Journal of Financial Economics 95 (2010) 107–127

Contents lists available at ScienceDirect

Journal of Financial Economics

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

Market liquidity, asset prices, and welfare ☆

Jennifer Huang [a,*], Jiang Wang [b,c,d]

[a] McCombs School of Business, University of Texas, Austin, TX 78712, USA
[b] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 12412, USA
[c] National Bureau of Economic Research, USA
[d] China Academy of Financial Research, China

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form
13 May 2008
Accepted 14 August 2008
Available online 4 September 2009

JEL classification:
D53
G12
E58
D5

Keywords:
Liquidity
Asset prices
Welfare
Central bank policy

ABSTRACT

This paper represents an equilibrium model for the demand and supply of liquidity and its impact on asset prices and welfare. We show that, when constant market presence is costly, purely idiosyncratic shocks lead to endogenous demand of liquidity and large price deviations from fundamentals. Moreover, market forces fail to lead to efficient supply of liquidity, which calls for potential policy interventions. However, we demonstrate that different policy tools can yield different efficiency consequences. For example, lowering the cost of supplying liquidity on the spot (e.g., through direct injection of liquidity or relaxation of ex post margin constraints) can decrease welfare while forcing more liquidity supply (e.g., through coordination of market participants) can improve welfare.

1. Introduction

Liquidity is of critical importance to the stability and the efficiency of financial markets. The lack of it has often been blamed for exacerbating market crises such as the 1987 stock market crash, the 1998 near collapse of the hedge fund Long Term Capital Management (LTCM), and the current upheaval in the credit market.[1] Yet much less consensus exists about what market liquidity is, what determines it, and how it affects asset prices and welfare. Views become even more divergent when it comes to appropriate policies with respect to liquidity, such as

☆ Part of this work was done during Jiang Wang's visit at the Federal Reserve Bank of New York as a resident scholar. We thank Anat Admati, Tobias Adrian, Franklin Allen, Markus Brunnermeier, Douglas Diamond, Joe Hasbrouck, Nobu Kiyotaki, Arvind Krishnamurthy, Pete Kyle, Arzu Ozoguz, Lasse Pedersen, Paul Pfleiderer, Michel Robe, Jacob Sagi, Suresh Sundaresan, Sheridan Titman, Andrey Ukhov, Dimitri Vayanos, S. "Vish" Viswanathan, Pierre-Olivier Weill, and participants at the FDIC and JFSR seventh annual Bank Research Conference, the third annual IIROC-DeGroote Conference, the 2009 American Finance Association Meeting and seminars at Boston University, Emory University, Columbia University, Federal Reserve Bank of New York, Hong Kong University of Science and Technology, London School of Economics, Nanyang Technological University, National University of Singapore, Princeton University, Stanford University, Stockholm School of Economics, Texas A & M University, University of Chicago, University of Michigan, and the University of Texas at Austin for comments and suggestions. Support from Morgan Stanley (Equity Market Microstructure Grant, 2006) (Huang and Wang), the National Natural Science Foundation of China (Project no. 70440420490, Wang), and JP Morgan Academic Outreach Program (Wang) are gratefully acknowledged.

* Corresponding author.
E-mail address: jennifer.huang@mccombs.utexas.edu (J. Huang).

[1] The Report of the Presidential Task Force on Market Mechanisms, the Review of Financial Market Events in Autumn 1998 by the Committee on the Global Financial System, and the Global Financial Stability Report by the International Monetary Fund provide an overview of events in 1987, 1998, and 2007, respectively.

0304-405X/$ - see front matter
doi:10.1016/j.jfineco.2008.08.008

lowering barriers of entry in securities trading, setting margin and capital requirements for broker-dealers, coordinating market participants, and supplying liquidity during crises. The ongoing debate on the interventions by the central banks and the US Treasury to inject liquidity into the market during the current credit market crisis is an excellent case in point. The purpose of this paper is to present a simple theoretical framework to facilitate the discussions on these issues.

We start with the observation that the lack of full participation in a market is at the heart of illiquidity. Imagine a situation in which all potential buyers and sellers are constantly present in the market and can trade without constraints and frictions, i.e., fully participate. Then all agents face the full demand and supply at all times and security prices depend only on the fundamentals such as payoffs and preferences. To the extent that illiquidity reflects forces beyond these fundamentals, a market with full participation can be considered perfectly liquid. Thus, illiquidity arises only when frictions prevent full participation of all agents.

To capture this notion of illiquidity in a simple way, we assume that agents face participation costs that prevent them from constant, active, and unfettered participation in the market. We then develop an equilibrium model of both liquidity demand and supply in the presence of such costs. The endogenous demand for liquidity arises when participation costs prevent potential buyers and sellers with matching trading needs from coordinating their trades. The same costs also hinder the supply of liquidity. As a result, purely idiosyncratic shocks can cause infrequent but large deviations in prices from the fundamentals. Moreover, we show that, in general, market forces fail to achieve efficient supply of liquidity. However, different policy interventions can lead to divergent consequences. For example, direct injection of liquidity when it is in shortage can reduce welfare, while coordinated supply of liquidity by market participants can improve welfare. We also show that different costs of market presence give rise to distinctively different market structures and price and volume behavior, and the welfare consequences of the same policy interventions heavily depend on the structure of the market.

To model the need for and the provision of liquidity in a unified framework, we start with an economy in which agents face both idiosyncratic and aggregate risks. The desire to share the idiosyncratic risks gives rise to their need to trade in the asset market. By definition, idiosyncratic risks sum to zero across all agents. Thus, underlying trading needs are always perfectly matched among agents.

When market presence is costless, all agents stay in the market at all times. The market price adjusts to coordinate all buyers and sellers. Buy and sell orders, driven by idiosyncratic risks, are always in balance. In this case, asset prices are fully determined by the fundamentals, in particular, the level of aggregate risk, and are independent of agents' idiosyncratic trading needs.

When market presence is costly, however, not all agents are in the market at all times. We assume that agents can participate in the market in two ways: either incur an ex ante cost to be a market maker and then trade constantly, or pay a spot cost to trade after learning about their trading needs. Such a cost structure is motivated by the market structure we observe: A subset of agents (such as dealers, trading desks, and hedge funds) maintain a constant market presence and act as market makers, while most agents (such as the majority of individual and institutional investors, whom we refer to as traders) enter the market only when they need to trade. By the cost of market presence we intend to capture not only the costs of being in the market, but also any costs associated with raising needed capital or adjusting existing positions, in other words, any costs or hurdles that prevent the free flow of capital in the market.

As they trade only infrequently, traders are forced to bear certain idiosyncratic risk. This extra risk makes them less risk tolerant and less willing to hold their share of the aggregate risk. For traders receiving an additional idiosyncratic risk in the same direction as the aggregate risk, they are further away from their desired position and thus are more eager to trade. Consequently, more of them enter the market than those with the opposite idiosyncratic risk (which partially offsets their exposure to the aggregate risk). Thus, despite perfectly matching trading needs, traders fail to coordinate their trades, leading to order imbalances.

The endogenous order imbalances exhibit several distinctive properties. First, they are always in the same direction as the impact of the aggregate risk on asset demand, as traders with higher than average risk are more likely to enter the market. Second, order imbalances are always of significant magnitudes when they occur. This is because, for small idiosyncratic shocks, gains from trading are small and all traders choose to stay out of the market. It is only with sufficiently large idiosyncratic shocks that gains from trading exceed participation costs for some traders, leading to the mismatch in their trades. The resulting order imbalance is also large. Third, the magnitude of possible order imbalances depends on the level of the aggregate risk, which affects the asymmetry in trading gains between different traders.

By endogenizing the order imbalance, we are able to characterize the impact of liquidity on asset prices. In particular, purely idiosyncratic shocks can generate aggregate liquidity needs and cause price to deviate from its fundamental value. Moreover, the impact of liquidity on price is in the same direction as that of the aggregate risk and is of significant magnitude. Consequently, it leads to high price volatility and fat tails.

Under exogenous liquidity demand, Grossman and Miller (1988) find that higher costs of market making lead to lower levels of liquidity in the market and more volatile prices. We show that, when liquidity demand is endogenously determined, it becomes interdependent with liquidity supply and prices are not necessarily more volatile in less liquid markets.

In particular, we obtain two different market structures. Only when the cost of market making is below a threshold do we have the usual market structure in which liquidity is supplied by market makers. When the cost of market making exceeds this threshold, a different market structure emerges: No market makers are in the market, and all liquidity is supplied by traders themselves on the spot.

Under such a market structure, the liquidity supply is extremely low but so is the observed need for liquidity. Traders choose to stay out of the market most of the time. They enter only when shocks are large and participation is sufficiently symmetric. In this case, prices become less volatile. In such a market, conventional measures of price impact fails to be informative about liquidity. Instead, the lack of trading volume properly reveals the low level of liquidity. Thus, our results also provide a theoretical justification for incorporating trading volume into measures of market liquidity.[2]

In our model, trading and liquidity provision generate externalities. A trader's participation in the market also benefits his potential counterparties, and a market maker's supply of liquidity helps all potential traders. We show that, in general, market mechanism fails to properly internalize these externalities and thus leads to inefficient supply of liquidity in the market. Such an inefficiency leaves room for policy interventions. However, given the endogenous nature of both liquidity demand and supply, we show that different policy choices can lead to surprising consequences. On the one hand, the overall welfare of the economy can be improved by forcing all agents to pay the participation cost. In this case, the extra liquidity generated by broad participation yields benefits for all agents, which can outweigh the extra costs they pay. On the other hand, in a market with insufficient liquidity supply, decreasing participation costs (in particular, the cost to enter the market on the spot) can reduce welfare. This is because lowering the cost to enter the market on the spot reduces the incentive to be in the market a priori, i.e., to become a market maker. The level of liquidity in the market then decreases, which hurts everyone, including those who now pay lower costs.

During market crises, such as the 1998 LTCM debacle and the current credit market upheaval, central banks have resorted to relaxing their lending conditions, e.g., by cutting the rates charged and broadening the collateral accepted, to increase liquidity into the market. This can be interpreted as cutting the cost of spot market participation in our model. Government agencies, such as the New York Federal Reserve Bank in the case of LTCM crisis and the US Treasury in the case of current credit market crisis, have also coordinated market participants to collectively supply pools of liquidity. Such an action is related to the forced spot participation in our analysis. Similarly, regulations such as designated market makers and high capital requirements can be interpreted as increasing ex ante participation in our model. By relying on an equilibrium setting in which both the demand and the supply of liquidity are endogenously determined, we are able to identify the sources of market inefficiencies and examine the overall welfare implications of various policy tools under different circumstances.

The paper proceeds as follows. Section 2 describes the basic model. Section 3 solves for the intertemporal equilibrium of the economy. In Section 4, we examine how the need for liquidity affects asset prices and trading volume. Section 5 describes the endogenous determination of liquidity provision in the market and how it influences prices and volume. In Section 6, we consider the welfare implications of liquidity need and provision. Section 7 further explores the policy implications of our analysis. Section 8 gives a more detailed discussion on the related literature, and Section 9 concludes. The Appendix contains all the proofs.

2. The model

We construct a simple model that captures two important elements in analyzing liquidity, the need to trade and the cost to trade. We are parsimonious in the description of the model and return at the end of this section to provide more discussion of the model, especially motivations for its different components.

2.1. Securities market

The economy has three dates, 0, 1 and 2. A competitive securities market consists of two securities, a riskless bond, which is also used as the numeraire, and a risky stock. The bond yields a sure payoff of 1 at date 2. The stock yields a risky dividend D at date 2, which has a mean of zero and a volatility of σ.

2.2. Agents

A continuum of agents of measure 1 exists in the economy, with identical preferences and zero initial holdings of the traded securities. Each agent i receives a nontraded payoff N^i at date 2, which is correlated with the payoff of the stock. Depending on their nontraded payoff, agents fall into two equally populated groups, denoted by a and b. All agents in group i, $i=a,b$, receive the same nontraded payoff

$$N^i = Y^i u, \tag{1}$$

where Y^a and Y^b have independent and identical distributions and are both independent of u. For simplicity, we use i to refer to an individual agent as well as agents in group i, where $i=a,b$.

Summing over all agents' nontraded payoff yields the aggregate nontraded payoff

$$\int_i N^i = \frac{1}{2}(Y^a+Y^b)u. \tag{2}$$

Let $Y \equiv \frac{1}{2}(Y^a+Y^b)$ and $Z \equiv \frac{1}{2}(Y^a-Y^b)$. We can rewrite each agent's nontraded payoff as

$$N^i = (Y+\lambda^i Z)u, \tag{3}$$

where $\lambda^a=1$, $\lambda^b=-1$ and Y and Z are uncorrelated.[3] Thus, Y gives the aggregate exposure to the nontraded risk and $\lambda^i Z$ gives the idiosyncratic exposure. By definition, agents'

[2] For empirical evidence on the role of volume in measuring liquidity, see, for example, Campbell, Grossman, and Wang (1993), Brennan, Chordia, and Subrahmanyam (1998), and Amihud (2002).

[3] The covariance between Y and Z is $\mathrm{Cov}[Y,Z]=\mathrm{Cov}[\frac{1}{2}(Y^a+Y^b), \frac{1}{2}(Y^a-Y^b)]=\frac{1}{4}(\mathrm{Var}[Y^a]-\mathrm{Var}[Y^b])=0$.

idiosyncratic exposures sum to zero. For simplicity, we assume that Y, Z, and u are jointly normal with zero mean and volatility of σ_Y, σ_Z, and σ_u, respectively. In addition, we let $u = D$.[4]

Agents first receive information about their nontraded payoff at date 1. In particular, they observe Y, λ^i, and a signal S about Z:

$$S = Z + \varepsilon, \tag{4}$$

where ε is a noise in the signal, normally distributed with a volatility of $\sigma_\varepsilon > 0$.

In the absence of idiosyncratic risks (i.e., when $Z = 0$), all agents are identical and they have no trading needs. In the presence of idiosyncratic risks (i.e., when $Z \neq 0$), however, agents want to share these risks. In particular, given the correlation between the nontraded payoff and the stock payoff, they want to adjust their stock positions to hedge their nontraded risk. Thus, agents' idiosyncratic risks give rise to their trading needs.

An agent's preference is described by an expected utility function over his terminal wealth. For tractability, we assume that he exhibits constant absolute risk aversion. In particular, agent i has the following utility function:

$$-e^{-\alpha W^i}, \tag{5}$$

where W^i denotes his terminal wealth and α is the absolute risk aversion. We further require

$$\alpha^2 \sigma^2 (\sigma_Y^2 + \sigma_Z^2) < 1 \tag{6}$$

to guarantee a bounded expected utility in the presence of nontraded payoffs.

2.3. Participation costs

At date 0, all agents are identical and thus need not trade. For simplicity, we allow them to trade in the market at no cost. Agents' trading needs arise at date 1 after they observe their risk exposures (Y, λ^i, and S). To trade at date 1, an agent has to pay a cost. He can either pay a cost c_m at date 0 before learning about his own trading needs, which allows him to trade at any time, or wait until after observing his shocks and pay a cost c to trade in the market if he chooses.

Those who pay the ex ante cost are in the market at all times, ready to trade with others. We call them market makers, denoted by m. Those who only pay the spot cost when they trade are called traders, denoted by n. Traders demand liquidity when they cannot meet their own trading needs and market makers provide it in these circumstances. In actual markets, institutional or individual investors usually behave as traders in our model while dealers and hedge funds serve as market makers. By explicitly modeling the choice of becoming a trader or a market maker, we fully endogenize the need for liquidity as well as its supply. This allows us to examine the pricing and welfare implications of liquidity in a full equilibrium setting.

2.4. Time line

For the economy defined above, we now detail the sequence of events, agents' actions, and the corresponding equilibrium. At date 0, agents first trade in the market to establish their initial position θ_0^i and the equilibrium stock price P_0. Given that they are identical, the equilibrium is reached at $\theta_0^i = 0$.

Each agent then decides if he wants to pay the cost c_m to become a market maker. Let η_m^i denote his choice, with $\eta_m^i = 1$ for being a market maker and $\eta_m^i = 0$ for not. A participation equilibrium determines the fraction of agents who become market makers, which we denote by μ.

At date 1, agents learn about their nontraded risks and decide whether to pay a cost c to enter the market to trade. Let η^i denote the entry choice of agent i, with $\eta^i = 1$ for entry and $\eta^i = 0$ for no entry. Because market makers are already in the market, they need not pay c. That is, $\eta^i = 0$ for all market makers. For traders, this entry decision depends on their draw of λ^i, the signal S on the magnitude of the idiosyncratic risk, and the aggregate risk Y. The participation equilibrium of traders at date 1 determines the fraction of each group that chooses to enter the market, which we denote by $\omega \equiv \{\omega^a, \omega^b\}$.

After the traders' participation decisions, all market makers and participating traders trade in the market to choose their stock holdings. Let $\theta_1^i(\eta_m^i, \eta^i)$ denote the stock shares held by a group-i agent (whose participation decisions are η_m^i and η^i, respectively) after trading at date 1. Hence, $\theta_1^i(1,0)$ denotes the holding of a group-i market maker and $\theta_1^i(0,1)$ denotes the holding of a participating group-i trader. For the nonparticipating traders, $\eta_m^i = \eta^i = 0$ and $\theta_1^i(0,0) = \theta_0^i = 0$. The trading among the market makers and the participating traders determines the market equilibrium at date 1 and the stock price P_1. For simplicity, we assume that agents observe Z when they trade after the participation decisions at date 1. Thus, there is no more need to trade afterward.[5]

Given his participation decisions η_m^i and η^i and his stock holding $\theta_1^i(\eta_m^i, \eta^i)$ at date 1, agent i's terminal wealth W^i is given by

$$W^i = -\eta_m^i c_m - \eta^i c + \theta_1^i (D - P_1) + N^i, \tag{7}$$

where N^i is his nontraded payoff given in Eq. (3).

Summarizing the description above, Fig. 1 illustrates the time line of the economy.

2.5. Discussions of the model

In this subsection, we provide additional discussions and motivations about several important features of the model. The two key ingredients of the model are the need to trade and the cost to trade in the market. Little justification is necessary for modeling agents' trading

[4] We only need the correlation between u and D to be nonzero. The qualitative nature of our results are independent of the sign and the magnitude of the correlation. To fix ideas, we set it to 1.

[5] Alternatively, we can assume that Z is realized at date 2. Our results remain qualitatively the same, but the solution becomes more tedious.

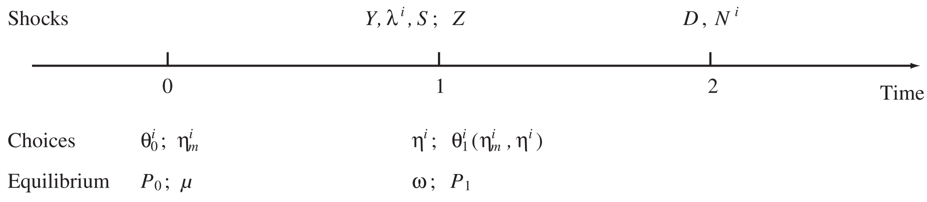

Fig. 1. The time line of the economy.

needs, given the large trading volume observed in the market. To model trading needs, we must allow for certain forms of heterogeneity among agents. For example, trading can arise from heterogeneity in endowments (e.g., Diamond and Verrecchia, 1981; Wang, 1994), preferences (e.g., Dumas, 1992; Wang, 1996), or beliefs (e.g., Harris and Raviv, 1993; Detemple and Murthy, 1994). Our modeling choice of heterogeneity in agents' endowments in the form of nontraded payoffs is mainly for tractability. Agents thus trade for risk-sharing motives. Our main results are not sensitive to this particular choice.

Another key component of our model is the cost to participate in the market. This cost is intended to capture in a reduced-form manner any frictions that prevent agents from constant, active, and unfettered participation in the market. The lack of such a full participation is at the heart of illiquidity and distinguishes it from other fundamentals.

An extensive literature exists on the nature of these costs and its significance. For example, Merton (1987) points out that most agents are prevented from active market presence due to costs of gathering and processing information, devising trading strategies and support systems, and raising capital.[6] Shleifer and Vishny (1997) argue that, even for agents who are actively participating in the market, capital constraints often limit their abilities to take on large positions.[7]

For instance, typical market makers such as trading desks and hedge funds all have limited capital, which is costly and time-consuming to raise but hard to maintain in needy times. Most institutional investors face external and internal constraints such as regulations and risk controls, which limit their flexibility in choosing asset allocations and risk budgets. Thus, the participation cost in our model should be interpreted broadly as costs or hurdles that hinder the free flow of capital in the market place, in addition to the direct costs of physical presence and information processing.

Mounting empirical evidence suggests that these costs not only exist but also can be substantial. For example, Coval and Stafford (2007) find that selling by financially distressed mutual funds leads to significantly depressed prices for the stocks sold, which persist over multiple quarters before recovery. This effect occurs despite the fact that these stocks are widely held by other mutual funds that are not suffering outflows. Mitchell, Pedersen, and Pulvino (2007) examine several markets such as convertible bonds and mergers and acquisitions, in which hedge funds actively pursue pricing anomalies. They show that, when hedge funds in a particular market face large redemptions, prices deviate significantly from the fundamentals. Capital returns only slowly, leaving the price deviations persist for long periods of time. The persistence of large price deviations caused by liquidity events implies that significant costs exist in preventing instantaneous capital flow or participation.

In our model, we further recognize that, in an intertemporal setting, the magnitude of participation costs depends on the time scale over which agents establish market presence. For costs of the same nature, e.g., costs of gathering and processing information or raising capital, they can be substantially higher when less time is allowed. If we interpret c and c_m in the model as these same costs of participation, paid on the spot and ex ante, respectively, it is reasonable to assume that c is higher than c_m.

If, however, the nature of the ex ante and spot costs are different, c_m can be higher than c. For example, if c_m is the cost to set up operations to become a market maker while c is merely the cost of occasional trading, then we would expect c_m to be much higher than c. In this case, however, the market maker expects to trade many times down the road. He has to weigh the total cost c_m with the total benefit from all his future trades. For a trader, he weighs the cost c for each of his trade. If a market maker trades frequently, as he should, on a per trade basis, his cost should be lower.[8] Because our model has only one trading cycle, the costs c_m and c should be interpreted as costs for each trade. Thus, we expect $c_m < c$.

Our use of the term "market makers" is broader than its most common use. In addition to designated dealers in a market, we include agents who maintain an active presence in the market and provide liquidity as market makers such as trading desks and hedge funds.

More capital in a market tends to reduce the risk aversion of marginal investors (see, e.g., Grossman and Vila, 1992) and thus improves the supply of liquidity. In our setting, all agents have constant risk aversion and the amount of capital each of them has does not matter. But the participation of more agents brings in more capital and lowers the effective risk aversion of market makers as a group (which is their average risk aversion divided by the total number of them). In this sense, the number of market

[6] See also Brennan (1975), Hirshleifer (1988), Leland and Rubinstein (1988), and Chatterjee and Corbae (1992).

[7] See also Gromb and Vayanos (2002) and Brunnermeier and Pedersen (2009), among others, for the impact of capital constraints on liquidity supply and asset prices.

[8] Otherwise, potential market makers are strictly better off trading only on the spot and no one would choose to become a market maker.

makers in our model is effectively playing the same role as the amount of total capital in the market.

In addition, the assumption that Z is not fully observed at the time of participation decision is important in our model. It implies that agents do not anticipate to trade away all their future idiosyncratic risks if they participate. As shown in Lo, Mamaysky, Wang (2004), in a fully intertemporal setting, agents always expect to bear certain idiosyncratic risks because they trade infrequently. By assuming partial information on Z when deciding on participation, we capture this dynamic aspect in a simple setting. Otherwise, the model becomes effectively static. As long as Z is realized after the participation decision, the exact timing of its revelation is unimportant.

3. Equilibrium

We solve for the equilibrium in three steps. First, taking as a given agents' initial stock holdings θ_0^i, the fraction μ of market makers, and the participation decision of traders, we solve for the stock market equilibrium at date 1. Next, we solve for individual traders' participation decisions and the participation equilibrium, given the market equilibrium at 1. Finally, we solve for individual agents' decision to become market makers and their equilibrium population μ as well as the stock market equilibrium at date 0.

3.1. Equilibrium with costless participation

We start with the special case of no participation costs, i.e., $c_m = c = 0$. This case serves as a benchmark when we examine the impact of participation costs on liquidity and market behavior.

In this case, agents are indifferent between being market makers or traders, i.e., any $\mu \in [0,1]$ is an equilibrium. They are in the market at all times, i.e., $\omega^a = \omega^b = 1$. The equilibrium price and agents' equilibrium stock holdings are

$$P_0 = 0, \quad \theta_0^i = 0,$$

$$P_1 = -\alpha\sigma^2 Y, \quad \theta_1^i = -\lambda^i Z, \tag{8}$$

where $i = a, b$.

The initial price of the stock is $P_0 = 0$ because its expected dividend is normalized to zero and it is in zero net supply. Because the nontraded payoff is perfectly correlated with the stock payoff, the aggregate (per capita) risk exposure Y is equivalent to an aggregate supply shock for the stock and thus affects its price at date 1. The aggregate risk, however, does not affect agents' share holdings in equilibrium.

Their idiosyncratic risk exposure $\lambda^i Z$, meanwhile, affects individual holdings. Agents' stock holdings are given by $-\lambda^i Z$, which reflects their hedging demand to offset their idiosyncratic risk exposure. Because agents' underlying trading needs are perfectly matched ($\lambda^a = -\lambda^b$), so are their trades when they are all in the market. In this case, there is no need for liquidity. The market is perfectly liquid in the sense that trading has no price impact. Stock prices do not depend on the idiosyncratic shock Z.

3.2. Stock market equilibrium at date 1

We now present equilibrium with participation costs, starting with the market equilibrium at date 1. Assume a population μ of agents becomes market makers. The remaining population $1-\mu$ is evenly split between group-a and -b traders, with $\omega = \{\omega^a, \omega^b\}$ fraction of each trader group participating. Together with Y and Z, μ and ω define the state of the economy at date 1. We introduce

$$\delta \equiv \begin{cases} \frac{1}{2}(1-\mu)(\omega^a - \omega^b)/[\mu + \frac{1}{2}(1-\mu)(\omega^a + \omega^b)] \\ \quad \text{for } \mu > 0 \text{ or } \omega > 0, \\ \lambda^i \quad \text{for } \mu = \omega = 0, \end{cases} \tag{9}$$

as a measure of asymmetry in participation between the two groups of traders. When $\mu > 0$ or $\omega > 0$, the numerator gives the net population imbalance between the two trader groups and the denominator is the total population in the market. When $\mu = \omega = 0$, no agent is in the market other than the agent under consideration (in group i), and δ is defined as the limiting ratio when $\mu = 0$, $\omega^{-i} = 0$, and $\omega^i \to 0$. Because ω^a and ω^b are bounded in $[0,1]$, we have $\delta \in [-\overline{\delta}, \overline{\delta}]$, where

$$\overline{\delta} = \frac{1-\mu}{1+\mu} \tag{10}$$

gives the maximum amount of participation asymmetry between the two trader groups.

Taking μ and δ as given, we solve the market equilibrium at date 1.

Proposition 1. *The equilibrium stock price at date 1 is*

$$P_1 = -\alpha\sigma^2 Y - \alpha\sigma^2 \delta Z, \tag{11}$$

and the equilibrium stock holdings of market makers and participating traders are

$$\theta_1^i = \delta Z - \lambda^i Z, \tag{12}$$

where $i = a, b$.[9]

Contrasting to the benchmark case when participation is costless and symmetric between the two trader groups, both individual holding and the equilibrium price now have an extra term related to δZ. When $\delta \neq 0$, the participation of the two groups of traders is asymmetric. The buy and sell orders are no long perfectly matched. The order imbalance leads to an additional net risk exposure, which is δZ on a per capita basis. All participating agents equally share this risk and increase their holding by δZ. The idiosyncratic shock Z now affects the equilibrium price as Eq. (11) shows. Thus, even though traders face offsetting shocks, asymmetry in their participation can give rise to a mismatch in their trades and cause the price to change in response to these shocks.

So far, we have taken traders' participation rate ω and the resulting δ as given. In the next subsection, we show that when individual participation decisions are made

[9] When $\mu = \omega = 0$, no agent is in the market and the market equilibrium allows a range of prices. Choosing the specific price in the proposition does not affect the overall equilibrium.

endogenously, asymmetric participation occurs as an equilibrium outcome.

3.3. Traders' optimal participation decisions at date 1

Given the stock market equilibrium at date 1, we now solve the participation equilibrium of traders in two steps. First, taking as a given the participation decision of other traders, we derive the optimal participation policy of an individual trader. Next, we find the competitive equilibrium for traders' participation decisions.

At the time of their participation decisions, all traders have a stock holding of $\theta_0^i = 0$ $(i = a, b)$. Moreover, they observe Y, λ^i, and a signal S on Z. We denote by X the expectation of Z conditional on signal S, σ_x^2 the variance of X, and σ_z^2 the variance of Z conditional on S. Then,

$$X \equiv \mathrm{E}[Z|S] = \beta S, \quad \sigma_x^2 \equiv \mathrm{Var}[X] = \beta\sigma_Z^2,$$
$$\sigma_z^2 \equiv \mathrm{Var}[Z|S] = (1-\beta)\sigma_Z^2, \tag{13}$$

where $\beta \equiv \sigma_Z^2/(\sigma_Z^2+\sigma_\varepsilon^2)$. Under normality, X is a sufficient statistic for signal S. Thus, we use X to denote agents' information about the magnitude of the idiosyncratic risk.

For trader i, let J_P^i and J_{NP}^i denote his indirect utility function given his decision to participate (P) or not to participate (NP), respectively. Under constant absolute risk aversion, trader i's indirect utility function takes the form of $J = -I(\cdot)e^{-\alpha W}$, where W is his wealth and $I(\cdot)$ depends on the initial stock holding θ_0^i, market condition δ, and nontraded risk exposure Y, X, and λ^i (see Appendix). The net gain from participation for group-i traders can be defined as the certainty equivalence gain in wealth,

$$g(\theta_0^i; Y, X, \lambda^i; \delta) \equiv -\frac{1}{\alpha}\ln\frac{J_P^i}{J_{NP}^i}, \quad i = a, b. \tag{14}$$

The minus sign on the right-hand side adjusts for the fact that J_P^i and J_{NP}^i are negative. Proposition 2 describes individual traders' optimal participation policy.

Proposition 2. *The net gain from participation for trader i is*

$$g(\theta_0^i; Y, X, \lambda^i; \delta) \equiv g_1(\theta_0^i; Y, X, \lambda^i; \delta) + g_2(\lambda^i; \delta) - c,$$
$$i = a, b, \tag{15}$$

where

$$g_1(\cdot) \equiv \frac{\alpha\sigma^2(1-k\lambda^i\delta)^2}{2(1-k)[1-k+k(1-\lambda^i\delta)^2]}\,(\theta_0^i - \hat{\theta}^i)^2,$$
$$g_2(\cdot) \equiv \frac{1}{2\alpha}\ln\left[1+\frac{(1-\lambda^i\delta)^2 k}{(1-k)}\right], \tag{16}$$

and

$$\hat{\theta}^i \equiv -\frac{1-\lambda^i\delta}{1-k\lambda^i\delta}(kY+\lambda^i X), \quad k \equiv \alpha^2\sigma^2\sigma_z^2. \tag{17}$$

He participates if and only if $g(\cdot) > 0$.[10]

When $\mu = \omega = 0$, $g(\cdot) = -c < 0$ for both traders. Without any agent in the market at date 1, a trader has no one to trade with if he chooses to participate and he ends up with the same stock position except that he is now c dollars poorer. Hence, he never participates.

When $\mu > 0$ or $\omega > 0$, a trader can benefit from trading. His net gain from participation consists of three terms, $g_1(\cdot)$, $g_2(\cdot)$, and $-c$. The first term, $g_1(\cdot)$, represents the expected trading gain in response to his current shocks. We can interpret $\hat{\theta}^i$ as trader i's desired holding after the shocks. Unless $\theta_0^i = \hat{\theta}^i$, he expects a positive net gain from trading. The second term, $g_2(\cdot)$, captures the expected trading gain from offsetting future shocks to nontraded risks. This term depends only on the market condition δ and k, which is proportional to future trading needs as captured by σ_z^2. The last term, $-c$, reflects the cost of participation.

For future convenience, we define

$$g^i(\delta; Y, X) \equiv g(0; Y, X, \lambda^i; \delta), \quad i = a, b, \tag{18}$$

by substituting in the initial holding $\theta_0^i = 0$. In general, trading gains are asymmetric between the two trader groups. This is true even when participation is symmetric (i.e., when $\delta = 0$), because

$$g^i(0; Y, X) = \frac{\alpha\sigma^2}{2(1-k)}(\hat{\theta}^i)^2 - \frac{1}{2\alpha}\ln(1-k) - c, \tag{19}$$

where $\hat{\theta}^i = -(kY + \lambda^i X)$. Clearly, $g^a \neq g^b$ (except for $Y = 0$ or $X = 0$), and $g^a \geq g^b$ whenever Y and X have the same sign.

To understand this asymmetry, we first consider the special case when $X = 0$. With zero current idiosyncratic shocks, all agents (market makers and traders) receive equal share of the aggregate risk. However, given the future idiosyncratic shocks, as represented by Z, traders still desire to trade. In particular, the prospect of bearing these risks makes them effectively more risk averse. Consequently, they prefer to bear less of the aggregate risk. Their desired position becomes $\hat{\theta}^i = -kY$, which is different from their initial position $\theta_0^i = 0$. Hence, traders would like to sell the stock to unload k fraction of their exposure to the aggregate risk. This desire is independent of the realization of the idiosyncratic shock X.

When $X \neq 0$, the desire to partially unload the aggregate risk is combined with the desire to unload their idiosyncratic risks. For those traders whose idiosyncratic shock $\lambda^i X$ is in the same direction as the aggregate shock Y, their initial position ($\theta_0^i = 0$) is further away from their desired position $\hat{\theta}^i = -(kY + \lambda^i X)$. For example, when Y and X have the same sign, $\hat{\theta}^a = -(kY + X)$ is further away from 0 than $\hat{\theta}^b = -(kY - X)$. The gain from trading, which is proportional to $(\hat{\theta}^i)^2$, is then larger for group-a traders than for group-b traders.[11] We thus have the following result: When participation in the market is costly, the gains from trading

[10] Parameter restriction in Eq. (6) guarantees that $k < 1$.

[11] In general, the gain from trading also depends on the initial position θ_0^i. In a setting such as ours, θ_0^i is always different from $\hat{\theta}^i$ because the latter depends on the current shocks while the former does not. In a stationary setting similar to ours, Lo, Mamaysky, and Wang (2004) show that the gain from trading is asymmetric around the optimal holding due to the fact that traders trade infrequently.

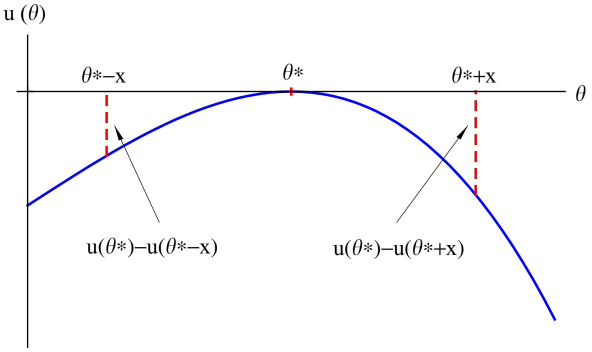

Fig. 2. Asymmetry in utility gain from costly trading.

are in general asymmetric between traders with perfectly matching trading needs. In addition, the gains are larger for those traders with idiosyncratic shocks in the same direction as the aggregate shock.

We shall emphasize that the asymmetry in trading gains is a general phenomenon. To see this, let $u(\theta)$ denote the utility from holding θ and θ^* be the optimal holding. Then, $u'(\theta^*)=0$. For a small deviation $x=\theta-\theta^*$ from the optimum, we can drop the higher order terms from the Taylor expansion and obtain the gain from trading as $u(\theta^*)-u(\theta^*+x)\simeq -u''(\theta^*)x^2/2$, which is the same for an opposite deviation $-x$. When trading is costless, traders constantly maintain the optimal position, and the gains from trading for traders with small offsetting shocks are always the same. This symmetry breaks down when trading is costly. Facing a cost, traders no longer trade constantly. They trade only when the deviation from the optimal is sufficiently large. As Fig. 2 illustrates, the trading gain is no longer symmetric for finite deviations from the optimum because $u(\theta^*)-u(\theta^*+x)\neq u(\theta^*)-u(\theta^*-x)$ for a finite x. Hence, as long as trading is infrequent, the gains from trading become different between traders with perfectly offsetting trading needs.

The result that trading gains are larger for traders receiving more (than average) risks is also fairly robust. It only requires traders to become effectively more risk-averse when faced with unhedged idiosyncratic risks. As Kimball (1993) shows, all preferences with standard risk aversion exhibit such a behavior.[12]

3.4. Participation equilibrium for traders at date 1

Given the asymmetric participation decisions of the two groups of traders, we show in Proposition 3 that the participation equilibrium is also asymmetric.

Proposition 3. *A participation equilibrium for traders exists. When Y and X have the same sign, the equilibrium (ω^a,ω^b) is given by*

(A) *For $g^b(0;Y,X)\le g^a(0;Y,X)\le 0$, $\omega^a=\omega^b=0$;*
(B) *For $g^a(0;Y,X)\ge g^b(0;Y,X)\ge 0$, $\omega^a=\omega^b=1$; or*
(C) *Otherwise, either $\omega^a=1$ and $\omega^b\in[0,1)$ or $\omega^a\in(0,1)$ and $\omega^b=0$, and $\omega^a>\omega^b$.*

When Y and X have opposite signs, the equilibrium (ω^a,ω^b) is given by exchanging subscripts a and b in cases (A)–(C). Moreover, the above equilibrium is unique when $\mu>0$. When $\mu=0$, there also exists an autarky equilibrium with $\omega^a=\omega^b=0$ for all Y and X, which is Pareto dominated by the above equilibrium.

We consider only the nondominated equilibrium when $\mu=0$ in future discussions. When X and Y have the same sign, we know from Eq. (19) that group-a traders enjoy larger gains from trading when the participation is symmetric ($\delta=0$). As a result, in equilibrium more group-a traders are entering the market than group-b traders, causing an order imbalance.

Fig. 3, Panel A illustrates the states, i.e., realizations of X and Y, for which there is no participation of traders (region A), full participation (region B), and asymmetric participation (region C). For any given level of the aggregate risk, Y, asymmetric participation occurs for a range of X with finite values. Fig. 3, Panel B plots δ, the degree of asymmetry in participation between the two groups of traders, for different values of Y and X. For any given Y, the range of X over which asymmetry occurs ($\delta\neq 0$) in Panel B corresponds exactly to the intersection of a horizontal line at this Y level and region C in Panel A.

3.5. Participation equilibrium for market makers at date 0

Up until now, the population of market makers μ is taken as given. We now study how it is determined in equilibrium. Our analysis shows that costly participation gives rise to mismatch in trades between traders with perfectly matching trading needs. The resulting order imbalance (or the need for liquidity) thus calls for market makers to supply liquidity. The market makers have to pay the participation cost ex ante. In return, they benefit from supplying liquidity by absorbing order imbalances in the market at favorable prices. When the benefit dominates, agents want to become market makers. But the benefit diminishes as the population of market makers increases and competition intensifies. An equilibrium population of market makers (or an equilibrium level of liquidity supply) is reached when the cost and benefit balance out.

To solve for the equilibrium level of liquidity supply, we first compute the value function of individual agents who choose to become market makers (J^m) or traders (J^n), for a given population of market makers. In particular, we have

$$J^m(\mu,c_m)\equiv \mathrm{E}[J^i_P|c^i=c_m],\quad J^n(\mu,c)\equiv \mathrm{E}[\max\{J^i_P,J^i_{NP}\}|c^i=c], \tag{20}$$

where the expectation is over the realizations of Y, X, and λ^i, and the indirect utility functions J^i_P and J^i_{NP} are defined in Section 3.3.

[12] Standard risk aversion is defined as the class of utility functions that exhibit both decreasing absolute risk aversion (DARA) and decreasing absolute prudence. In our setting, the underlying utility function, with constant absolute risk aversion, does not exhibit standard risk aversion, but the indirect utility function, i.e., the value function, does.

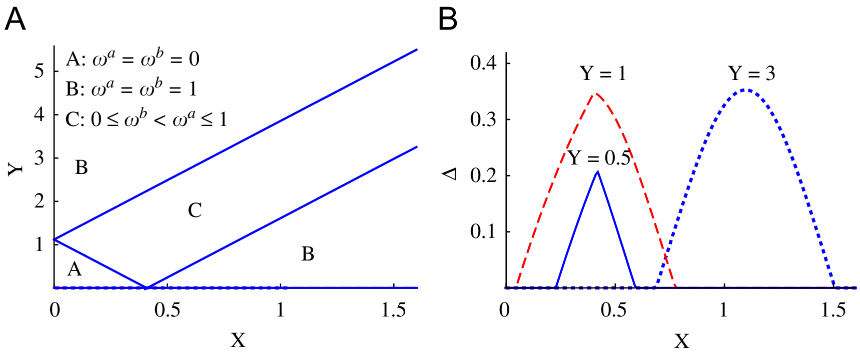

Fig. 3. Participation equilibrium. Panel A illustrates the participation equilibrium in the $Y>0$ and $X>0$ quadrant. The other quadrants can be obtained by symmetry. Region A represents states of no participation ($\omega^a = \omega^b = 0$); region B represents states of full participation ($\omega^a = \omega^b = 1$); region C represents states with asymmetric participation ($\omega^a > \omega^b$). Panel B illustrates the degree of asymmetry in participation between the two groups of trades, δ, for different values of Y and X. The market maker population is fixed at $\mu = \frac{1}{3}$. Parameters are set at the following values: $\alpha = 4$, $\sigma = 0.25$, $\sigma_z = 0.7$, $\sigma_\varepsilon = 1.2$, $\sigma_y = 0.7$, and $c = 0.09$.

The participation equilibrium for market makers is reached if one of the following three conditions is satisfied: (1) all agents choose to become market makers, i.e., $\mu = 1$ and $J^m(1, c_m) \ge J^n(1, c)$; (2) for some $\mu \in (0,1)$, agents are indifferent between being a market maker or a trader, i.e., $J^m(\mu, c_m) = J^n(\mu, c)$, and the fraction of agents choosing to become market makers is exactly μ; or (3) no agent chooses to become a market maker, i.e., $\mu = 0$ and $J^m(0, c_m) \le J^n(0, c)$. Lemma 1 is useful in obtaining the equilibrium population of market makers.

Lemma 1. *For any given population of market makers μ, there exists a unique $\kappa(\mu) \in [0, c]$ such that $J^m(\mu, \kappa) = J^n(\mu, c)$. Moreover, $\kappa(\mu)$ strictly decreases with μ for $\mu \in (\underline{\mu}, 1]$ and remains constant for any $\mu \in [0, \underline{\mu}]$, where*

$$\underline{\mu} \equiv \max\{0, \min\{\sqrt{4k/[(e^{2\alpha c} - 1)(1-k)]} - 1, 1\}\}. \tag{21}$$

The quantity $\kappa(\mu)$ is the break-even cost for an agent to become a market maker, taking as given the existing population of market makers μ. The second part of the lemma states that the benefit of becoming a market maker diminishes as the total population of market makers increases but could remain constant for sufficiently small μ.

The participation equilibrium of traders at date 1 is given in Proposition 4.

Proposition 4. *Let $\bar{c}_m \equiv \kappa(0)$, $\underline{c}_m \equiv \kappa(1)$, and $\kappa^{-1}(\cdot)$ be the inverse function of $\kappa(\cdot)$ defined in Lemma 1. The equilibrium population of market makers μ is determined as*

(i) $\mu = 1$, *if* $c_m < \underline{c}_m$;

(ii) $\mu = \kappa^{-1}(c_m) \in (\underline{\mu}, 1]$ *if* $\underline{c}_m \le c_m < \bar{c}_m$;

(iii) *any* $\mu \in [0, \underline{\mu}]$, *if* $c_m = \bar{c}_m$;

and (iv) $\mu = 0$, *if* $c_m > \bar{c}_m$. (22)

Except when $c_m = \bar{c}_m$, the equilibrium is unique. Moreover, as c_m approaches $\bar{c}_m$ from below, μ changes drastically with c_m. In particular, for $\underline{\mu} > 0$, μ drops discretely from $\underline{\mu}$ to 0. For $\underline{\mu} = 0$, $\partial\mu/\partial c_m = -O(e^{1/\mu^2})$, that is, μ decreases to 0 at an exponential rate.

Thus, in terms of equilibrium liquidity supply, the market exhibits two distinctive regimes. For $c_m < \bar{c}_m$, $\mu > 0$ and a finite amount of liquidity is supplied by market makers. For $c_m \ge \bar{c}_m$, however, $\mu = 0$ and zero liquidity is supplied by market makers. Moreover, the equilibrium market making capacity μ is not robust at low levels. When $\underline{\mu} > 0$, a discrete drop occurs in μ from $\underline{\mu}$ to 0 as the cost goes from slightly below $\bar{c}_m$ to slightly above. When $\underline{\mu} = 0$, even though there is no discrete drop, μ decreases to 0 at exponential speed for small μ. In both cases, low levels of μ are not sustainable in equilibrium. A slight increase in c_m can shift the equilibrium into a state with no market makers.

We conclude the solution of the equilibrium with Proposition 5, including the market equilibrium at date 0.

Proposition 5. *When $c_m < \bar{c}_m$, a unique equilibrium exists in which $P_0 = 0$, $\theta_0^i = 0$, and $\mu > 0$. When $c_m > \bar{c}_m$, there is a stationary equilibrium with $P_0 = 0$, $\theta_0^i = 0$, $\mu = 0$, and $\omega > 0$. When $c_m = \bar{c}_m$, there are multiple equilibria with different values of μ, which are Pareto equivalent.*

3.6. Properties of the equilibrium

The equilibrium obtained above exhibits several striking features. First, despite the fact that the two trader groups have perfectly matching trading needs, their actual trades are not matched when participation in the market is costly. A set of traders could bring their orders to the market while traders with offsetting trading needs are absent, creating an imbalance of orders and a need for liquidity. Second, the order imbalance causes the stock price to adjust to induce the market makers to absorb it. As a result, the stock price depends not only on the fundamentals (i.e., its expected future payoffs and the aggregate risk), but also on idiosyncratic shocks market participants face. Third, the market making capacity, determined endogenously in equilibrium, exhibits two distinctive regimes, one at a finite level and another at zero, depending on the costs of trading and market making.

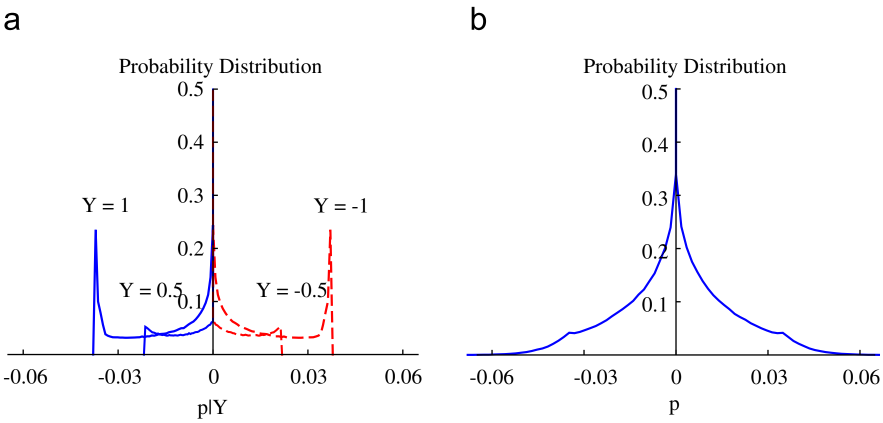

Fig. 4. Impact of illiquidity on price. Panel A reports the probability distribution of the liquidity impact p, given different values of the aggregate exposure Y. The solid lines are for $Y=1$ and 0.5, and the dotted lines are for $Y=-1$ and -0.5. Panel B reports the unconditional probability distribution of p. In both panels, the value at $p=0$ represents the total probability mass and at everywhere else represents the probability density. The market maker fraction is fixed at $\mu=\frac{1}{3}$. Parameters are set at the following values: $\alpha=4$, $\sigma=0.25$, $\sigma_z=0.7$, $\sigma_\varepsilon=1.2$, $\sigma_y=0.7$, and $c=0.09$.

4. Price and volume

As self-interest fails to coordinate traders' costly participation, perfectly matching trading needs give rise to unbalanced buy and sell orders. The sign and the magnitude of the order imbalance depend on the asymmetry in traders' participation δ and their idiosyncratic shock Z. In fact, we can define

$$q \equiv -\delta Z \tag{23}$$

to be the (normalized) order imbalance at date 1. At the time of participation decision, the expected order imbalance is $\mathrm{E}[-\delta Z|Y,X]=-\delta X$, which is mostly determined by δ, the asymmetry in participation between traders.

The endogenous order imbalance exhibits two interesting properties. First, it is often zero; but whenever it is nonzero, it has large magnitudes. For small values of Y and X (in region A), which represent most likely states, the gains from trading are small and no trader enters the market. As stated in Proposition 3 and shown in Fig. 3, the order imbalance is zero and there is no need for liquidity. Only for sufficiently large Y and X (in region C) do some traders start to participate in the market. Their asymmetric participation leads to an order imbalance, which is also of significant sizes.

Second, the order imbalance is always in the same direction as the impact of the aggregate shock on the demand of the stock. For example, when $Y>0$, the aggregate nontraded risk is positive, which is equivalent to an extra endowment of the stock, and the stock demand decreases. From Proposition 3 and Fig. 3, δX is positive in this case and the expected order imbalance is negative, further decreasing the demand. The reason the order imbalance always exacerbates the impact of the aggregate shock is that traders whose idiosyncratic shock is in the same direction as the aggregate shock Y always have higher trading gains and are more likely to enter the market. We thus summarize our main results on the endogenous need of liquidity as follows.

Result 1. *The endogenous order imbalance arises in significant magnitudes when it occurs. Moreover, it is always in the same direction as the impact of aggregate risk on asset demand.*

The need for liquidity affects prices. From Eq. (11), we see that the equilibrium stock price consists of two components, the fundamental value, $-\alpha\sigma^2, Y$, and a component driven by liquidity needs,

$$p \equiv -\alpha\sigma^2\delta Z. \tag{24}$$

We focus on this liquidity component. As mismatched trades give rise to order imbalances and the need for liquidity in the market, the stock price has to adjust to attract the market makers to provide liquidity and to accommodate the order imbalance. The price deviation p is driven by agents' idiosyncratic shocks and arises only when participation is costly.

For convenience, we consider the expected value of p conditional on Y and X, which we refer to as the average liquidity impact on price. From Eq. (24), the average liquidity impact is simply proportional to the expected order imbalance and exhibits the same properties. It depends on idiosyncratic shocks, and such a dependence is mostly for shocks of finite sizes. These properties lead to interesting predictions about price and return distributions.

Fig. 4, Panel A plots the probability distribution of the liquidity impact p given a level of aggregate risk Y, and Fig. 4, Panel B plots the unconditional probability distribution of p. The discrete nature of the liquidity needs gives rise to the high likelihood of large price movements. The liquidity impact is always zero under costless participation, which corresponds to a probability mass of 1 at $p=0$. Hence, Panels A and B clearly demonstrate that prices of the stock can significantly deviate away from its fundamental value, leading to additional variability and fat tails in the price. These deviations are caused by a surge in the liquidity need in the market, which is driven by idiosyncratic shocks among agents.[13] Thus, we have Result 2.

Result 2. *The impact of liquidity increases the price volatility of the stock and leads to fat tails in its returns.*

In addition to its impact on price, we can examine how liquidity affects the level of trading volume in equilibrium, which is given by

$$V \equiv \frac{1}{2}(1-\mu)\sum_{i=a,b}\omega^i|\delta Z-\lambda^i Z|+\frac{1}{2}\mu\sum_{i=a,b}|\delta Z-\lambda^i Z|. \tag{25}$$

In the absence of participation costs, the volume is simply $V=|Z|$. In the presence of participation costs, the volume is lower.

An exogenous order imbalance is the starting point for most models of market liquidity such as those in market microstructure analysis (e.g., Ho and Stoll, 1980; Glosten and Milgrom, 1985). By studying the need and the supply of liquidity in a unified framework, we show that the endogenous need for liquidity exhibits distinctive properties, including its highly nonlinear dependence on idiosyncratic shocks and its correlation with the aggregate risk. These properties lead to interesting implications on equilibrium prices and volume.

5. Equilibrium liquidity

The impact of liquidity needs on asset prices clearly depends on the amount of liquidity available in the market, which is supplied by market makers. Thus, the population of market makers measures the ex ante supply of liquidity.[14] In our setting, this is determined endogenously. Two factors are important in determining the equilibrium level of liquidity, the ex ante cost to be a market maker c_m and the spot cost c to jump in the market when needed. The cost c affects the potential need for liquidity and thus the benefit to supply liquidity as a market maker. We now consider how these two factors influence the equilibrium level of liquidity.

5.1. Supply of liquidity

Fig. 5 reports the equilibrium population of market makers μ as a function of their cost c_m, given traders' participation cost c. Consistent with Proposition 4, when

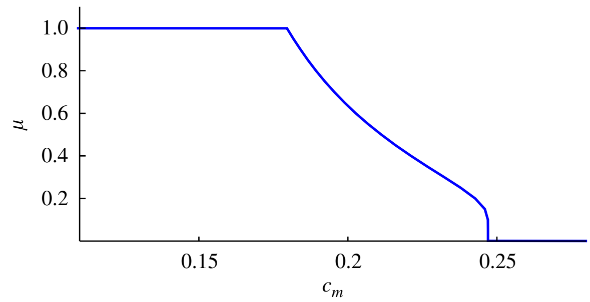

Fig. 5. Equilibrium population of market makers. The figure reports the population of market makers μ as a function of ex ante cost c_m. The spot participation cost for traders is set at $c=0.4$. Other parameters are set at the following values: $\alpha=4$, $\sigma=0.25$, $\sigma_z=0.7$, $\sigma_\varepsilon=1.2$, and $\sigma_y=0.7$.

c_m is small, i.e., less than $\underline{c}_m=0.179$, all agents choose to become market makers and $\mu=1$. When c_m is large, i.e., more than $\overline{c}_m=0.247$, no agent chooses to become a market maker and $\mu=0$. For in-between values of c_m, the fraction of market makers μ decreases as c_m increases. For the set of parameter values in the figure, $\underline{\mu}$ in Eq. (21) is zero. By Proposition 4, no discrete change occurs in μ as c_m approaches $\overline{c}_m$. However, in the figure, it appears that the value of μ drops from about 0.09 to 0 at $c_m=0.247$. This drastic change in μ is consistent with the extreme sensitivity of μ to c_m at small μ (of order $O(e^{1/\mu^2})$) described in the proposition.

The drastic decrease in μ indicates that low levels of market making capacity is in general not robust. A slight increase in the cost of supplying liquidity pushes the market into an equilibrium with no market makers. This result is driven by the externality in ex ante liquidity provision. As c_m increases, there are fewer market makers and traders expect to trade more with each other. This forces the participation decisions of the two groups of traders to become more correlated and their trades to become better matched. Better matching in their trades reduces potential order imbalances and further diminishes the need for market makers. Such an interaction between endogenous liquidity needs and endogenous liquidity provision makes low levels of liquidity provision ($\mu<0.09$ in the above example) unsustainable, as Fig. 5 illustrates. We summarize this result as follows.

Result 3. *When both the need and the supply of liquidity are determined endogenously, the level of ex ante supply of liquidity is not robust at low levels.*

Our result contrasts with that of Grossman and Miller (1988), in which the benefit for market makers decreases smoothly with their total population and the number of market makers decreases gradually as the cost increases. The difference comes from how liquidity needs are modeled. They take the liquidity need as exogenously given. We model the liquidity need endogenously, together with the endogenous liquidity supply by the market makers. We show that, as the supply decreases, the need for liquidity observed in the market also decreases, leading to a low liquidity equilibrium.

[13] The impact of liquidity on prices, which is driven purely by idiosyncratic shocks, implies that the average stock price carries a corresponding risk premium in addition to the aggregate risks, which we can refer to as a liquidity premium. Moreover, the impact of liquidity increases with the absolute level of aggregate risk, as Fig. 4, Panel A shows. As a result, the liquidity premium also increases with the level of aggregate risk. For a more detailed discussion about liquidity and its premium, see Huang and Wang (2009).

[14] Given the assumption of constant absolute risk aversion, each market maker's investment in the stock is independent of his wealth. Therefore, the total population of market makers also reflects the amount of capital they put in the stock market.

5.2. Two market structures: dealer market and trader market

The two regimes, one with market makers (when $c_m \le \overline{c}_m$) and the other with no market makers (when $c_m > \overline{c}_m$), correspond to two different market structures. Because the role of market making is often acclaimed by dealers, we refer to the market with market makers as a dealer market and the market without market makers as a trader market. We now consider how these two markets behave.

Fig. 6 reports the volatility of the liquidity component in price p and the average trading volume for different values of c_m, both of which exhibit different behavior under the two market structures. For $c_m \le \overline{c}_m$, which equals 0.247, we have the dealer market. Under this market structure, the supply of liquidity decreases as c_m increases, leading to an increase in the price impact, as measured by σ_p, and a decrease in the trading volume. For $c_m > \overline{c}_m$, we have the trader market, in which traders trade only among themselves. No liquidity is supplied by market makers. Because no one chooses to pay the cost c_m, neither the price nor the volume depends on the level of c_m. The participation of traders with offsetting trading needs can still be asymmetric in some states. The price adjusts to clear the market, giving rise to a positive σ_p. The benefit from participation is drastically reduced in the absence of market makers, and the average trading volume is very low (at about 0.007 in the figure).

Comparing the two market structures, we make two additional observations. First, even though a drastic drop in μ is evident at $c_m = \overline{c}_m = 0.247$ in Fig. 5, no discrete change occurs in price volatility. In fact, the volatility remains constant beyond a threshold level of $c_m = 0.238 < \overline{c}_m$. The reason for this result is as follows. When μ decreases, a given order imbalance has a larger impact on price. However, the large price impact also reduces the chance of order imbalances. In particular, traders with lower trading gains participate more to act as market makers, while traders with higher trading gains reduce their participation in anticipation of the low market making capacity. Although the equilibrium participation rate of each trader group varies with μ, the difference in their participation rates, δ, is maintained at a level such that the marginal group is indifferent between participating or not. The resulting price impact becomes independent of μ.

Second, while the literature usually associates higher volatility with lower liquidity in the market, our analysis shows that it is important to incorporate volume into the description of liquidity. Although, in a partial equilibrium analysis, the lack of ex ante liquidity supply usually leads to large price volatility, our example clearly indicates that volatility alone can be misleading. While the level of σ_p remains the same for all costs $c_m > 0.238$, the market structure is different for $0.238 \le c_m \le 0.247$ (the dealer market) and $c_m > 0.247$ (the trader market) and so is the level of liquidity. This can be seen from the different level of trading volume between the two markets. The average volume is significantly higher in the dealer market ($\mathrm{E}[V] > 0.1$) than in the trader market ($\mathrm{E}[V] = 0.007$). The reason that σ_p does not necessarily increase as liquidity drops is that traders optimally stay out of the market most of the time. The need for liquidity that arrives at the market can be low given the lack of its ex ante supply.

5.3. Demand of liquidity

Given the importance of the interaction between the demand and supply of liquidity, we now take c_m as given and examine how the cost of spot participation c affects the need for liquidity and the resulting equilibrium. Fig. 7, Panel A plots the equilibrium level of liquidity μ for different values of c. For small values of c, everyone can jump into the market on the spot at relatively low cost and thus no one chooses to become a market maker (i.e., $\mu = 0$). The equilibrium is a trader market. As c reaches a critical value of 0.281, the market maker fraction μ increases significantly and the market becomes a dealer market. The critical value of the spot participation cost, 0.281, is higher than the cost to become a market maker, which is set to $c_m = 0.2$. The reason for this difference is clear. Spot participation allows agents not to pay the cost in the event of low ex post trading needs. The value of this option is offset only when the cost of ex ante participation is significantly lower. As c keeps increasing, more agents choose to become market makers (i.e., μ increases with c). When c becomes sufficiently high (greater than 0.510), all agents become market makers and μ is always 1.

Fig. 7, Panel B demonstrates how the price impact of liquidity, as measured by σ_p, varies with the spot

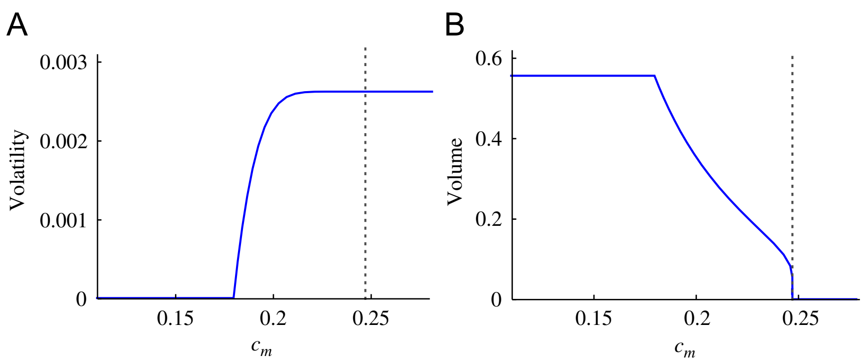

Fig. 6. Price volatility and volume. Panel A reports the volatility of liquidity component σ_p, and Panel B reports the average trading volume $\mathrm{E}[V]$ as functions of the ex ante cost c_m. The vertical dotted lines mark the point of $c_m = 0.247$, above which $\mu = 0$. The spot participation cost for traders is set at $c = 0.4$. Other parameters are set at the following values: $\alpha = 4$, $\sigma = 0.25$, $\sigma_z = 0.7$, $\sigma_\varepsilon = 1.2$, and $\sigma_y = 0.7$.

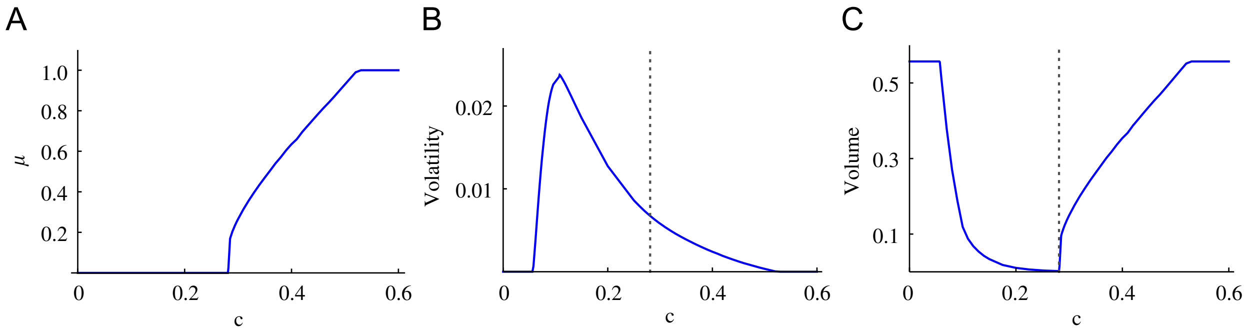

Fig. 7. Equilibrium and the cost of spot participation c. Panels A, B, and C report how equilibrium liquidity supply μ, price impact of liquidity σ_p and trading volume depend on c, respectively. The vertical dotted lines mark the point of $c=0.281$, below which $\mu=0$. The cost to become a market maker is fixed at $c_m=0.2$. Other parameters are set at the following values: $\alpha=4$, $\sigma=0.25$, $\sigma_z=0.7$, $\sigma_\varepsilon=1.2$, and $\sigma_y=0.7$.

participation cost. When $c \le 0.281$, we have a trader market ($\mu=0$). Surprisingly, even within this market structure, the price volatility is not monotonic in c. For very small c, all agents participate, leading to perfectly matched trades and no need for liquidity. Consequently, $\sigma_p=0$. As c increases, asymmetric participation occurs between traders. The stock price has to adjust to balance the buyers and the sellers. The increasing price volatility reflects an increase in participation asymmetry and a need for liquidity. When c increases further, the price volatility σ_p becomes decreasing with c. It is misleading, however, to interpret the reduction in σ_p as an indication of an improving market liquidity. Similar to the result of Fig. 6, this is due to the endogeneity of liquidity needs. An increase in c reduces spot liquidity, which forces traders to enter the market more symmetrically and reduces the observed need for liquidity. The much steeper drop in trading volume in Fig. 7, Panel C confirms the reduction in market liquidity. We summarize the result as follows.

Result 4. *When the need for liquidity is endogenous, a less liquid market could exhibit lower observed price impact of liquidity as traders refrain from trading, accompanied by lower trading volume.*

When c reaches a critical value, 0.281 in the figure, the market switches to a dealer market ($\mu > 0$). As Fig. 7, Panel A indicates, further increase of c encourages more agents to become market makers. Fig. 7, Panels B and C show that the price volatility continues the decreasing trend as the participation cost increases, while the volume starts to increase with the participation cost. Therefore, both price volatility and volume reflect an increasing market liquidity as the participation cost increases. This result is counterintuitive and is driven by the fact that higher ex post costs encourage agents to participate ex ante (and provide liquidity). We phrase the following result in terms of decreasing participation costs to be consistent with policy discussions.

Result 5. *When both the demand and supply of liquidity are endogenous, lowering the cost of spot participation can reduce market liquidity by discouraging agents to participate ex ante.*

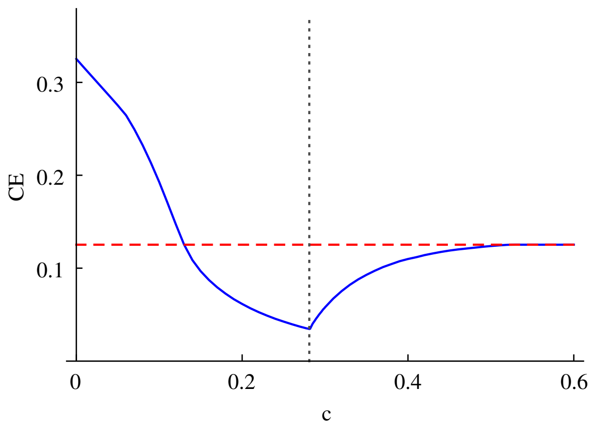

Fig. 8. Welfare and the cost of spot participation c. The solid line reports the certainty equivalent gain CE from optimal participation (hence $CE=CE^n=CE^m$) as a function of spot participation cost c. The horizontal dashed line marks the level of CE at $c=0.6$. The vertical dotted line marks the point of $c=0.281$, below which $\mu=0$. The cost to become a market maker is fixed at $c_m=0.2$. Other parameters are set at the following values: $\alpha=4$, $\sigma=0.25$, $\sigma_z=0.7$, $\sigma_\varepsilon=1.2$, and $\sigma_y=0.7$.

This result reflects the negative liquidity externality when agents withdraw from the market.

6. Externality and welfare of liquidity

In this section, we consider the welfare implications of the externality from trading. We measure an agent's welfare by his certainty equivalence gain from participating in the market. Using the value functions of market makers and traders in Eq. (20), we can define the certainty equivalence gain as $CE^i \equiv -(1/\alpha)\ln J^i/J_{NP}$, for $i=m,n$, where $J_{NP}=\mathrm{E}[J^n_{NP}]$ is the value function of an agent who never participates. Because all agents are ex ante identical and have the choice of becoming a market maker or a trader, their ex ante welfare is also identical, which is given by $CE \equiv \max\{CE^n, CE^m\}$.

In Fig. 8, we plot CE for different values of c (the solid line) when both liquidity demand (δ) and supply (μ) are determined endogenously. In the absence of any externalities, one might expect the welfare to decrease with c.

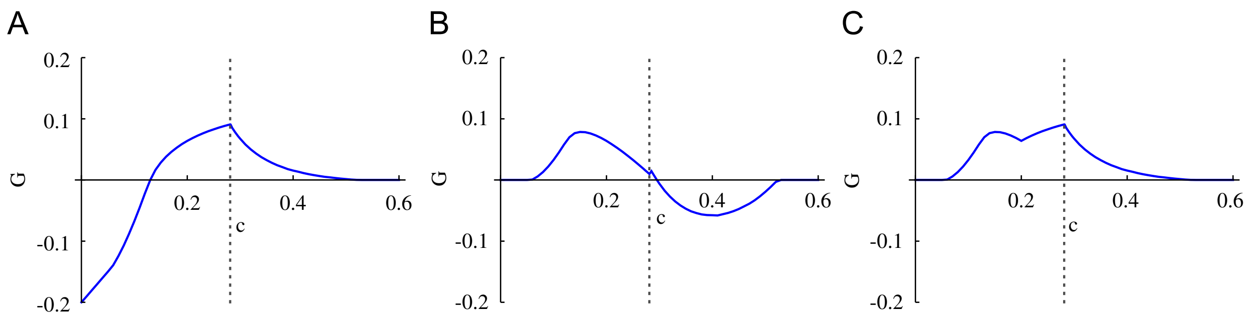

Fig. 9. The welfare improvement of forced participation. The figure reports the change in the certainty equivalent wealth G as a function of spot participation cost c. Panel A reports the case of ex ante intervention, in which all agents are forced to pay the ex ante cost c_m. Panels B and C report the case of spot intervention, in which all potential traders are forced to pay the spot cost c. The forced participation comes as a surprise in Panel B and is fully anticipated in Panel C. The vertical dotted lines mark the point of $c=0.281$, below which $\mu=0$. The cost to become a market maker is fixed at $c_m=0.2$. Other parameters are set at the following values: $\alpha=4$, $\sigma=0.25$, $\sigma_z=0.7$, $\sigma_\varepsilon=1.2$, and $\sigma_y=0.7$.

Fig. 8 clearly indicates that the opposite can be true. That is, when the spot participation cost increases, agents' welfare can improve. The dashed line is a horizontal line that marks the CE level at $c=0.6$. It is higher than (or equal to) the CE for all $c>0.13$, indicating that agents are better off at $c=0.6$ than at $0.13<c<0.6$. This is a surprising result, which arises from the externality generated by those who supply liquidity by becoming market makers.

To see how liquidity provision influences welfare, we note from Fig. 7, Panel A that the market structure changes around $c=0.281$. The population of market makers increases steeply from zero to about 0.17 as c increases from slightly below 0.281 to slight above. The increase in the population of market makers increases the liquidity supply in the market, which enhances the welfare of all agents as Fig. 8 demonstrates. Moreover, the point at which μ becomes positive ($c=0.281$) coincides with the point at which the welfare of agents starts to increase with c.

At $c=0.281$, the market structure changes from a trader market to a dealer market. Although the population of market makers increases drastically from 0 to 0.17, the change in the welfare level of the economy is smooth at this point. The continuity in welfare, however, does not imply that the change in market structure is immaterial. Fig. 8 demonstrates a clear regime shift at this point. A discrete change in the relation occurs between welfare and primitives of the economy such as the cost of spot participation. In particular, when $c<0.281$, decreasing the spot participation cost does not change the market structure (μ remains 0) and always improves welfare. When $c>0.281$, however, decreasing the spot participation cost reduces ex ante liquidity provision and hence decreases welfare. Such change in the properties of the market has important policy implications. We summarize this result as follows.

Result 6. *When both the demand and the supply of liquidity are determined endogenously, lowering the cost of spot participation can have the adverse effect of reducing welfare.*

This result suggests that, in the presence of trading externality, the market equilibrium (in which agents optimally choose whether and when to enter the market) can be suboptimal. In particular, the equilibrium supply of liquidity can be inefficient. To illustrate this point, we show that agents' welfare in the market equilibrium can be improved by simply forcing agents to pay the participation cost to be in the market. The forced participation can be carried out either ex ante or on the spot. In the former case all agents are forced to pay cost c_m and become market makers, while in the latter case all traders are forced to pay cost c to participate in the market on the spot. We consider these two cases separately.

We define the welfare gain under forced participation as the difference between the welfare levels in the equilibrium under forced participation and under optimal participation,

$$G \equiv CE_{\mathrm{P}} - CE, \tag{26}$$

where CE_{P} is defined as the CE in the forced participation equilibrium.

We first consider the case of forced ex ante participation. Fig. 9, Panel A reports the welfare gain G in this case. When c is very small, forcing agents to pay the high cost c_m, which is set at 0.2, is clearly inefficient and reduces welfare. For the range of $0.13<c<0.20$, the gain $G>0$, indicating that forcing all agents to pay cost c_m can improve welfare, even though all agents have the option to pay a lower cost c in the spot market in the competitive equilibrium. The improvement in welfare reflects the fact that each agent's participation brings additional liquidity to the market and thus improves the welfare of others. In the competitive market equilibrium, an agent is not sufficiently compensated for such a social benefit. Thus, their individual decisions can be different from what is socially optimal. In the equilibrium under forced participation, enough gains are generated to be shared equally among all agents, which can outweigh the extra costs paid. In summary, we have the following result.

Result 7. *Individual participation choices can lead to insufficient liquidity supply in the market and the resulting welfare loss can outweigh total participation costs.*

We now consider the case of forced spot participation, in which all traders pay cost c to participate, independent of their trading needs. We first consider the welfare gain, holding the population of market makers the same as that under the competitive equilibrium. This is equivalent to assuming that the forced participation is unanticipated so that agents do not adjust their decisions to become market makers in the first place. Fig. 9, Panel B shows the welfare gain in this situation.[15] For $c<0.057$, agents optimally choose to always participate as traders, yielding the same outcome as in the forced participation equilibrium. Hence $G=0$. For c ranges from 0.057 to 0.295, forced spot participation improves welfare. From Fig. 7, Panel A, the equilibrium population of market makers is small for this range of c. Forcing spot participation improves market liquidity significantly, leading to the welfare gain.

In the case of forced spot participation, if agents are allowed to adjust their ex ante participation decisions in anticipation of forced participation, their welfare is further increased. They rationally choose to pay the spot cost c for all $c<c_m$ and to pay the ex ante cost c_m for all $c>c_m$. Thus, the welfare gain G, given in Fig. 9, Panel C, is simply the maximum of the G's in Fig. 9, Panels A and B. In this case, the gain G is always positive, indicating that forced spot participation always improves social welfare when all agents rationally anticipate the policy. The gain is driven mainly by the increased ex ante liquidity provision. Thus, we have the following result.

Result 8. *Forcing agents to participate in the market can improve social welfare, especially if it encourages ex ante liquidity provision.*

Despite the simplicity of our setting, the mechanism we have identified for a market failure in coordinating costly liquidity provision is general: Each agent not only benefits from his own trades but also brings liquidity to the market. Bearing the full cost alone, each agent might not be able to efficiently internalize the benefit he creates for the market. As a result, the traders' participation decisions, while optimal at the individual level, could well be socially suboptimal.

The literature shows that markets might not always achieve efficient outcomes when frictions are present. For example, Diamond (1982) examines markets in which trading is conducted through a search process and shows that the resulting equilibrium can be inefficient. Pagano (1989) and Allen and Gale (1994) show the possibility of Pareto-dominated equilibria in markets with ex ante participation costs. Our results are different in nature. These papers focus on the multiplicity of equilibria and the Pareto inefficiency of some of these equilibria relative to others. We focus on the equilibrium that is not dominated, and our results are on its inefficiency (as stated in Section 3.4, we ignore the dominated equilibrium). Our welfare comparison is between equilibrium under different primitives such as c and c_m (i.e., between different economies), not between different equilibria of a given economy.

[15] In this situation, the ex ante welfare of market makers and traders is no longer the same. G is then defined as the population weighted average of market makers' and traders' ex ante welfare measured in certainty equivalence.

7. Policy implications

Liquidity in the market, especially at the time of crises, has been an important issue for regulators and policy makers. For example, during unusual times, such as the LTCM crisis in 1998, the days around Y2K, the time after the September 11, 2001, terrorist attacks, and the recent subprime mortgage crisis, the Federal Reserve Bank took direct actions to ensure sufficient level of liquidity in the market when needed. These actions range from the coordination of major dealers in providing liquidity (e.g., for the LTCM crises) to the direct injection of liquidity (e.g., for Y2K, September 11, and the subprime crisis). The current surge of the hedge fund industry also raises new challenges. On the one hand, facing fewer constraints than most existing financial intermediaries, hedge funds often play the role of market makers and supply liquidity. On the other hand, the risk taking nature of their business tends to put hedge funds in volatile situations especially when crises hit. Increasing concerns have arisen about their impact on market stability if they become liquidity constrained themselves. Tightening margins and restricting exposures of major banks to hedge funds have been proposed as preventive measures to restrain potential liquidity crunches.

Arguments have been presented both for and against these actions and proposals. But a comprehensive theoretical foundation for these policy discussions remains lacking. Although a detailed policy discussion is not the focus of this paper, our model nonetheless provides a useful framework to consider the determinants of market liquidity and to examine the welfare impact of certain intervention policies. A full analysis of the model's policy implications is beyond the scope of this paper, so our discussion below is only exploratory.

Our theory predicts that lowering the cost of ex ante participation in general increases liquidity supply and welfare. Therefore, policies that lower the entry cost and restrictions for dealers and market makers are welfare improving. To the extent that hedge funds perform the market making role, relaxing their margin constraints could decrease the cost for them to maintain their constant presence and improve market liquidity. However, we find that lowering the cost of spot participation does not necessarily increase liquidity supply and welfare, especially if it is anticipated by market participants. This suggests that an anticipated government policy of relaxing margin constraints or injecting liquidity during crises is not always optimal. It tends to reduce the incentive for agents to establish themselves as market makers and thus lowers the level of liquidity supplied by the market.[16]

Our discussion is based on the interpretation of spot liquidity injection as lowing the cost of spot participation

[16] Before Y2K, the Federal Reserve Bank of New York sold loan options to depository institutions and Treasury bond dealers (Special Liquidity Facility and Special Financing Facility) to guarantee sufficient liquidity during the Y2K transition. This is in the spirit of a state-contingent liquidity injection considered here. Interested readers are referred to Sundaresan and Wang (2009) for a more detailed account of the Y2K options and the market behavior during that time.

to lure those who are holding back to jump in. In the time of crisis, however, liquidity injection often takes the form of relaxing capital constraints. Although in our model capital plays no explicit role in agents' behavior, the population of market makers plays the same role as the total amount of capital in the market in terms of affecting the overall risk taking capacity. From this perspective, increasing capital is equivalent to adding more market makers ex post, which can reduce the profit for existing market makers. This effect is similar to that of lowering the spot participation cost in the model, which encourages traders with offsetting trading needs to provide liquidity and to compete away the profit for existing market makers. In both cases, the spot liquidity reduces the ex ante incentive to become a market maker (or to stock capital).

If, however, the capital is targeted directly at the existing market makers, then it becomes a subsidy to them. The resulting impact can be complex, depending on factors such as how the capital is raised and distributed. Suppose, for example, that the capital is distributed evenly among market makers free of charge. Then, this capital injection amounts to a government handout to existing market makers and can induce more agents to become market makers. However, this subsidy needs to be paid, say, through an ex ante tax over all agents. The net effect depends on the trade-off between the gain from more market makers (and more liquidity) and the cost to induce them. Suppose, however, the liquidity is offered to the market makers through a market mechanism, such as the new credit facilities the US Fed offered to banks and security firms during the current credit crisis. The same market failure arises in which market makers choose to buy inefficient amounts of liquidity.

Our findings by no means rule out the possibility of positive intervention during crises. If, instead, the government can coordinate traders to participate in the market in the event of severe liquidity shortage, liquidity and welfare can be improved under certain circumstances.[17] In general, our theory suggests that mechanisms that resemble forced spot participation (e.g., coordination of trading), especially if they are anticipated by the market, are better at improving liquidity than those that resemble subsidized spot participation (e.g., direct injection of liquidity or relaxation of ex post margin constraints). The reason is that agents do not expect to gain by waiting for spot participation, and hence the anticipation of future interventions does not hinder their ex ante liquidity provision motive.

Our analysis also shows that policy implications can be different under different market structures. For example, as shown in Fig. 8, while lowering the spot participation cost can improve welfare in a trader market (when $c<0.281$), it decreases welfare in a dealer market (when $c\geq 0.281$).

[17] During the LTCM crisis, the Federal Reserve Bank of New York facilitated the formation of a consortium of investment banks, which provided the new capital to prevent the hedge fund from collapsing.

8. Related literature

The literature on liquidity and its impact on the securities market is extensive. In this section, we discuss those works that are closest to this paper. Most of the previous work has focused on the supply of liquidity, taking its demand as exogenous. The theory on market microstructure, which studies the actual trading process, starts with an exogenous order flow process and examines how market makers provide liquidity by accommodating order imbalances (e.g., Ho and Stoll, 1980; Stoll, 1985; Glosten and Milgrom, 1985; Kyle, 1985). Grossman and Miller (1988) further point out that it is costly for market makers to maintain market presence. They analyze how these costs determine the level of liquidity supply and its impact on prices under exogenous liquidity shocks. In this paper, we show that the same costs give rise to the need for liquidity in the first place. By explicitly modeling the endogenous need for liquidity, we obtain important insights on how it behaves, how it interacts with the supply of liquidity in equilibrium, and how liquidity affects prices and welfare.

Our paper expands the work of Grossman and Miller (1988), Pagano (1989), and Allen and Gale (1994). By observing that the same participation cost causes the need for liquidity in the first place, we fully endogenize the liquidity need (or order imbalance). Instead of relying on exogenous liquidity shocks at the aggregate level, we show how liquidity need arises from purely idiosyncratic shocks. This allows us to gain additional insights into its properties, which can be different from those assumed for exogenous liquidity shocks. It also allows us to examine how the demand and supply of liquidity interact with each other in equilibrium, leading to different market structures and different relations between liquidity and price behavior. It further allows us to study how liquidity affects welfare.

The model we use shares many features with the model of Lo, Mamaysky, and Wang (2004), who consider the impact of fixed transactions costs on trading volume and the level of asset prices. The main difference is that we focus on the possible imbalance in liquidity needs and its impact on prices while they do not. They allow the cost to be allocated endogenously so that the trades of different market participants are always synchronized in equilibrium and there is no order imbalance and net liquidity need. As we show in this paper, order imbalance leads to changes in liquidity needs and instability in asset prices.

A closely related paper is Huang and Wang (2009), which uses a similar setting to arrive at endogenous liquidity need. The main differences are twofold. In Huang and Wang (2009), the supply of liquidity is taken as given while analyzing the demand for liquidity. In this paper, we also endogenize the supply of liquidity. As we have shown, the interaction between the two, when both are endogenous, has a fundamental influence on the behavior of liquidity in the market. Second, Huang and Wang (2009) focus on the impact of liquidity on prices. In this paper, we focus on market structure, welfare, and policy implications concerning liquidity. It is also for this purpose that we have to endogenize liquidity supply in a unified setting. At a more technical level, the aggregate risk is assumed to be positive and constant in Huang and Wang (2009). This is needed in modeling assets with positively supply such as the equity market. For our purpose, we do not need this restrictive assumption.

In our model, costs to transact in the market take the simple form of participation costs. The organization of the market still takes the form of a centralized exchange. This is a reasonable description for major securities markets, such as the New York Stock Exchange or Chicago Mercantile Exchange, but less so for others, such as over-the-counter (OTC) markets for long term options and corporate bonds. For these OTC markets, costs to transact could take different forms. For example, Duffie, Garleanu, and Pedersen (2005) solve for equilibrium prices in an OTC market with search and bargaining among market participants.[18]

Our paper is also related to a growing literature that studies the welfare implications of different market structures. Brusco and Jackson (1999) show that competitive ex post trading reduces the incentive for agents to participate ex ante to become market makers and can lead to Pareto inefficiency. They argue that giving market makers ex post market power can increase their ex ante participation and improve social welfare.

9. Conclusion

In this paper, we show that frictions such as participation costs can induce imbalances in agents' trades even when their trading needs are perfectly matched. Each trader, when arriving at the market, faces only a partial demand and supply of the asset. The mismatch in the timing and the size of trades creates temporary order imbalance and the need for liquidity, which causes asset prices to deviate from the fundamentals. By endogenously determining both the demand and supply of liquidity, we are able to show that purely idiosyncratic liquidity shocks can affect prices, introducing additional price volatility. The price deviations always amplify the price impact of aggregate shocks and is of large sizes whenever they occur, leading to fat tails in returns.

Moreover, we find that traders optimally refrain from participating in less liquid market, leading to lower observed liquidity needs. As a result, prices do not necessarily exhibit higher liquidity impact or higher volatility in less liquid markets, rendering it necessary to incorporate trading volume into measures of market liquidity.

Finally, we show that partial participation in the market by a subset of traders can have important welfare implications. In particular, the withdrawal of a subset of traders from the market reduces market liquidity, which further reduces the incentive for others to participate in the market. The fact that participating agents cannot fully internalize the benefit from their liquidity provision leads to suboptimal provision of liquidity despite the optimizing behavior at the individual level.

This inefficiency in the market mechanism leaves room for policy intervention. However, the design of efficient intervention is far from obvious as it affects the demand and supply of liquidity in intricate ways. For example, lowering the cost of supplying liquidity on the spot (e.g., through direct injection of liquidity or relaxation of ex post margin constraints) can decrease welfare by reducing the profit opportunities for market makers and thus the ex ante incentive for them to be there. However, forcing more liquidity supply (e.g., through coordination of market participants) during times of crises can improve welfare. The key distinction is that agents do not expect to be subsidized during crises, and hence the anticipation of future interventions does not hinder their ex ante incentive to supply liquidity.

[18] The literature that utilizes the search framework to model financial market transactions includes Rubinstein and Wolinsky (1987), Gale (1987), and Vayanos and Wang (2007).

Appendix A

Proof of Proposition 1. Participating agent i maximizes his expected utility over his terminal wealth W_2^i, defined in Eq. (7). Integrating over the distribution of D, we have the following:

$$\max_{\theta_1^i} -e^{-\alpha[-c^i+\theta_0^i(P_1-P_0)+\theta_1^i(-P_1)-\frac{1}{2}\alpha\sigma^2(\theta_1^i+Y+\lambda^i Z)^2]}. \tag{27}$$

The optimal holding is obtained by solving the first-order condition with respect to θ_1^i:

$$\theta_1^i = -P_1/(\alpha\sigma^2) - Y - \lambda^i Z, \quad i = a, b. \tag{28}$$

Given initial holding $\theta_0^i = 0$ and (ω^a, ω^b), the market clearing condition at time 1_+ is

$$\tfrac{1}{2}\mu(\theta_1^a + \theta_1^b) + \tfrac{1}{2}(1-\mu)(\omega^a\theta_1^a + \omega^b\theta_1^b) = 0. \tag{29}$$

Substituting θ_1^i into Eq. (28) and the definition of δ in Eq. (9) yields the equilibrium price P_1. The optimal holding in the proposition is obtained by substituting the equilibrium price P_1 back into Eq. (28). □

Proof of Proposition 2. To calculate J_P^i, we substitute $\theta_0^i = 0$, the equilibrium P_1, and θ_1^i into Eq. (27) and integrate over Z conditional on Y, X, and λ^i, which yields

$$J_P^i(\cdot) = -\frac{1}{\sqrt{1-k+k(1-\lambda^i\delta)^2}} \times e^{-\alpha[-c^i-(\alpha\sigma^2/2(1-k))(Y+\lambda^i X)^2+(\alpha\sigma^2(1-\lambda^i\delta)^2/2(1-k)[1-k+k(1-\lambda^i\delta)^2])(kY+\lambda^i X)^2]}. \tag{30}$$

To calculate J_{NP}^i, we set $\theta_1^i = \theta_0^i$ and $c^i = 0$ in Eq. (27) and integrate over Z conditional on Y, X, and λ^i:

$$J_{NP}^i(\cdot) = -\frac{1}{\sqrt{1-k}} e^{-\alpha[-(\alpha\sigma^2/2(1-k))(Y+\lambda^i X)^2]}. \tag{31}$$

Substituting J_P^i and J_{NP}^i into Eq. (14) yields the trading gain $g(\cdot)$. Clearly, $J_P^i > J_{NP}^i$ if and only if $g^i(\cdot) > 0$. □

Proof of Proposition 3. For brevity, we denote $g^i(\delta) \equiv g^i(\delta; Y, X)$. We prove the result when X and Y have the same sign. The case of different signs can be proved by switching the indexes a and b. □

Lemma A.1. *The gains $g^a(\delta)$ strictly decreases with δ and $g^b(\delta)$ strictly increases with δ.*

Proof. Using Eq. (15), we compute the partial derivative of $g^i(\cdot)$ with respect to δ,

$$\frac{\partial g^i(\delta)}{\partial \delta} = -\lambda^i(1-\lambda^i\delta)\left[\frac{\alpha\sigma^2(kY+\lambda^i X)^2}{(d^i)^2} + \frac{k}{\alpha d^i}\right], d^i \equiv 1-k+k(1-\lambda^i\delta)^2. \tag{32}$$

Because $k>0$, $d^i>0$, $\delta<1$, and $\lambda^a=-\lambda^b=1$, we have $\partial g^a/\partial\delta<0$ and $\partial g^b/\partial\delta>0$. □

Lemma A.2. *When $\delta=0$, $\gamma(0)\geq g^b(0)$.*

Proof.

$$g^i(0) = \frac{\alpha\sigma^2}{2(1-k)}(kY+\lambda^i X)^2 - \frac{1}{2\alpha}\ln(1-k) - c, \quad i=a,b.$$

Whenever X has the same sign as Y, we have $g^a(0)\geq g^b(0)$. □

From Lemma A.2, the state space has three regions: (A) $0\geq g^a(0)\geq g^b(0)$, (B) $g^a(0)\geq g^b(0)\geq 0$, and (C) $g^a(0)>0>g^b(0)$, which correspond to the three cases in the proposition. In region A, we can show that $\omega^a=\omega^b=0$ is the unique equilibrium. If instead $\omega^a>\omega^b$, then $\delta>0$ and $\gamma(\delta)<\gamma(0)\leq 0$ from Lemma A.1 and the condition for region A. Hence, some group-a traders exit and ω^a decreases. Similarly, if $\omega^a<\omega^b$, then $\delta<0$ and $g^b(\delta)<g^b(0)\leq 0$. Group-b traders exit and ω^b decreases. Hence, in equilibrium, $\omega^a=\omega^b$ and $\delta=0$. Because both $g^i(0)\leq 0$, $\omega^a=\omega^b=0$ is the unique equilibrium. Similarly, in region B, we can show that $\omega^a=\omega^b=1$ and $\delta=0$ is the unique equilibrium. $g^a(0)=g^b(0)=0$ is included in both regions A and B. In fact, any $\omega^a=\omega^b\in[0,1]$ is a solution. We do not separate out this case for conciseness, as it occurs only for a single realization of X and Y.

In region C, we consider three subcases based on $g^i(\overline{\delta})$, where $\overline{\delta}\equiv(1-\mu)/(1+\mu)$ is the maximum possible δ in (9) (because ω^a and ω^b are bounded in $[0,1]$).

(i) If $\gamma(\overline{\delta})>0>g^b(\overline{\delta})$, then Lemma A.1 yields $\gamma(\delta)\geq\gamma(\overline{\delta})>0>g^b(\overline{\delta})\geq g^b(\delta)$ for any feasible δ. Thus, $\omega^a=1$ and $\omega^b=0$ is the unique equilibrium, and $\delta=\overline{\delta}$.

(ii) If $\gamma(\overline{\delta})>0$ and $g^b(\overline{\delta})>0$, then there exists a unique $s^b\in(0,\overline{\delta})$ that solves $g^b(s^b)=0$. (Lemma A.1 and $g^b(0)<0$ in region C.) Because $\gamma(\delta)\geq\gamma(\overline{\delta})\geq 0$ for any feasible δ, we always have $\omega^a=1$ in equilibrium. Let

$$\hat{\omega}^b \equiv \frac{\frac{1-\mu}{2}(1-s^b)-\mu s^b}{\frac{1-\mu}{2}(1+s^b)},$$

then for any $\omega^b>\hat{\omega}^b$, $\delta<s^b$ and $g^b(\delta)<g^b(s^b)=0$, some group-b traders stop participating and ω^b decreases. For any $\omega^b<\hat{\omega}^b$, $\delta>s^b$ and $g^b(\delta)>g^b(s^b)=0$, and more group-b traders participate and ω^b increases. Hence, $\omega^a=1$, $\omega^b=\hat{\omega}^b\in[0,1)$, and $\delta=s^b$ is the unique equilibrium.

(iii) If $\gamma(\overline{\delta})\leq 0$, there exists a unique $s^a\in(0,\overline{\delta}]$ that solves $g^a(s^a)=0$. If $g^b(s^a)\leq 0$, then a similar argument to case (ii) shows that

$$\omega^a=\hat{\omega}^a\equiv\frac{\mu s^a}{\frac{1-\mu}{2}(1-s^a)}\in(0,1]$$

and $\omega^b=0$ is the unique equilibrium and $\delta=s^a$. If $g^b(s^a)>0$, there exists a unique $s^b\in(0,s^a)$ that solves $g^b(s^b)=0$. Because $\gamma(s^b)>\gamma(s^a)=0$, $\omega^a=1$, $\omega^b=\hat{\omega}^b\in[0,1)$, and $\delta=s^b$ is the unique equilibrium.

We now consider the case of $\mu=0$. First, $\omega^a=\omega^b=0$ is always an equilibrium. Assume the equilibrium belief is $\omega^a=0$, then $J^b_P=J^b_{NP}e^{\alpha c}<J^b_{NP}$ and $\omega^b=0$ is the only equilibrium outcome. Similarly, a belief of $\omega^b=0$ leads to a unique equilibrium of $\omega^a=0$. Second, in the above positive participation equilibrium, because $kY+X$ can be arbitrarily large, $g^a(0)>0$ is always possible. Hence, region A does not cover the full state space and we have $\omega^a=1$ for at least some realizations of X and Y. Whenever $\omega^a=1$, the trading gain $g^a\geq 0$. Because $g^a=0$ when $\omega^a=\omega^b=0$, the equilibrium without participation is always Pareto dominated by the one with participation.

Proof of Lemma 1. We first prove the existence and uniqueness of κ. From Eqs. (20) and (30), we have

$$\frac{\partial J^m(\mu,c_m)}{\partial c_m}=\alpha J^m<0. \tag{33}$$

Also, we show that $J^m(\mu,c)\leq J^n(\mu,c)\leq J^m(\mu,0)$, where the first inequality is because of Eq. (20) and the fact that $J^i_P\leq\max\{J^i_P,J^i_{NP}\}$, and the second inequality is because $J^m(\mu,0)=J^n(\mu,0)>J^n(\mu,c)$ for any $c\leq 0$. Hence, there exists a unique $\kappa\in[0,c]$ such that $J^m(\mu,\kappa)=J^n(\mu,c)$.

To show that κ decreases with μ, we take derivative of $J^m(\cdot)=J^n(\cdot)$ with respect to μ on both sides:

$$\frac{\partial J^m(\mu,\kappa)}{\partial\mu}+\frac{\partial J^m(\mu,\kappa)}{\partial\kappa}\frac{\partial\kappa}{\partial\mu}=\frac{\partial J^n(\mu,c)}{\partial\mu}. \tag{34}$$

Given Eq. (33), we only need to calculate $\partial J^m/\partial\mu$ and $\partial J^n/\partial\mu$ to sign $\partial\kappa/\partial\mu$.

Following the proof of Proposition 3, we separate the state space (X,Y) into five regions: (A) $\omega^a=\omega^b=0$, (B) $\omega^a=\omega^b=1$, (C_1) $\omega^a=1$, $\omega^b=0$, (C_2) $\omega^a\in(0,1)$, $\omega^b=0$, and (C_3) $\omega^a=1$, $\omega^b\in(0,1)$. Regions A and B are the same as those in Proposition 3, and combining regions C_1, C_2 and C_3 yields region C. Let $G^i\equiv J^i_P-J^i_{NP}$, then $G^i=J^i_{NP}(e^{-\alpha g^i(\delta)}-1)$, where $g^i(\delta)\equiv g^i(\cdot)$ in Eq. (15). Thus, $G^i>0$ if and only if $g^i>0$, which occurs only if $\omega^i=1$. Hence, we can write $J^n(\cdot)$ as J_{NP} plus the gains from trading in regions with $\omega^i=1$. That is,

$$J^n(\mu,c)=\tfrac{1}{2}\,\mathrm{E}[J^a_{NP}+J^b_{NP}]+4\times\tfrac{1}{2}(\mathrm{E}_{\{B,C_1,C_3\}}[G^a]+\mathrm{E}_B[G^b]), \tag{35}$$

where the factor $\frac{1}{2}$ reflects averaging over realizations of λ^i and the factor 4 reflects the symmetric gain in the four quadrants while we focus only on the $X>0, Y>0$ quadrant.

To calculate $\partial J^n(\mu,c)/\partial\mu$, note that J^i_{NP} is clearly independent of μ. Because $g^i(\cdot)$ depends on μ only through δ, so does G^i. Moreover, in regions A and B, $\delta=0$ and is clearly independent of μ. In regions C_2 and C_3, δ solves either $g^a(\delta)=0$ or $g^b(\delta)=0$ and is also independent of μ. Therefore, G^i depends on μ only in region C_1, in which $\delta=\overline{\delta}$. Let N^0 be the boundary of any region N, then

$$\frac{\partial \mathrm{E}_N[G^i]}{\partial\mu} = G^i(N^0)\frac{\partial N^0}{\partial\mu} + \mathrm{E}_N\left[\frac{\partial G^i}{\partial\mu}\right]. \tag{36}$$

Hence the second term is nonzero only in region C_1. To calculate the first term, note that $N=\{B,C_1,C_3\}$ for agent a. From the proof of Proposition 3, the boundary N^0 is $g^a(\delta)=0$. Hence, $G^a(N^0)=0$. For agent b, $N=\{B\}$ and the boundary is $g^b(0)=0$, which is independent of μ. Hence, $\partial N^0/\partial\mu=0$. So the first term of Eq. (36) is always 0. Therefore,

$$\frac{\partial J^n(\mu,c)}{\partial\mu} = 2\mathrm{E}_{C_1}\left[\frac{\partial G^a}{\partial\mu}\right]. \tag{37}$$

Similarly, we write J^m as J_{NP} plus trading gains and apply Eq. (36) to calculate $\partial J^m(\mu,c)/\partial\mu$.

$$J^m(\mu,c_m) = \tfrac{1}{2}\,\mathrm{E}[J^a_{NP}+J^b_{NP}] + 4\times\tfrac{1}{2}\mathrm{E}_N[G^a+G^b]|_{c^i=c_m},$$
$$N=\{A,B,C_1,C_2,C_3\}. \tag{38}$$

Because N is the full space, the first term of Eq. (36) is also zero. Hence,

$$\frac{\partial J^m(\mu,c_m)}{\partial\mu} = 2\left(\mathrm{E}_{C_1}\left[\frac{\partial G^a}{\partial\mu}\right] + \mathrm{E}_{C_1}\left[\frac{\partial G^b}{\partial\mu}\right]\right). \tag{39}$$

Combining Eqs. (33), (34), (37) and (38), we have

$$\frac{\partial\kappa}{\partial\mu} = -\frac{2}{\alpha J^m}\,\mathrm{E}_{C_1}\left[\frac{\partial G^b}{\partial\mu}\right] \le 0, \tag{40}$$

where the inequality follows from $J^m<0$, $\partial G^b/\partial\mu = \partial G^b/\partial\delta\,\partial\overline{\delta}/\partial\mu$, and $\partial G^b/\partial\delta>0$ (from Lemma A.1) and $\partial\overline{\delta}/\partial\mu<0$ (from the definition of $\overline{\delta}$).

The condition for the strict inequality can be derived in three steps. First, given the strict negativity of $\partial G^b/\partial\mu$, the inequality is strict if and only if there exists a region C_1. Second, there exists a region C_1 in which $g^a(\overline{\delta})>0>g^b(\overline{\delta})$ if and only if $g^b_2<c$, where g^b_2 is in Eq. (44). We plug $\overline{\delta}$ into Eq. (15) to derive the following trading gains in region C_1,

$$g^a(\overline{\delta}) = g^a_1(X+kY)^2 + g^a_2 - c \tag{41}$$

$$g^a_1 \equiv \frac{2\alpha\sigma^2\mu^2}{(1-k)[(1+\mu)^2 - k(1-\mu)(1+3\mu)]},$$
$$g^a_2 \equiv \frac{1}{2\alpha}\ln\left[1+\frac{4k\mu^2}{(1-k)(1+\mu)^2}\right] \tag{42}$$

$$g^b(\overline{\delta}) = g^b_1(X-kY)^2 + g^b_2 - c \tag{43}$$

$$g^b_1 \equiv \frac{2\alpha\sigma^2}{(1-k)[(1+\mu)^2 + k(1-\mu)(3+\mu)]},$$
$$g^b_2 \equiv \frac{1}{2\alpha}\ln\left[1+\frac{4k}{(1-k)(1+\mu)^2}\right]. \tag{44}$$

To prove this second step, note that if $g^b_2\ge c$, then $g^b(\overline{\delta})\ge 0$ and there does not exist region C_1. If $g^b_2<c$, $g^b(\overline{\delta})<0$ for X close enough to kY, while it is always possible to find sufficiently large $X+kY$ such that $g^a(\overline{\delta})>0$. Hence, there always exists a region C_1. Third, $g^b_2<c$ if and only if $\mu>\underline{\mu}$, defined in Eq. (21). If $\underline{\mu}\in(0,1)$, the result holds because $\underline{\mu}$ solves $g^b_2=c$ and g^b_2 strictly decreases in μ. If $\underline{\mu}=0$, then $g^b_2(0)\le c$. Hence, for any $\mu>\underline{\mu}=0$, $g^b_2<g^b_2(0)\le c$, proving the result. Similarly, if $\underline{\mu}=1$, the result holds because for any $\mu<\underline{\mu}=1$, $g^b_2>g^b_2(1)\ge c$. $\square$

Proof of Proposition 4. If $c_m<\underline{c}_m\equiv\kappa(1)$, then for any $\mu\le 1$, we have $c_m<\kappa(\mu)$ by Lemma 1. Hence, $J^m(\mu,c_m)>J^m(\mu,\kappa(\mu))=J^n(\mu,c)$, where the equality is the definition of $\kappa(\mu)$. Thus, equilibrium is reached only when $\mu=1$. Similarly, if $c_m>\overline{c}_m\equiv\kappa(0)$, we have $c_m>\kappa(\mu)$ and $J^m(\mu,c_m)<J^m(\mu,\kappa(\mu))=J^n(\mu,c)$ for any $\mu\ge 0$, and $\mu=0$ is the unique equilibrium.

If $c_m=\overline{c}_m\equiv\kappa(0)$, we have $\kappa(0)=\kappa(\mu)$ for any $\mu\in[0,\underline{\mu}]$ and $\kappa(0)>\kappa(\mu)$ for any $\mu>\underline{\mu}$ from Lemma 1. At any $\mu>\underline{\mu}$, $J^m(\mu,c_m)=J^m(\mu,\kappa(0))<J^m(\mu,\kappa(\mu))=J^n(\mu,c)$, and μ decreases in equilibrium. At any $\mu\in[0,\underline{\mu}]$, $J^m(\mu,c_m)=J^m(\mu,\kappa(0))=J^m(\mu,\kappa(\mu))=J^n(\mu,c)$. Hence, any $\mu\in[0,\underline{\mu}]$ is an equilibrium. As a special case, if $\underline{\mu}=0$, then $\mu=0$.

If $\underline{c}_m\le c_m<\overline{c}_m$, we can show that $\mu=\kappa^{-1}(c_m)$ is the unique equilibrium. Because $c_m<\overline{c}_m=\kappa(0)=\kappa(\underline{\mu})$, we have $\kappa^{-1}(c_m)>\underline{\mu}$ from Lemma 1. For any $\mu<\kappa^{-1}(c_m)$, we have $\kappa(\mu)>c_m$ and $J^m(\mu,c_m)>J^m(\mu,\kappa(\mu))=J^n(\mu,c)$. Hence, μ increases in equilibrium. Similarly, for any $\mu>\kappa^{-1}(c_m)$, we have $\kappa(\mu)<c_m$ and $J^m(\mu,c_m)<J^m(\mu,\kappa(\mu))=J^n(\mu,c)$ and μ decreases in equilibrium. As a result, $\mu=\kappa^{-1}(c_m)$ is the unique equilibrium.

We now derive the speed of decrease in optimal μ when $\underline{\mu}=0$, especially for small μ. Because $\mu=\kappa^{-1}(c_m)$ in this case, we have $\partial\mu/\partial c_m = 1/(\partial\kappa/\partial\mu)$. From Eq. (40), both the size of region C_1 and the value of $\partial G^b/\partial\mu$ in region C_1 affect $\partial\kappa/\partial\mu$.

We first bound the size of region C_1. From Eq. (21), $\underline{\mu}=0$ requires $g^b_2(0)\le c$. Combining with Eqs. (41)–(44), we have $g^a_2(\mu)\le g^b_2(\mu)\le g^b_2(0)\le c$ for any μ. Define $g^a_3\equiv\sqrt{(c-g^a_2)/g^a_1}$ and $g^b_3\equiv\sqrt{(c-g^b_2)/g^b_1}$. When $X,Y>0$, the condition $g^a(\overline{\delta})>0>g^b(\overline{\delta})$ (for region C_1) requires

$$X>-kY+g^a_3 \quad\text{and}\quad kY-g^b_3<X<kY+g^b_3, \tag{45}$$

which requires $Y>(1/2k)(g^a_3-g^b_3)$. From Eqs. (41) to (44), when $\mu\to 0$, $g^a_3=\mathrm{O}(1/\mu)$ and $g^b_3=\mathrm{O}(1)$. Hence,

$$P_{C_1}\equiv \mathrm{Prob}[X,Y\in C_1] < \mathrm{Prob}[Y>(g^a_3-g^b_3)/(2k)]$$
$$= 1-\Phi[\mathrm{O}(1/\mu)] = \mathrm{O}(e^{-1/\mu^2}) \tag{46}$$

gives the size of region C_1, where $\Phi(\cdot)$ is the cumulative normal density.

Next, we bound the term $\partial G^b/\partial\mu$ in Eq. (40) within region C_1. From the definition of G^i,

$$\frac{\partial G^b}{\partial\mu} = \alpha J^b_{NP} e^{-\alpha g^b(\overline{\delta})} \frac{2(1+\mu)(1-k)}{(1+\mu)^2 + k(1-\mu)(3+\mu)} \times \left[g^b_1(X-kY)^2 + \frac{2k}{\alpha(1-k)(1+\mu)^2} \right]. \tag{47}$$

Because $g^b(\overline{\delta})<0$ in region C_1, from Eqs. (41) to (44), we have $0 \le g^b_1(X-kY)^2 \le c - g^b_2$. Thus, there exists positive constants F_1, F_2 such that $-F_1 < \partial G^b/\partial\mu < -F_2$, and $\mathrm{E}_{C_1}[\partial G^b/\partial\mu] \in (-F_1 P_{C_1}, 0)$. Combining this bound with Eq. (40), we have $\partial\kappa/\partial\mu = -\mathrm{O}(e^{-1/\mu^2})$. Thus, $\partial\mu/\partial c_m = -\mathrm{O}(e^{1/\mu^2})$ for small μ. $\square$

Proof of Proposition 5. From Proposition 4, when $c_m < \overline{c}_m$, the equilibrium for market makers is unique and $\mu > 0$. Taking μ as given, we derive the first-order condition for a market maker, using Eq. (30):

$$\frac{\partial J^m}{\partial \theta^i_0} = \mathrm{E}\left[\frac{1}{2} \left(\frac{\partial J^a_P}{\partial \theta^i_0} + \frac{\partial J^b_P}{\partial \theta^i_0} \right) \middle| c^i = c_m \right], \tag{48}$$

where

$$\frac{\partial J^i_P}{\partial \theta^i_0} = -\alpha J^i_P(-P_0) + \alpha J^i_P D^i_P,$$

$$D^i_P \equiv \frac{\alpha\sigma^2[k\delta^2\theta^i_0 + \delta X + (1 - k\lambda^i\delta + k\delta^2)Y]}{1 - k + k(1-\lambda^i\delta)^2}, \quad i = a, b. \tag{49}$$

Given Proposition 3 and the symmetry between group-a and -b traders, we have

$$\delta(X, Y) = \delta(-X, -Y) = -\delta(X, -Y) = -\delta(-X, Y).$$

At $\theta^i_0 = 0, J^i_P(X, Y) = J^i_P(-X, -Y)$ and $D^i_P(X, Y) = -D^i_P(-X, -Y)$. Thus, $\mathrm{E}[J^i_P D^i_P | \theta^i_0 = 0] = 0$, and Eq. (48) simplifies to $\partial J^m/\partial\theta^i_0|_{\theta^i_0=0} = -\alpha(-P_0)\,\mathrm{E}[(J^a_P + J^b_P)/2 | c^i = c_m]$. Market clearing requires that $\partial J^m/\partial\theta^i_0|_{\theta^i_0=0} = 0$. Hence, $P_0 = 0$ and $\theta^i_0 = 0$ is the unique equilibrium.

When $c_m > \overline{c}_m$, from Proposition 4, $\mu = 0$ is the unique equilibrium. From Proposition 3, we know that the autarky equilibrium for traders with $\omega^a = \omega^b = 0$ is Pareto dominated by the equilibrium with participation. In the positive participation equilibrium, δ is still well defined, and all the above derivation applies. Hence, $P_0 = 0$ and $\theta^i_0 = 0$ is still the equilibrium.

When $c_m = \overline{c}_m$, from Proposition 4, there are multiple equilibria for $\mu \in [0, \underline{\mu})$ when $\underline{\mu} > 0$. From the proof of Lemma 1, we know that $\mu \le \underline{\mu}$ is the necessary and sufficient condition for $g^b_2 \ge c$, which is the necessary and sufficient condition to rule out the existence of region C_1. From Eqs. (37) and (39), we see that the utility for both traders and market makers is independent of μ in the absence of region C_1, which coincides with the above condition for multiple equilibria. Hence, even though there are multiple equilibria for μ when $c_m = \overline{c}_m$, the welfare level remains constant across these equilibria. Similar to the $c_m > \overline{c}_m$ case, if $\mu = 0$, there exist an additional autarky equilibrium, which is Pareto dominated by all the positive participation equilibria.

References

Allen, F., Gale, D., 1994. Limited market participation and volatility of asset prices. American Economic Review 84 (4), 933–955.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

Brennan, M.J., 1975. The optimal number of securities in a risky portfolio when there are fixed costs of transacting: theory and some empirical results. Journal of Financial and Quantitative Analysis 10, 483–496.

Brennan, M.J., Chordia, T., Subrahmanyam, A., 1998. Alternative factor specifications, security characteristics, and the cross-section of expected returns. Journal of Financial Economics 49 (3), 345–373.

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

Brusco, S., Jackson, M.O., 1999. The optimal design of a market. Journal of Economic Theory 88, 1–39.

Campbell, J.Y., Grossman, S.J., Wang, J., 1993. Trading volume and serial correlation in stock returns. Quarterly Journal of Economics 108 (4), 905–939.

Chatterjee, S., Corbae, D., 1992. Endogenous market participation and the general equilibrium value of money. Journal of Political Economy 100 (3), 615–646.

Coval, J., Stafford, E., 2007. Asset fire sales (and purchases) in equity markets. Journal of Financial Economics 86, 479–512.

Detemple, J., Murthy, S., 1994. Intertemporal asset pricing with heterogeneous beliefs. Journal of Economic Theory 62, 294–320.

Diamond, D.W., Verrecchia, R.E., 1981. Information aggregation in a noisy rational expectations economy. Journal of Financial Economics 9, 221–235.

Diamond, P.A., 1982. Aggregate demand management in search equilibrium. Journal of Political Economy 90 (5), 881–894.

Duffie, D., Garleanu, N., Pedersen, L.H., 2005. Over-the-counter markets. Econometrica 73 (6), 1815–1847.

Dumas, B., 1992. Dynamic equilibrium and the real exchange rate in a spatially separated world. Review of Financial Studies 5, 153–180.

Gale, D., 1987. Limit theorems for markets with sequential bargaining. Journal of Economic Theory 43, 20–54.

Glosten, L.R., Milgrom, P., 1985. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 14, 71–100.

Gromb, D., Vayanos, D., 2002. Equilibrium and welfare in markets with financially constrained arbitrageurs. Journal of Financial Economics 66, 361–407.

Grossman, S.J., Miller, M.H., 1988. Liquidity and market structure. Journal of Finance 38, 617–633.

Grossman, S.J., Vila, J.-L., 1992. Optimal investment strategies with leverage constraints. Journal of Financial and Quantitative Analysis 27, 151–168.

Harris, M., Raviv, A., 1993. Differences of opinion make a horse race. Review of Financial Studies 6 (3), 473–506.

Hirshleifer, D., 1988. Residual risk, trading costs, and commodity futures risk premia. Review of Financial Studies 1 (2), 173–193.

Ho, T., Stoll, H.R., 1980. On dealer markets under competition. Journal of Finance 35 (2), 259–267.

Huang, J., Wang, J., 2009. Liquidity and market crashes. Review of Financial Studies 22 (7), 2607–2643.

Kimball, M.S., 1993. Standard risk aversion. Econometrica 61 (3), 589–611.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1336.

Leland, H., Rubinstein, M., 1988. Comments on the market crash: six months after. The Journal of Economic Perspectives 2 (3), 45–50.

Lo, A., Mamaysky, H., Wang, J., 2004. Asset prices and trading volume under fixed transactions costs. Journal of Political Economy 112 (5), 1054–1090.

Merton, R.C., 1987. A simple model of capital market equilibrium with incomplete information. Journal of Finance 42 (3), 483–510.

Mitchell, M., Pedersen, L.H., Pulvino, T., 2007. Slow moving capital. American Economic Review 97, 215–220.

Pagano, M., 1989. Endogenous market thinness and stock price volatility. Review of Economic Studies 56 (2), 269–288.
Rubinstein, A., Wolinsky, A., 1987. Middlemen. Quarterly Journal of Economics 102, 581–594.
Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. Journal of Finance 52, 35–55.
Stoll, H.R., 1985. Market making and the changing structure of the securities industry. In: Amihud, Y., Ho, T., Schwartz, R. (Eds.), Alternative Views of Market MakingLexington Books chap, pp. 67–92.
Sundaresan, S., Wang, Z., 2009. Y2K options and the liquidity premium in treasury markets. Review of Financial Studies 22, 1021–1056.
Vayanos, D., Wang, T., 2007. Search and endogenous concentration of liquidity in asset markets. Journal of Economic Theory 136 (1), 66–104.
Wang, J., 1994. A model of competitive stock trading volume. Journal of Political Economy 102, 127–168.
Wang, J., 1996. The term structure of interest rates in a pure exchange economy with heterogeneous investors. Journal of Financial Economics 41 (1), 75–110.