also exclude very short-term contracts, with less than ten days before maturity, from consideration. The expected jumps in aggregate volatility can make even "similar volatility time" different. To deal with the problem, I try to narrow attention to contracts over the 2016-2017 years, when the Chinese market was relatively stable, as it is shown in Figure 35 .

Observed option contracts might have different types and carry a different risk for the agents, but the fundamental value is driven by the same factors for all of them. These are characteristics related to the underlying ETF and the interest rate. Together they define the fundamental value of each of the contracts. Since there are many contracts for which market price is observed at the same time, I have more than enough observations and can fit the necessary risk components. The residual from the over-identified model is a gift because it allows us to separate the fundamental value of the asset and the market price. The deviation of the latter from the former can be considered as the result of frictions defined by the behavior of market makers.

To be more concrete, let us assume that the fundamental value of every contract is defined by the Black-Scholes (BS) formula:

$$\begin{aligned} C_t(T,K;\sigma_{t,T},S_t,D_{t,T}) &= \Phi(d_+)S_t - \Phi(d_-)D_{t,T}K \\ Pu_t(T,K;\sigma_{t,T},S_t,D_{t,T}) &= (\Phi(d_+)-1)\cdot S_t - (\Phi(d_-)-1)\cdot D_{t,T}K \end{aligned} \tag{4}$$

where $C_t(T,K,\sigma_{t,T},S_t)$, $Pu_t(T,K,\sigma_{t,T},S_t)$ are prices of call and put options respectively with strike K and maturity date T. They are measured as the average mid-quote price over a time interval around t. S_t is the spot price of ETF,$D_{t,T}$ is the discount factor from t to T, $\Phi(\cdot)$ is is the cumulative distribution function of the standard normal distribution, and $d_\pm = \frac{1}{\sigma_{t,T}}\left(\log\left(\frac{S_t}{D_{t,T}K}\right) \pm \frac{1}{2}\sigma_{t,T}\right)$.

Using cross-section of contracts one can evaluate agents' supposition about parameters of the ETF and discount rate at time t, for example, using the non-linear least method (NLS):

$$\begin{aligned} \min_{S_t,\sigma_{t,T},D_{t,T}} \Bigg\{ \sum_{K\in\mathcal{K}_t} &(C_t(T,K) - \Phi(d_+)S_t + \Phi(d_-)D_{t,T}K)^2 + \\ &(Pu_t(T,K) + \Phi(-d_+)S_t - \Phi(-d_-)D_{t,T}K)^2 \Bigg\}, \end{aligned}$$

where $\mathcal{K}_t$ is the set of available strikes at time t.[19] One can also note that $\sigma_{t,T}$ must be equal to $\sigma\sqrt{T-t}$, what can reduce number of parameters, but since $T-t$ can be measured non-linearly, I save time superscript to allow better fitting of the model. Though this NLS approach works well, I stopped on three step procedure described below which turned out to be less sensitive to measurement errors related to rare outliers in the mid-quote price.

The first step relies on the Call-Put parity. Taking into a group all contracts with same expiration date T and matching call and put contracts with the same strikes there should be the linear relationship for different strikes:

$$C_t(T,K) - Pu_t(T,K) = S_t - K\cdot D_{t,T} + \varepsilon_{K,T} \qquad T\in\mathbb{T}_t \tag{5}$$

Though no-arbitrage condition requires $\varepsilon_{K,T}$ to be zero, in practice, they are not. The reason is not necessarily existing of arbitrage opportunities[20], but the imprecise definition of the correct price by the mid-quote price. If the spread is wide, the difference between the mid-quote price and fundamental value, which must be somewhere between Ask and Bid, can be substantial. Also, note that the prices I use are not the spot

[19] in the sample at every time t call option with strike K is traded iff call option with strike K is also traded

[20] which rarely actually exists

prices, but averages in the interval, that can also introduce additional noise in the estimates.

Thus, running the regression for fixed T, one can identify suspicious residuals $\varepsilon_{K,T}$ and contracts, which may spoil the estimate of discount rate and price of ETF. At the second step, I rerun the same regression ignoring strikes K for which $\varepsilon_{K,T}$ was anomaly large by absolute value.[21]. Doing so, I get estimated $S_t(T)$ and $D_{t,T}$ and an almost perfect fit for call-put parity. Figure 15b shows that it is smaller than ten basis points for almost all observations, which is neglectable given that the median value of the spread is around 18bp, with 80%-quantile equal to 106bp. Note that estimate $S_t(T)$ must be equal to the same number, price of the ETF S_t, for each T, in practice there can be small fluctuation because of reasons discussed above. On top of that, for very long-term contracts Call options tend to be relatively cheaper.[22] Take the average of those to define the price of the stock at t.

Finally, having good estimate for price of stock S_t and discount rate $D_{t,T}$, $T \in \mathbb{T}_t$, one can get implied volatility $\sigma_{t,T}$ using NLS:

$$\min_{\sigma_{t,T}} \Big\{ \sum_{K \in \mathcal{K}_t} (C_t(T,K) - \Phi(d_+)S_t - \Phi(d_-)D_{t,T}K)^2 + \\ (Pu_t(T,K) - \Phi(-d_+)S_t - \Phi(-d_-)D_{t,T}K)^2 \Big\}.$$

Figure 15b shows that the three-step procedure with the final estimates $\hat{S}_t$, $\{\hat{\sigma}_{t,T}, \hat{D}_{t,T}\}_{T \in \mathbb{T}_t}$ worked reasonably well in my sample. Define $\hat{P}_{k,t}$ as the price of the contract k based on BS formula (4) and the estimated parameters. We can see that the difference between $\hat{P}_{k,t}$ and $P_{k,t}$ is not higher than twenty basis points in the most cases, which is an excellent result given the distribution of spreads. The residual variation in the difference is also good for us because it can be used for estimation of the demand curve based on the shift in the price.

Suppose that for similar contracts in similar times the arrival of buy orders to market makers is described as in the model by a Poisson process with the same intensity $A \cdot \exp\{-\kappa \cdot (Ask_{k,s}^{mm,*} - P_{k,s}^0)\}$, where $Ask_{k,s}^{mm,*}$ and $P_{k,s}^0$ are the best ask price provided by the market makers and fundamental value of the contract at spot time s. Then if $Ask_{k,s}^{mm,*} - P_{k,s}^0$ stays constant over a period of time t then total number of buy orders from outside during t, $B_{k,t}$, has Poisson distribution with parameter $\lambda = A \cdot \exp\{-\kappa \cdot (Ask_{k,t}^* - P_{k,t}^0)\} \cdot dt$. Using maximum-likelihood estimation (MLE) method one can estimate parameters A and κ:

$$\max_{A,\kappa} \prod_{t \in \tilde{\mathcal{T}}} \prod_{k \in \tilde{\mathcal{C}}_t} \frac{\lambda(t,k)^{B_{k,t}}}{B_{k,t}!} e^{-\lambda(t,k)}, \quad \lambda(t,k) = \exp\{\log(Adt) - \kappa(Ask_{k,t}^* - P_{k,t}^0)\},$$

[21] the formal criteria is to ignore if $|\varepsilon_{K,T}|$ larger than 99.5%-quantile of all $\varepsilon_{K,T}$ residuals
[22] add the plot if have time

where $\tilde{\mathcal{T}}$ is the set of "similar times" and $\tilde{\mathcal{C}}_t$ is the set of similar contracts. The same approach must work for the estimation of parameters on sell order demand parameters. The only difference is using $P^0_{k,t} - Ask^*_{k,t}$ in front of κ.

We do not know the fundamental value $P^0_{k,t}$, but can use the BS estimated value $\hat{P}_{k,t}$ as a proxy. Hence, the only tricky question is to define similar times and similar contracts. We have already prepared the ground for the last, splitting contracts into different baskets. Defining similar times is a less trivial question because their many dimensions in which demand can be different. At this point, I simplify the analysis splitting the times by quarters of the year to address possible seasonality in demands and by implied volatilities of the ETF to address variation in demand based on aggregate uncertainty.

Tables 2 and 8 reports values for Parameters $\log(Adt)$ and κ I received for different type of contracts based on Poisson regression. We can see that they vary quite a lot over quarters and time-to-maturity. For the paper, I take the average values for A and κ over 2016. In my sample, $A = 1686 = \exp(1.95) \cdot 240$, where multiplication by 240 comes as a number of trading minutes over a day and $\kappa = 238$.

	Contract	Time to Maturity	Volatility	$\log(Adt)$	κ
1	PUT	Long	Low	2.01	315.93
2	PUT	Long	High	-0.35	21.49
3	CALL	Long	Low	1.61	12.33
4	CALL	Long	High	0.16	73.25
5	PUT	Short	Low	3.55	432.58
6	PUT	Short	High	1.08	31.78
7	CALL	Short	Low	3.37	500.54
8	CALL	Short	High	1.35	65.67

Table 2: **Estimated on buy orders parameters, A and κ.** High volatility times are the times when implied volatility for the stock is higher than its median value.The first two columns specify the type of the contract within $\mathcal{B}_1$. I report for in-the-money contracts only.

The only parameter which I cannot estimate using the data is the preference parameter of the agents. For the main specification, I use $\gamma = 2$, which is quite small risk-aversion. Scaling γ will scale the spread, respectively.

There are pitfalls in the way I match models and data. First of all, the data suggests more than just two market makers. Extension of the model to many market makers is out of the scope of the paper. At the same time, not that many market makers decide to be active during a day in a specific contract, and the situation is not too far from a duopoly game.

Next, the dynamic of the orders can be different in the eyes of market makers. For example, the arrival of orders can be autocorrelated, what can be modeled by making A time-varying. That can change the dynamic equilibrium, increasing competition of agents: losing a chance to win on some orders today might mean that tomorrow there will be no chance to make a profit.

Another issue is that buy and sell demands must not be identical. The model can be solved numerically for different A and κ for buy and sell orders, but it does not allow any more an approximation with tractable properties.

7 Non-cooperative equilibrium

7.1 Characterization of the non-cooperative equilibrium

By definition, a non-cooperative equilibrium is memoryless. The actions and expected utilities of agents depend on the current states only. These are time to maturity, agents' inventory, cash balances and price of the asset.

Suppose the economy is at the state $S_t = s$, $q_t = q$, and $X_t = x$. Over a short horizon dt the inventory can change only for one agent[23], i.e., q_{t+dt} can be one of $\{q, q \pm e_1, q \pm e_2\}$, where $e_1 = [1,0]$ and $e_2 = [0,1]$. Inventory of an agent can decrease/increase if and only if she posts the best ask/bid quote δ_t^a/δ_t^b. The following notation will be convenient : $q^{i,a} = q - e_i$ and $q^{i,b} = q + e_i$.

The non-cooperative equilibrium can be characterized by the family of functions $u_i(t,x,q,s)$, $i = 1,2$ which follows a system of Hamilton–Jacobi–Bellman (HJB) partial differential equations:

$$\partial_t u_i(t,x,q,s) + \frac{1}{2}\sigma^2 \partial_{ss}^2 u_i(t,x,q,s) + \sum_{z=a,b} \mathbb{U}_i^z(t,x,q,s) = 0 \tag{6}$$

with the final condition

$$u_i(T,x,q,s) = -\exp(-\gamma(x_i + q_i s)). \tag{7}$$

$\mathbb{U}_i^a(t,x,q,s)$ and $\mathbb{U}_i^b(t,x,q,s)$ are instant utilities from the fact that inventories of agents can change as a result of their trading activity. For the most common case, when both agents can post buy and sell orders,

[23] probability of arrival of two orders is infinitesimal, i.e., it has order $\bar{o}(dt)$.

$|q_i| < \bar{Q}$ and $|q_j| < \bar{Q}$, $j \neq i$,

$$
\begin{aligned}
&\text{for } z = a,b, \quad \mathbb{U}_i^z(t,x,q,s) = \\
&\sup_{\delta_{i,s}^z} \Bigg(\lambda^z\left(\delta_s^{\min,z}\right)\left[u_i\left(t,x^{i,z},q^{i,z},s\right) - u_i(t,x,q,s)\right] \cdot \left[\mathbb{I}\{\delta_{i,s}^z < \delta_{j,s}^z\} + \frac{1}{2}\mathbb{I}\{\delta_{i,s}^z = \delta_{j,s}^z\}\right] \; + \\
&\lambda^z\left(\delta_s^{\min,z}\right)\left[u_i\left(t,x^{j,z},q^{j,z},s\right) - u_i(t,x,q,s)\right] \cdot \left[\mathbb{I}\{\delta_{i,s}^z > \delta_{j,s}^z\} + \frac{1}{2}\mathbb{I}\{\delta_{i,s}^z = \delta_{j,s}^z\}\right] \Bigg),
\end{aligned}
\tag{8}
$$

where $\mathbb{I}$ is an indicator function, $\delta_s^{\min,z} = \min\{\delta_{i,s}^z, \delta_{j,s}^z\}$, $x^{i,a} = x + e_i \cdot \delta^{\min,a}$ and $x^{i,b} = x - e_i \cdot \delta^{\min,b}$. Look at the appendix for careful definition of $\mathbb{U}_i^z(t,x,q,s)$ at every boundary case, i.e., when $|q_1| = \bar{Q}$ or $|q_2| = \bar{Q}$.

An agent receives utility not only from a possible adjustment of her inventory but also from a jump in inventories of her rival. That makes the competition strategic. A jump in inventories of the rival makes her compete differently in the future. For example, if the rival has non-negative inventory and she gets a buy order increasing her inventory, then starting the next moment, she might compete less aggressively on the bid quote.

CARA utility function guarantees an equilibrium which does not depend on the cash balance and the current asset price which might play just the role of scaling parameter for the agents' utilities,

$$u_i(t,x,q,s) = -\exp(-\gamma(x_i + q_i s)) \cdot \exp(\gamma \cdot \theta_i(t,q)),$$

which allows to rewrite HJB as

$$\partial_t \theta_i(t,q) + \frac{1}{2}\gamma q_i^2 \sigma^2 + \sum_{z=a,b} \Theta_i^z(t,q) = 0, \tag{9}$$

$\Theta_i^z(t,q) = \gamma^{-1} \cdot \mathbb{U}_i^z(t,x,q,s) \cdot (u_i(t,x,q,s))^{-1}$, with the boundary condition

$$\theta_i(T,q) = 0 \qquad \forall q.$$

$\theta_i(t,q)$ has an interpretation of the cost which agent i would be ready to pay to sell all her inventory at the current fundamental value and leave the market. Therefore, θ_i measures disutility of agent from the current inventory distribution. Clearly, $\theta_i(t,q)$ is non-positive when $q_i = 0$ since agent has always an option of not-trading. As becomes clear later, in the long-run $\theta_i(t,q)$ is always negative, because agents can do positive profit over the time.

If agent i were alone in the market she would post orders with the monopolistic reservation price

$$\delta_i^{z**} = \delta_i^{z**}(t,q) = \ln\left(1+\frac{\gamma}{k}\right)^{1/\gamma} + \theta_i(t,q^{i,z}) - \theta_i(t,q) \tag{10}$$

In the competitive environment, if the agent wants to win an order providing the best quote, she must post the price such that posting a price below this value is unprofitable for her competitor. Hence, her reservation price must be smaller than the following number:

$$\delta_i^{z***} = \delta_i^{z***}(t,q) = \theta_j(t,q^{j,z}) - \theta_j(t,q^{i,z}) \tag{11}$$

At the same time, the reservation price must not be too low, such that the agent does not receive negative utility from being best on the quote. That imposes the boundary from below:

$$\delta_i^{z****} = \delta_i^{z****}(t,q) = \theta_i(t,q^{i,z}) - \theta_i(t,q^{j,z})$$

The next proposition summarizes these three facts to get the optimal quote.

Proposition 1. *In equilibrium, when both agents can post orders ($q_i < \bar{Q}$ for bid, $q_i > -\bar{Q}$ for asks)*

$$\delta_i^{z,*} = \delta_i^{z*}(t,q) = \max\{\min\{\delta_i^{z**},\delta_i^{z***}\},\delta_i^{z****}\}, \quad i=1,2, \quad z=a,b. \tag{12}$$

There is no closed-form solution for the problem, but one can solve agents' quotes and their utility functions backward from $t=T$. At this point, it is unclear which combinations of inventories lead to lower prices and who are the agents providing the liquidity on Ask and Bid. The intuition might tell us that the agent with the lowest/highest inventories should be more aggressive on Bid/Ask order since the agent is eager to reduce her inventory risk. Nevertheless, I will show that this intuition does not always hold. Moreover, the solution is non-trivial when $\bar{Q}$ is sufficiently small.

7.2 "Resting" market maker.

In the subsection, I will show the case when there is a market maker, which is the most active on both ask and bid orders. In opposite, her rival will "rest" waiting for the time when the inventories of first agent change. We get this result assuming no cooperation between the agents. The "resting" behavior is a natural outcome of the equilibrium when agents strategically prevent their rivals from the competition.

We might expect that agent i provides the best bid (ask) price when her inventories are such that $q_i <$

$(>)q_j$ because the buying (selling) exposure to the inventory risk is lower for her than for her competitor. Let us call it *monotonic principle*. It is illustrated in Figure 36a for a toy-model where there are 5×5 states. Two agents are called Ina and Joe. Every cell of the plotted table corresponds to a particular inventory distribution. $(Name1, Name2)$ inside the cell means that Name1 suggests the best ask price and Name2 suggests the best bid price. In all states under (over) the diagonal, Ina is the most active on ask (bid) and Joe is the most active on bid (ask).

Surprisingly, the monotonic principle can be violated when the limit on trading for an agent with inventories at capital constraint exists. Let us call the situation when $|q_i| = \bar{Q}$ a 'corner' for agent i. Agent j, $j \neq i$, can benefit from having her rival in the corner since agent i is prevented from trading on one side of the spread, leaving monopoly power to agent j.

Consider a situation when both agents are out of the corner, $-\bar{Q} < q_i < q_j$, but their inventories q_i and q_j are sufficiently close to the capital constraint $\bar{Q}$. If agent i sells a few more units of the asset, then she cannot compete on selling order with j leaving her monopolistic power on "Sell". That creates incentives for agent j to "drive her rival to the corner" $-\bar{Q}$, i.e., let her be the best on Ask quote. We already see that agent j violates the monotonic principle. Moreover, agent j may not just "rest" but compete with i more aggressively on the Bid quote to make agent i staying in the corner more likely.

Figure 16 shows a simple example with $\bar{Q} = 2$, which is constructed in the way to show this phenomenon.[24] I use a numerical solution with the parametrization from the previous section. The table shows which agent provides the best ask price at every inventory case, and the plot shows the ask reservation prices in some of the states.[25] If the monotonic principle held, then one should have seen agent 1 posting the best ask price in all states below the diagonal. However, agent 2 is more active in many of the states. That violates the monotonic principle.

The most substantial ask price is achieved when agents are in different "corners" of inventories, and both act as monopolists on different sides of the spread. When $q = [2, -2]$ agent 1 is clearly the agent which provides the Ask and agent 2 provides the best Bid. At state $q = [2, -1]$ where agent 2 purchased one unit of the asset, the situation changes dramatically. Ask price goes down because agent 2 starts competing on both sides of the spread. Moreover, agent 1 ends up "resting". Agent 2 values being the monopolist on Bid so much that she is ready to pay some cost to hold agent 2 in the corner. Agent 1 "lays down her arms" and rests until agent 2 gets into one of the corners being prevented from the competition on one of the sides of the spread.

[24]In appendix, Figure 36b shows all the decisions players make.

[25]bid prices can be analyzed similarly, since the model is symmetric around $[0, 0]$

The state $q = [1, -1]$ violates the monotonic principle twice. Both agents are active on the side of the spread, which will make their inventories even larger. One observes that every agent drives rival to the corner competing with her aggressively. That is why the reservation ask price (as bid price) is even negative at the states.

A similar behavior, which is harder to illustrate in one picture, can be observed for larger capital constraints values of $\bar{Q}$. Even for very large $\bar{Q}$ when one of the agents is close to the capital constraints another will try to keep her there. Dynamically, the picture is the following: agents start competing in a normal way satisfying monotonic principle which will be discussed in next section. At some point, one of them gets dangerously close to the capital constraint. Her competitor starts playing on both sides of spread while she "lays down her arms". She waits when her competitor will get close to capital constraint. Then they compete until the moment one of them leaves the capital constraint starting being a monopolist on one side of the spread and preventing the rival from wining another side of spread.

Though previous papers, e.g., Gromb and Vayanos (2002), discussed the role of financial constraints on liquidity provision, to best of my knowledge this is the first paper which discussed this dynamic competition effect of the constraints. It is also one of a few explanations of why the same market maker can be the most active both on bid and ask side and why agents with highly negative/positive inventories can continue selling/buying the asset; see Figure 14. Finally, one should not take capital constraints as a necessary condition to get this type of equilibrium. Even if the limits do not exist, the similar effects could be received if I assume a more convex inventory cost so that if agents achieve some level of inventory saturation, they stop competing on one of the spread sides.

7.3 Market Makers under Relaxed Capital Constraints

When the inventory constraint $\bar{Q}$ is sufficiently large, the required time to drive rival to the corner is longer, and the cost of competing with her on the wrong side of the spread can prevail. Therefore, the behavior of market makers will start satisfying the monotonic principle: agent i provides the best bid (ask) price when her inventories are smaller (bigger) than an inventory of her rival, $q_i < (>) q_j$. The force will move inventories of the market makers to the same number. In this subsection, I will discuss the properties of the spread and utilities of the agents in this "normal" environment when the time to maturity is sufficiently large. Assume in the subsection that $\bar{Q} = \infty$ and T is large.

Let us start from the case when $q_1 = q_2$ or $q_1 = q_2 \pm 1$. These states, when the difference in inventories is not larger than one unit, are absorbing. "Normal" market makers with $q_1^{old} = q_2^{old} \pm 1$ can move only to the states with $q_1^{new} = q_2^{new} = q_1^{old}$ or $q_1^{new} = q_2^{new} = q_2^{old}$. At the same time, since it is possible to get a series of the same orders, i.e., only Buys or only Sells, the average of the inventories can move more than just one unit of asset. As in the model with a monopolistic market maker, there is a small probability of going too far from $[0,0]$. When agents' inventories are too negative, ask and bid prices decrease, shifting external demand towards sells, what drives inventories back to zero.

Hence, when the time to maturity is ample, the dynamics are the following. Agents reduce their inventory imbalances converging to the same inventory size, then with similar inventory converge to zero-inventory state, where they spend most of the time fluctuating about it.

We are primarily interested in the market liquidity conditional on inventory imbalance and the size of aggregate inventory. To analyze it, I need to find the utility functions of agents. Unfortunately, the model does not have a closed-form solution. Nevertheless, using Taylor expansion for the utility function, one can guess and verify that at the states around $\bar{q}$, $q_i, q_j \in \{\bar{q}-k, \ldots, \bar{q}, \bar{q}+k\}$, when k is sufficiently small, utility functions of agents are defined, by

$$\begin{aligned}\theta_1(t,Q,Q-k) = \theta_{0,0} + a_2 \cdot Q^2 + a_4 \cdot Q^4 + c_2 \cdot k^2 + c_4 \cdot k^4 + b_{1,1} \cdot Q \cdot k + \\ b_{3,1} \cdot Q^3 \cdot k + b_{1,3} \cdot Q \cdot k^3 + b_{2,2} \cdot Q^2 \cdot k^2 + d \cdot (T-t) + \theta_1^r(t,Q,k),\end{aligned} \tag{13}$$

where $\theta_1^r(t,Q,k)$ is the residual term which can be ignored under relatively small Q^4, k^4 and large $T-t$. The complete rigorous approximation exercise is outside the scope of this paper, but one can take a look at the fitting of the approximation into numerical solution and discuss a few facts that one can extract conditional on the approximation. As one can see in Table 10 , the fitting is excellent based on 12201 numerical solution points around $(0,0)$[26]

$c \cdot (T-t)$ is a term that captures future profits, which market makers do on the arriving orders in the long run. When $T-t$ is sufficiently large after getting into absorbing states, agents do fixed expected profit per unit of time, that is why the term is linear in time. The time trend is the same for all states because from any initial state model gets into the state with $q = [Q,Q]$ for some Q in relatively short finite time. The profit which market makers do on the way to the absorbing state is captured by the rest terms which are independent of time.

Quadratic term $a \cdot Q^2$ captures the inventory risk of the agents. It does not depend on the size of the imbalance k. Nevertheless, the imbalance is critical because it creates better trading opportunities for agents on the different sides of the spread. That is reflected in other terms which include imbalance k.

[26] in this table I fit similar form for $\theta_1(t,Q,k)$ which is trivially can be transformed into the form (13)

First, let us look at the states where inventories for both agents are the same. The optimal prices are linear functions in the inventory level Q[27]:

$$\delta^{a,*}(-Q,-Q) = \delta^{b,*}(Q,Q) = (\theta_{1,0} - \theta_{0,1})(1+2Q) = (1+2Q)\left(a_2 + \frac{1}{2}a_4 + (b_{1,1} + b_{1,3})\right) \tag{14}$$

agents increase the ask and bid prices when their inventory goes down/up. Therefore, the mid-price of the asset is affected by aggregate inventories of agents. At the same time, spread does not depend on the level of inventories. The fourth order approximation is just a positive constant:[28]

$$Spread(Q,Q) = 2(b_{1,1} + b_{1,3}) + 2a_2 + a_4 + O(Q^4) \tag{15}$$

This prediction falls in line with the models of a monopolistic market maker, as in Stoll (1978), Avellaneda and Stoikov (2008). Note that HJB equation (9) guarantees that θ is linear in σ^2 what translates into the linearity of the spread in the uncertainty level.

Though agents have the same inventories spread is not equal neither to 0 nor to the double value of risk one extra unit of asset carries, i.e., agents do not compete in Bertrand fashion.Agent 1 prefers state $q = [0,1]$ to state $q = [1,0]$, i.e., $\theta_{1,0} > \theta_{0,1}$, because having similar trading opportunities there is less risk when $q_1 = 0$ comparative to $q_1 = 1$. Thus, the spread is positive. The repeated interactions and moving on different sides of spread sometimes allow agents to extract positive utility from trading.

Let us take a look at the numerical solution, which does not assume the linear approximation. Figure 17 shows us the values of $\theta(t,q)$, which has inverse relationship with utility function for different inventories of agents q[29]. First of all, one can see that for all values of q, $\theta(t,q)$ is a linear function of time in the long-run. When inventory is larger, more time is needed to achieve the linear limit. The slope is negative because the agent can make a profit on trading. The larger absolute value[30] of agent's inventory is associated with lower utility because of the inventory risk. Because of the risk, when both agents have very high inventory, their θ-s are positive even after ten days, i.e., the agents suffer serious disutility. The larger inventory of rival causes larger utility, e.g., $\theta_1(t,5,5) > \theta(t,5,40)$, because the agent can make a higher profit on one of the spread's sides. Finally, $q(t,k,-k) \approx q(t,0,0)$. That says restriction on inter-dealer trading does not make a huge difference in the equilibrium dynamics.

[27] Using (13) one can get $\theta_{1,0} - \theta_{0,1} = b_{3,1} + b_{1,1} + b_{1,3} + b_{2,2} + b_{3,1} + a_4 + a_2 + (2Q^2 - 2Q + 1) + Q^2 b_{3,1} - Q$, $\delta^{a,*} = b_{1,1} + b_{1,3} + b_{2,2} + b_{3,1} + a_2 + a_4 Q\,(2b_{1,1} + 2b_{1,3} + 2b_{2,2} + 3b_{3,1} + 2a_2 + 4a_4) + O(Q^2)$

[28] use $Spread(Q,Q) = \delta^{b,*}(Q,Q) + \delta^{a,*}(Q,Q) = 2\,(b_{1,0} + b_{2,2} + b_{3,0} + d_{1,0} + a_2 + a_4) + 2q^2\,(3b_{3,0} + 6a_4) + O(Q^4)$

[29] recall, $\log(-u(t,q,x)) = -\gamma(x_1 + q_1 S) + \theta(t,q)$, hence larger θ is related to smaller u.

[30] $\theta(t,q)$ is symmetric in q around zero $[0,0]$.

Figure 18 reports the utilities ten days before maturity. Panel (a) shows what happens when q_2 is fixed and q_1 changes. It can be considered as $\theta(t, \bar{q}+k, \bar{q})$, when $\bar{q} = q_1$. The blue line, which is $\theta_{k,0}$ in the approximation, increases quadratically when $|k|$ goes up. The optimum is having no inventories when rival has no inventories. Nevertheless, when rival's inventory goes up, the optimum shifts in opposite direction. The reason is that agent can make profit on the other side of spread. Panel (b) illustrates how $\theta(T, \bar{q}, \bar{q}+k)$ changes when aggregate inventory $\bar{q}$ changes. For comparison, I add approximation (13), which works very well almost coinciding with the numerical solution. As one can see, independent of k the quadratic dependence on $\bar{q}$ is unchanged.

Figure 19 shows how prices change in time before maturity conditional on agents' inventories. The prices converge to fundamental value of the asset when there is little time left before maturity because agents start competing in Bertrand fashion. The largest bid price happens where both agents have high inventory of the asset, $q = [100, 100]$. Price decreases when inventory of one of the agents goes down, $q = [100, 0]$. It is different from intuition suggested in Ho and Stoll (1983) that the second best buyer, i.e., agent with highest inventory, must set the bid price. Comparing $q = [100, 0]$ with $q = [100, 100]$, inventory of the second best buyer does not change, but price goes down substantially because agent 2 starts being less reluctant to buy. It does not mean that agent 1 does not affect the bid price, she still prevents monopolistic behavior for agent 2 and, as one can see comparing cases $q = [100, 0]$ and $q = [150, 0]$, does it better when has smaller inventory. Case $[0, -50]$ shows us that bid price can be even smaller than fundamental value of the asset if in average agents want to reduce their negative inventories. Finally, case $[150, -150]$ confirms that competition is the core for defining the price. The agent with inventory $q = -150$ must be very willing to buy more units of asset, but since she does not meet aggressive competition from her rival with $q = 150$, she still can charge some positive fee, $\delta^b(150, -150) > 0$. Clearly, similar analysis can be done on Ask side of spread.

As observers, we cannot know the fundamental value of an asset, hence the only observed liquidity measure in the model is the width of the bid-ask spread. Figure 19b confirms finding from equilibrium approximation. Spread is positive for all values of inventories but it is minimal when agents are identical. Spread does not depend on aggregate inventory of market makers. That makes inseparable lines for $q = [0, 0]$ and $q = [100, 100]$. The width of the spread in .25 basis points is too small comparative to what is observed. When one increases inventory imbalance, spread goes up. The size of 2.5 basis points at $q = [150, -150]$ is already something what can be observed in the data. Further increasing of inventory imbalance can make spread significantly larger.

8 Non-Markov Equilibrium

In the equilibrium discussed in the previous section, inventories of the agents converge to each other. At the same time, it follows from Figures 11a and 11b that they diverge substantially for different traders. That can be described by disagreement or private values from specific contracts for the market makers, as in Du and Zhu (2017) and Kyle et al. (2017). In this section, I provide an alternative explanation. Market makers can cooperate and do not compete with each other on different sides of spread.

Let us look at states where the market makers' inventories have opposite sign and similar by absolute values, in Figure 17. The states are not significantly worse[31] than the state with zero inventories. That creates fertile soil for collusion because even if one of the agents defect and collusion breaks, both agents are not likely to be in a very bad state.

Previously I assumed that market makers play the game that does not depend on pre-history. Now let us suppose they can depend and build a grim-trigger like equilibrium in the model. Consider the case when agent 1 acts as a monopolist seller even if her inventory is negative and agent 2 acts as a monopolist buyer even if her inventory is positive. What can prevent an agent j from competing with agent i on ask quote? For example, it can be the guarantee that i does not compete with j on bid quote. Suppose that if someone of agents deviates from this rule and starts competing with another agent then the players return to standard equilibrium discussed in the previous section.

Let us investigate when cooperation is sustainable. Denote the set of inventories at time t such that agent starts acting as discussed in the previous section by $\check{\mathcal{Q}}(t)$. Suppose that parameters are such that $\check{\mathcal{Q}}(T-0)$ is an empty set, i.e., there is no incentive to deviate when there is little time before maturity.[32]

Denote utility of the agents by $\check{v}_i(t,q) = \exp\{\check{\theta}_i(t,q)\}$ with $\check{\theta}_i(t,q) = \theta_i(t,q)$ for $q \in \check{\mathcal{Q}}(t)$ and when $q \notin \check{\mathcal{Q}}$ following this dynamic equation:

$$\partial_t \check{\theta}_i(t,q) + \frac{1}{2}\gamma^2 q_i^2 \sigma^2 + \sum_{z=a,b} \check{\Theta}_i^z(t,q) = 0, \tag{16}$$

where

$$\check{\Theta}_i^z(t,q,\delta_{i_z}^{z**}) = \lambda\left(\delta_{i_z}^{z**}\right)\left[\exp\left\{-\gamma\cdot\delta_{i_z}^{z**}\cdot\mathbb{I}\{i=i_z\} + \check{\theta}_i(t,q^{i_z,z}) - \check{\theta}_i(t,q)\right\} - 1\right],$$
$$z_1 = a,\ z_2 = b, i_a = 1, i_b = 2.$$

[31] they can be, but the difference is tiny

[32] the positive side of continuous-time is that this can be guaranteed even if there is a finite horizon because in any small interval there exists an infinite number of interactions between agents

and as before

$$\delta_i^{z**} = \delta_i^{z**}(t,q) = -\frac{1}{\gamma}\cdot\ln\left(\frac{k}{(k+\gamma)}\right) - \frac{\check{\theta}_i(t,q)-\check{\theta}_i(t,q^{i,z})}{\gamma}.$$

Simplifying, the monopolist's profit is

$$\check{\Theta}_i(t,q,\delta_1^{z_i**}) = -\lambda\,(\delta_1^{z_i**})\,\frac{\gamma}{k+\gamma}.$$

Hence the states q where the agent 1 will want to defect and start competing on bid are

$$\sum_{z=a,b}\check{\Theta}_1^z(t,q) \geq \sum_{z=a,b}\Theta_i^z(t,q).$$

These are the states where the inventory risk starts being too high dominating all benefits from being a monopolist.

Let me illustrate the monopolistic case when the model parameters are the same as before. Surprisingly, the market makers choose never deviate except when they both achieve the inventory constraint. The reason is similar to discussed in the "resting market maker" section. Agents want their competitors to move over inventories as far as possible. Then when one of them defects, the rival's inventory will allow being more monopolistic. Figure 20 illustrates the difference in terms of liquidity between cooperative and non-cooperative equilibrium. First of all, the spread increases significantly. Now, one can have 80bp-spread even if agents have zero-inventories. The spread is higher than the average observed in the data and does not vary a lot when inventories change. That is very different from the non-cooperative case where the distribution of inventories had a primary impact on liquidity. That confirms that heterogeneity acts through the distortion of strategic incentives of agents, but not through the actual change in inventory cost.

Figure 20b reports the ratio of spread in cooperative equilibrium to non-cooperative equilibrium. It explodes when there is not much time before maturity, because competition drives non-cooperative spread to zero. On the longer horizons, the ratio decreases for higher inventory imbalances. That is because in non-cooperative equilibrium agents act more and more like monopolists when the imbalance grows up.

I discuss a very specific type of cooperative equilibrium. There can be many options of collusive equilibria which will create different equilibria outcome. For example, the discussed equilibrium will be always broken once the agents achieve the "corners". Nevertheless, if agents are allowed for inter-dealer trading then they will just periodically cancel their positions without defection. A more realistic version is when there are many assets and market makers choose specific contracts in which they provide liquidity starting to trade in all contracts if someone defects.

The cooperative equilibria despite the reach set of possible outcomes will be always subject to critique. First of all, there is always lack of discipline, which one of the plenty equilibria do the market makers play? Another issue is that, in reality, in anonymous centralized market it can be hard to support all these cooperative equilibria. It is easy to read some noisy demand shock as the defection of the competitor and trigger "a war". At the same time, since modern market makers operate on a very high speed they can potentially read the actions of others to some extend and implicitly come to a sort of cooperative equilibria specializing on specific contracts and specific sides of spread.

9 Testing Predictions of the model

9.1 Aggregate Liquidity and Inventory Distribution

First, let me take look into a more causal relationship between inventories and liquidity comparative to discussed in Section 4. The usual problem with establishing casual relationship between liquidity and any descriptive variable are the existence of omitted variables. For example, both liquidity and the inventory positions of market makers are both affected by the demand shocks. Inventory position cannot react immediately on such shocks, but market makers could potentially anticipate them accumulating position in advance.

There are two potential ways to deal with the problem. First approach could be to instrument unanticipated shocks. This requires deep knowledge of Chinese market and I left the approach for the future work. Alternative way is to compare similar contracts for which it could be harder to predict demand shocks.

The beneficial side of the environment is that one can observe quite similar contracts at the same time. For example, consider two call-option contracts with the same maturity date for which strikes are different by .05 RBN. Given that daily fluctuations in the price of the underlying ETF has the same order around .02RBN, They are unlikely to have a big predicted difference in demand at some time, but even if there is the difference one can try to capture it by the fixed effect on the contracts over longer periods of times and their characteristics such as Greeks.

This logic leads us to the following regression specification

$$Y_{k,t} = \alpha_{t,B(k)} + \alpha_{k,\tilde{t}(k,t)} + \beta_D D_{k,t} + \beta_X X_{k,t} + \varepsilon_{k,t}, \tag{17}$$

where dependent $Y_{k,t}$ is a liquidity measure, the variable of interest $D_{k,t}$ is a measure built on inventory distributions, $X_{k,t}$ are the controls for the contract. $\alpha_{t,B(k)}$ is the fixed effect for all contracts in the same basket $B(k)$. It captures liquidity shifts for similar contracts at the same time. $\tilde{t}(k,t)$ is a longer period than t which includes t and $\alpha_{k,\tilde{t}(k,t)}$ is the fixed effect for the contract over the period.

The distribution measures under consideration must capture both average inventory position of all market makers, denote it $Q^{mm}_{k,t}$, and size of individual deviations from the average. For the last, I firstly use the maximal and minimal inventory positions of the market makers, $Q^{mm,max}_{k,t}$ and $Q^{mm,min}_{k,t}$, because, as was described in section 4, these are good predictors of being active on the best bid and ask. Next, the model predicts that the liquidity provision of the maximal market makers should depend on the competitors, since the maximal guys may want to prevent them from competition. That is why I include the sizes of the second largest (smallest) positions $Q^{mm,max,2}_{k,t}$ ($Q^{mm,min,2}_{k,t}$). Based on these first- and second- best inventories i define

$$\begin{aligned} Q^{max-min}_{k,t} &= Q^{max}_{k,t} - Q^{min}_{k,t} \\ Q^{max-min,2}_{k,t} &= Q^{max,2}_{k,t} - Q^{min,2}_{k,t} \\ \Delta Q^{max,2}_{k,t} &= Q^{max}_{k,t} - Q^{max,2}_{k,t} \end{aligned}$$

As an alternative to the measures could be measures that capture aggregate heterogeneity of the agents, such as second and larger moments of the inventory distribution. I use the measures such as,

$$Q^{d}_{k,t} = \sum_{i=1}^{N} \Big(Q_{i,k,t} \times \mathbb{I}\{Q_{i,k,t} > Q^{med}_{k,t}\} - Q_{i,k,t} \times \mathbb{I}\{Q_{i,t} < Q^{med}_{t}\} \Big),$$

where $Q^{med}_{k,t}$ is the median inventory of market makers in asset k at t. Alternatively, $Q^{med}_{k,t}$ can be substituted by quantiles of inventory distribution, $Q^{q25}_{k,t}$ and $Q^{q75}_{k,t}$.

Analysis in the section is done over 10-minutes intervals. I exclude first 30 minutes of the trading day, 10 minutes before and after the mid-day break, and 10 minutes before the end of trading day. It is motivated by anomaly high spread over the time illustrated in Figure 25.

Let us start from analyzing the most popular liquidity measure, the bid-ask spread. Our model predicts that spread does not depend on aggregate level of inventory, but might depend on the distribution of inventories across the agents. Table **??** confirms both statements. Hence, it can be seen in all the specifications that one cannot find dependence of the spread on absolute value of aggregate inventory of agents[33]. At the same time, the difference in inventories of marginal market makers causes spread to shrink. This prediction is opposite to the finding of the relaxed constraint case where imbalances caused higher spread. The reason is probably that we have many market makers. Hence, maximal market makers set prices conditional on competition faced from market makers with the second largest positions. That tells us that spread should

[33] the same is true if absolute value is removed

shrink when inventories of second-largest position holders diverge. That is what the second column of the table shows.

Next, let us study deviations of the ask and bid prices, $Ask_{i,t} - \hat{P}_{k,t}$ and $\hat{P}_{k,t} - Bid_{i,t}$, from the fundamental value of the contract, which is defined as the deviation from the Black-Scholes price. Tables 4 - 5 confirm that the average position of market makers create substantial deviations from the fundamental value. As before, one can see that when inventory of the second largest market maker goes up the gap between ask price and fundamental value shrinks. Similarly, for the bid price low inventories of the second second minimal market maker push bid price up.

9.2 Individual level regression

The data can be used to describe liquidity provision at individual level. Accounts are different, but one can use variation in their liquidity provision across similar contracts. The key specification has the following form:

$$Y_{i,k,t} = \alpha_{i,B(k),t} + \nu_{i,k,around\ t} + f(Q_{i,k,t}) + g(distribution(Q_{k,t})) + \beta_X X_{i,k,t} + \varepsilon_{i,k,t}, \tag{18}$$

where $\alpha_{i,B(k),t}$ is a fixed effect which captures liquidity provision of market makers in a set of similar contracts at the time. $\nu_{i,k,around\ t}$ captures average activity of the agent in the contract over a longer period. $f(Q_{i,k,t})$ is a function of the individual inventories, $g(distribution(Q_{k,t}))$ function of inventory distributionover all market makers.

Let us consider 1-minute intervals. Define $BestAsk_{i,k,t}$ as the average time agent i spent on the best ask of contract k at t. This measure of agents' activity usually distributed non-uniformly with high mass around 60 and 0 seconds. To deal with the discontinuity let us build dummy. Define $\overline{BestAsk}_{i,B(k),day(t)}$ as the average time agent i spends on the best ask quote for contract k in basket $B(k)$ during a day of t. Let us introduce the dependent variable:

$$Y_{i,k,t} := \mathbb{I}\{BestAsk_{i,k,t} - \overline{BestAsk}_{B(k),day(t)} > 0\}$$

If an agent is active in the contract, then she will most likely be active over a more prolonged period than the average. Ideally, I would like to run logit regression on specification (18), but since the data is huge and I wish to include many fixed effects, even estimating of simple logit requires extra effort and time. The convergence is unstable and long. Getting robust standard errors is even a bigger problem. That is why I report simple linear specification at this point.

To capture potential non-linearities in the impact of inventory positions on liquidity, I use both linear terms and concave terms such as $sign(Q_{i,k,t}) \times \log(1 + |Q_{i,k,t}|)$. The convex terms are statistically and economically insignificant under the fixed effects.

Tables 6 - 7 report the results. Agents reduce activity in the bid-price when they hold a larger inventory position. Similarly, agents are less active in ask price when their inventories are low. That falls in line with the idea that market makers mean-revert their positions. At the same time, economically, the effect is weak. One standard deviation increase in the inventory measure $sign(Q_{i,k,t}) \times \log(1 + |Q_{i,k,t}|)$ increase probability to be active roughly just by 2.5%. The effect of average inventories is not statistically significant in 2015 for both Bid and Ask and 2017 for Bid. Nevertheless, it has the correct sign and economically is substantial. In 2017 the deviation in the average inventory of market makers can increase the probability of being active by 10%.

Behind the size of market makers' inventories, the individual decision must be affected by the distance to the extreme competitors. When the distance to maximal (minimal) market maker shrinks, the agent must compete more on ask (bid). It is right in the whole sample, but the effect disappears once one focuses on the sub-periods. Though the sign is correct, economically, the effect is weak. One standard deviation in the distance may help to describe from 0 to 1% of probability to be active.

10 Conclusions

This paper studies the behavior of oligopolistic market makers. The empirical evidence from the Chinese option market suggests that oligopolistic market makers do not compete aggressively with each other. Market makers specialize in different contracts charging substantial spread. We show that market makers can "rest" providing no liquidity on both sides of the bid-ask spread. Nonetheless, they act on both of them sometimes. There is a positive relationship between the absolute value of the inventory position and liquidity provision.

A tractable model of duopoly market making with a single risky asset is examined. Risk-averse market-makers act strategically, taking into consideration both their own and the competitors' inventories. They realize that their actions affect the future behavior of their competitors as well as their temporary profits and risk exposure. Looking forward, market makers decide to compete or not on each side of the spread separately. Many equilibria are possible. I discuss both Markov and non-Markov equilibria.

The properties of the non-cooperative equilibrium depend on the tightness of the capital constraint. If the inventory bounds can be achieved in a relatively short time, then the *monotonic principle* can be violated, for example, agents with the lowest inventories can be more active on sell orders than their competitor.

If the inventory constraint relaxed, then the monotonic principle holds, and equilibrium can be well approximated. The inventory imbalance is the main reason for variation in liquidity. Higher distance between agents' inventories allows agents to act more as the monopolists. Agents never compete in Bertrand fashion but act more aggressively when maturity approaches.

Finally, I discuss a grim-trigger cooperative equilibrium. Market makers act as monopolists on different sides of spread until one of them defect. The bid-ask spread increases to monopolistic values, but it does not change much when imbalance increases.

Next, I test if the data speak in favor of one of the specific types of competition. I look at the dependence of aggregate level liquidity on the distribution of inventories. I employ cross-section similarity of some contracts to separate the effect of lag-inventories on liquidity provision. The findings are mixed. On the one hand, I can see that the average position of market makers moves the ask and bid prices in the direction of helping mean-revert market makers positions. The maximal and minimal inventories affect the prices as well. Similarly, the second largest and second smallest inventory positions are essential, while the distance from them to the extreme positions are not statistically and economically significant. Probably, active market makers set their prices based on the competition from other market makers, as the model suggests.

Finally, the paper studies account-level liquidity provision based on the inventory. The decision of a market maker to be active in a specific contract depends on both her inventory and inventories of her competitors.

References

Ait-Sahalia, Y. and M. Saglam (2017). High frequency market making: Implications for liquidity. Available at SSRN 2908438.

Amihud, Y. and H. Mendelson (1980). Dealership market: Market-making with inventory. Journal of Financial Economics 8(1), 31–53.

Amihud, Y. and H. Mendelson (1986). Asset pricing and the bid-ask spread. Journal of financial Economics 17(2), 223–249.

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

Biais, B., T. Foucault, and S. Moinas (2015). Equilibrium fast trading. Journal of Financial Economics 116(2), 292–313.

Brunnermeier, M. K. and L. H. Pedersen (2009, Jun). Market liquidity and funding liquidity. Review of Financial Studies 22(6), 2201–2238.

Budish, E., P. Cramton, and J. Shim (2015a, Nov). The high-frequency trading arms race: Frequent batch auctions as a market design response. The Quarterly Journal of Economics 130(4), 1547–1621.

Budish, E., P. Cramton, and J. Shim (2015b). The high-frequency trading arms race: Frequent batch auctions as a market design response. The Quarterly Journal of Economics 130(4), 1547–1621.

Carrion, A. (2013, Nov). Very fast money: High-frequency trading on the nasdaq. Journal of Financial Markets 16(4), 680–711.

Cartea, Á., S. Jaimungal, and J. Ricci (2014). Buy low, sell high: A high frequency trading perspective. SIAM Journal on Financial Mathematics 5(1), 415–444.

Cespa, G. and X. Vives (2013). The welfare impact of high frequency trading. Technical report, Working paper.

Comerton-Forde, C., T. Hendershott, C. M. Jones, P. C. Moulton, and M. S. Seasholes (2010, Feb). Time variation in liquidity: The role of market-maker inventories and revenues. The Journal of Finance 65(1), 295–331.

Demsetz, H. (1968). The cost of transacting. The quarterly journal of economics 82(1), 33–53.

Du, S. and H. Zhu (2014). Welfare and optimal trading frequency in dynamic double auctions. Technical report, National Bureau of Economic Research.

Du, S. and H. Zhu (2017, Jan). What is the optimal trading frequency in financial markets? The Review of Economic Studies.

Fuchs, W. and A. Skrzypacz (2017). Timing design in the market for lemons.

Glosten, L. R. and P. R. Milgrom (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of financial economics 14(1), 71–100.

Gromb, D. and D. Vayanos (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. Journal of financial Economics 66(2-3), 361–407.

Guéant, O., C.-A. Lehalle, and J. Fernandez-Tapia (2013, Sep). Dealing with the inventory risk: a solution to the market making problem. Mathematics and Financial Economics 7(4), 477–507.

Harris, L. (2013). What to do about high-frequency trading.

Ho, T. and H. R. Stoll (1981, Mar). Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics 9(1), 47–73.

Ho, T. S. and H. R. Stoll (1983). The dynamics of dealer markets under competition. The Journal of finance 38(4), 1053–1074.

Hong, H. and J. Wang (2000). Trading and returns under periodic market closures. The Journal of Finance 55(1), 297–354.

Johnson, T., M. Liang, and Y. Liu (2016, Nov). What drives index options exposures? Review of Finance, rfw061.

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

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

Kyle, A. S. (1985). Continuous auctions and insider trading. Econometrica: Journal of the Econometric Society, 1315–1335.

Kyle, A. S., A. A. Obizhaeva, and Y. Wang (2017). Smooth trading with overconfidence and market power. The Review of Economic Studies 85(1), 611–662.

Kyle, A. S. and W. Xiong (2001). Contagion as a wealth effect. The Journal of Finance 56(4), 1401–1440.

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

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

Reiss, P. C. and I. M. Werner (1998). Does risk sharing motivate interdealer trading? The Journal of Finance 53(5), 1657–1703.

Rosu, I. (2006). Multi-stage game theory in continuous time. pp. 14.

Sannikov, Y. and A. Skrzypacz (2014). Dynamic trading: Price inertia, front-running and relationship banking preliminary and incomplete.

Stoll, H. R. (1978). The supply of dealer services in securities markets. The Journal of Finance 33(4), 1133–1151.

Swan, P. and P. J. Westerholm (2006). Market architecture and global exchange efficiency: One design need not fit all stock sizes.

Vayanos, D. (1999a). Strategic trading and welfare in a dynamic market. The Review of Economic Studies 66(2), 219–254.

Vayanos, D. (1999b, Apr). Strategic trading and welfare in a dynamic market. Review of Economic Studies 66(2), 219–254.

Tables

Dependent Variable:			Spread$_{k,t}$		
$\log(1+Q_{k,t}^{mm,max-min})$	-0.474*				
	(0.253)				
$\log(1+Q_{k,t}^{mm,max-min,2})$		-0.503*			
		(0.300)			
$\log(1-Q_{k,t}^{mm,min})$			-0.234**		
			(0.099)		
$\log(1+Q_{k,t}^{mm,max})$			-0.011		
			(0.086)		
$\log(1-Q_{k,t}^{mm,min,2})$				-0.090	-0.097
				(0.073)	(0.073)
$\log(1+Q_{k,t}^{mm,max,2})$				-0.070	-0.072
				(0.095)	(0.093)
$\log(1+\Delta Q_{k,t}^{mm,max,2})$					-0.012
					(0.061)
$\log(1-\Delta Q_{k,t}^{mm,min,2})$					-0.055
					(0.052)
$\lvert Q_{k,t}^{mm} \times 10^{-3}\rvert$	-0.001	0.008	-0.031	-0.038	-0.040
	(0.022)	(0.025)	(0.022)	(0.027)	(0.026)
$\Delta_{k,t}$	12.26**	12.12**	12.28**	11.89**	11.92**
	(5.116)	(5.133)	(5.135)	(5.160)	(5.162)
$\mathcal{V}_{k,t}$	-1.401	-1.394	-1.416	-1.492	-1.497
	(1.469)	(1.469)	(1.481)	(1.512)	(1.513)
$\Gamma_{k,t}$	0.781	0.783	0.769	0.816	0.815
	(0.564)	(0.565)	(0.567)	(0.590)	(0.590)
$\Theta_{k,t}$	18.85	18.89	18.01	16.69	16.57
	(57.86)	(57.84)	(58.67)	(60.72)	(60.73)
$\rho_{k,t}$	0.206	0.207	0.206	0.217*	0.217*
	(0.126)	(0.126)	(0.127)	(0.129)	(0.129)
$\lambda_{k,t}$	-0.035**	-0.035**	-0.035**	-0.035**	-0.035**
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Contract×10 DAYS FE	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes
Observations	2,654,835	2,654,835	2,649,326	2,619,130	2,619,130
R^2	0.92984	0.92984	0.92987	0.92994	0.92994
Within R^2	0.00156	0.00156	0.00159	0.00163	0.00164

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses.
Signif Codes: ***: 0.01, **: 0.05, *: 0.

Table 3: Aggregate market-level spread conditional on MMs' inventories

Dependent Variable:			$Ask_{i,t} - \hat{P}_{k,t}$			
Years	All	All	All	2015	2016	2017
$\log(1+Q_{k,t}^{mm,max,2})$	-0.548***	-0.556***		-0.471*	-0.271***	-0.249***
	(0.089)	(0.089)		(0.247)	(0.065)	(0.051)
$\log(1+\Delta Q_{k,t}^{mm,max,2})$	0.057		0.088*	0.323**	0.031	0.090***
	(0.053)		(0.052)	(0.150)	(0.043)	(0.035)
$\log(1+Q_{k,t}^{mm,max-min,2})$			-0.854*			
			(0.504)			
$Q_{k,t}^{mm} \times 10^{-3}$	-0.654***	-0.647***	-0.793***	-5.257***	-0.554***	-0.508***
	(0.046)	(0.046)	(0.049)	(0.764)	(0.054)	(0.033)
$\Delta_{k,t}$	9.831***	9.820***	10.55***	12.87*	11.29***	0.530
	(2.481)	(2.485)	(2.434)	(6.699)	(2.603)	(1.621)
$\mathcal{V}_{k,t}$	-7.247***	-7.248***	-7.241***	-10.89***	-6.620***	-7.076***
	(1.231)	(1.232)	(1.226)	(2.485)	(1.009)	(0.479)
$\Gamma_{k,t}$	1.577**	1.577**	1.595**	22.32***	2.169*	-0.233
	(0.777)	(0.777)	(0.761)	(5.914)	(1.221)	(0.379)
$\Theta_{k,t}$	113.9**	113.8**	113.6**	173.3***	130.9***	24.71
	(49.51)	(49.52)	(50.41)	(65.36)	(50.29)	(43.02)
$\rho_{k,t}$	-0.123	-0.124	-0.128	-0.080	0.015	-0.074
	(0.104)	(0.104)	(0.104)	(0.208)	(0.114)	(0.051)
$\lambda_{k,t}$	0.010	0.009	0.008	0.993**	0.011	-0.031
	(0.028)	(0.028)	(0.027)	(0.481)	(0.045)	(0.020)
Contract×10 DAYS FE	yes	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes	yes
Date-Clustered SE	yes	yes	yes	yes	yes	yes
Observations	2,624,993	2,624,993	2,654,835	912,394	1,023,615	688,984
R^2	0.93255	0.93255	0.93186	0.91712	0.95611	0.95657
Within R^2	0.04757	0.04754	0.0471	0.05049	0.09889	0.33087

*: One-way clustered (on Date) standard-errors in parentheses.

Signif Codes: ***: 0.01, **: 0.05, *: 0.

Table 4: Aggregate market-level deviation of Ask Price.

Dependent Variable:	$\hat{P}_{k,t} - Bid_{i,t}$					
Years	All	All	All	2015	2016	2017
Variables						
$\log(1 - Q_{k,t}^{mm,min,2})$	-0.549***	-0.554***		-0.293	-0.366***	-0.076
	(0.132)	(0.130)		(0.385)	(0.108)	(0.063)
$\log(1 - \Delta Q_{k,t}^{mm,min,2})$	0.024		0.043	0.329*	-0.082	-0.007
	(0.066)		(0.065)	(0.190)	(0.055)	(0.041)
$\log(1 + Q_{k,t}^{mm,max-min,2})$			0.837			
			(0.514)			
$Q_{k,t}^{mm} \times 10^{-3}$	0.697***	0.694***	0.808***	5.884***	0.525***	0.541***
	(0.052)	(0.050)	(0.052)	(0.848)	(0.061)	(0.034)
$\Delta_{k,t}$	2.425	2.443	2.394	16.73**	-3.831	0.402
	(3.231)	(3.236)	(3.217)	(7.474)	(2.412)	(1.881)
$\mathcal{V}_{k,t}$	6.503***	6.501***	6.540***	13.40***	6.272***	6.687***
	(1.150)	(1.150)	(1.150)	(2.080)	(0.977)	(0.476)
$\Gamma_{k,t}$	-1.494**	-1.494**	-1.498**	-45.45***	-2.403**	0.411
	(0.620)	(0.620)	(0.612)	(7.325)	(1.075)	(0.391)
$\Theta_{k,t}$	-88.59*	-88.64*	-87.00	-189.3***	-126.1***	-21.35
	(53.05)	(53.07)	(53.24)	(67.01)	(44.03)	(45.02)
$\rho_{k,t}$	0.321***	0.321***	0.322***	0.339*	-0.031	0.083
	(0.105)	(0.105)	(0.105)	(0.196)	(0.110)	(0.053)
$\lambda_{k,t}$	-0.024	-0.023	-0.025	-1.091**	0.003	0.037*
	(0.026)	(0.026)	(0.025)	(0.478)	(0.039)	(0.019)
Contract×10 DAYS FE	yes	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes	yes
Date-Clustered SE	yes	yes	yes	yes	yes	yes
Observations	2,648,972	2,648,972	2,654,835	921,359	1,032,429	695,184
R^2	0.92422	0.92422	0.92417	0.91348	0.94176	0.95287
Within R^2	0.03739	0.03738	0.03697	0.05982	0.08146	0.31877

*: One-way clustered (on Date) standard-errors in parentheses.

Signif Codes: ***: 0.01, **: 0.05, *: 0.

Table 5: Aggregate market-level deviation of Bid Price.

Dependent Variable:	$Y_{i,k,t} := \mathbb{I}\{BestBid_{i,k,t} - \overline{BestBid}_{B(k),day(t)} > 0\}$			
Years	All	2015	2016	2017
$sign(Q_{i,k,t}) \times \log(1+\|Q_{i,k,t}\|)$	-0.0063***	-0.0035***	-0.0066***	-0.0086***
	(0.0016)	(0.0011)	(0.0019)	(0.0023)
$Q_{i,k,t}$	0.0013	-0.0169	-0.0036	0.0081
	(0.0062)	(0.0153)	(0.0106)	(0.0079)
$Q_{k,t}^{mm,mean}$	0.0381**	0.0501	0.0736***	0.0232
	(0.0178)	(0.0512)	(0.0254)	(0.0239)
$\log(1+Q_{i,k,t} - Q_{k,t}^{mm,min})$	0.0020**	-2.887×10^{-5}	0.0031**	0.0029
	(0.0009)	(0.0008)	(0.0014)	(0.0020)
$Q_{i,k,t} - Q_{k,t}^{mm,min}$	0.0008	-0.0017	0.0009	0.0014
	(0.0021)	(0.0061)	(0.0028)	(0.0029)
$\log(1+Q_{k,t}^{mm,max} - Q_{i,k,t})$	-0.0025**	-0.0012	-0.0022	-0.0039**
	(0.0010)	(0.0010)	(0.0016)	(0.0018)
$Q_{k,t}^{mm,max} - Q_{i,k,t}$	0.0003	0.0008	-0.0045	0.0035
	(0.0028)	(0.0085)	(0.0039)	(0.0041)
Fixed Effects:				
Account × Contract × 10 DAYS	yes	yes	yes	yes
Account × Time × Basket	yes	yes	yes	yes
Clustered SE:				
Date	yes	yes	yes	yes
Account × Maturity	yes	yes	yes	yes
Observations	21,869,649	7,942,185	8,709,012	5,218,452
R^2	0.493	0.468	0.519	0.496

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

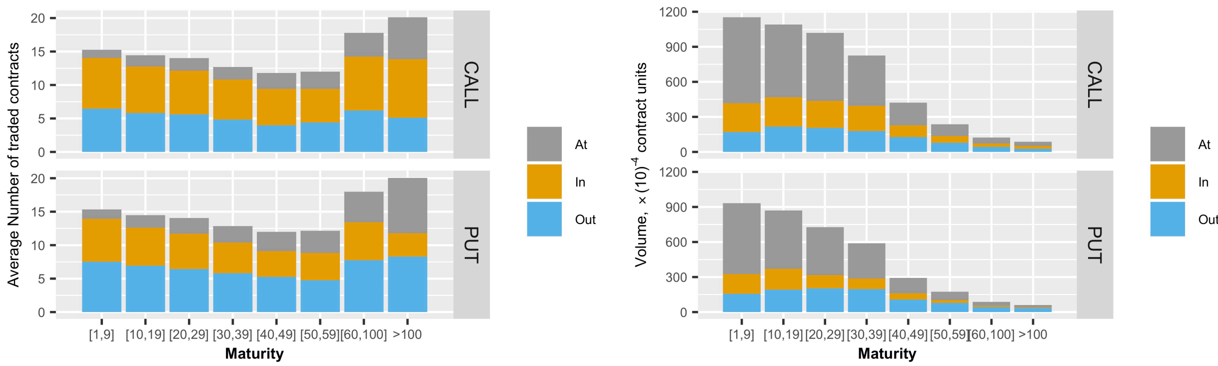
Table 6: Account-Level Regression. Activity on Bid.

Dependent Variable:	$Y_{i,k,t} := \mathbb{I}\{BestAsk_{i,k,t} - \overline{BestAsk}_{B(k),day(t)} > 0\}$			
Years	All	2015	2016	2017
$sign(Q_{i,k,t}) \times \log(1 + \lvert Q_{i,k,t}\rvert)$	0.0068***	0.0031***	0.0074***	0.0101***
	(0.0016)	(0.0010)	(0.0018)	(0.0022)
$Q_{i,k,t}$	-0.0031	-0.0017	0.0006	-0.0098
	(0.0061)	(0.0239)	(0.0101)	(0.0081)
$Q^{mm,mean}_{k,t}$	-0.0392**	-0.0339	-0.0637**	-0.0370*
	(0.0183)	(0.0372)	(0.0310)	(0.0204)
$\log(1 + Q^{mm,max}_{k,t} - Q_{i,k,t})$	0.0018*	5.275×10^{-6}	0.0015	0.0019
	(0.0010)	(0.0009)	(0.0016)	(0.0022)
$Q^{mm,max}_{k,t} - Q_{i,k,t}$	0.0042***	0.0040	0.0058	0.0042*
	(0.0016)	(0.0053)	(0.0041)	(0.0024)
$\log(1 + Q_{i,k,t} - Q^{mm,min}_{k,t})$	-0.0029***	-0.0014**	-0.0029**	-0.0033*
	(0.0009)	(0.0007)	(0.0013)	(0.0019)
$Q_{i,k,t} - Q^{mm,min}_{k,t}$	-0.0011	0.0010	-0.0026	-0.0009
	(0.0021)	(0.0068)	(0.0025)	(0.0025)
Fixed Effects:				
Account × Contract × 10 DAYS	yes	yes	yes	yes
Account × Time × Basket	yes	yes	yes	yes
Clustered SE:				
Date	yes	yes	yes	yes
Account × Maturity	yes	yes	yes	yes
Observations	21,576,105	7,805,889	8,610,381	5,159,835
R^2	0.49252	0.46664	0.51941	0.49399

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

Table 7: Account-Level Regression. Activity on Ask.

Figures



(a) Average Number of Traded contracts (b) Average Volume in contract units

Figure 1: Option contracts, summary

The upper chart corresponds to Call options, while the lower chart corresponds to Put options. Each column represents contracts with different maturities. The legend next to the plot indicates the mapping between colors and moneyness, with a few At-the-money contracts represented in grey. The average number of contracts increases as maturity approaches zero, reflecting the occasional issuance of new short-term contracts. The last two columns cover longer periods, with a higher average number of these contracts due to observations in the last months, where numerous long-term contracts were issued and did not have the chance to become short-term contracts in our sample. Panel B illustrates the average trading volume during a 10-minute interval for each type of contract, measured in contract units. At-the-money options emerge as the most actively traded. In the Appendix, the same Figure 21 is reported using unified baskets $\mathcal{B}_2^u$, aggregating all contracts into At- vs. Not At-the-money. The difference between volume in contract units and RBN (Renminbi)-volume is presented in Figure 22 .

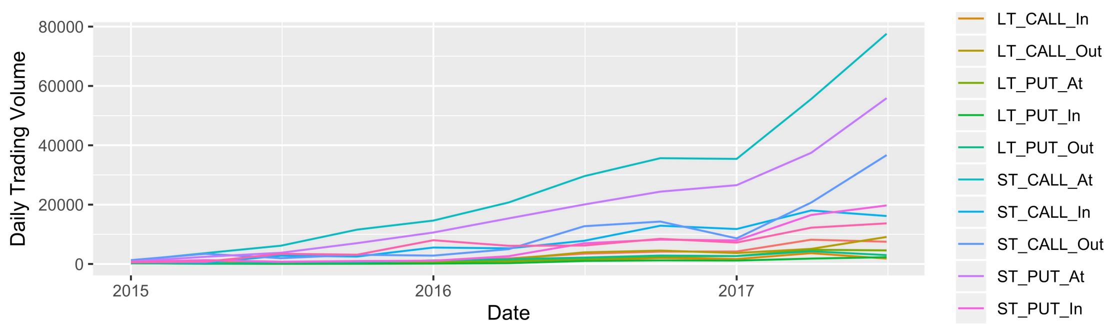
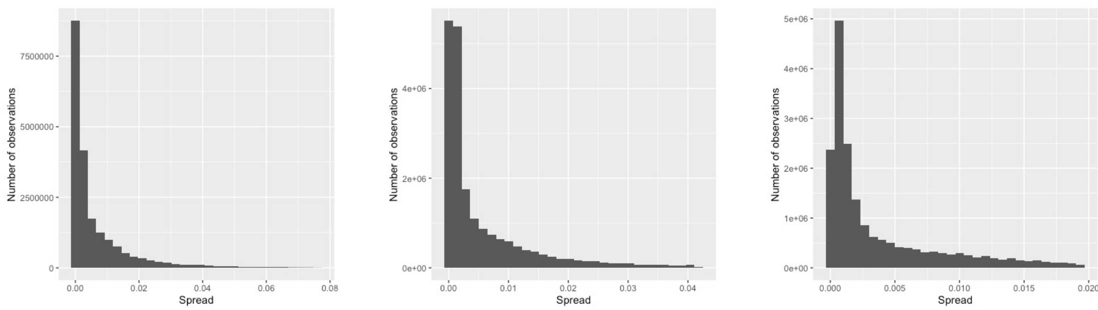


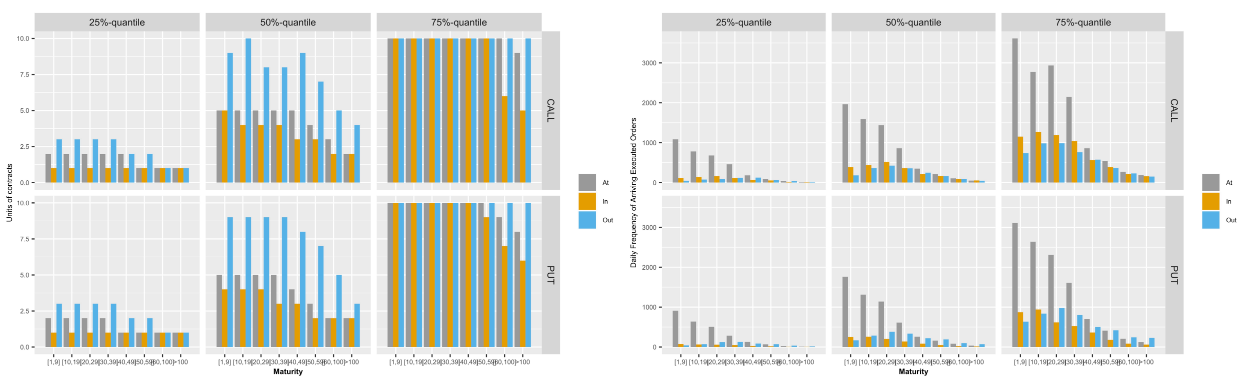
Figure 2: Trading Volume for $B \in \mathcal{B}_1$, quarterly.

Figure 29 presents the same plot while excluding Short-Term At-the-money Call and Put, and Short-Term Out-of-the-money Call baskets.



(a) $q = 99.5\%$ (b) $q = 97.5\%$ (c) $q = 90\%$

Figure 3: Observed Spread winsorized at different upper quantile q



(a) Average size of an order (b) Average Daily Frequency of orders

Figure 4: **Executed Orders Summary.** Panels (a) and (b) report a summary over executed orders ' distribution. Every column of the panels corresponds to a specific quantile in the distribution of the variable of interest. The leftmost and rightmost columns correspond to 25% and 75% quantile, respectively. The middle column reports the median values of the variable.

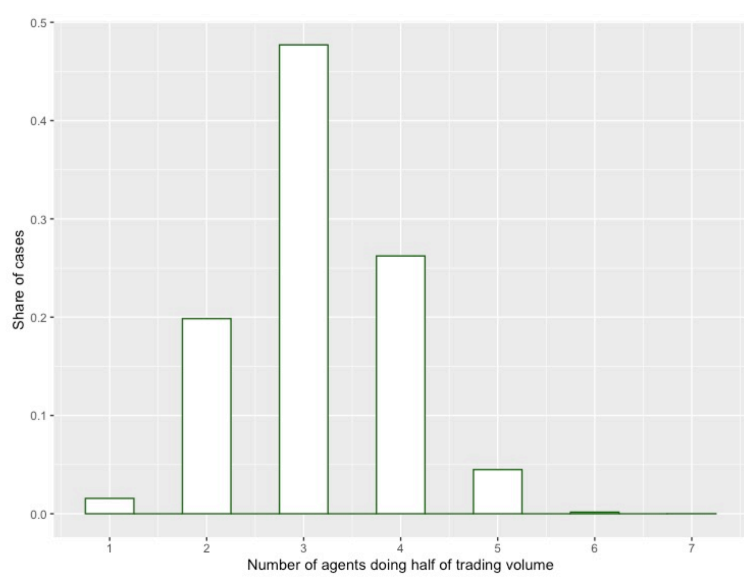
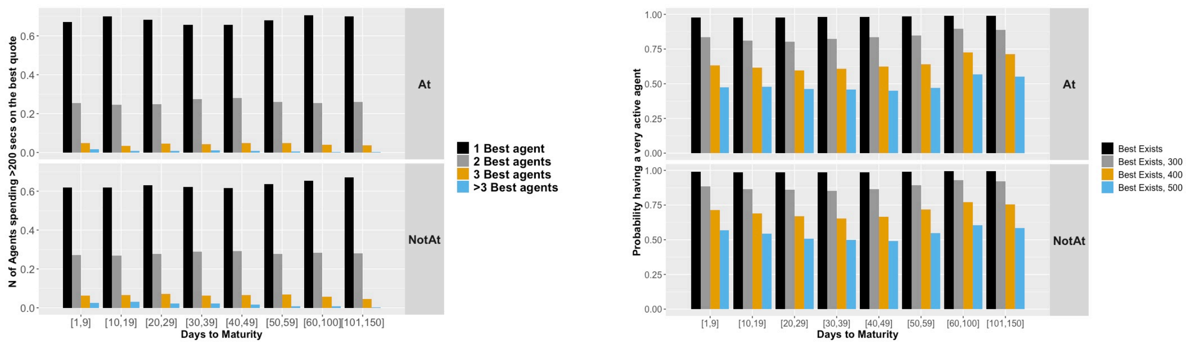


Figure 5: Is trading volume equally split between market makers?



(a) Number of “leading” MMs

(b) Probability of having “leading” MMs

Figure 6: **Existence of “leading” market makers.** Panel (a) defines an agent to be “leading” if she spends at least one-third of ten-minutes interval on the best quote. The bars show the probabilities of having one, two, three, or more “leading” agents. Panel (b) reports probability of having a “leading” market maker conditional on different definitions of being “leading”. The black bar corresponds to baseline definition: agent spends at least $t = 200$ seconds on the best quote. Other bars correspond to $t = 300, 400, 500$ respectively.

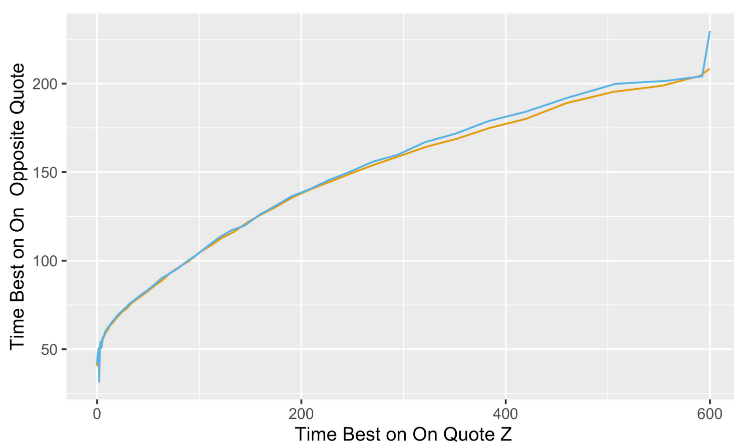
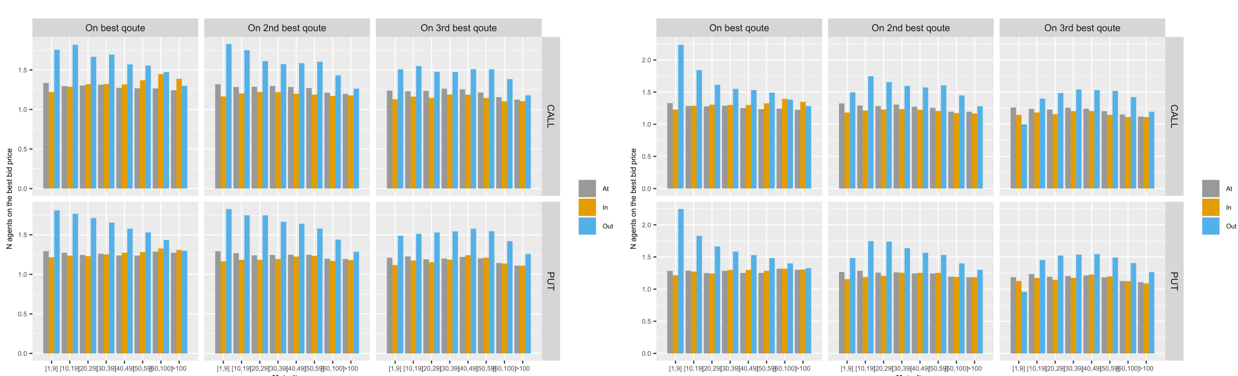


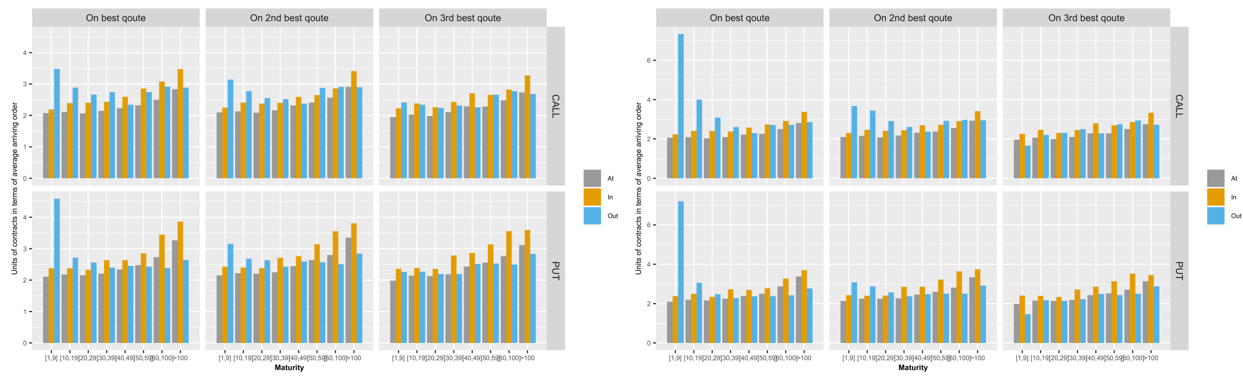
Figure 7: **Is agent active on both sides of spread?** The plot shows the average time on the best quote conditional on average time on the opposite best quote. For the blue/orange line conditional time is for quote Z =Ask/Bid. The length of the time interval is ten minutes. The axes units are seconds.



(a) Ask

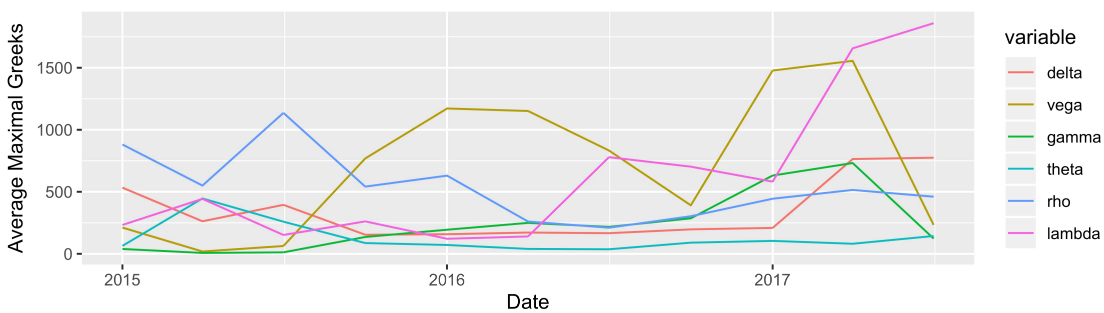
(b) Bid

Figure 8: Number of the agents with the best ask and the best bid

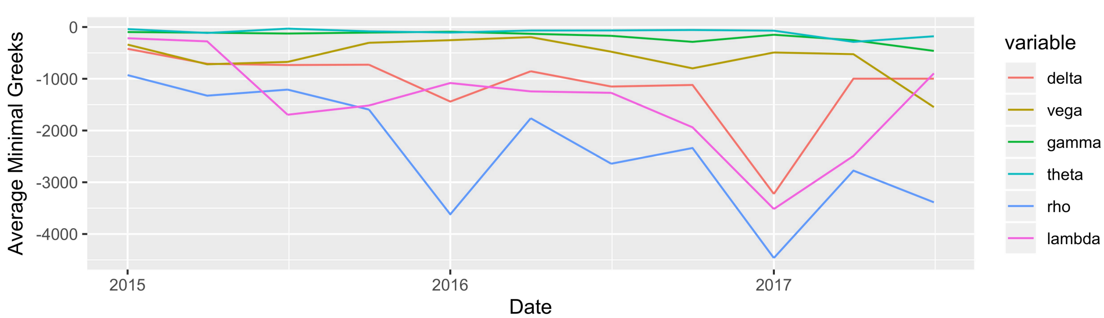


(a) Ask (b) Bid

Figure 9: Volume on the best quotes normalized by average size of executed orders

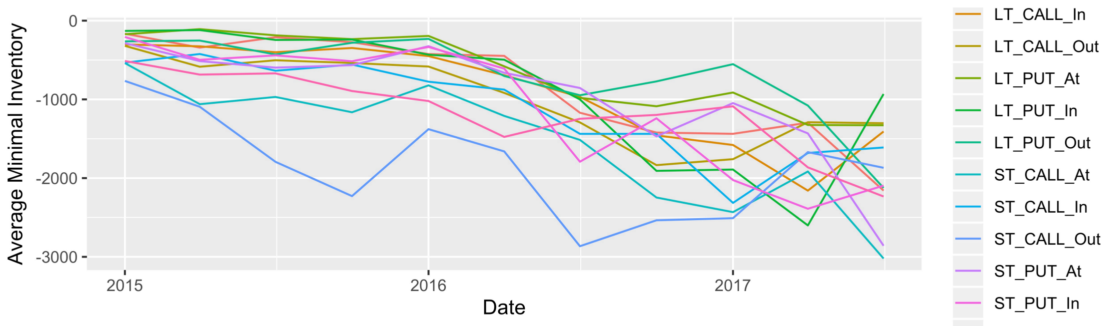


(a) Maximal Portfolio Greeks

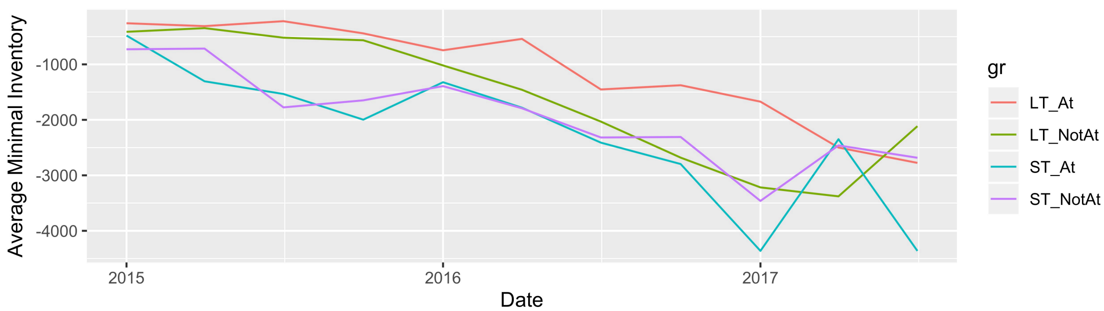


(b) Minimal Portfolio Greeks

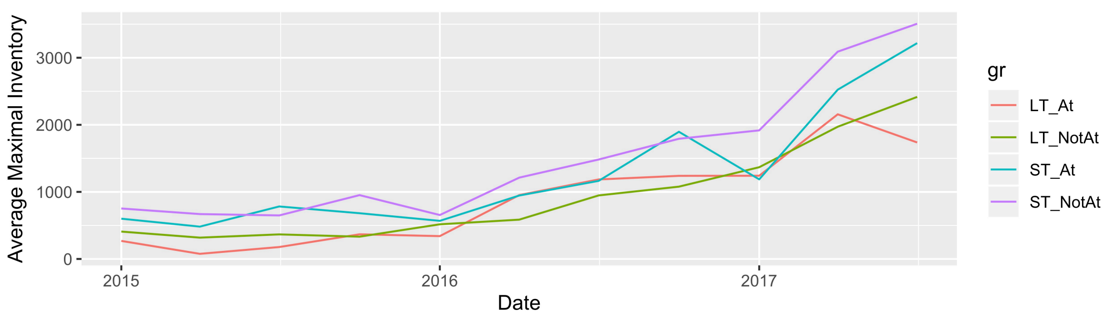
Figure 10: **Maximal and Minimal Portfolio Greeks held by MMs.** The reported values are defined as the following. For a portfolio of every market maker, we calculate its greeks. Take the portfolio with a maximum (minimum) specific greek for every day. Take the average over a quarter. Greeks are scaled to feet the same plot. All variables are scaled by the inverse of the number of options in one contract, i.e., 10^{-4}. On top of that, delta is scaled by 10^{-1}. Vega and Gamma are scaled by 10^{-2}. Rho and Lambda are scaled by 10^{-4}. Theta is scaled by 10.



(a) Average Minimal Inventory for $B \in \mathcal{B}_1$

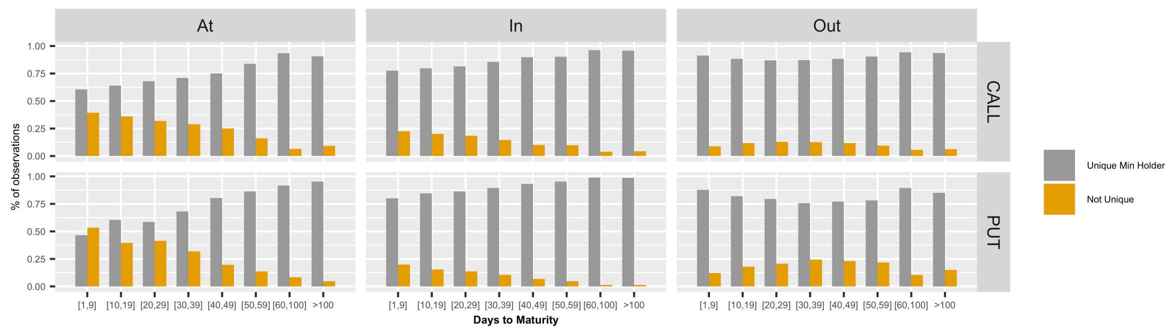


(b) Average Minimal Inventory, Q_t^u by basket $B \in \mathcal{B}_1^u$

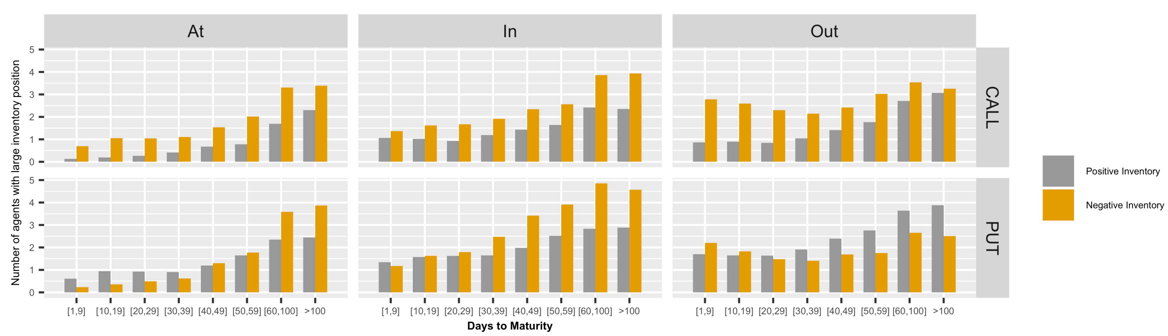


(c) Average Maximal Inventory, Q_t^u by basket $B \in \mathcal{B}_1^u$

Figure 11: **Minimal and Maximal Inventories by Contract groups.** The reported value is the averaged over similar contracts, i.e. from the same basket, and quarters maximal (minimal) inventory position in the contract. Maximal (minimal) inventory position in the contract is just the maximum (minimum) inventory in the contract over the set of market makers.



(a) Share of observations for which there is a unique market maker with a minimum level of inventories. An agent is unique if her inventories are smaller than inventories of all other agents at least by average trading volume per 10 minutes interval. Figure 31 in the Appendix reports similar results for a share of observations with maximum inventories. Figure 32 shows probabilities for having one or two minimal/maximal market makers.



(b) Average number of agents which hold substantial positive or negative inventory position. Position is defined to be a *substantial* if its absolute value is larger than an average trading volume over 30 minutes for this type of contracts.

Figure 12: Agents with substantial inventories

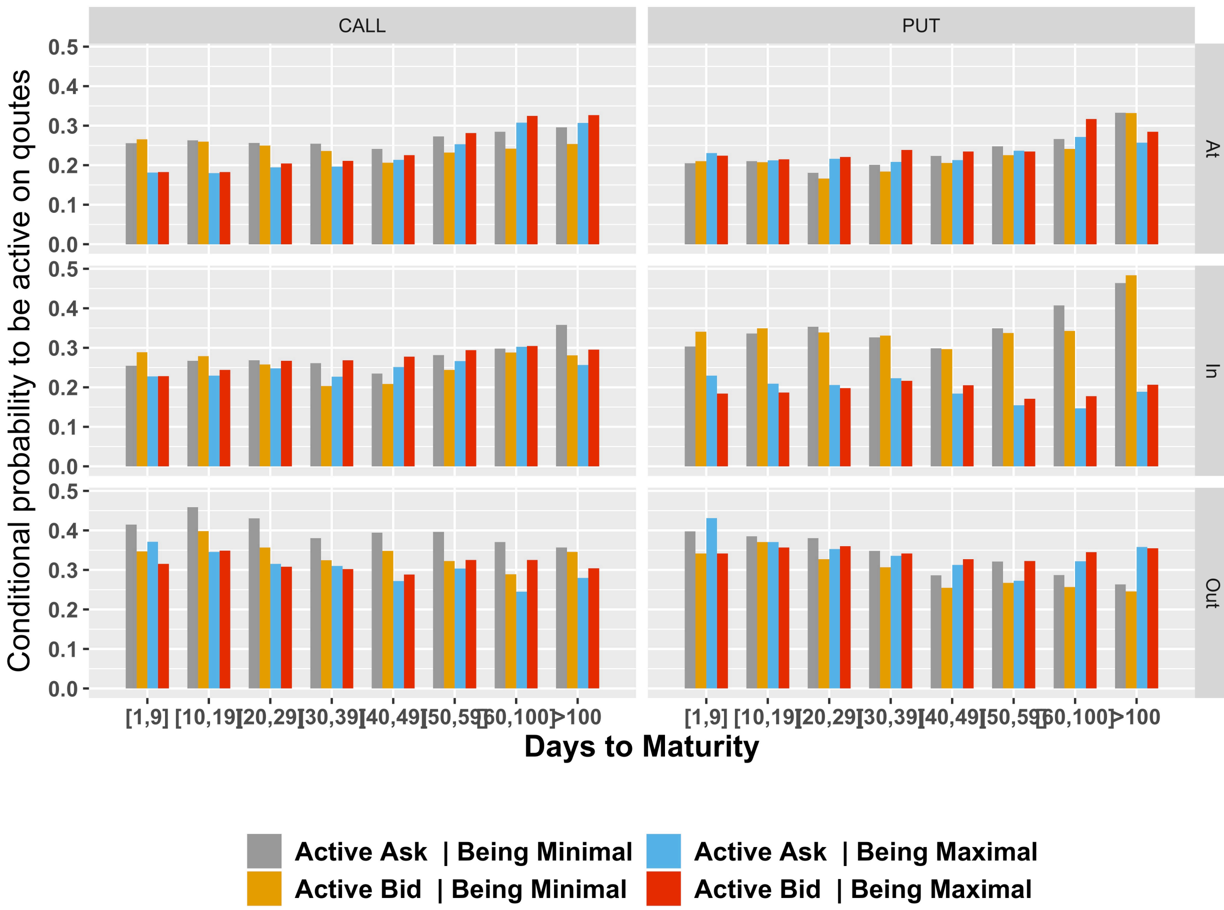


Figure 13: Probability of being an active market maker in a contract conditional on being maximal or minimal market maker (in the same contract). An active on ask/bid market maker is defined as the market maker spending at least 5 minutes on the best or second best ask/bid prices.

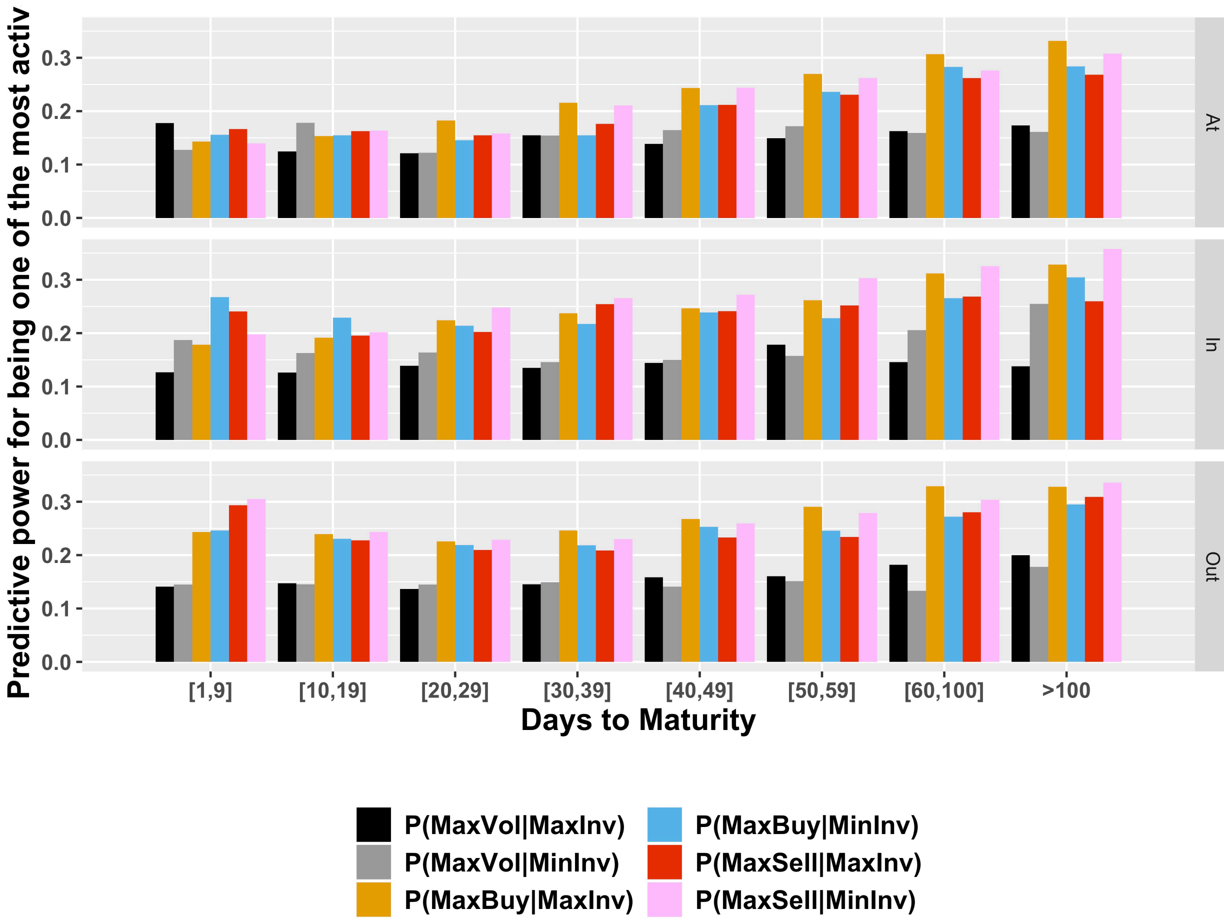
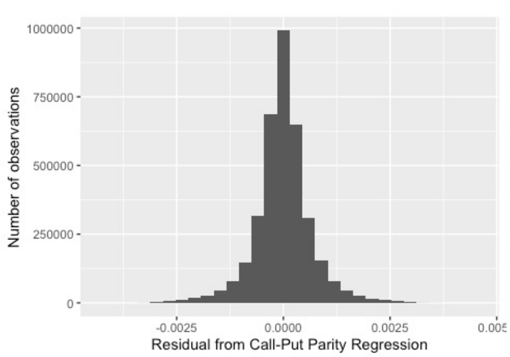
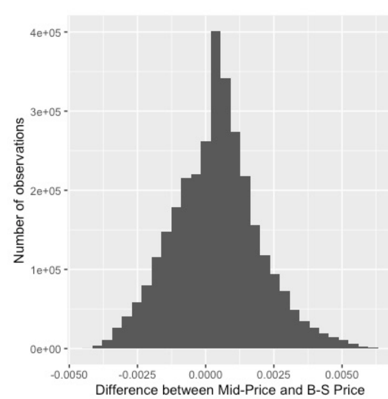


Figure 14: Probability of being the agent with highest trading volume, buy or sell, conditional on having the largest/smallest inventory in the contract, $Q_{i,k,t}$. The interval of a measurement is 10 minutes.



(a) $\varepsilon_{K,T}$ from (5)



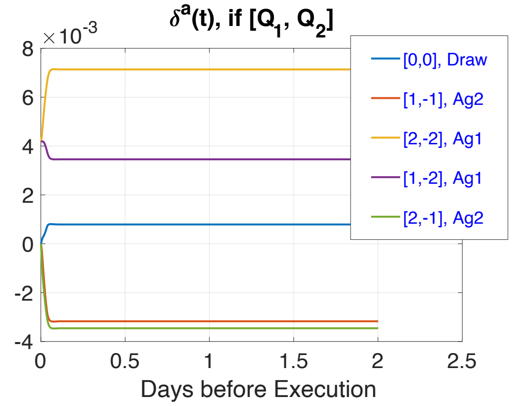
(b) $P_{k,t} - \hat{P}_{k,t}$

Figure 15

$t \in \mathcal{T}^1$. Data is winsorized at 0.5% level from each side

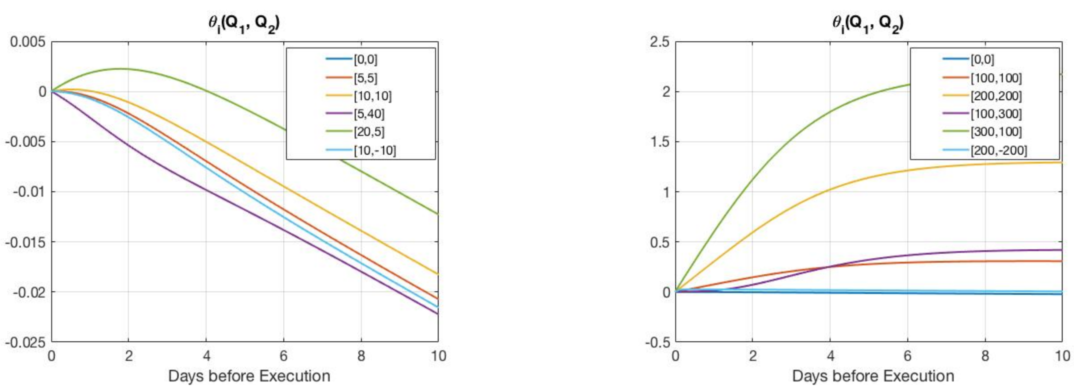
No Ask $[-2,-2]$	Ag2	Ag2	Ag2	Ag2
Ag1	Draw	Ag1	Ag1	Ag1
Ag1	Ag2	Draw $[0,0]$	Ag2	Ag1
Ag1	Ag2	Ag1	Draw	Ag1
Ag1 $[2,-2]$	Ag2 $[2,-1]$	Ag2	Ag2	Draw $[2,2]$

(a) Who posts the best Ask



(b) Ask conditional on inventories

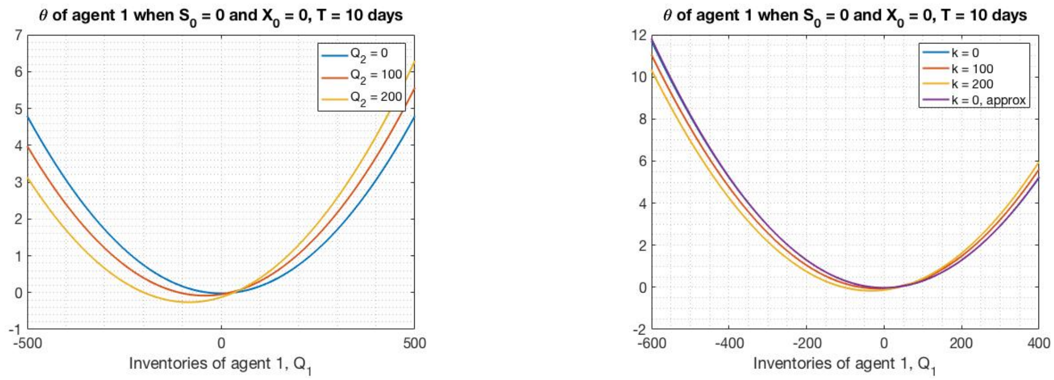
Figure 16: Each cell of the table (a) corresponds to a state with inventory $q = [q_1, q_2]$, $q_i \in \{-2, -1, 0, 1, 2\}$. Rows correspond to q_1 ordered by size with $q_1 = -2$ as a top row and $q_1 = 2$ as a bottom row. Columns correspond to q_2 with $q_2 = -2$ as the leftmost column and $q_2 = 2$ as the rightmost column. A cell shows who posts the best ask price at the state. “Draw” means that agents post the same ask price. Surprisingly, even in the case when $q_1 > q_2$ agent 2 can qoute better Ask price. For example, if $q_1 = 2$ and $|q_2| \neq 2$, agent 2 provides both best bid price and best ask price. The reason why agent 2 provides better price then agent 1 is because she wants to keep the monopoly power on the best bid. Plot (b) shows “Sell” price at different states $q = [q_1, q_2]$. x-axis denotes dates before the maturity. Almost immediately prices stabilize at some fixed levels independent of time to maturity. Parameters used: $S = 0$, $A = 392$, $k = 40.2$, $\gamma = 2$, $\sigma = 0.002$. k, σ are in daily units.



(a) Small values of Q and k.

(b) Large values of $\bar{q}$ and k.

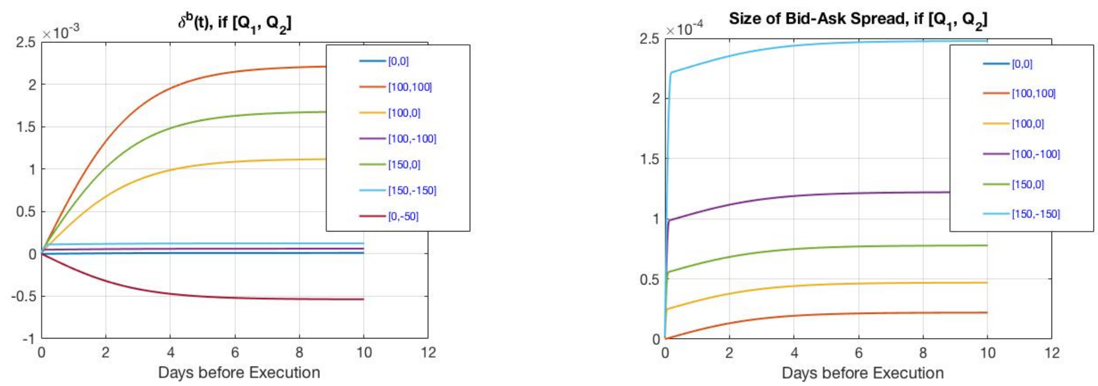
Figure 17: $\theta_1(t, q_1, q_2)$ for different values of inventory q_1 and imbalances $k = q_2 - q_1$. For any q, $\theta_1(t, q)$ achieves linear limit with downward slope. Less time is needed for small inventories. The blue and light blue lines almost coincide on the plots, i.e., $q(t, k, -k) \approx q(t, 0, 0)$.



(a) Change q_1, q_2 fixed

(b) Change q_1, k fixed

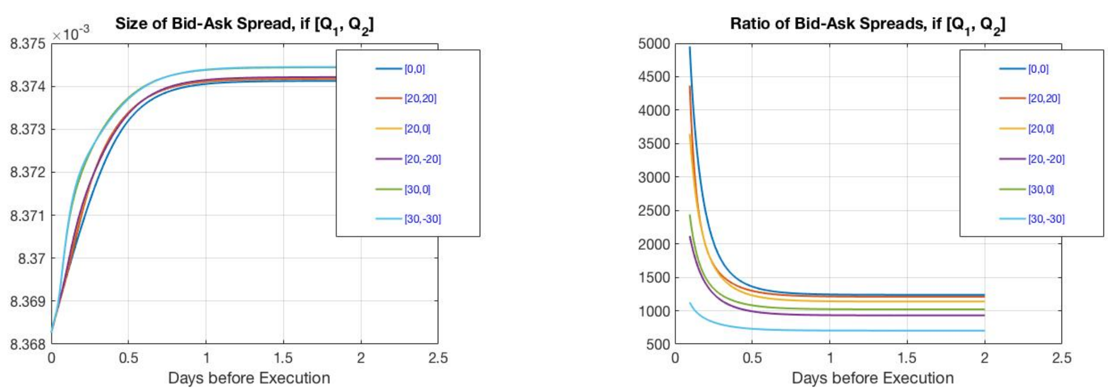
Figure 18: $\theta_1(T, q_1, q_2)$ for different values of inventory q_1 and q_2. Panel *a* shows how θ_1 changes when q_2 is fixed and q_1 moves. Panel *b* shows $\theta_1(T, q_1, q_1 + k)$ when k is fixed. Purple line corresponds to solution of (13), for calibrated parameters $\theta_{0,0}$ and a. $T = 10$ days.



(a) $\delta^{b,*}(t, q_1, q_2)$

(b) $Spread(t, q_1, q_2)$

Figure 19: Panel (a) shows dependence of difference between fundamental value and best bid price, $\delta^{b,*}(r, q_1, q_2)$, on time before maturity for different values of agents' inventories q_1 and q_2. Panel (b) reports observed spread, $\delta^{a,*}(r, q_1, q_2) + \delta^{b,*}(r, q_1, q_2)$.



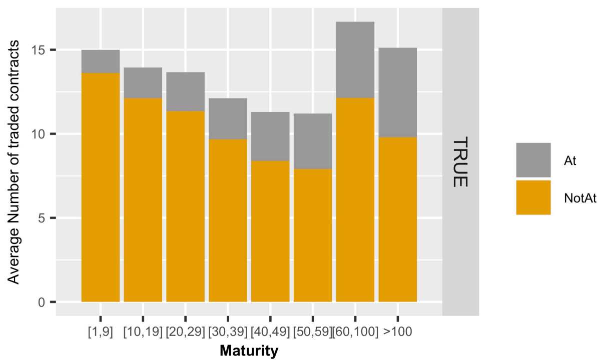
(a) Cooperative Spread

(b) Ratio of Spreads

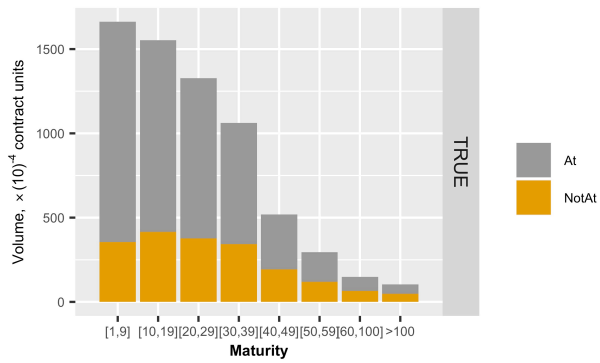
Figure 20: Plot (a) shows spread in cooperative equilibrium for different values of q. Plot (b) shows the ratio of spread in cooperative equilibrium to spread in non-cooperative equilibrium. $T = 2$ days. $\bar{Q} = 60$.

11 Appendix

11.1 Supplementary Plots and Tables

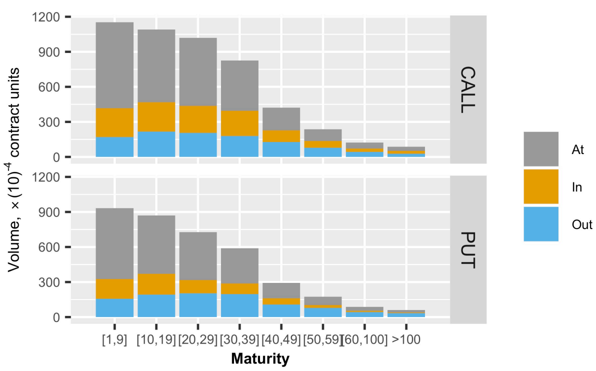


(a) Average Number of Traded contracts

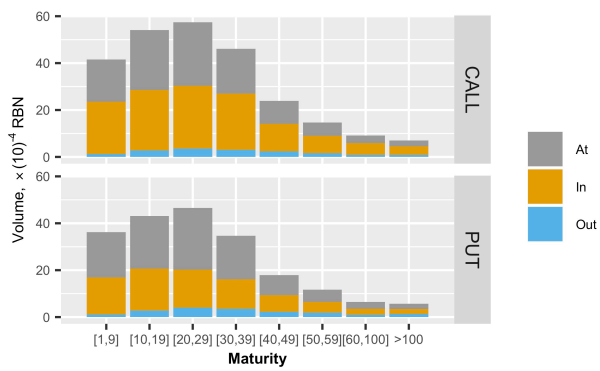


(b) Average Volume in contract units

Figure 21: Option contracts, summary. Each column corresponds to contracts with different maturities. The legend next to the plot specifies mapping between colors and moneyness. Contract at the money are those which have $|Delta| \in (.375, .625)$. Other contracts are classified as not at the money. All Panel B shows average trading volume during a 10 minutes interval for each type of contracts. Volume is measured in contract units. At-the-money options are the most actively traded options.



(a) Contract units



(b) RBN

Figure 22: Volume in terms of contract units vs Volume in terms of Chinese currency RBN. The value is the average over a ten-minutes interval for the continuous trading session.

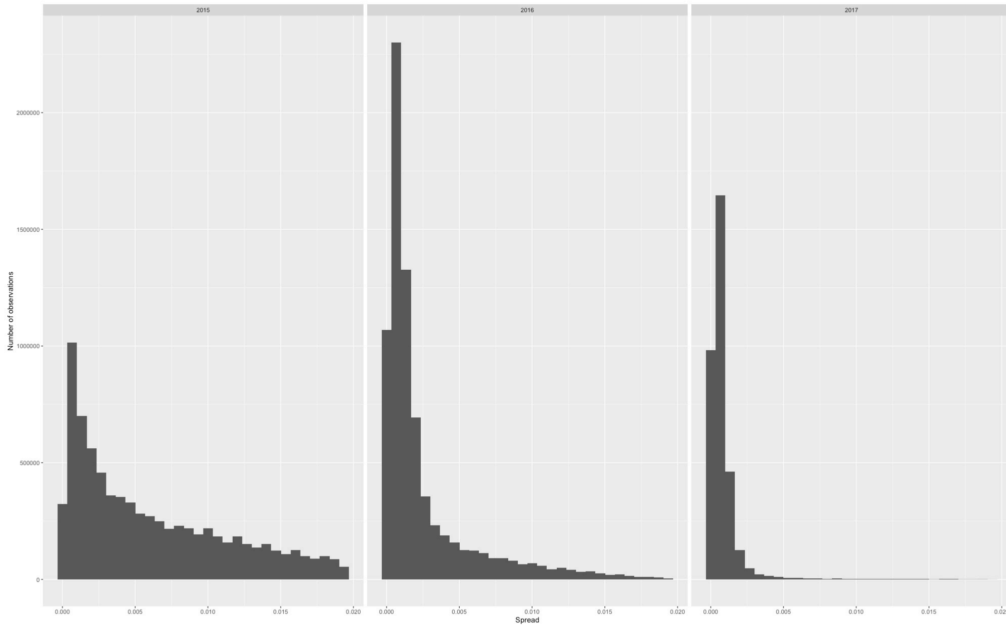


Figure 23: Observed Spread winsorized at different upper quantile q

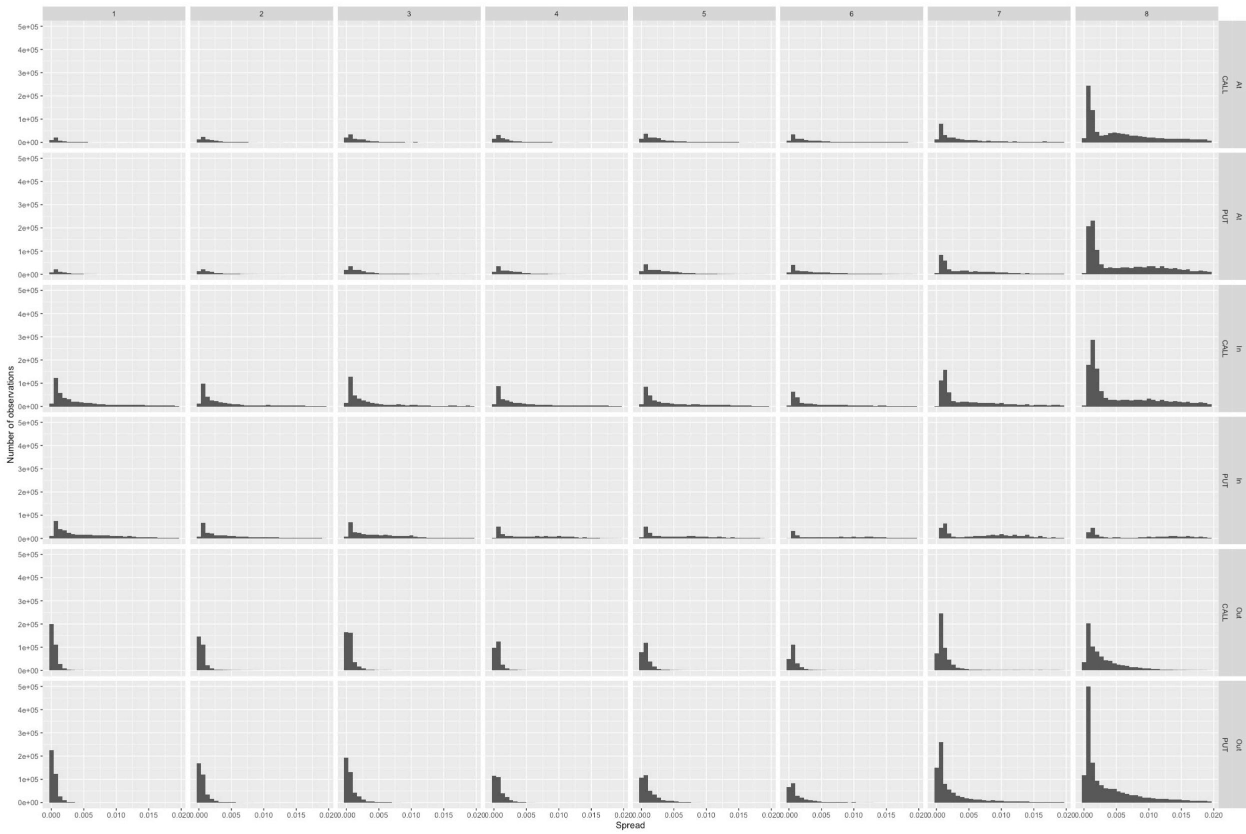


Figure 24: **Spread by Basket $\mathcal{B}_2$.** Observed Spread by Basket $\mathcal{B}_2$ winsorized at 90% quantile. Histograms from the left to the right corresponds to contracts different by time to maturity with the shortest-term on the left. Odd/even rows of the panel corresponds to Call/Put contracts. Top/mid/bottom two rows of the panel correspond to At-/In-/Out- the money contracts.

11.2 Full HJB System

11.2.1 Variables

- $x = [x_1, x_2] \in \mathbb{R}^2$
- $q = [q_1, q_2] \in \mathcal{Q}^2$, where $\mathcal{Q} = \{-Q, -Q+1, \ldots, 0, \ldots Q\}$ for $Q \in \mathbb{N}$
- For $i = 1,2$ define basis vectors in $\mathbb{R}^2$ by e_i.
- For $i = 1,2$: $\quad q^{i,b} = q + e_i$, $q^{i,a} = q - e_i$.
- For $i = 1,2$: $\quad x^{i,b} = q - e_i \cdot \delta^{\min,z}$, $x^{i,a} = x + e_i \cdot \delta^{\min,z}$.
- Other state variables are the current value of stock and time to maturity, $s, t \in \mathbb{R}$.
- Parameters, $\sigma, a, k, \gamma \in \mathbb{R}_+$

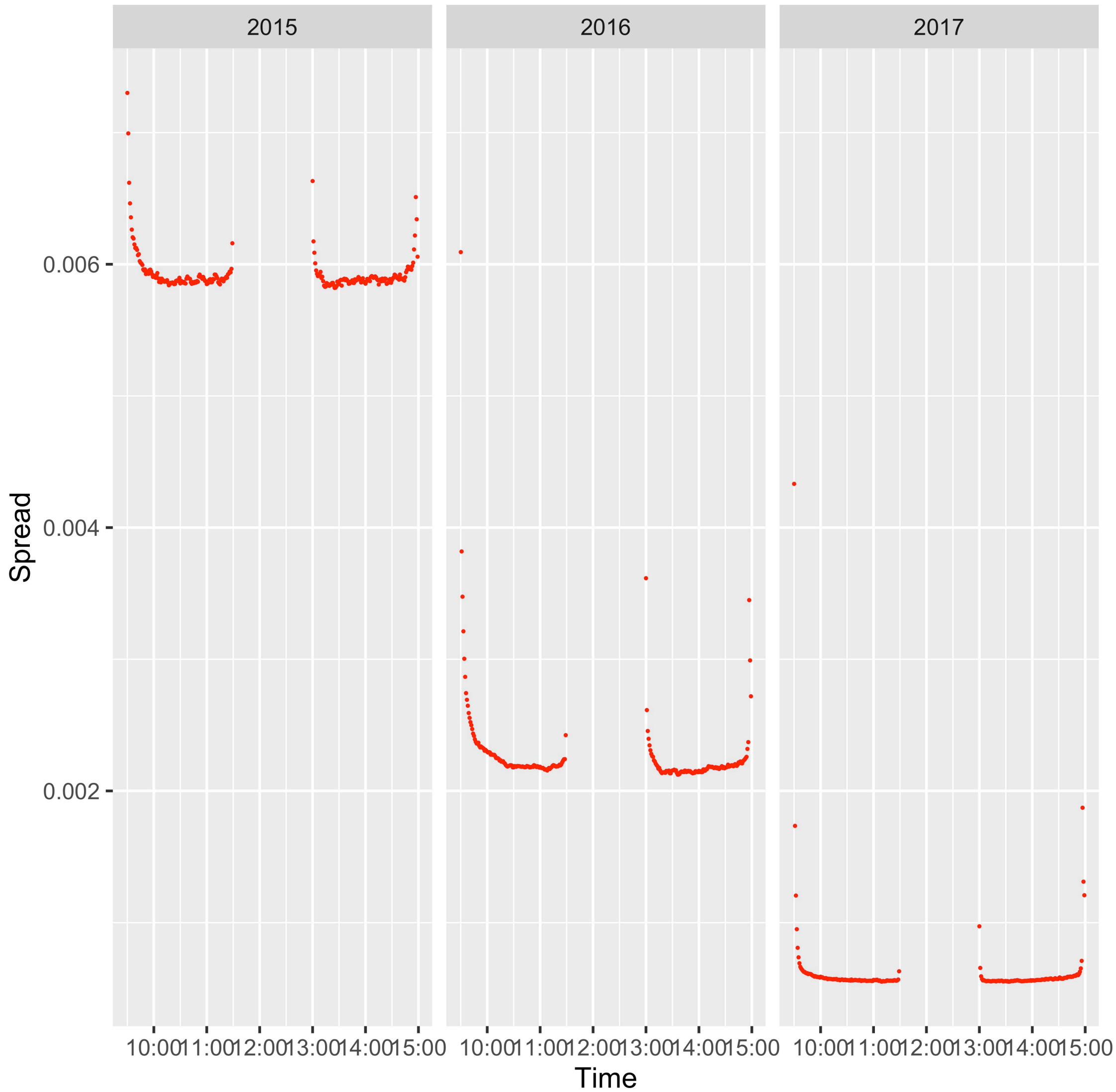


Figure 25: Intra-day Spread

- Functions $\lambda^z(\delta) = Ae^{-k\delta}$, $z = \{a, b\}$ (i.e. the same function for now)

11.2.2 Equilibrium with no cooperation

Firstly, let us define $\mathbb{U}_i^z(t, x, q, s)$ for each discrete parameter $q = [q_1, q_2]$.

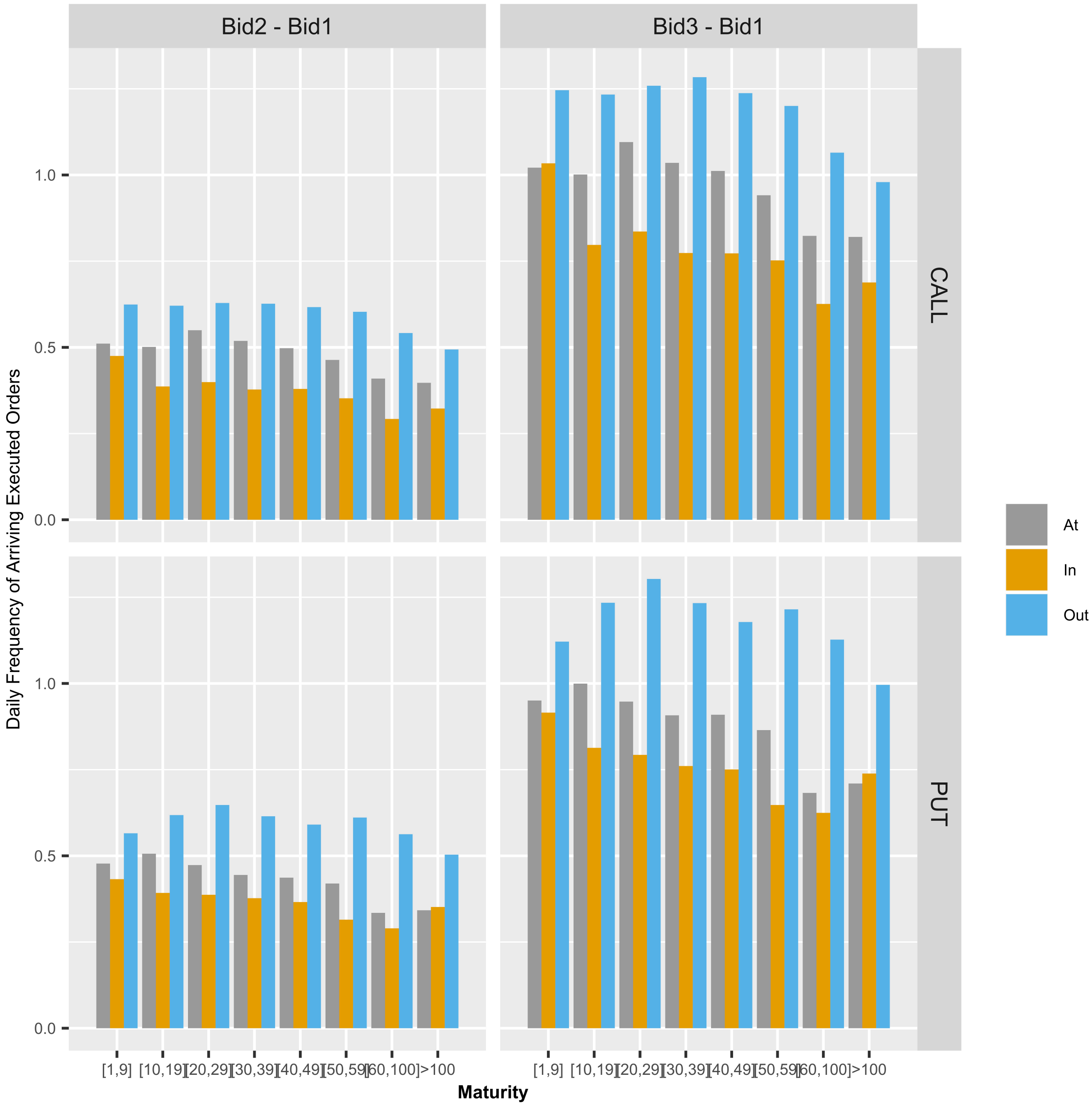


Figure 26: **Difference between the best bid price and the Nth best bid price.** The difference is normalized by the average spread of the considered contract at the time.

For most of the states $-Q < q_i < Q$ and $-Q < q_j < Q$, $j \neq i$

$$\begin{aligned}
&\mathbb{U}_i^z(t,x,q,s) = \\
&\sup_{\delta_i^z} \Bigg(\lambda^z \left(\delta^{\min,z}\right) \left[u_i\left(t,x^{i,z},q^{i,z},s\right) - u_i(t,x,q,s)\right] \cdot \left[\mathbb{I}\{\delta_i^z < \delta_j^z\} + \frac{1}{2}\mathbb{I}\{\delta_i^z = \delta_j^z\}\right] \; + \\
&\lambda^z \left(\delta^{\min,z}\right) \left[u_i\left(t,x^{j,z},q^{j,z},s\right) - u_i(t,x,q,s)\right] \cdot \left[\mathbb{I}\{\delta_i^z > \delta_j^z\} + \frac{1}{2}\mathbb{I}\{\delta_i^z = \delta_j^z\}\right] \Bigg)
\end{aligned} \tag{19}$$

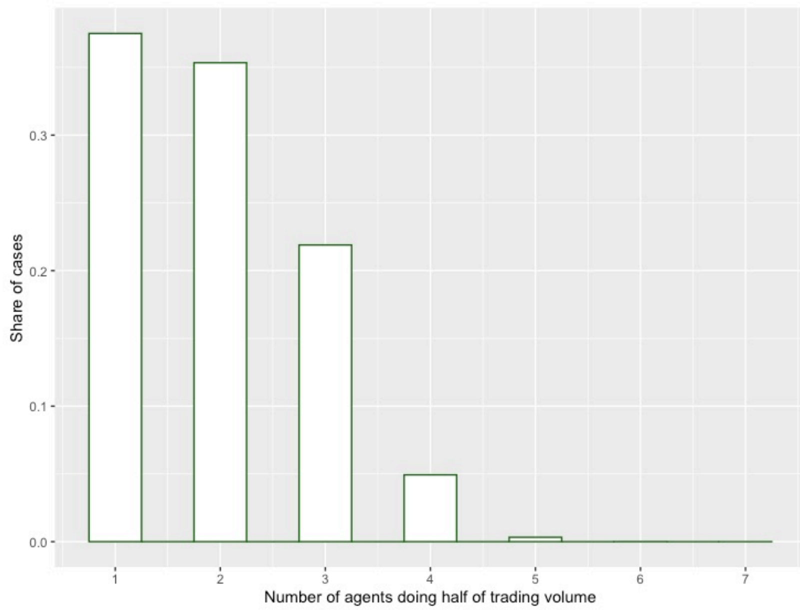


Figure 27: **Is trading volume equally split between market makers?** This picture reports how many agents are responsible for at least half of trading volume over a 10 minutes interval.

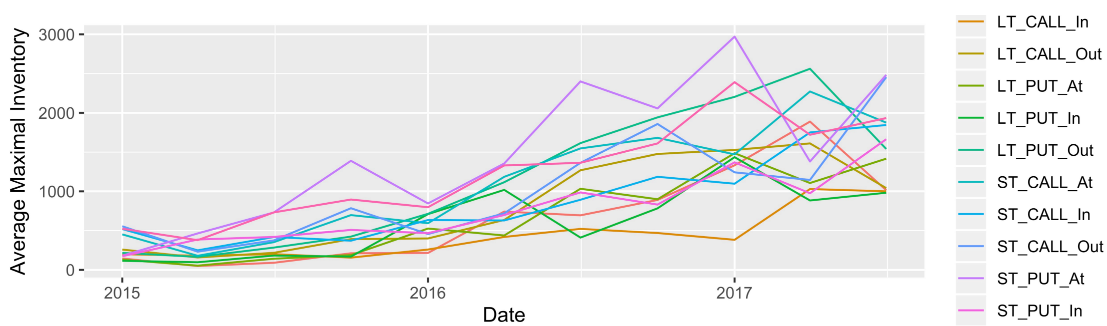


Figure 28: Average Maximal Inventory for $B \in \mathcal{B}_1$

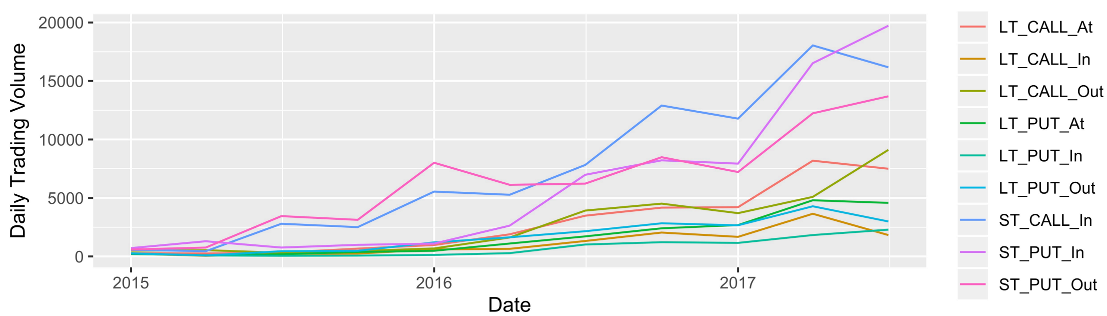


Figure 29: Trading Volume for $B \in \mathcal{B}_1$ excluding Short Term At- Call and Put, and ST Out- Call baskets.

where $\delta_i^{\min,z} = \min\{\delta_1^z, \delta_2^z\}$, $\lambda^z\left(\delta^{\min,z}\right) = Ae^{-k\delta^{\min,z}}$.

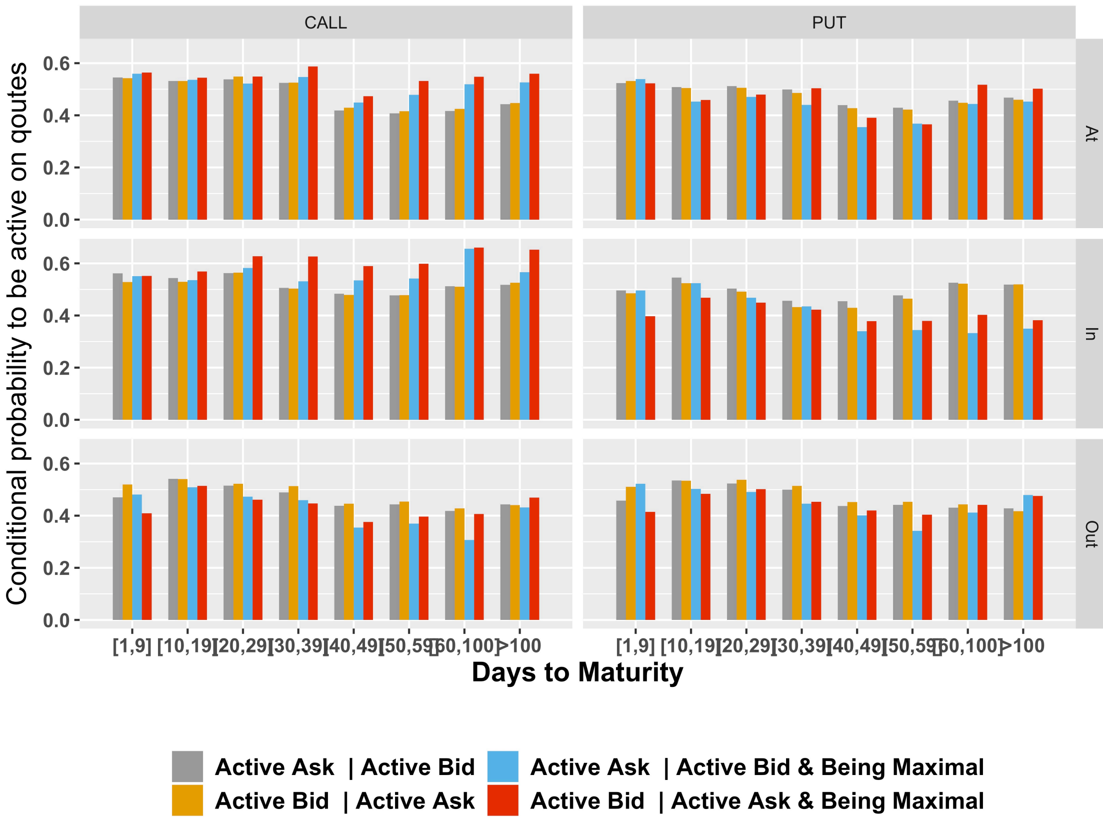


Figure 30: Probability to be active on both sides of spread.

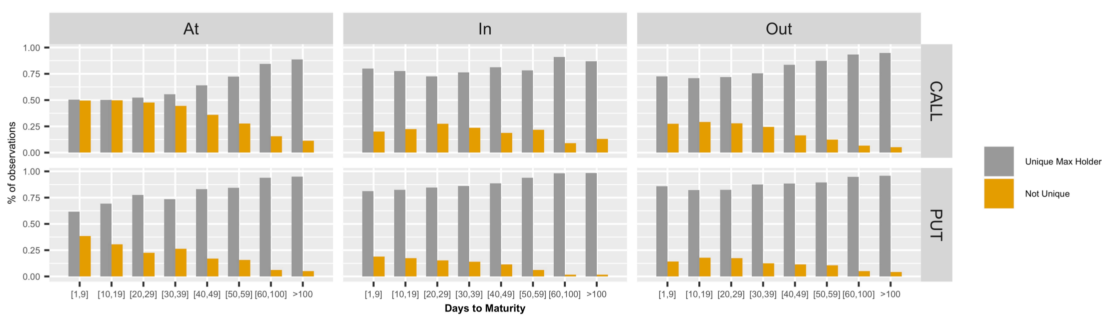


Figure 31: Share of observations for which there is a unique market maker with a maximum level of inventories. An agent is unique if her inventories are larger than inventories of all other agents at least by average trading volume per 10 minutes interval.

For $q_i = \bar{Q}$ and $-\bar{Q} < q_j \leq Q$, $j \neq i$ (cannot buy, $\delta_i^b = \infty$)

$$\mathbb{U}_i^b(t,x,q,s) = \lambda^b\left(\delta_j^b\right)\left[u_i\left(t,x^{j,b},q^{j,b},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^a(t,x,q,s) = \text{as in (19)}$$

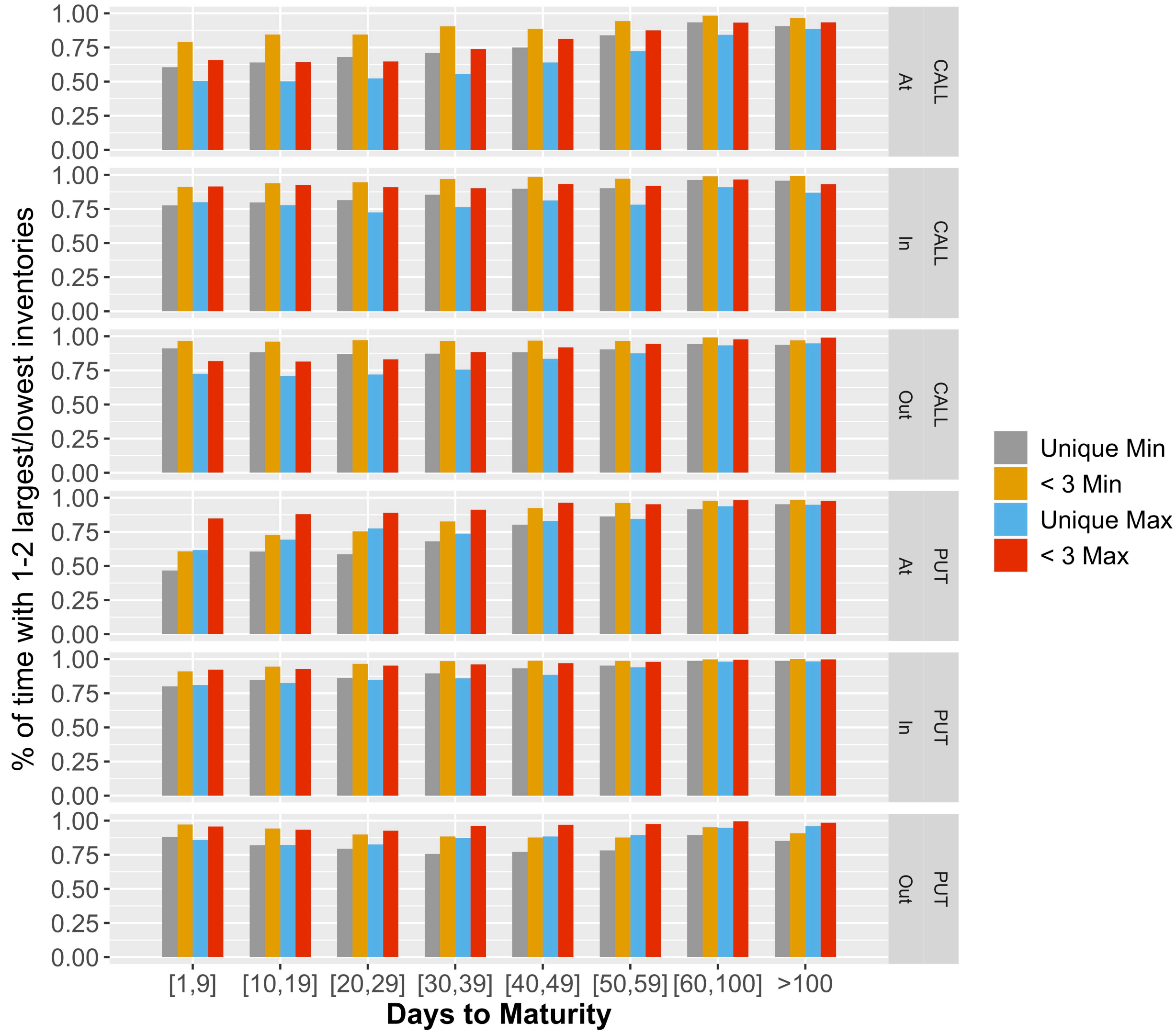


Figure 32: Share of observations for which there is a unique or at most two minimal/maximal market makers.

For $q_i = -\bar{Q}$ and $-\bar{Q} < q_j < \bar{Q}$, $j \neq i$ (cannot sell)

$$\mathbb{U}_i^a(t,x,q,s) = \lambda^a\left(\delta_j^a\right)\left[u_i\left(t,x^{j,a},q_i^{j,a},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^b(t,x,q,s) = \text{as in (19)}$$

For $q_j = \bar{Q}$ and $-\bar{Q} < q_i < \bar{Q}$, $j \neq i$

$$\mathbb{U}_i^b(t,x,q,s) = \sup_{\delta_i^b} \lambda^b\left(\delta_i^b\right)\left[u_i\left(t,x^{i,b},q^{i,b},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^a(t,x,q,s) = \text{as in (19)}$$

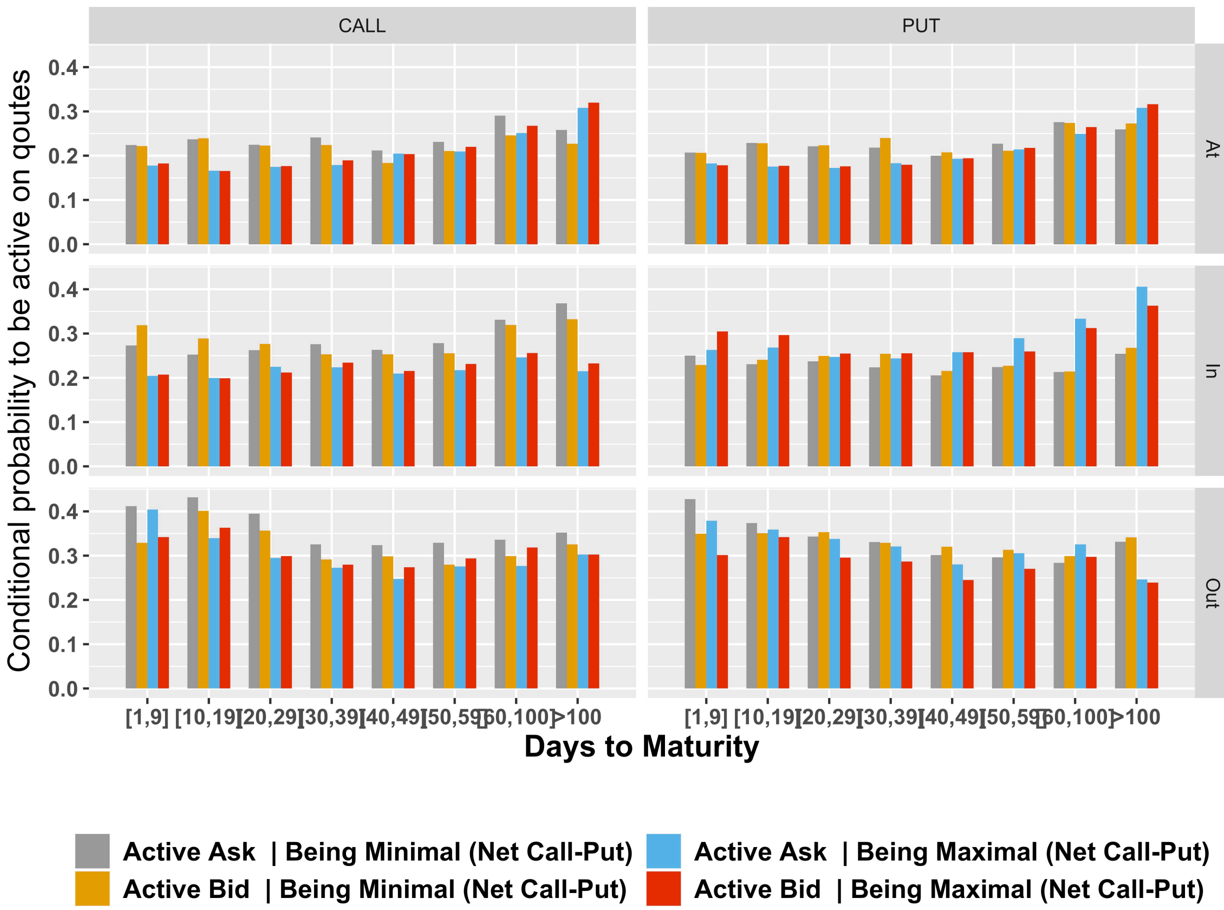


Figure 33: Probability of being an active market maker in a contract conditional on being maximal or minimal market maker (in the same contract). An active on ask/bid market maker is defined as the market maker spending at least 5 minutes on the best or second best ask/bid prices.

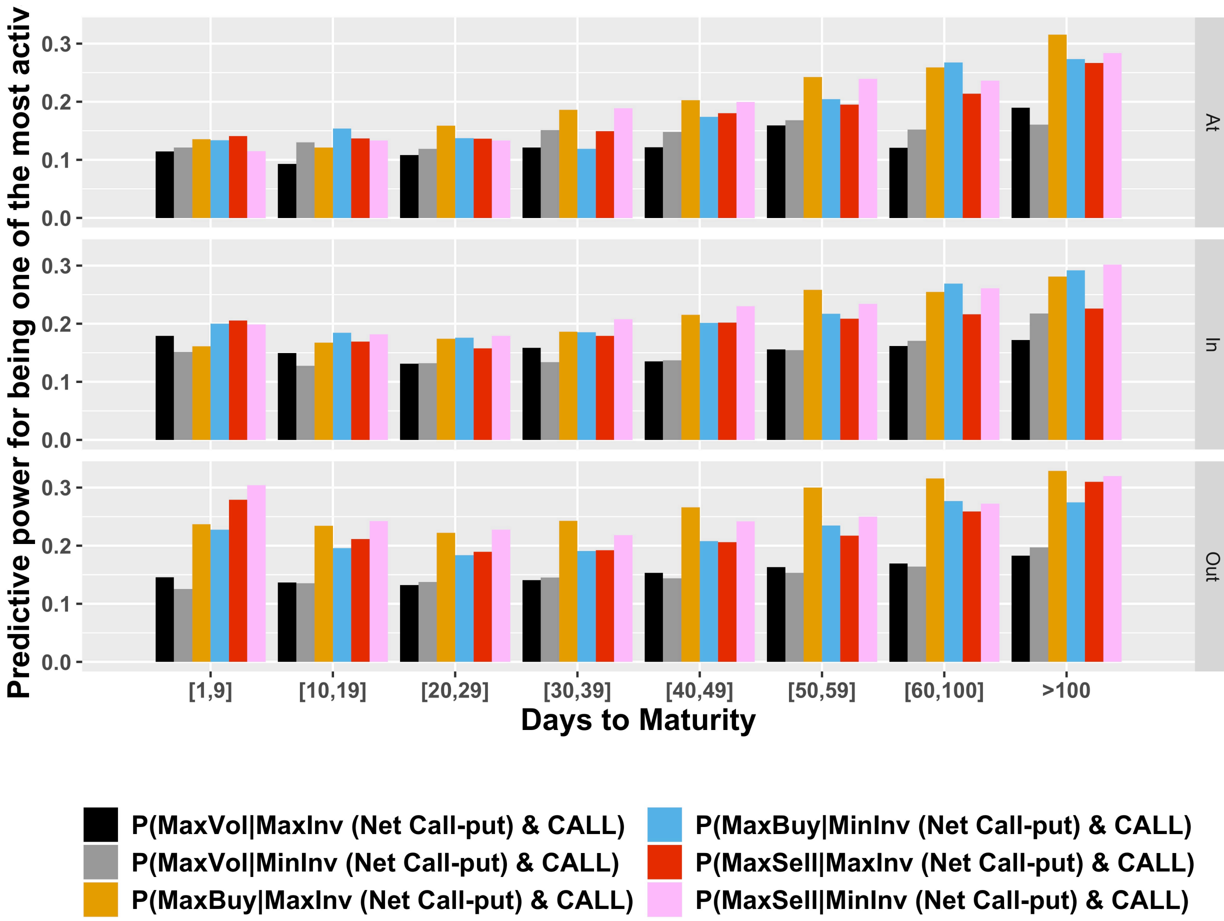


Figure 34: Probability of being the agent with highest trading volume, buy or sell, in a call option with strike K and maturity T conditional on having the largest/smallest difference between inventories in call and put with these strikes and maturities, $Q^{u}_{i,(K,T),t}$. The interval of a measurement is 10 minutes.

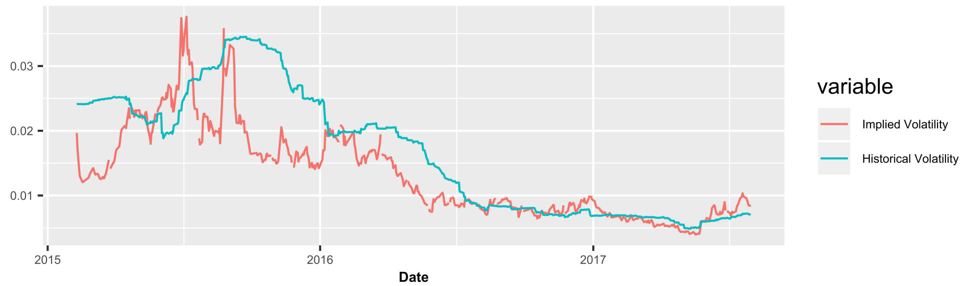
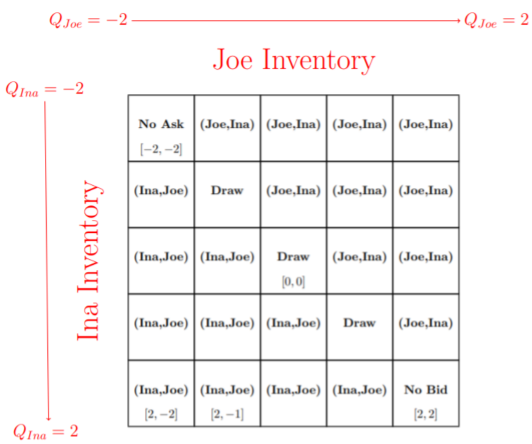
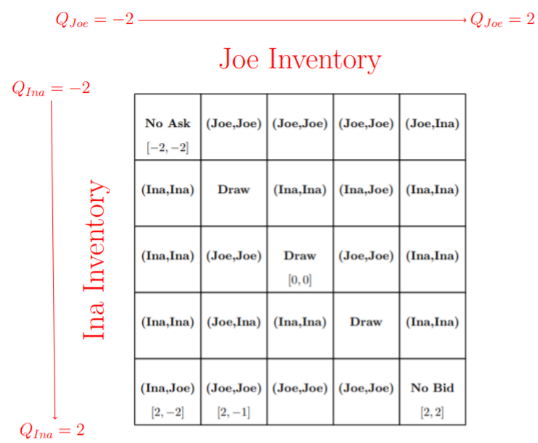


Figure 35: The plot reports daily measures of volatility in the market. Implied volatility is build based on the mid-prices of the options. Historical volatility is the volatility of ETF over the preceding 90 days.



(a) Monotonic Princple



(b) Tight Capital Constraint

Figure 36: The plots shows who is the most active agent on (Ask, Bid) at every inventory state. Inventories vary in $Q \in \{-2,-1,0,1,2\}$ for every agent. Left panel shows the case when monotonic principle holds (for instance, if the capital constraints are relaxed). The right plot reports the tight capital constraint case when $\overline{Q} = 2$ and parameters are estimated in Section 6.

For $q_j = -\bar{Q}$ and $-\bar{Q} < q_i < \bar{Q}$, $j \neq i$

$$\mathbb{U}_i^a(t,x,q,s) = \sup_{\delta_i^a} \lambda^a\left(\delta_i^a\right)\left[u_i\left(t,x^{i,a},q^{i,a},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^b(t,x,q,s) = \text{as in (19)}$$

	Contract	Time to Maturity	Quarter	$\log(Adt)$	$-\kappa$
1	PUT	2	1	2.75	-75.49
2	PUT	2	2	4.09	-577.62
3	PUT	2	3	3.02	-110.23
4	PUT	2	4	3.65	-302.60
5	CALL	2	1	3.45	-748.58
6	CALL	2	2	3.19	-375.50
7	CALL	2	3	2.88	-36.36
8	CALL	2	4	3.01	-74.54
9	PUT	3	1	3.16	-569.98
10	PUT	3	2	3.11	-157.09
11	PUT	3	3	2.61	-133.54
12	PUT	3	4	3.06	-268.54
13	CALL	3	1	3.34	-784.47
14	CALL	3	2	2.91	-203.22
15	CALL	3	3	2.70	-95.10
16	CALL	3	4	3.10	-425.61
17	PUT	4	1	3.03	-583.45
18	PUT	4	2	2.35	-336.93
19	PUT	4	3	1.70	-23.88
20	PUT	4	4	2.99	-495.00
21	CALL	4	1	3.07	-494.66
22	CALL	4	2	2.76	-299.57
23	CALL	4	3	2.22	-297.81
24	CALL	4	4	2.50	-184.48
25	PUT	5	1	2.06	-357.52
26	PUT	5	2	2.25	-530.96
27	PUT	5	3	1.34	-12.17
28	PUT	5	4	2.77	-405.78
29	CALL	5	1	2.36	-269.51
30	CALL	5	2	2.17	-327.45
31	CALL	5	3	1.84	-111.39
32	CALL	5	4	1.69	-54.22
33	PUT	6	1	1.51	-245.07
34	PUT	6	2	2.02	-514.72
35	PUT	6	3	0.76	-81.31
36	PUT	6	4	2.13	-391.23
37	CALL	6	1	1.73	-304.32
38	CALL	6	2	1.47	-406.16
39	CALL	6	3	1.16	-81.60
40	CALL	6	4	1.60	-298.35

Table 8: **Estimated parameters A and κ based on In-the-money contracts for buy orders.** Note that $-\kappa$ is the coefficient from poisson regression and it has negative sign as the model requires.

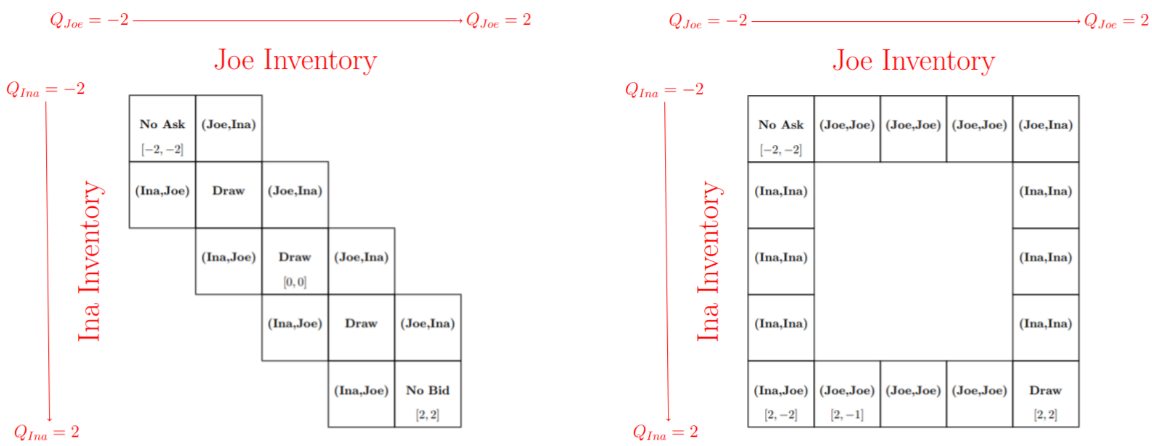
For $q_i = -\bar{Q}$ and $q_j = \bar{Q}$, $j \neq i$

$$\mathbb{U}_i^b(t,x,q,s) = \sup_{\delta_i^b} \lambda^b\left(\delta_i^b\right)\left[u_i\left(t,x^{i,b},q^{i,b},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^a(t,x,q,s) = 0$$

	Contract	Time to Maturity	$\log(Adt)$	κ
1	PUT	Long	-0.57	54.66
2	CALL	Long	0.19	70.63
3	PUT	Short	0.42	64.27
4	CALL	Short	0.92	97.64

Table 9: Estimated on sell orders parameters, A and κ. Results for in-the-money contracts.



(a) Monotonic Princple

(b) Tight Capital Constraint

Figure 37: Absorbing set of states under monotonic principle and tight capital constraints. The illustrated states can be reached in finite time. Then, the inventory can jump from the illustrated states to their illustrated neighbors. Probability of leaving the set of states is zero.

For $q_j = -\bar{Q}$ and $q_i = \bar{Q}$, $j \neq i$

$$\mathbb{U}_i^a(t,x,q,s) = \sup_{\delta_i^a} \lambda^a\left(\delta_i^a\right)\left[u_i\left(t,x^{i,a},q^{i,a},s\right) - u_i(t,x,q,s)\right]$$

$$\mathbb{U}_i^b(t,x,q,s) = 0$$

For $q_j = q_i$ and $|q_i| = \bar{Q}$, $j \neq i$

$$\mathbb{U}_i^a(t,x,q,s) = \mathbb{I}\{q_i = \bar{Q}\} \cdot (\text{as in } (19))$$

$$\mathbb{U}_i^b(t,x,q,s) = \mathbb{I}\{q_i = -\bar{Q}\} \cdot (\text{as in } (19))$$

The system of equations for $q \in \mathcal{Q}^2$

$$\partial_t u_i(t,x,q,s) + \frac{1}{2}\sigma^2 \partial_{ss}^2 u_i(t,x,q,s) + \sum_{z=a,b} \mathbb{U}_i^z(t,x,q,s) = 0 \tag{20}$$

with the final condition

$$u_i(T,x,q,s) = -\exp(-\gamma(x_i + q_i s)) \tag{21}$$

Estimated Coefficients:	Estimate	SE	tStat	pValue
$s_{0,0}$	-0.00038145	1.4281×10^{-8}	-26710	0
$b_{1,1} \times 10^2$	1.5146×10^{-4}	2.0678×10^{-9}	73249	0
$a_2 \times 10^2$	0.00020171	2.506×10^{-09}	80490	0
$c_2 \times 10^2$	-4.8481×10^{-5}	2.506×10^{-09}	-19346	0
$a_4 \times 10^4$	-1.0074×10^{-10}	1.0471×10^{-10}	-0.96212	0.33601
$b_{3,1} \times 10^4$	-1.6713×10^{-8}	9.1331×10^{-11}	-182.99	0
$c_4 \times 10^4$	1.0911×10^{-8}	1.0471×10^{-10}	104.2	0
$b_{1,3} \times 10^4$	-4.2522×10^{-8}	9.1331×10^{-11}	-465.58	0
$b_{2,2} \times 10^4$	4.7531×10^{-8}	8.9663×10^{-11}	530.11	0

Number of observations: 10201, Error degrees of freedom: 10192
Root Mean Squared Error: 5.23e-07
R-squared: 1, Adjusted R-Squared 1
F-statistic vs. constant model: 1.92e+10, p-value = 0

$$\theta_1(T,Q,k) = \theta_{0,0} + a_2 \cdot Q^2 + a_4 \cdot Q^4 + c_2 \cdot k^2 + c_4 \cdot k^4 + b_{1,1} \cdot Q \cdot k +$$
$$b_{3,1} \cdot Q^3 \cdot k + b_{1,3} \cdot Q \cdot k^3 + b_{2,2} \cdot Q^2 \cdot k^2 + \theta_1^r(T,Q,k),$$

Table 10: Excellent Fit of 4th order approximation into numerical solution for θ. $\gamma = 2$, $\kappa = 238$, $A = 1686.9$, $\bar{Q} = 300$, $\sigma = 3 \times 10^{-3}$ for numerical solution. I use points with $Q \in \{-50,-49,\dots,50\}$ and $k \in \{-50,-49,\dots,50\}$, $T = 6$. The approximation has the following form.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Spread (bp)	3,453,370	20.332	29.011	1	3.000	22.018	102.123
Log Spread (bp)	3,453,370	−7.072	1.309	−9.210	−8.112	−6.118	−4.584
$\log(1+Q_{k,t}^{mm,max-min})$	3,453,370	0.943	0.536	0.000	0.489	1.332	2.966
$\log(1+Q_{k,t}^{mm,max-min,2})$	3,453,370	1.365	0.673	0.000	0.803	1.884	3.320
$\lvert Q_{k,t}^{mm}\times 10^{-3}\rvert$	3,453,370	2.401	3.272	0.000	0.291	3.084	26.815
$\log(1+\lvert Q_{k,t}^{mm}\rvert)$	3,453,370	0.908	0.743	0.000	0.255	1.407	3.326
$\log(1+Q_{k,t}^{mm,max})$	3,448,123	6.051	1.439	0.000	5.198	7.112	9.297
$\log(1-Q_{k,t}^{mm,min})$	3,452,939	6.341	1.303	0.000	5.602	7.238	9.510
$\log(1+\Delta Q_{k,t}^{mm,max,2})$	3,453,370	5.116	1.695	0.000	4.043	6.382	9.156
$\log(1+\Delta Q_{k,t}^{mm,min,2})$	3,453,370	5.262	1.586	0.000	4.331	6.394	9.484

Table 11: Summary for variables in Aggregate Market-Level regressions

Dependent Variable:	Spread					
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
(Intercept)	46.08***	50.29***	75.57***	44.95***	34.37***	27.35***
	(1.632)	(1.625)	(2.008)	(1.732)	(1.765)	(1.706)
$\lvert Q_{k,t}^{mm} \times 10^{-3}\rvert$	14.47***	15.38***	8.384***	5.225***	5.867***	6.602***
	(0.830)	(0.861)	(0.838)	(0.577)	(0.521)	(0.550)
$\lvert Q_{k,t}^{mm} \times 10^{-3}\rvert^2$	-0.415***	-0.391***	-0.273***	-0.169***	-0.195***	-0.203***
	(0.027)	(0.027)	(0.027)	(0.020)	(0.018)	(0.018)
$\log(1+\lvert Q_{k,t}^{mm}\rvert)$	-59.18***	-53.46***	-26.80***	-16.85***	-13.63***	-17.26***
	(2.819)	(2.678)	(2.619)	(1.744)	(1.526)	(1.494)
$Q_{k,t}^{mm,max-min}$		0.661***	1.331	-5.743***	1.422	2.151
		(0.175)	(1.893)	(1.473)	(1.228)	(1.428)
$Q_{k,t}^{mm,max-min,2}$		-3.407***	9.608***	12.28***	7.167***	4.815***
		(0.161)	(1.179)	(0.901)	(0.706)	(0.854)
$\log(1+Q_{k,t}^{mm,max-min})$			0.355	11.77***	-4.493	-8.981**
			(5.768)	(4.161)	(3.638)	(3.778)
$(Q_{k,t}^{mm,max-min})^2$			-0.090	0.259***	-0.073	-0.108
			(0.082)	(0.071)	(0.055)	(0.066)
$\log(1+Q_{k,t}^{mm,max-min,2})$			-59.39***	-54.00***	-35.56***	-19.72***
			(4.818)	(3.641)	(3.052)	(3.009)
$(Q_{k,t}^{mm,max-min,2})^2$			-0.198***	-0.318***	-0.144***	-0.123***
			(0.034)	(0.030)	(0.021)	(0.028)
$\Delta_{k,t}$				-8.225***	-6.009***	-1.253
				(1.852)	(1.693)	(1.732)
$\mathcal{V}_{k,t}$				-3.780***	-4.234***	-4.246***
				(0.211)	(0.233)	(0.269)
$\Gamma_{k,t}$				-0.902***	-2.718***	-5.848***
				(0.241)	(0.383)	(0.589)
Θ				161.4***	204.5***	213.1***
				(6.277)	(7.114)	(7.678)
ρ				0.438***	0.510***	0.563***
				(0.023)	(0.020)	(0.020)
λ				0.042***	0.047***	0.013
				(0.008)	(0.009)	(0.014)
Greeks $\times$ $\lvert Q_{k,t}^{mm} \times 10^{-3}\rvert$					YES	YES
Greeks $\times$ Heterogeneity						YES
$Q_{k,t}^{mm,max-min,2} \times \mathcal{V}$						0.604***
						(0.060)
$Q_{k,t}^{mm,max-min,2} \times \Gamma$						0.497***
						(0.086)
$Q_{k,t}^{mm,max-min,2} \times \Theta$						-15.62***
						(2.701)
$Q_{k,t}^{mm,max-min} \times \rho$						0.026***
						(0.007)
$Q_{k,t}^{mm,max-min,2} \times \rho$						-0.084***
						(0.006)
Fit statistics						
Observations	2,662,047	2,662,047	2,662,047	2,654,835	2,654,835	2,654,835
R^2	0.26795	0.32751	0.41174	0.6568	0.70097	0.72664
Adjusted R^2	0.26795	0.32751	0.41173	0.6568	0.70096	0.72664

Clustered (order_date) standard-errors in parentheses. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

For Greeks × Heterogeneity only significant terms are reported below

Table 12: Simple OLS with no fixed effects.

Dependent Variable:	$Ask_{i,t} - \hat{P}_{k,t}$					
Years	All	All	All	2015	2016	2017
$\log(1+Q^{mm,max,2}_{k,t})$	-0.548***	-0.556***		-0.471*	-0.271***	-0.249***
	(0.089)	(0.089)		(0.247)	(0.065)	(0.051)
$\log(1+\Delta Q^{mm,max,2}_{k,t})$	0.057		0.088*	0.323**	0.031	0.090***
	(0.053)		(0.052)	(0.150)	(0.043)	(0.035)
$\log(1+Q^{mm,max-min,2}_{k,t})$			-0.854*			
			(0.504)			
$Q^{mm}_{k,t} \times 10^{-3}$	-0.654***	-0.647***	-0.793***	-5.257***	-0.554***	-0.508***
	(0.046)	(0.046)	(0.049)	(0.764)	(0.054)	(0.033)
$\Delta_{k,t}$	9.831***	9.820***	10.55***	12.87*	11.29***	0.530
	(2.481)	(2.485)	(2.434)	(6.699)	(2.603)	(1.621)
Contract×10 DAYS FE	yes	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes	yes
Option Greeks Controls	yes	yes	yes	yes	yes	yes
Observations	2,624,993	2,624,993	2,654,835	912,394	1,023,615	688,984
R^2	0.93255	0.93255	0.93186	0.91712	0.95611	0.95657
Within R^2	0.04757	0.04754	0.0471	0.05049	0.09889	0.33087

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

Table 13: Deviations of Best Ask Price from the fundamental value. The observations are contracts per minute.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Seconds Best on Ask	21,590,739	56.345	129.773	0	0	24.1	600
$\mathbb{I}\{BestBid_{i,k,t} - \overline{BestBid}_{B(k),day(t)} > 0\}$	21,590,739	0.191	0.393	0	0	0	1
$\log(1-Q_{i,k,t}+Q^{mm,max}_{k,t})$	21,590,739	5.564	2.124	0.000	4.804	7.004	9.838
$\log(1+Q_{i,k,t}-Q^{mm,min}_{k,t})$	21,590,739	5.663	2.109	0.000	5.004	7.038	9.838
$\lvert Q_{i,k,t}-Q_{i,k,t-1}\rvert$	21,590,739	3.697	16.681	0	0	0	2,097
$Q_{i,k,t}$	21,590,739	−28.164	600.738	−13,520	−72	32	10,911
$sign(Q_{i,k,t}) \times \log(1+\lvert Q_{i,k,t}\rvert)$	21,590,739	−0.442	4.325	−10	−4.3	3.5	9
$Q^{mm,mean}_{k,t}$	21,590,739	−0.026	0.207	−1.361	−0.084	0.026	1.577

Table 14: Summary for Account-Level Regression

Dependent Variable:	$Y_{i,k,t} := \mathbb{I}\{BestBid_{i,k,t} - \overline{BestBid}_{B(k),day(t)} > 0\}$					
Year	2015		2016		2017	
Type	CALL	PUT	CALL	PUT	CALL	PUT
$sign(Q_{i,k,t}) \times \log(1+\|Q_{i,k,t}\|)$	-0.0035***	-0.0069***	-0.0088***	-0.0036***	-0.0063***	-0.0083***
	(0.0011)	(0.0018)	(0.0024)	(0.0011)	(0.0020)	(0.0024)
$Q_{i,k,t}$	-0.0276	0.0012	0.0065	0.0013	-0.0050	0.0112
	(0.0208)	(0.0096)	(0.0097)	(0.0234)	(0.0143)	(0.0079)
$Q_{k,t}^{mm,mean}$	0.0581	0.0557**	0.0217	0.0341	0.0835**	0.0207
	(0.0542)	(0.0230)	(0.0306)	(0.0702)	(0.0380)	(0.0227)
$\log(1+Q_{i,k,t}-Q_{k,t}^{mm,min})$	-0.0010	-0.0036**	-0.0034	-0.0014	-0.0004	-0.0046**
	(0.0013)	(0.0018)	(0.0023)	(0.0013)	(0.0016)	(0.0019)
$Q_{k,t}^{mm,max} - Q_{i,k,t}$	0.0009	4.989×10^{-5}	0.0038	0.0008	-0.0077	0.0040
	(0.0096)	(0.0045)	(0.0040)	(0.0097)	(0.0053)	(0.0046)
$\log(1+Q_{i,k,t}-Q_{k,t}^{mm,min})$	-0.0011	0.0031*	0.0033	0.0012	0.0031**	0.0026
	(0.0016)	(0.0016)	(0.0025)	(0.0010)	(0.0015)	(0.0019)
$Q_{i,k,t} - Q_{k,t}^{mm,min}$	-9.515×10^{-6}	0.0006	0.0028	-0.0054	-0.0003	-0.0009
	(0.0078)	(0.0029)	(0.0039)	(0.0095)	(0.0050)	(0.0034)
Fixed Effects:						
Account × Contract × 10 DAYS	yes	yes	yes	yes	yes	yes
Account × Time × Basket	yes	yes	yes	yes	yes	yes
Clustered SE:						
Date	yes	yes	yes	yes	yes	yes
Account × Maturity	yes	yes	yes	yes	yes	yes
Observations	4,012,227	4,343,814	2,614,059	3,929,958	4,365,198	2,604,393
R^2	0.464	0.525	0.508	0.471	0.512	0.481

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses. Signif Codes: ***: 0.01, **: 0.05, *: 0.1

Table 15: Account-Level Regression. Activity on Ask.

Dependent Variable:	Probability to be active on Ask			
Model:	(1)	(2)	(3)	(4)
Variables				
$\log(1+Q_{k,t}^{mm,max})$	0.5769***	-0.7173	1.2949***	0.3118**
	(0.2059)	(0.5960)	(0.3617)	(0.1325)
$\log(1+Q_{k,t}^{mm,min})$	-0.0775	0.0479	-0.0518	-0.1469
	(0.0790)	(0.2150)	(0.0838)	(0.1230)
$\log(1+\Delta Q_{k,t}^{mm,max,2})$	-0.1733	0.7230*	-0.5131**	-0.0152
	(0.1084)	(0.4306)	(0.2047)	(0.0663)
$Q_{k,t}^{mm} \times 10^{-3}$	-0.0461***	8.806×10^{-5}	-0.0771***	-0.0378***
	(0.0133)	(0.0451)	(0.0213)	(0.0135)
$\Delta_{k,t}$	0.4517*	0.4983	0.2889	0.5171
	(0.2342)	(0.3031)	(0.2672)	(0.3734)
Fixed-Effects				
Contract $\times$ Account$\times$ T1:T8	Yes	Yes	Yes	Yes
Fit statistics				
Observations	1,600,032	729,244	571,423	298,391
Squared Correlation	0.134	0.115	0.156	0.128
Pseudo R^2	0.12357	0.10975	0.14169	0.11913
BIC	1,924,920.58	839,187.66	691,680.73	372,607.31

Table 16: Logit regression. Probability to be active on Ask for a maximal market maker. The standard errors are clustered on Date and Account×Group.

Dependent Variable:	$\hat{P}_{k,t} - Bid_{i,t}$					
Years	All	All	All	2015	2016	2017
$\log(1 - Q_{k,t}^{mm,min,2})$	-0.549***	-0.554***		-0.293	-0.366***	-0.076
	(0.164)	(0.162)		(0.480)	(0.133)	(0.074)
$\log(1 - \Delta Q_{k,t}^{mm,min,2})$	0.024		0.043	0.329	-0.082	-0.007
	(0.088)		(0.086)	(0.242)	(0.068)	(0.055)
$Q_{k,t}^{mm} \times 10^{-3}$	0.697***	0.694***	0.808***	5.884***	0.525***	0.541***
	(0.065)	(0.064)	(0.065)	(1.019)	(0.079)	(0.046)
$\Delta_{k,t}$	2.425	2.443	2.394	16.73*	-3.831	0.402
	(4.002)	(4.006)	(3.985)	(8.812)	(3.531)	(2.817)
$\mathcal{V}_{k,t}$	6.503***	6.501***	6.540***	13.40***	6.272***	6.687***
	(1.248)	(1.249)	(1.247)	(2.303)	(1.082)	(0.582)
$\Gamma_{k,t}$	-1.494**	-1.494**	-1.498**	-45.45***	-2.403**	0.411
	(0.699)	(0.699)	(0.690)	(8.072)	(1.174)	(0.472)
$\Theta_{k,t}$	-88.59	-88.64	-87.00	-189.3***	-126.1***	-21.35
	(57.44)	(57.45)	(57.62)	(72.88)	(47.39)	(52.21)
$\rho_{k,t}$	0.321***	0.321***	0.322***	0.339	-0.031	0.083
	(0.115)	(0.115)	(0.115)	(0.211)	(0.121)	(0.067)
$\lambda_{k,t}$	-0.024	-0.023	-0.025	-1.091**	0.003	0.037*
	(0.030)	(0.030)	(0.029)	(0.534)	(0.044)	(0.022)
$\log(1 + Q_{k,t}^{mm,max-min,2})$			0.837			
			(0.668)			
Contract×10 DAYS FE	yes	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes	yes
Date-Clustered SE	yes	yes	yes	yes	yes	yes
Observations	2,648,972	2,648,972	2,654,835	921,359	1,032,429	695,184
R^2	0.92422	0.92422	0.92417	0.91348	0.94176	0.95287
Within R^2	0.03739	0.03738	0.03697	0.05982	0.08146	0.31877

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses.
Signif Codes: ***: 0.01, **: 0.05, *: 0.1

Table 17: Bid Two way

Dependent Variable:			$Ask_{i,t} - \hat{P}_{k,t}$			
Years	All	All	All	2015	2016	2017
$\log(1+Q_{k,t}^{mm,max,2})$	-0.548***	-0.556***		-0.471	-0.271***	-0.249***
	(0.116)	(0.116)		(0.322)	(0.091)	(0.071)
$\log(1+\Delta Q_{k,t}^{mm,max,2})$	0.057		0.088	0.323	0.031	0.090*
	(0.070)		(0.070)	(0.200)	(0.057)	(0.048)
$\log(1+Q_{k,t}^{mm,max-min,2})$			-0.854			
			(0.667)			
$Q_{k,t}^{mm} \times 10^{-3}$	-0.654***	-0.647***	-0.793***	-5.257***	-0.554***	-0.508***
	(0.059)	(0.059)	(0.062)	(0.942)	(0.074)	(0.044)
$\Delta_{k,t}$	9.831***	9.820***	10.55***	12.87	11.29***	0.530
	(3.384)	(3.389)	(3.329)	(8.543)	(3.751)	(2.657)
$\mathcal{V}_{k,t}$	-7.247***	-7.248***	-7.241***	-10.89***	-6.620***	-7.076***
	(1.303)	(1.303)	(1.296)	(2.656)	(1.115)	(0.583)
$\Gamma_{k,t}$	1.577*	1.577*	1.595*	22.32***	2.169	-0.233
	(0.848)	(0.848)	(0.828)	(6.577)	(1.327)	(0.460)
$\Theta_{k,t}$	113.9**	113.8**	113.6**	173.3**	130.9**	24.71
	(54.58)	(54.59)	(55.01)	(72.59)	(54.13)	(50.25)
$\rho_{k,t}$	-0.123	-0.124	-0.128	-0.080	0.015	-0.074
	(0.113)	(0.113)	(0.112)	(0.222)	(0.125)	(0.065)
$\lambda_{k,t}$	0.010	0.009	0.008	0.993*	0.011	-0.031
	(0.032)	(0.032)	(0.031)	(0.540)	(0.050)	(0.024)
Contract×10 DAYS FE	yes	yes	yes	yes	yes	yes
Group×Time FE	yes	yes	yes	yes	yes	yes
Date-Clustered SE	yes	yes	yes	yes	yes	yes
Observations	2,624,993	2,624,993	2,654,835	912,394	1,023,615	688,984
R^2	0.93255	0.93255	0.93186	0.91712	0.95611	0.95657
Within R^2	0.04757	0.04754	0.0471	0.05049	0.09889	0.33087

*: Two-way clustered (Date and Contract×10 DAYS) standard-errors in parentheses.

Signif Codes: ***: 0.01, **: 0.05, *: 0.1