NBER WORKING PAPER SERIES

MARKET LIQUIDITY — THEORY AND EMPIRICAL EVIDENCE

Dimitri Vayanos
Jiang Wang

Working Paper 18251
http://www.nber.org/papers/w18251

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2012

Forthcoming in the Handbook of the Economics of Finance, edited by George Constantinides, Milton Harris, and Rene Stulz. We thank Bruno Biais, Joost Driessen, Denis Gromb, Terrence Hendershott and Ronnie Sadka for very helpful comments. Financial support from the Paul Woolley Centre at the LSE is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Market Liquidity — Theory and Empirical Evidence
Dimitri Vayanos and Jiang Wang
NBER Working Paper No. 18251
July 2012
JEL No. D42,D53,D82,D83,G01,G11,G12,G14

ABSTRACT

In this paper we survey the theoretical and empirical literature on market liquidity. We organize both literatures around three basic questions: (a) how to measure illiquidity, (b) how illiquidity relates to underlying market imperfections and other asset characteristics, and (c) how illiquidity affects expected asset returns. Using a unified model from Vayanos and Wang (2010), we survey theoretical work on six main imperfections: participation costs, transaction costs, asymmetric information, imperfect competition, funding constraints, and search---and for each imperfection we address the three basic questions within that model. We review the empirical literature through the lens of the theory, using the theory to both interpret existing results and suggest new tests and analysis.

Dimitri Vayanos
Department of Finance, OLD 3.41
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM
and CEPR
and also NBER
d.vayanos@lse.ac.uk

Jiang Wang
MIT Sloan School of Management
100 Main Street, E62-614
Cambridge, MA 02142
and NBER
wangj@mit.edu

1 Introduction

Financial markets deviate, to varying degrees, from the perfect-market ideal in which there are no impediments to trade. Trade impediments reduce the liquidity that markets offer. A large and growing literature studies market liquidity and its properties. Theoretical papers trace illiquidity, i.e., the lack of liquidity, to underlying market imperfections such as asymmetric information, different forms of trading costs, and funding constraints. They also study how imperfections affect expected asset returns. Empirical papers estimate measures of illiquidity, some of which are derived from theoretical models, and relate them to asset characteristics and asset returns.

In this paper we survey the theoretical and empirical literature on market liquidity. We organize the survey around three basic questions: (a) how to measure illiquidity, (b) how illiquidity relates to underlying market imperfections and other asset characteristics, and (c) how illiquidity affects expected asset returns. We use these three questions as a common framework to organize both the theoretical and the empirical literature.

Theoretical papers study a variety of market imperfections, relying on different modeling assumptions. For example, papers on trading costs typically assume life-cycle or risk-sharing motives to trade, while papers on asymmetric information often rely on noise traders. Some papers on asymmetric information further assume risk-neutral market makers who can take unlimited positions, while papers on other imperfections typically assume risk aversion or position limits. Instead of surveying this literature in a descriptive manner, we use a common, unified model from Vayanos and Wang (2010) to study all the market imperfections that we consider, and for each imperfection we address the three basic questions within that model. Our model generates many of the key results shown in the literature, and serves as a point of reference for surveying other results derived in different or more complicated settings. Using a common model to study all imperfections makes the presentation more streamlined and pedagogical. It also helps better understand the effects of the imperfections since these can be compared holding constant other assumptions such as trading motives and risk attitudes.

Empirical papers often do not take a view as to the specific market imperfection that causes illiquidity. Some of the measures of illiquidity that they compute are derived in theoretical models, while other measures are more intuitive or heuristic. We survey the empirical literature through the lens of the theory, using the theory to both interpret existing results and suggest new tests and analysis. Our survey of the empirical literature is more limited in scope than of the theoretical literature: we goal is not to be comprehensive but to link the main empirical methodologies and findings with the theory.

We survey the theoretical literature in Section 2. Our model has three periods, $t = 0, 1, 2$. In Periods 0 and 1, risk-averse agents can trade a riskless and a risky asset that pay off in Period 2. In Period 0, agents are identical so no trade occurs. In Period 1, agents can be one of two types. Liquidity demanders receive an endowment correlated with the risky asset's payoff, and need to trade to share risk. They can trade with liquidity suppliers, who receive no endowment. Agents learn whether or not they will receive the endowment in an interim period $t = 1/2$. While we model heterogeneity through endowments, our analysis would be similar for other types of heterogeneity, e.g., different beliefs or investment opportunities. Market imperfections concern trade in Period 1. We determine how the imperfections affect measures of illiquidity in Period 1. We also determine the effect of the imperfections on ex-ante expected returns as of Period 0: how does the expected return that agents require to buy the risky asset in Period 0 depends on the imperfections that they anticipate to face in Period 1? We consider six imperfections, studied extensively in the theoretical literature:

1. **Participation costs:** In the perfect-market benchmark, all agents are present in the market in all periods. Thus, a seller, for example, can have immediate access to the entire population of buyers. In practice, however, agents face costs of market participation, e.g., monitor market movements and be ready to trade. To model costly participation, we assume that agents must incur a cost to trade in Period 1. Consistent with the notion that participation is an ex-ante decision, we assume that agents must decide whether or not to incur the cost in Period 1/2, i.e., after learning whether or not they will receive an endowment but before observing the price in Period 1. A related imperfection is that of entry costs, e.g., learning about an asset. The cost would then concern buying the asset in Period 0.

2. **Transaction costs:** In addition to costs of market participation, agents typically pay costs when executing transactions. Transaction costs drive a wedge between the buying and selling price of an asset. They come in many types, e.g., brokerage commissions, transaction taxes, and bid-ask spreads. Some types of transaction costs can be viewed as a consequence of other market imperfections, while other types, such as transaction taxes, can be viewed as more primitive. We assume that transaction costs concern trade in Period 1. The difference with participation costs is that the decision whether or not to incur the transaction costs is contingent on the price in Period 1.

3. **Asymmetric information:** In the perfect-market benchmark, all agents have the same information about the payoff of the risky asset. In practice, however, agents can have different information because they have access to different sources of information or have

different abilities to process information from the same source. To model asymmetric information, we assume that some agents observe in Period 1 a private signal about the asset payoff. We assume that these agents are the liquidity demanders. This assumption, which in our model is without loss of generality, allows us to determine how the supply of liquidity is influenced by suppliers' concern about trading against better-informed agents.

4. **Imperfect competition:** In the perfect-market benchmark, agents are competitive and have no effect on prices. In many markets, however, some agents are large relative to others in the sense that they can influence prices, either because of their size or because of their information advantage. We model imperfect competition by assuming that some agents can exert market power in Period 1. We focus on the case where liquidity demanders behave as a single monopolist, but also consider more briefly monopolistic behavior by liquidity suppliers. We consider both the case where liquidity demanders have no private information on asset payoffs, and so information is symmetric, and the case where they observe a private signal.

5. **Funding constraints:** Agents' portfolios often involve leverage, i.e., borrow cash to establish a long position in a risky asset, or borrow a risky asset to sell it short. In the perfect-market benchmark, agents can borrow freely provided that they have enough resources to repay the loan. But as the Corporate Finance literature emphasizes, various frictions can limit agents' ability to borrow and fund their positions. We derive a funding constraint by assuming that agents cannot pledge some of their future income. Because our focus is on how the funding constraint influences the supply of liquidity, we impose it on liquidity suppliers only, i.e., assume that the lack of commitment concerns only them.

6. **Search:** In the perfect-market benchmark, the market is organized as a centralized exchange. Many markets, however, have a more decentralized form of organization. For example, in over-the-counter markets, investors negotiate prices bilaterally with dealers. Locating suitable counter-parties in these markets can take time and involve search. To model decentralized markets, we assume that agents do not meet in a centralized exchange in Period 1, but instead must search for counterparties. When a liquidity demander meets a supplier, they bargain bilaterally over the terms of trade.

We consider two measures of illiquidity. The first measure is lambda, defined as the regression coefficient of the return between Periods 0 and 1 on liquidity demanders' signed volume in Period 1. This measure characterizes the price impact of volume, which has a transitory and a permanent component. The second measure is price reversal, defined as minus the autocovariance of returns. This measure characterizes the importance of the transitory component in

price, which in our model is entirely driven by volume. Lambda and price reversal have been derived in theoretical models focusing on specific market imperfections, and have been widely used in empirical work ever since.

We determine how each imperfection affects lambda, price reversal, and ex-ante expected returns. Many of these effects have been derived in the literature, albeit in a less systematic and unified manner. We highlight the links with the literature, and use more generally our model to organize and survey it. Many models in the literature can be viewed as enrichments of our model in terms of, e.g., information structure, agent characteristics, and dynamics.

Deriving the effects of the imperfections within a unified model delivers new insights. We show, for example, that most imperfections raise lambda, but fewer raise price reversal. Thus, lambda is a more accurate measure of the imperfections. Intuitively, lambda measures the price impact per unit trade, while price reversal concerns the impact of the entire trade. Market imperfections generally raise the price impact per unit trade, but because they also reduce trade size, the price impact of the entire trade can decrease. We show additionally that imperfections do not always raise expected returns. The literature has shown this result for some imperfections; we examine its validity across all imperfections and identify those under which it is more likely to hold.

We survey the empirical literature in Section 3. We begin by presenting various measures of illiquidity. Some measures, such as lambda and price reversal, are motivated by theory, while others, such as bid-ask spread, market depth, turnover and trade size, are more intuitive or heuristic. We also discuss ways to measure the illiquidity of an aggregate asset class rather than of a single asset. We next examine how the measures of illiquidity vary across assets and over time, how they relate to asset characteristics such as supply and volatility, and how they relate to each other. We finally examine the link between illiquidity measures and expected asset returns. Some of the work in that area links the level of illiquidity to expected returns. More recent work also allows for the possibility that illiquidity is a priced risk factor, which affects expected returns through its comovement with returns.

Throughout our survey of the empirical literature, we emphasize the links with the theory. For example, how closely do empirical measures of illiquidity reflect the underlying market imperfections? How can the theory account for the variation of the illiquidity measures across assets and over time? How can it account for the relationships between the illiquidity measures and asset characteristics or expected returns? As we hope to demonstrate, the theory can shed new light on existing empirical results and suggest new tests and analysis. For example, the theory implies that the effectiveness of a particular measure of illiquidity, in terms of reflecting

the underlying market imperfection, depends on the imperfection. This could help explain why some measures are more successful than others in capturing illiquidity and its relationship with expected asset returns in some markets. Using those measures, and controlling for additional factors suggested by theory, could yield sharper empirical tests.

Liquidity effects can manifest themselves over different time horizons. The market microstructure literature focuses on short horizons, from minutes or hours to days or weeks. At the same time, recent work on the limits of arbitrage finds that flows can affect returns even at the longer horizons used in asset-pricing analysis, e.g., months, quarters or years. We view both horizons as relevant for the purposes of our survey—provided that the price movements under consideration are temporary departures from fundamental value caused by flows. Our model can accommodate both horizons simply by changing the length of a "period." The horizon that we mostly have in mind, and emphasize in the empirical part of this survey, is that of days, weeks or months. We do not rule out longer horizons, although it can be more difficult to identify transitory price movements at such horizons. On the other hand, our model is not well-suited for very short horizons of minutes or hours because we intentionally omit the level of market detail that is required to study the corresponding effects.

Our survey does not cover some important issues, either because they represent open questions on which research so far has been limited, or because covering them would detract from our main focus. Nevertheless, it is important to recognize these issues, both to put our survey in perspective and to outline promising areas for future research. A first issue, mentioned above, is the horizon of liquidity effects. In our model horizon is exogenous and corresponds to the length of a period. It should be derived endogenously, however, through a more detailed description of market imperfections and agents' trading needs.

A second issue concerns the interactions between market imperfections. Most of the theoretical literature considers one imperfection at a time and thus does not allow for interactions. Our model also does not cover interactions, except between imperfect competition and asymmetric information. Other interactions, such as between funding constraints and asymmetric information, are interesting and have received some attention in the literature.

A related but more fundamental issue concerns the underlying economic causes of the imperfections and the ways in which imperfections are linked. Following much of the literature, we treat each imperfection as primitive. Yet, some imperfections could be the consequence of other more fundamental ones. For example, some types of transaction costs can be viewed as a consequence of other imperfections, such as participation costs or asymmetric information. Moreover, if, for example, participation costs are costs to monitor market information, then

costly participation could be linked to asymmetric information. Asymmetric information could also underlie the contracting frictions that give rise to funding constraints. Endogenizing some market imperfections from more fundamental frictions could further streamline, clarify and deepen the study of market liquidity. In particular, various forms of informational problems could be the underlying economic cause for various forms of imperfections.

A fourth issue concerns the institutional context. A large fraction of trading activity in financial markets is generated by specialized financial institutions, and these institutions can be important suppliers or demanders of liquidity. Following much of the literature, we model instead liquidity suppliers and demanders as individuals, thus ignoring contracting frictions and other institutional complexities. (We only consider such frictions briefly in the context of funding constraints.) The liquidity shock in our model could result from institutional frictions, but only in reduced form. The importance of financial institutions in affecting asset prices is emphasized in a rapidly growing literature on the limits of arbitrage.

Related to the institutional context is the issue of market design. While we consider ways in which markets deviate from the Walrasian ideal, we do not study market design in depth. The market microstructure literature studies various dimensions of market design and shows that they can affect market performance. Such dimensions include whether liquidity is supplied by dedicated market makers or an open limit-order book, whether limit orders are visible to all traders, whether transactions are disclosed to all traders after they are executed, etc. While we survey some of that work, we conduct our analysis at a more aggregate level with less market detail, so that we can derive some key effects within a tractable unified model. The downside is that our model is not well suited for very short horizons such as seconds or minutes, e.g., to compute empirical measures of market depth and resiliency at such horizons. Our model is also not well suited for addressing the benefits of different market designs.

Finally, we do not perform any analysis of welfare or policy (even though our model could be used for that purpose as well). For example, we do not examine how imperfections affect the welfare of different agents and what policy actions could mitigate these effects. We survey, however, some papers that consider welfare issues.

Our survey is related to both market microstructure and asset pricing. We emphasize fundamental market imperfections covered in the market microstructure literature, but abstract away from the level of market detail often adopted in that literature. At the same time, we study how market imperfections affect expected asset returns—an asset-pricing exercise. Surveys with greater focus on market microstructure include the book by O'Hara (1995) for the theory, the article by Hasbrouck (2007) for the empirics, and the articles by Madhavan

(2000), Biais, Glosten, and Spatt (2005), and Parlour and Seppi (2008) for both theory and empirics. Amihud, Mendelson, and Pedersen (2005) survey theoretical and empirical work on market liquidity and asset-pricing effects. They mainly focus on transaction costs and not on the other market imperfections that we consider. Gromb and Vayanos (2010a) survey the theoretical literature on the limits of arbitrage.

2 Theory

We organize our discussion of the theoretical literature on market liquidity using the unified model of Vayanos and Wang (2010, 2011). The model nests a variety of market imperfections studied in the literature and shows how they relate to empirical measures of liquidity. It also shows how market imperfections affect ex-ante expected returns.

There are three periods, $t = 0, 1, 2$. The financial market consists of a riskless and a risky asset that pay off in Period 2. The riskless asset is in supply of B shares and pays off one unit of a consumption good per share with certainty. The risky asset is in supply of $\bar{\theta}$ shares and pays off D units per share, where D has mean $\bar{D}$ and variance σ^2. Using the riskless asset as the numeraire, we denote by S_t the risky asset's price in Period t, where $S_2 = D$.

There is a measure one of agents, who derive utility from consumption in Period 2. Utility is exponential,

$$-\exp(-\alpha C_2),$$

where C_2 is consumption in Period 2, and $\alpha > 0$ is the coefficient of absolute risk aversion. We denote agents' wealth in Period t by W_t. Wealth in Period 2 is equal to consumption, i.e., $W_2 = C_2$. Agents are identical in Period 0, but become heterogeneous and trade in Period 1. Because all agents have the same exponential utility, there is no preference heterogeneity. We instead introduce heterogeneity through agents' endowments.

A fraction π of agents receive an endowment of the consumption good in Period 2, and the remaining fraction $1 - \pi$ receive no endowment. The endowment is $z(D - \bar{D})$ per receiving agent, where z has mean zero and variance σ_z^2 and is independent of D. Since the endowment is correlated with D, it generates a hedging demand. When, for example, $z > 0$, the endowment exposes agents to the risk that D will be low, and agents hedge against that risk by selling the risky asset. Agents learn whether or not they will receive the endowment in an interim period $t = 1/2$, and those who will receive the endowment observe z in Period 1. Thus, agents learn whether or not they will need to trade before learning the exact size of their desired trade.

We assume that the endowment is perfectly correlated with D for simplicity; what matters for our analysis is that the correlation is non-zero. For tractability, we assume that D and z are normal. While we model heterogeneity through endowments, our analysis would be similar for other types of heterogeneity, e.g., different beliefs or investment opportunities.

In equilibrium, agents receiving an endowment initiate trades with others to share risk. Because the agents initiating trades can be thought of as consuming market liquidity, we refer to them as liquidity demanders and denote them by the subscript d. Moreover, we refer to z as the liquidity shock. The agents who receive no endowment accommodate the trades of liquidity demanders, and hence supply liquidity. We refer to them as liquidity suppliers and denote them by the subscript s.

Because liquidity suppliers require compensation to absorb risk, the trades of liquidity demanders affect prices. Therefore, the price in Period 1 is influenced not only by the asset payoff, but also by the liquidity demanders' trades. Our measures of liquidity, defined in Section 2.1, are based on the price impact of these trades.

The assumptions introduced so far describe our model's perfect-market benchmark, to which we subsequently add market imperfections. We maintain the perfect-market assumption in Period 0 when determining the ex-ante effect of the imperfections, i.e., how the anticipation of imperfections in Period 1 impacts the price in Period 0. Imperfections in Period 0 are, in fact, not relevant in our model because agents are identical in that period and there is no trade.

We can give two interpretations to our model. Under the first interpretation, the set of agents is the entire set of households in an economy. The only liquidity shocks that can then have non-trivial price impact are those large enough to be comparable to the size of the economy. Under the second interpretation, the set of agents is the subset of households who participate in a specific market. Liquidity shocks can then have non-trivial price impact even when they are small relative to the size of the economy; all that is needed is that they are comparable to the size of the set of agents participating in that market. That set can be smaller than the entire set of households in the economy because of participation costs. While we consider participation costs as a market imperfection (Section 2.2), they can be viewed as implicit in the perfect-market benchmark under the second interpretation.[1]

[1]An additional imperfection that is implicit in the perfect-market benchmark is that agents cannot write contracts in Period 0 contingent on whether they are a liquidity demander or supplier in Period 1. Thus, the market in Period 0 is incomplete in the Arrow-Debreu sense. If agents could write complete contracts in Period 0, they would not need to trade in Period 1, in which case liquidity would not matter. Complete contracts are infeasible in our model because whether an agent is a liquidity demander or supplier is private information.

2.1 Perfect-Market Benchmark

In this section we describe the equilibrium in our model's perfect-market benchmark. Detailed derivations and proofs of the results in this and subsequent sections are in Vayanos and Wang (2010, 2011). Agents' demand functions for the risky asset in Period 1 are

$$\theta_1^s = \frac{\bar{D} - S_1}{\alpha\sigma^2}, \tag{1a}$$

$$\theta_1^d = \frac{\bar{D} - S_1}{\alpha\sigma^2} - z. \tag{1b}$$

Liquidity suppliers are willing to buy the risky asset as long as its price S_1 in Period 1 is below the expected payoff $\bar{D}$, and are willing to sell otherwise. Liquidity demanders have a similar price-elastic demand function, but are influenced by the liquidity shock z. When z is positive, for example, liquidity demanders are willing to sell because their endowment is positively correlated with the asset.

Market clearing requires that the aggregate demand equals the asset supply $\bar{\theta}$:

$$(1-\pi)\theta_1^s + \pi\theta_1^d = \bar{\theta}. \tag{2}$$

Substituting (1a) and (1b) into (2), we find

$$S_1 = \bar{D} - \alpha\sigma^2\left(\bar{\theta} + \pi z\right). \tag{3}$$

The price S_1 decreases in the liquidity shock z. When z is positive, for example, liquidity demanders are willing to sell, and the price must drop so that the risk-averse liquidity suppliers are willing to buy.

Agents choose holdings θ_0 of the risky asset in Period 0 to maximize the ex-ante expected utility

$$U \equiv (1-\pi)U^s + \pi U^d,$$

where U^s denotes the interim utility of becoming a liquidity supplier in Period 1/2, and U^d that of becoming a liquidity demander. The solution to this maximization problem coincides with the aggregate demand in Period 0, since all agents are identical in that period and are in measure one. In equilibrium, aggregate demand has to equal the asset supply $\bar{\theta}$, and this determines the price S_0 in Period 0. The price is

$$S_0 = \bar{D} - \alpha\sigma^2\bar{\theta} - \frac{\pi M}{1-\pi+\pi M}\Delta_1\bar{\theta}, \tag{4}$$

where (M, Δ_1) are positive constants. The first term in (4) is the asset's expected payoff in Period 2, the second term is a discount arising because the payoff is risky, and the third term is a discount due to illiquidity (i.e., low liquidity). The risk discount is independent of the parameter σ_z^2 that measures the size of liquidity shocks, and is non-zero even when there are no shocks, i.e., $\sigma_z^2 = 0$. The illiquidity discount is instead increasing in σ_z^2, and is zero when $\sigma_z^2 = 0$.

The illiquidity discount is the product of two terms. The first term, $\frac{\pi M}{1-\pi+\pi M}$, can be interpreted as the risk-neutral probability of being a liquidity demander: π is the true probability, and M is the ratio of marginal utilities of wealth of demanders and suppliers, where utilities are interim in Period 1/2. The second term, $\Delta_1 \bar{\theta}$, is the discount that an agent would require conditional on being a demander. Since the illiquidity discount lowers the asset price in Period 0, it raises the expected return

$$E(R) = \bar{D} - S_0$$

from buying the asset in Period 0 and holding it until it pays off in Period 2. From now on, we refer to $E(R)$ simply as the asset's expected return.

Intuitively, the illiquidity discount arises because agents are concerned in Period 0 that the endowment they might receive in Period 1 will increase their existing risk exposure. Therefore, if they are less able to hedge in Period 1, they are less willing to take risk in Period 0, and they require a larger discount to buy the asset. A discount arises even in the perfect-market benchmark because while risk sharing is optimal, hedging affects the price. Market imperfections affect the discount because they affect agents' ability to hedge.

We construct two measures of illiquidity in Period 1, both based on the price impact of the liquidity demanders' trades. The first measure, to which we refer as lambda or price impact, is the coefficient of a regression of the asset's return, $S_1 - S_0$, between Periods 0 and 1 on the signed volume, $\pi(\theta_1^d - \bar{\theta})$, of liquidity demanders in Period 1:

$$\lambda \equiv \frac{\text{Cov}\left[S_1 - S_0, \pi(\theta_1^d - \bar{\theta})\right]}{\text{Var}\left[\pi(\theta_1^d - \bar{\theta})\right]}. \tag{5}$$

Intuitively, when λ is large, trades have large price impact and the market is illiquid. Eq. (3) implies that the return between Periods 0 and 1 is

$$S_1 - S_0 = \bar{D} - \alpha\sigma^2\left(\bar{\theta} + \pi z\right) - S_0. \tag{6}$$

Eqs. (1b) and (3) imply that the signed volume of liquidity demanders is

$$\pi(\theta_1^d - \bar{\theta}) = -\pi(1-\pi)z. \tag{7}$$

Eqs. (5)-(7) imply that

$$\lambda = \frac{\alpha\sigma^2}{1-\pi}. \tag{8}$$

Price impact λ is higher when agents are more risk-averse (α large), the asset is riskier (σ^2 large), or there are fewer liquidity suppliers ($1-\pi$ small).

Since the signed volume of liquidity demanders is minus that of liquidity suppliers, λ is also minus the regression coefficient of the return between Periods 0 and 1 on suppliers' signed volume in Period 1:

$$\lambda = -\frac{\text{Cov}\left[S_1 - S_0, (1-\pi)(\theta_1^s - \bar{\theta})\right]}{\text{Var}\left[(1-\pi)(\theta_1^s - \bar{\theta})\right]}. \tag{9}$$

The supplier-based definition of λ can be easier to implement empirically than the equivalent demander-based definition. This is because an important class of suppliers in some markets are designated market makers, and information on their trades is often available.

The second measure of illiquidity is based on the autocovariance of returns. The liquidity demanders' trades in Period 1 cause the price to deviate from fundamental value, while the two coincide in Period 2. Therefore, returns exhibit negative autocovariance, and more so when trades have large price impact. We use minus autocovariance

$$\gamma \equiv -\text{Cov}\left(S_2 - S_1, S_1 - S_0\right), \tag{10}$$

as a measure of illiquidity, and refer to it as price reversal. Eqs. (3), (6), (10) and $S_2 = D$ imply that

$$\gamma = -\text{Cov}\left[D - \bar{D} + \alpha\sigma^2\left(\bar{\theta} + \pi z\right), \bar{D} - \alpha\sigma^2\left(\bar{\theta} + \pi z\right) - S_0\right] = \alpha^2\sigma^4\sigma_z^2\pi^2. \tag{11}$$

Price reversal γ is higher when agents are more risk-averse, the asset is riskier, there are more liquidity demanders (π large), or liquidity shocks are larger (σ_z^2 large).

The measures λ and γ have been defined in models focusing on specific market imperfections, and have been widely used in empirical work ever since. We emphasize their basic properties below, leaving a more detailed discussion of the measures and their empirical estimation to Section 3.

Kyle (1985) defines λ in a model where an informed insider trades with uninformed market makers and noise traders. In Kyle, market makers are risk neutral, and trades affect prices only because they contain information. Thus, the price impact, as measured by λ, reflects the amount of information that trades convey, and is permanent because the risk-neutral market makers set the price equal to their expectation of fundamental value. In general, as in our model, λ has both a transitory and a permanent component. The transitory component, present even in our perfect-market benchmark, arises because liquidity suppliers are risk averse and require a price movement away from fundamental value to absorb a liquidity shock. The permanent component arises only when information is asymmetric (Sections 2.4 and 2.5) for the same reasons as in Kyle.

Roll (1984) links γ to the bid-ask spread, in a model where market orders cause the price to bounce between the bid and the ask. Grossman and Miller (1988) link γ to the price impact of liquidity shocks, in a model where risk-averse liquidity suppliers must incur a cost to participate in the market. In both models, price impact is purely transitory because information is symmetric. In our model, price impact has both a transitory and a permanent component, and γ isolates the effects of the transitory component. Note that besides being a measure of imperfections, γ provides a useful characterization of price dynamics: it measures the importance of the transitory component in price arising from temporary liquidity shocks, relative to the random-walk component arising from fundamentals.

2.2 Participation Costs

In the perfect-market benchmark, all agents are present in the market in all periods. Thus, a seller, for example, can have immediate access to the entire population of buyers. In practice, however, agents face costs of market participation. Such costs include buying trading infrastructure or membership of a financial exchange, having capital available on short notice, monitoring market movements, etc. To model costly participation, we assume that agents must incur a cost c to trade in Period 1. Consistent with the notion that participation is an ex-ante decision, we assume that agents must decide whether or not to incur c in Period 1/2, after learning whether or not they will receive an endowment but before observing the price in Period 1. (The price depends on the liquidity shock, which is observed only in Period 1.) If the decision can be made contingent on the price in Period 1, then c is a fixed transaction cost rather than a participation cost. We consider transaction costs as a separate market imperfection.

We first describe the equilibrium, then examine how participation costs affect the illiquidity measures and the expected return, and finally survey the literature on participation costs. We

look for an equilibrium where all liquidity demanders participate, but only a fraction $\mu > 0$ of liquidity suppliers do. Such an equilibrium exists under two sufficient conditions, which we assume from now on. First, the participation cost c is below a threshold $\hat{c}$, so that liquidity suppliers are willing to participate. Second, the measure of demanders is less than that of suppliers, i.e., $\pi < 1/2$. Intuitively, when demanders are the short side of the market, they stand to gain more from participation, and can therefore cover the participation cost if suppliers do.[2] We focus on that case since we are interested in examining how participation costs can affect the supply of liquidity.

Market clearing requires that the aggregate demand of participating agents equals the asset supply held by these agents. Since in equilibrium agents enter Period 1 holding $\bar{\theta}$ shares of the risky asset, market clearing takes the form

$$(1-\pi)\mu\theta_1^s + \pi\theta_1^d = [(1-\pi)\mu + \pi]\,\bar{\theta}. \tag{12}$$

Agents' demand functions are as in Section 2.1. Substituting (1a) and (1b) into (12), we find that the price in Period 1 is

$$S_1 = \bar{D} - \alpha\sigma^2 \left[\bar{\theta} + \frac{\pi}{(1-\pi)\mu + \pi} z\right]. \tag{13}$$

We derive the participation decision of liquidity suppliers by comparing the interim utility U^s of a participating supplier in Period 1/2 to the interim utility U^{sn} of a non-participating supplier. If the participation cost c is below a threshold $\underline{c}$, then all suppliers participate ($U^s > U^{sn}$ and $\mu = 1$). If c is above $\underline{c}$ and below a larger threshold $\bar{c}$, then suppliers are indifferent between participating or not ($U^s = U^{sn}$), and only some participate ($0 < \mu < 1$). Increasing c within that region reduces the fraction μ of participating suppliers, while maintaining the indifference condition. This is because with fewer participating suppliers, competition becomes less intense, enabling the remaining suppliers to cover their increased participation cost. Finally, if c is above $\bar{c}$, then no suppliers participate ($U^s < U^{sn}$ and $\mu = 0$). The price in Period 0 is given by the same equation (4) as in the perfect-market benchmark, but for different constants (M, Δ_1).

We next examine how participation costs impact the illiquidity measures and the expected

[2]For $c \leq \hat{c}$ and $\pi \leq 1/2$, only two equilibria exist: the one described above and the one where no agent participates. The same is true for π larger but close to 1/2.

return. Proceeding as in Section 2.1, we can compute price impact λ and price reversal γ:

$$\lambda = \frac{\alpha\sigma^2}{(1-\pi)\mu}, \tag{14}$$

$$\gamma = \frac{\alpha^2\sigma^4\sigma_z^2\pi^2}{[(1-\pi)\mu+\pi]^2}. \tag{15}$$

Both measures are inversely related to the fraction μ of participating liquidity suppliers. We can also show that the illiquidity discount is inversely related to μ.

We derive comparative statics with respect to c, and consider only the region $c > \underline{c}$, where the measure μ of participating suppliers is less than one. This is without loss of generality: in the region $c \leq \underline{c}$, where all suppliers participate, prices are not affected by the participation cost and are as in the perfect-market benchmark. When $c > \underline{c}$, an increase in the participation cost lowers μ, and therefore raises price impact, price reversal and the illiquidity discount. Since the illiquidity discount increases, so does the asset's expected return.

Result 1 *An increase in the participation cost c raises price impact λ, price reversal γ, and the asset's expected return $E(R)$.*

The idea that participation in financial markets is costly and hence limited dates back to Demsetz (1968). Demsetz (1968) studies the provision of immediacy, i.e., immediate execution of trades. He argues that supplying immediacy is costly but there is a demand for it. Because of the costs of supplying immediacy, only a subset of agents will choose to supply it, and they will be compensated from the price concessions they will earn from the demanders of immediacy. Demsetz (1968) identifies the suppliers of immediacy with market makers, and their compensation with the bid-ask spread.

A subsequent literature models price formation in the presence of market makers. In most of that literature, market makers are assumed to be the only suppliers of immediacy and to receive an exogenous flow of orders from the demanders of immediacy. The literature determines the bid-ask spreads chosen by market makers as a function of the process of order arrival, the degree of competition between market makers, and the inventory and risk aversion of market makers. Examples are Garman (1976), Stoll (1978b), Amihud and Mendelson (1980), Ho and Stoll (1980, 1981, 1983), Cohen, Maier, Schwartz, and Whitcomb (1981) and Mildenstein and Schleef (1983). Because of the focus on market makers' inventory, that literature is often referred to as the inventory literature.

Most of the inventory literature takes the market structure as exogenous, e.g., assumes

an exogenous number of market makers. One exception is Stoll (1978b), who endogenizes the number of market makers in the spirit of Demsetz (1968), taking the costs of supplying immediacy to be fixed costs of processing orders. Grossman and Miller (1988) perform a similar exercise, but emphasize more explicitly costs of market participation. Their setting is closely related to ours: a number of liquidity suppliers choose whether or not to participate in a market, and those choosing to participate pay a cost and can absorb an uncertain and exogenous order by liquidity demanders. The analysis of Grossman and Miller (1988) is closely related to the equilibrium in Periods 1/2 and 1 in our model. Grossman and Miller (1988) emphasize additionally that the bid-ask spread has drawbacks as a measure of liquidity, and suggest the use of price reversal instead. They show that price reversal increases in participation costs, consistent with our Result 1.

Grossman and Miller (1988) do not derive the effect of participation costs on ex-ante expected returns because they do not introduce our Period 0. They introduce, however, two periods after our Period 1: a Period 2 in which an offsetting liquidity shock arrives in the market, and a Period 3 in which the asset pays off. This captures the important idea that a liquidity shock experienced by some agents is absorbed first by a small set of market participants (the liquidity suppliers in Period 1) and then by a larger set of agents who gradually arrive in the market in response to the shock (the traders with the offsetting liquidity shock in Period 2). The idea that some agents arrive gradually into a market in response to profitable opportunities has received attention recently in the search literature reviewed in Section 2.7. Duffie (2010) and Suominen and Rinne (2011) model a similar effect in a centralized market. They assume that some agents arrive into a market infrequently with liquidity shocks. These shocks are absorbed by market makers present in the market in all periods, and by other infrequent traders arriving in future periods who can trade with market makers.

Huang and Wang (2009) study how participation costs affect both the demand for immediacy, which Grossman and Miller (1988) treat as exogenous, and the supply. They assume that liquidity shocks are opposite across agents and so do not affect the price in the absence of participation costs. Participation costs lower the price because sellers are more willing to participate than buyers. The intuition is that sellers receive a larger risky endowment, and are hence more concerned about the risk that an additional shock will leave them with a large risk exposure. This effect of participation costs on ex-ante expected returns is closely related to the one that we derive in Period 0 of our model. Huang and Wang (2010) employ a similar framework as in Huang and Wang (2009) to study welfare questions. They show, in particular, that the market can provide less liquidity than the social optimum.

The costs of market participation in our model concern Period 1, which is after agents have bought the asset. Costs to participate in the market in Period 0 and to buy the asset can be interpreted as entry costs, e.g., learning about the asset. Goldsmith (1976), Mayshar (1979) and Merton (1987) show that entry costs induce agents to under-invest and under-diversify, and typically reduce asset prices. Entry costs would have a similar effect in our model: they would render agents less willing to buy the asset in Period 0, and hence would lower the Period 0 price. Mankiw and Zeldes (1991) conjecture that limited investor participation in the stock market can render stocks cheaper relative to bonds, explaining the equity premium puzzle of Mehra and Prescott (1985). Basak and Cuoco (1998) show that when some investors cannot participate in the stock market, stocks' expected excess returns relative to bonds increase, and interest rates decrease.

Pagano (1989a) and Allen and Gale (1994) show that entry costs can result in multiple equilibria: high-volatility ones, where few agents enter the market for an asset, causing volatility to be high and entry to be undesirable, and low-volatility ones, where many agents enter. Key for the multiplicity in Pagano (1989a) is the feedback from asset prices to equity issuance by firms, and in Allen and Gale (1994) the heterogeneity between investors.[3]

2.3 Transaction Costs

In addition to costs of market participation, agents typically pay costs when executing transactions. Transaction costs drive a wedge between the buying and selling price of an asset. They come in many types, e.g., brokerage commissions, exchange fees, transaction taxes, bid-ask spreads, price impact. Some types of transaction costs can be viewed as a consequence of other market imperfections: for example, Section 2.2 shows that costly participation can generate price-impact costs. Other types of costs, such as transaction taxes, can be viewed as more primitive. We assume that transaction costs concern trade in Period 1. The difference with the participation costs of Section 2.2 is that the decision whether or not to incur the transaction costs is contingent on the price in Period 1.

We focus on the case where transaction costs are proportional to transaction size, and for simplicity assume that proportionality concerns the number of shares rather than the dollar value. Denoting by κ the cost per unit of shares traded and by θ_t the number of shares that an agent holds in Period $t = 0, 1$, proportional costs take the form $\kappa\,|\theta_1 - \theta_0|$. We assume that

[3] A different type of multiplicity arises when agents can choose between market venues to trade an asset. Agents prefer to trade in a venue where others are trading, and this causes concentration of trade in one venue (Pagano (1989b)). A related multiplicity result arises in our model because there exists one equilibrium in which there is market participation in Period 1 and one equilibrium in which no agent participates.

the liquidity shock z is drawn from a general distribution that is symmetric around zero with density $f(z)$; specializing to a normal distribution does not simplify the analysis.

We first describe the equilibrium, then examine how transaction costs affect the illiquidity measures and the expected return, and finally survey the literature on transaction costs. Transaction costs generate a bid-ask spread in Period 1. An agent buying one share pays the price S_1 plus the transaction cost κ, and so faces an effective ask price $S_1 + \kappa$. Conversely, an agent selling one share receives S_1 but pays κ, and so faces an effective bid price $S_1 - \kappa$. Because of the bid-ask spread, trade occurs only if the liquidity shock z is sufficiently large. Suppose, for example, that $z > 0$, in which case liquidity demanders value the asset less than liquidity suppliers. If liquidity suppliers buy, their demand function is as in Section 2.1 (Eq. (1a)), but with $S_1 + \kappa$ taking the place of S_1, i.e.,

$$\theta_1^s = \frac{\bar{D} - S_1 - \kappa}{\alpha\sigma^2}. \tag{16}$$

Conversely, if liquidity demanders sell, their demand function is as in Section 2.1 (Eq. (1b)), but with $S_1 - \kappa$ taking the place of S_1, i.e.,

$$\theta_1^d = \frac{\bar{D} - S_1 + \kappa}{\alpha\sigma^2} - z. \tag{17}$$

Since in equilibrium agents enter Period 1 holding $\bar{\theta}$ shares of the risky asset, trade occurs if there exists a price S_1 such that $\theta_1^s > \bar{\theta}$ and $\theta_1^d < \bar{\theta}$. Using (16) and (17), we can write these conditions as

$$\kappa < \bar{D} - S_1 - \alpha\sigma^2\bar{\theta} < \alpha\sigma^2 z - \kappa.$$

Therefore, trade occurs if $z > \frac{2\kappa}{\alpha\sigma^2} \equiv \hat{\kappa}$, i.e., the liquidity shock z is large relative to the transaction cost κ. The price S_1 can be determined by substituting (16) and (17) into the market-clearing equation (2). Repeating the analysis for $z < 0$, we find that trade occurs in Period 1 if $|z| > \hat{\kappa}$, and in that case the price is

$$S_1 = \bar{D} - \alpha\sigma^2 \left[\bar{\theta} + \pi z + \hat{\kappa}\left(\tfrac{1}{2} - \pi\right)\operatorname{sign}(z)\right]. \tag{18}$$

The effect of transaction costs on the price depends on the relative measures of liquidity suppliers and demanders. Suppose, for example, that $z > 0$. In the absence of transaction costs, liquidity demanders sell and the price drops. Because transaction costs deter liquidity suppliers from buying, they tend to depress the price, amplifying the effect of z. At the same time, transaction costs deter liquidity demanders from selling, and this tends to raise the price, dampening the

effect of z. The overall effect depends on agents' relative measures. If $\pi < 1/2$ (more suppliers than demanders), the impact on suppliers dominates, and transaction costs amplify the effect of z. The converse holds if $\pi > 1/2$. The price in Period 0 is given by the same equation (4) as in the perfect-market benchmark, but for different constants (M, Δ_1).

We next examine how transaction costs impact the illiquidity measures and the expected return. Because transaction costs deter liquidity suppliers from trading, they raise price impact λ. Note that λ rises even when transaction costs dampen the effect of the liquidity shock z on the price. Indeed, dampening occurs not because of enhanced liquidity supply, but because liquidity demanders scale back their trades.

Result 2 *Price impact λ is higher than without transaction costs ($\kappa = 0$).*

Defining price reversal γ involves the complication that for small values of z there is no trade in Period 1, and therefore the price S_1 is not uniquely defined. We define price reversal conditional on trade in Period 1. The empirical counterpart of our definition is that no-trade observations are dropped from the sample. Transaction costs affect price reversal both because they limit trade to large values of z, and because they impact the price conditional on trade occuring. The first effect raises price reversal. The second effect works in the same direction when transaction costs amplify the effect of z on the price, i.e., when $\pi < 1/2$.

Result 3 *Price reversal γ is increasing in the transaction cost coefficient κ if $\pi \leq 1/2$.*

Because transaction costs hinder trade in Period 1, a natural conjecture is that they raise the illiquidity discount. When, however, $\pi \approx 1$, transaction costs can lower the discount. The intuition is that for $\pi \approx 1$ liquidity suppliers are the short side of the market and stand to gain the most from trade. Therefore, transaction costs hurt them the most, and reduce the ratio M of marginal utilities of wealth of demanders and suppliers. This lowers the risk-neutral probability $\frac{\pi M}{1-\pi+\pi M}$ of being a demander, and can lower the discount. Transaction costs always raise the discount, and hence the asset's expected return, when $\pi \leq 1/2$.

Result 4 *The asset's expected return $E(R)$ is increasing in the transaction cost coefficient κ if $\pi \leq 1/2$.*

We can sharpen Results 3 and 4 by assuming specific distributions for the liquidity shock z. When, in particular, z is normal, transaction costs raise γ for all values of π, and numerical calculations suggest that they also raise the discount for all values of π.

Early papers on the effects of transaction costs are Amihud and Mendelson (1986) and Constantinides (1986). Constantinides (1986) derives the optimal investment policy of an infinitely lived agent, who can trade a riskless and a risky asset. The return of the riskless asset is constant over time, and that of the risky asset is *i.i.d.* The risky asset carries transaction costs, which are proportional to the dollar value traded. Because the agent has CRRA preferences, the optimal policy in the absence of transaction costs is to maintain a constant fraction of wealth invested the risky asset, as in Merton (1971). In the presence of transaction costs, the agent instead prevents this fraction from exiting an interval. When the fraction is strictly inside the interval, the agent does not trade. The agent incurs a small utility loss from transaction costs, even though he trades infinitely often in their absence. Intuitively, the derivative of the utility at the optimal policy is zero, and hence a deviation from that policy results in a second-order loss.

The solution of Constantinides (1986) is approximate because consumption is assumed to be an exogenous constant fraction of wealth. Davis and Norman (1990) provide an exact solution. Fleming, Grossman, Vila, and Zariphopoulou (1990), Dumas and Luciano (1991) do the same in the more tractable case where the agent consumes only at the end of his investment horizon. To eliminate horizon effects, they focus on the limit where the horizon converges to infinity. Liu and Loewenstein (2002) consider explicitly the finite-horizon case. Balduzzi and Lynch (1999), Lynch and Balduzzi (2000), Liu (2004), Jang, Koo, Liu, and Loewenstein (2007) and Lynch and Tan (2011) consider richer settings, involving multiple risky assets and predictable returns.

While Constantinides (1986) and the subsequent literature mainly emphasize portfolio optimization, they also explore implications for equilibrium asset prices. They do this by giving the agent a choice between two economies: one in which the risky asset carries transaction costs, and one in which it does not but its expected return is lower. They interpret the reduction in expected return that would make the agent indifferent between the two economies as an equilibrium effect of transaction costs. Whether this effect would arise in an explicit equilibrium model, in which agents can trade both assets, is doubtful; for example, the effect should depend on the scarcity of the more liquid asset, but asset supply is not taken into consideration. This exercise, however, provides an intuitive metric to express the utility loss from transaction costs.

Amihud and Mendelson (1986) build an equilibrium model, in which agents are risk neutral and have different investment horizons. Upon entering the economy, agents can invest in a set of assets that differ in transaction costs. Agents must sell their assets when they exit the economy, and exit rates are independent of age but can differ across agents. Assets with high transaction costs trade at a lower price in equilibrium. Moreover, they are held by agents with

long investment horizons, i.e., low exit rates, who can amortize the costs over a longer period. Each agent holds only one asset, the one maximizing expected return net of transaction costs amortized over the agent's horizon. The effect of transaction costs on asset prices is concave. Indeed, the price differential between one asset and its next closest in terms of transaction costs is determined by the "marginal" investor who is indifferent between the two assets. Since the marginal investor in the case of assets with high transaction costs has a long horizon and hence is less concerned about costs, the price differential between these assets is smaller than for low-cost assets.

Aiyagari and Gertler (1991) and Vayanos and Vila (1999) allow for intertemporal consumption and risk aversion in a setting with two riskless assets, one of which carries transaction costs. The latter paper shows that a decrease in the supply of the more liquid asset increases the liquidity premium, i.e., the premium investors are willing to pay to hold that asset. This is in the spirit of Amihud and Mendelson (1986): since the horizon of the marginal investor becomes shorter, the investor is more concerned about transaction costs. Heaton and Lucas (1996) assume that the asset carrying transaction costs is risky and agents trade to smooth labor income shocks. A calibration of their model reveals that transaction costs have small effects on asset prices unless agents face borrowing constraints.

Vayanos (1998) re-examines the effects of transaction costs in a setting with multiple risky assets and risk averse agents. Agents hold a diversified portfolio at all times, but when they need to reduce their risk exposure they sell disproportionately more of the assets with low transaction costs. Moreover, because transaction costs make agents less willing not only to buy but also to sell an asset, assets with high costs can trade at higher prices than assets with low costs. This result, which also holds in Period 1 of our model, cannot arise when agents are risk neutral or assets are riskless because of a "dominance" argument: since assets are perfect substitutes except for transaction costs, agents would not buy assets with high costs if these trade at higher prices than assets with low costs. Furthermore, the marginal-investor pricing derived in Amihud and Mendelson (1986) does not hold since agents hold diversified portfolios and hence are all marginal for an asset pair.

Huang (2003) assumes stochastic liquidation needs and two riskless assets, one of which carries transaction costs. He shows that transaction costs can generate a strict preference for diversification even though the assets are riskless. This is because returns net of transaction costs are risky: investing in the less liquid asset yields a low payoff if an agent needs to sell on short notice, and a high payoff otherwise. Lo, Mamaysky, and Wang (2004) assume that agents trade to share risk and have access to a riskless asset carrying no costs and a risky asset

carrying fixed costs, i.e., independent of transaction size. They show that transaction costs hinder risk sharing, as in Period 1 of our model, and this causes the price of the risky asset to decrease, as in Period 0 of our model.

More recent work on transaction costs emphasizes the time variation in these costs and in the liquidity premia per unit of the costs. Acharya and Pedersen (2005) assume that investors have a one-period horizon and transaction costs are stochastic. They show that part of the costs' price effect is through a risk premium. This is because transaction costs impact the covariance between an asset's return net of costs and the net return of the market portfolio. For example, if an asset's transaction costs increase when the costs of the market portfolio increase or when the market portfolio's dividends decrease, this adds to the asset's risk and causes the asset price to decrease. Beber, Driessen, and Tuijp (2012) examine the effects of stochastic transaction costs when investors differ in their horizons.

Vayanos (2004) explores time variation in investor horizons, assuming constant transaction costs. He assumes that investors are fund managers subject to withdrawals when their performance drops below a threshold, and that the volatility of asset dividends is time-varying. During volatile times, fund managers' horizons shorten because their performance is more likely to drop below the threshold. This causes liquidity premia per unit of transaction costs to increase. It also causes the market betas of assets with high transaction costs to increase precisely during the times when the market is the most risk averse.

Papers on time varying transaction costs and liquidity premia show that the traditional CAPM should be augmented by pricing factors relating to aggregate liquidity. These factors are aggregate transaction costs in Acharya and Pedersen (2005) and Beber, Driessen, and Tuijp (2012), and volatility (which correlates with liquidity premia) in Vayanos (2004).

Buss and Dumas (2011) and Buss, Uppal, and Vilkov (2011) develop numerical algorithms to solve dynamic general equilibrium models with transaction costs. Buss, Uppal, and Vilkov (2011) assume multiple risky assets and labor income shocks, and show that transaction costs have small price effects. Buss and Dumas (2011) show that deterministic transactions costs give rise to time-variation in measures of illiquidity, such as price impact and volume, and this variation can be a priced factor.

2.4 Asymmetric Information

In the perfect-market benchmark, all agents have the same information about the payoff of the risky asset. In practice, however, agents can have different information because they have

access to different sources of information or have different abilities to process information from the same source. We model asymmetric information through a private signal s about the asset payoff D that some agents observe in Period 1. The signal is

$$s = D + \epsilon \tag{19}$$

where ϵ is normal with mean zero and variance σ_ϵ^2, and is independent of (D, z). We assume that liquidity demanders, who observe the liquidity shock z in Period 1, are also the only ones to observe s. Assuming instead that liquidity suppliers are the only ones to observe s would get us back to symmetric information since each set of agents would infer the variable they do not observe from the price. Note that because liquidity suppliers are uninformed, our model determines how the supply of liquidity is influenced by suppliers' concern about trading against better-informed agents.

We first describe the equilibrium, then examine how asymmetric information affects the illiquidity measures and the expected return, and finally survey the literature on asymmetric information. The price in Period 1 incorporates the signal of liquidity demanders, and therefore reveals information to liquidity suppliers. To solve for equilibrium, we follow the standard rational expectations equilibrium (REE) procedure to conjecture a price function, i.e., a relationship between the price and the signal, then determine how agents use their knowledge of the price function to learn about the signal and formulate demand functions, and finally confirm that the conjectured price function clears the market. We conjecture a price function that is affine in the signal s and the liquidity shock z, i.e.,

$$S_1 = a + b(s - \bar{D} - cz) \tag{20}$$

for constants (a, b, c). For expositional convenience, we set $\xi \equiv s - \bar{D} - cz$. Agents use the price and their private information to form a posterior distribution about the asset payoff D. For a liquidity demander, the price conveys no additional information relative to observing the signal s. For a liquidity supplier, the only information is the price, which is equivalent to observing ξ. Agents' demand functions in Period 1 are as in Section 2.1, with the conditional distributions of D replacing the unconditional one, i.e.,

$$\theta_1^s = \frac{\mathrm{E}[D|S_1] - S_1}{\alpha\sigma^2[D|S_1]}, \tag{21a}$$

$$\theta_1^d = \frac{\mathrm{E}[D|s] - S_1}{\alpha\sigma^2[D|s]} - z. \tag{21b}$$

Substituting (21a) and (21b) into the market-clearing equation (2), we can write the latter as an affine equation in (s, z). Setting the coefficients of (s, z) and of the constant term to zero

yields a system of three equations in the constants (a, b, c) that characterize the price in Period 1. The price in Period 0 is given by the same equation (4) as in the perfect-market benchmark, but for different constants (M, Δ_1).

We next examine how asymmetric information impacts the illiquidity measures and the expected return. When some agents observe a private signal, this not only generates dispersion in information across agents, but also renders each agent more informed because the signal is partially revealed through the price. The improvement in each agent's information is not a distinguishing feature of asymmetric information: information can also improve if all agents observe a public signal. To focus on the dispersion in information, which is what distinguishes asymmetric information, we compare with two symmetric-information benchmarks: the no-information case, where information is symmetric because no agent observes the signal s, and the full-information case, where all agents observe s.

Result 5 *Price impact λ is higher under asymmetric information than under either of the two symmetric-information benchmarks. It also increases when the private signal (19) becomes more precise, i.e., when σ_ϵ^2 decreases.*

The comparison between the asymmetric-, no- and full-information cases is driven by an uncertainty and a learning effect. Price impact increases in the uncertainty faced by liquidity suppliers, measured by their conditional variance of the asset payoff. Because of this uncertainty effect, price impact tends to be lowest under full information, since liquidity suppliers observe the signal perfectly, next lowest under asymmetric information, since the signal is partially revealed to liquidity suppliers through the price, and highest under no information.

An additional source of price impact, present only under asymmetric information, is that liquidity suppliers seek to learn the signal from the price. Because, for example, liquidity suppliers attribute selling pressure partly to a low signal, they require a larger price drop to buy. This learning effect works in the same direction as the uncertainty effect when comparing asymmetric to full information, but in the opposite direction when comparing asymmetric to no information. Result 5 implies that in the latter comparison the learning effect dominates.

While price impact is unambiguously higher under asymmetric information, the same is not true for price reversal. Suppose for example, that $\pi \approx 1$, i.e., almost all agents are liquidity demanders (informed). Then the price processes under asymmetric and full information approximately coincide, and so do the price reversals. Since, in addition, liquidity suppliers face more uncertainty under no information than under full information, price reversal is highest under no information. If instead $\pi \approx 0$, i.e., almost all agents are liquidity suppliers (unin-

formed), then price impact λ converges to infinity (order $1/\pi$) under asymmetric information. This is because the trading volume of liquidity demanders converges to zero, but the volume's informational content remains unchanged. Because of the high price impact, price reversal is highest under asymmetric information.

Result 6 *Price reversal γ is higher under asymmetric information than under either of the two symmetric-information benchmarks if $\pi \approx 0$. If, however, $\pi \approx 1$, then price reversal is higher under no information than under asymmetric information.*

The analysis of the illiquidity discount involves an effect that goes in the direction opposite to the uncertainty effect. This is that information revealed about the asset payoff in Period 1 reduces uncertainty and hence the scope for risk sharing. Less risk sharing, in turn, renders agents less willing to buy the asset in Period 0 and raises the illiquidity discount. The negative effect of information on risksharing and welfare has been shown in Hirshleifer (1971). We derive the implications of the Hirshleifer effect for asset pricing: the reduced scope for risksharing in Period 1 lowers the asset price in Period 0 and raises the illiquidity discount.

Because of the Hirshleifer effect, the illiquidity discount under full information is higher than under no information—a comparison which is exactly the reverse than for the measures of illiquidity. The Hirshleifer effect implies that the illiquidity discount under asymmetric information should be between that under no and under full information. The discount under asymmetric information, however, is also influenced by the learning effect, which raises price impact, reduces the scope for risk sharing and hence raises the discount. The learning effect works in the same direction as the Hirshleifer effect when comparing asymmetric to no information, but in the opposite direction when comparing asymmetric to full information. Result 7 implies that in the latter comparison the learning effect dominates: the illiquidity discount, and hence the asset's expected return, is highest under asymmetric information than under either of the two symmetric-information benchmarks.

Result 7 *The asset's expected return $E(R)$ is higher under asymmetric information than under either of the two symmetric-information benchmarks.*

The analysis of REE with asymmetric information was pioneered by Grossman (1976). Grossman (1976) assumes that agents observe private signals about the payoff of a risky asset, which are of equal quality and independent conditional on the payoff. The equilibrium price of the risky asset reveals the average of agents' signals, which is a sufficient statistic for all the signals because of normality. Hence, the price aggregates information perfectly.

Grossman and Stiglitz (1980) assume that some agents observe a common signal about the payoff of a risky asset and the remaining agents observe no signal. Following some of the literature, we term this information structure "asymmetric information structure," and that in Grossman (1976) as "differential information structure." Grossman and Stiglitz (1980) allow additionally for the supply of the risky asset to be stochastic. With a deterministic supply, the price reveals perfectly the signal of the informed agents, and hence the uninformed can achieve the same utility as the informed. With a stochastic supply instead, the informed can achieve higher utility. The analysis of Grossman and Stiglitz (1980) is closely related to the equilibrium in Period 1 of our model, except that we introduce noise in the price through endowments rather than through the asset supply. Diamond and Verrecchia (1981) are first to use this modeling trick, and do so in a differential information model.

Grossman (1976) and Grossman and Stiglitz (1980) derive two basic paradoxes relating to information aggregation. Since the price in Grossman (1976) aggregates perfectly agents' private signals, agents should form their asset demand based only on the price and not on their signals. The paradox then is how can the price aggregate the signals. A second paradox is that if the price in Grossman and Stiglitz (1980) reveals perfectly the signal of the informed agents, then why would the informed be willing to commit resources to acquire their signal.

Both paradoxes can be resolved by introducing noise, e.g., through stochastic asset supply. Grossman and Stiglitz (1980) show that with stochastic supply, the informed can achieve higher utility than the uninformed and hence can have an incentive to acquire costly information. This has the important implication that markets cannot be fully efficient when information acquisition is costly because information will be acquired only when the price is not fully revealing.[4] Hellwig (1980) introduces stochastic supply in a differential information model, which generalizes Grossman (1976) by allowing for heterogeneity in signal quality and agent risk aversion. He shows that the price does not aggregate information perfectly, and hence agents have an incentive to use both the price and their private signal when forming their asset demand.

All papers mentioned so far assume that agents can trade one riskless and one risky asset over one period. Admati (1985) extends the analysis to multiple risky assets, while also allowing for a general correlation structure between asset payoffs, asset supplies, and agents' private signals. She shows that because signals about one asset are also informative about the payoff and supply of others, surprising phenomena can arise. For example, a high price of one asset, holding other prices constant, can cause agents to lower their expectation of that asset's payoff.

[4]This conclusion does not extend to settings with imperfect competition, as we point out in Section 2.5.

Grundy and McNichols (1989) and Brown and Jennings (1990) assume two trading periods and one risky asset. They show that uninformed traders learn about the asset payoff not only from current prices but also from past ones because prices are noisy signals of asset payoffs. The optimal strategy thus uses the entire price history, in a manner similar to strategies used by technical traders. Wang (1993) studies a continuous-time setting with one risky asset. He shows that uninformed agents behave as price chasers, buying following a price increase. He also shows that return volatility and price reversal can be highest under asymmetric information than under full or no information. The latter result is consistent with our Result 6. Wang (1994) employs a similar model to study the behavior of trading volume and its relationship with price changes. These papers assume an asymmetric information structure. He and Wang (1995) study the joint behavior of trading volume and prices under a differential information structure. Vives (1995) studies the speed at which prices aggregate information under a combined asymmetric-differential information structure, where some agents observe conditionally independent signals about the payoff of a risky asset and the remaining agents observe no signal.

Much of the literature on REE with asymmetric information focuses on the informativeness of prices, rather than on market liquidity. Market liquidity is instead emphasized in a subsequent literature which combines asymmetric information with strategic behavior or sequential arrival of traders. This literature was pioneered by Glosten and Milgrom (1985) and Kyle (1985), and is surveyed in the next section. Yet, even REE models with asymmetric information have implications for market liquidity. We derive such implications in the context of our model in Results 5 and 6. Moreover, Cespa and Foucault (2011) show that asymmetric information can generate liquidity spillovers: because asset payoffs are correlated, a drop in liquidity in one asset reduces the information available on other assets, hence reducing the liquidity of those assets.

A number of recent papers examine whether agents require higher expected returns to invest in the presence of asymmetric information. O'Hara (2003) and Easley and O'Hara (2004) show that prices and lower and expected returns higher when agents receive private signals than when signals are public. This comparison concerns Period 1 of our model, and reverses when using the alternative symmetric-information benchmark where no signals are observed. By contrast, we show (Result 7) that the price in Period 0 is lower under asymmetric information than under either symmetric-information benchmark. Comparing prices in Period 0 measures the ex-ante effect of the imperfection, i.e., what compensation do agents require to invest ex-ante knowing that they will face asymmetric information ex-post? Qiu and Wang (2010) derive similar results in an infinite-horizon model. Garleanu and Pedersen (2004) show in a model with risk-neutral agents and unit demands that asymmetric information can raise or lower prices, with the effect

being zero when probability distributions are symmetric—as is the case under normality. Ellul and Pagano (2006) show that asymmetric information lowers prices in a model of IPO trading.

2.5 Imperfect Competition

In the perfect-market benchmark, agents are competitive and have no effect on prices. In many markets, however, some agents are large relative to others in the sense that they can influence prices, either because of their size or because of their information advantage. We model imperfect competition by assuming that some agents can exert market power in Period 1. We focus on the case where liquidity demanders behave as a single monopolist, but also consider more briefly monopolistic behavior by liquidity suppliers. We emphasize the former case because it has received more attention in the literature. When liquidity suppliers behave monopolistically, imperfect competition obviously influences the supply of liquidity. More surprisingly, it can also influence liquidity supply when liquidity demanders behave monopolistically and suppliers do not.

We consider both the case where liquidity demanders have no private information on asset payoffs, and so information is symmetric, and the case where they observe the private signal (19), and so information is asymmetric. The second case nests the first by setting the variance σ_ϵ^2 of the signal noise to infinity. Hence, we treat both cases simultaneously, and compare imperfect competition to the competitive equilibrium with asymmetric information studied in Section 2.4.

The trading mechanism in Period 1 is that liquidity suppliers submit a demand function and liquidity demanders submit a market order, i.e., a price-inelastic demand function. Restricting liquidity demanders to trade by market order is without loss of generality: they do not need to condition their demand on price because they know all information available in Period 1.

We first describe the equilibrium when liquidity demanders behave monopolistically, and examine how imperfect competition affects the illiquidity measures and the expected return. We next treat more briefly the case where liquidity suppliers behave monopolistically, and finally survey the literature on imperfect competition.

We conjecture that the price in Period 1 has the same affine form (20) as in the competitive case, with possibly different constants (a, b, c). Given (20), the demand function of liquidity suppliers is (21a) as in the competitive case. Substituting (21a) into the market-clearing equation (2), yields the price in Period 1 as a function $S_1(\theta_1^d)$ of the liquidity demanders' market

order θ_1^d. Liquidity demanders choose θ_1^d to maximize the expected utility

$$-\mathrm{E}\exp\left\{-\alpha\left[W_1+\theta_1^d\left(D-S_1(\theta_1^d)\right)+z(D-\bar{D})\right]\right\}. \tag{22}$$

The difference with the competitive case is that liquidity demanders behave as a single monopolist and take into account the impact of their order θ_1^d on the price S_1. The optimal order of liquidity demanders satisfies

$$\theta_1^d=\frac{\mathrm{E}[D|s]-S_1(\theta_1^d)-\alpha\sigma^2[D|s]z+\hat{\lambda}\bar{\theta}}{\alpha\sigma^2[D|s]+\hat{\lambda}}, \tag{23}$$

where $\hat{\lambda}\equiv\frac{dS_1(\theta_1^d)}{d\theta_1^d}$. Eq. (23) determines θ_1^d implicitly because it includes θ_1^d in both the left- and the right-hand side. We write θ_1^d in the form (23) to facilitate the comparison with the competitive case. Indeed, the competitive counterpart of (23) is (21b), and can be derived by setting $\hat{\lambda}$ to zero. The parameter $\hat{\lambda}$ measures the price impact of liquidity demanders, and is closely related to the price impact λ. Because in equilibrium $\hat{\lambda}>0$, the denominator of (23) is larger than that of (21b), and therefore θ_1^d is less sensitive to changes in $\mathrm{E}[D|s]-S_1$ and z than in the competitive case. Intuitively, because liquidity demanders take price impact into account, they trade less aggressively in response to their signal and their liquidity shock.

Substituting (21a) and (23) into the market-clearing equation (2), and proceeding as in Section 2.4, we find a system of three equations in the constants (a,b,c) that characterize the price in Period 1. The price in Period 0 is given by the same equation (4) as in the perfect-market benchmark, but for different constants (M,Δ_1).

We next examine how imperfect competition by liquidity demanders impacts the illiquidity measures and the expected return.

Result 8 *Price impact λ is the same as under perfect competition when information is symmetric, and higher when information is asymmetric.*

When information is asymmetric, imperfect competition lowers liquidity, as measured by price impact, even though liquidity suppliers are competitive. The intuition is that when liquidity demanders take into account their effect on price, they trade less aggressively in response to their signal and their liquidity shock. This reduces the size of both information- and liquidity-generated trades (hence lowering b). The relative size of the two types of trades

(measured by c) remains the same, and so does price informativeness, measured by the signal-to-noise ratio. Monopoly trades thus have the same informational content as competitive trades, but are smaller in size. As a result, the signal per unit trade is higher, and so is the price impact λ of trades. Imperfect competition has no effect on price impact when information is symmetric because trades have no informational content.

An increase in information asymmetry, through a reduction in the variance σ_ϵ^2 of the signal noise, generates an illiquidity spiral. Because illiquidity increases, liquidity demanders scale back their trades. This raises the signal per unit trade, further increasing illiquidity. When σ_ϵ^2 reaches a lower bound $\hat{\sigma}_\epsilon^2$, illiquidity becomes infinite and trade ceases, leading to a market breakdown. For $\sigma_\epsilon^2 \leq \hat{\sigma}_\epsilon^2$, a linear equilibrium fails to exist.

While imperfect competition raises price impact λ, it lowers price reversal γ. Intuitively, price reversal arises because the liquidity demanders' trades in Period 1 cause the price to deviate from fundamental value. Under imperfect competition, these trades are smaller and so is price reversal.

Result 9 *Price reversal γ is lower than under perfect competition.*

Imperfect competition can lower or raise the illiquidity discount. Indeed, since liquidity demanders scale back their trades, they render the price less responsive to their liquidity shock. Therefore, they can obtain better insurance against the shock, and become less averse to holding the asset in Period 0. This effect drives the illiquidity discount, and hence the asset's expected return, below the competitive value when information is symmetric. When information is asymmetric, the comparison can reverse. This is because the scaling back of trades generates the spiral of increasing illiquidity, and this reduces the insurance received by liquidity demanders.

Result 10 *The asset's expected return $E(R)$ is lower than under perfect competition when information is symmetric, but can be higher when information is asymmetric.*

The case where liquidity suppliers behave monopolistically can be treated in a manner similar to the case where demanders do, so we provide a brief sketch. Suppose that demanders are competitive but suppliers behave as a single monopolist in Period 1. Since suppliers do not know the liquidity shock z and signal s, their trading strategy is to submit a price-elastic demand function (rather than a market order). Monopolistic behavior renders this demand function less price-elastic than its competitive counterpart (21a). The lower elasticity manifests

itself through an additive positive term in the denominator of the competitive demand (21a), exactly as is the case for liquidity demanders in (21b) and (23).

Because liquidity suppliers submit a less price-elastic demand function than in the competitive case, the trades of liquidity demanders have larger price impact. Hence, price impact λ and price reversal γ are larger than in the competitive case. The illiquidity discount is also larger because liquidity demanders receive worse insurance against the liquidity shock.

Two seminal papers on imperfect competition in financial markets and its relationship with asymmetric information are Kyle (1985,1989). Kyle (1989) assumes a combined asymmetric-differential information structure, where some agents observe conditionally independent signals about the payoff of a risky asset and the remaining agents observe no signals. Agents submit demand functions, as in competitive rational expectations equilibrium (REE), but the equilibrium concept is instead Nash equilibrium in demand functions, as in Wilson (1979) and Klemperer and Meyer (1989). Noise traders add a stochastic amount to the asset supply, preventing prices from being fully revealing. Because informed agents take into account their effect on price, they trade less aggressively in response to their signal. Imperfect competition thus reduces the size of information-based trades. Since it has no effect on liquidity-generated trades, which are initiated by the exogenous noise traders, it lowers price informativeness.

Kyle (1985) assumes a risk neutral insider who observes a private signal about the payoff of a risky asset and can trade with market makers and noise traders. The insider and the noise traders submit market orders, which are batched together and absorbed by market makers. Because the latter are risk neutral and competitive, they compete a la Bertrand and absorb all orders at a price equal to their conditional expectation of the asset payoff. Imperfect competition reduces price informativeness, as in Kyle (1989). An advantage of Kyle (1985) is that it is highly tractable and can be extended in many directions. One important extension, performed in Kyle (1985), is to allow trading to take place dynamically, over more than one period. The insider then reveals his information slowly over time, as revealing it quickly would subject him to a higher price impact in the early periods. In the continuous-time limit, the insider reveals his information in a way that exactly equates price impact across time.

The model of Kyle (1985) has been extended in many directions. A first extension is to introduce multiple insiders. Admati and Pfleiderer (1988) show that liquidity traders can concentrate their trades in the same period, to reduce price impact, and this effect can be amplified when there are multiple insiders. Holden and Subrahmanyam (1992) assume multiple insiders who receive a common signal about the payoff of a risky asset, and show that they

reveal it almost immediately in the continuous-trading limit because each insider tries to exploit his information before others do. Foster and Viswanathan (1996) assume multiple insiders who receive imperfectly correlated signals, and show that information revelation slows down because of a "waiting-game" effect, whereby each insider attempts to learn the others' signals. Back, Cao, and Willard (2000) formulate this problem in continuous time and show that information is not fully revealed in prices until the end of the trading session.

A second extension is to drop the noise traders and derive non-informational trading from utility maximization. Glosten (1989) generates non-informational trading through a random endowment received by the insider. We make the same assumption, and the equilibrium in our Period 1 is closely related to the one that Glosten (1989) derives in the case where market makers are competitive. Glosten (1989) assumes risk-neutral market makers; a paper even closer to our model is Bhattacharya and Spiegel (1991), which assumes that liquidity suppliers are risk averse.[5] Both papers find that the market breaks down when information asymmetry is severe. The mechanism is the same as in our model, and the key assumptions are that some liquidity demanders are informed and all are non-price-takers.[6] The idea that adverse selection can cause market breakdown dates back to Akerlof (1970).

Other extensions are to introduce risk aversion, non-normal probability distributions for the asset payoff, and a minimum trade size. Back (1992) shows that the result on the equalization of price impact across time extends to general payoff distributions. Holden and Subrahmanyam (1994) and Baruch (2002) show that a risk averse insider reveals his information faster than a risk neutral one because he is eager to reduce the uncertainty at which his trades will be executed. Back and Baruch (2004) assume that noise traders execute discrete transactions rather than trading continuously, in which case the insider must do the same so not to be revealed. They show that the insider follows a mixed strategy, and can trade in a direction opposite to his information in some cases.

A further extension is to change the information structure. Kyle (1985) assumes that the insider receives all his information in an initial period, and the information is announced publicly in a final period. Chau and Vayanos (2008) and Caldentey and Stacchetti (2010) show that when the insider receives a constant flow of new information over time, he chooses to reveal it infinitely fast in the continuous-trading limit. This result is in sharp contrast to Kyle (1985), where revelation is slow. Moreover, markets are arbitrarily close to efficiency and yet informed

[5]Both papers assume one trading period and do not derive effects on ex-ante expected returns, as we do.

[6]For example, market breakdown does not occur in Kyle (1985) because noise traders submit price-inelastic demands, which can be viewed as an extreme form of price taking.

traders earn abnormal profits, in sharp contrast to Grossman and Stiglitz (1980). Efficient markets and insider profits are not contradictory because continuous trading gives insiders the flexibility to earn profits even though they reveal each piece of new information within a very short interval.[7] Other models exploring insider trading with a flow of new information are Back and Pedersen (1998) and Guo and Ou-Yang (2010).

A final set of extensions examine issues relating to market design. For example, Chowhdry and Nanda (1991) study the competition between market venues. Fishman and Hagerty (1992), Leland (1992) and Repullo (1999) study whether insider trading is desirable or should be banned. Admati and Pfleiderer (1991) study "sunshine trading," whereby liquidity traders pre-announce their intention to trade, so to distinguish themselves from insiders and reduce their trading costs. Pagano and Roell (1996) and Naik, Neuberger, and Viswanathan (1999) study the effects of a transparency regulation requiring disclosure of all trades but of not traders' identities. Huddart, Hughes, and Levine (2001) consider instead a regulation requiring disclosure of trades by insiders. They show that the regulation speeds up information revelation and reduces insiders' profits. It also induces the insiders to trade less aggressively and follow a mixed strategy, trading occasionally in a direction opposite to their information. Buffa (2011) shows that because of the latter effect, the regulation can instead slow down information revelation when insiders are risk averse.

Kyle (1985) and much of the subsequent literature assume that the non-price-taking agents are insiders who receive private information about asset payoffs. In many cases, however, agents without such information affect prices simply because of the size of their trades. For example, trades by pension funds can exceed the average daily volume of many stocks, and are often triggered by reasons other than information, e.g., regulatory constraints. Vayanos (1999) assumes that large traders with no private information about asset payoffs receive random endowments over time and need to share risk. He shows that these agents break their trades into small pieces so to reduce price impact, and risk sharing is accomplished slowly even in the continuous-trading limit. What deters them from trading faster is that this will signal to the market that a larger trade is yet to come, and so price impact will be large. Vayanos (2001) shows that the presence of noise traders in this setting can accelerate trading and hence improve

[7] Jackson (1991) provides an alternative resolution of the Grossman and Stiglitz (1980) paradox within a static setting. He assumes that agents can acquire private signals at a cost and submit demand functions as in Kyle (1989). Unlike in Kyle (1989), there are no noise traders. The equilibrium price is fully revealing and yet agents have an incentive to acquire information. This is because information helps them predict their price impact, which the price does not reveal. Normal-linear models cannot generate this effect because price impact is a constant independent of information. For an analysis of information revelation without noise traders, normality and linearity, see also Laffont and Maskin (1990).

risk sharing.[8]

Large traders who trade over time to share risk are similar to durable-good monopolists. According to the Coase conjecture, the monopolists should trade infinitely fast in the continuous-trading limit. Trading occurs slowly in Vayanos (1999) because each trader is the only one to observe his endowment and hence his eagerness to share risk; if instead endowments are publicly observed, the Coase conjecture holds. DeMarzo and Urosevic (2006) consider a general setting where a large trader needs to share risk but can also take actions to increase asset payoffs, e.g., monitor the firm's managers. The trader's eagerness to share risk is public information. DeMarzo and Urosevic (2006) confirm the Coase conjecture in the case where asset payoffs are independent of the trader's actions. Rostek and Weretka (2011) study risk sharing in a dynamic setting where agents' endowments are public information. They decompose the price impact of a trade into a permanent component, due to the risk aversion of agents taking the other side, and a temporary one, due to their monopoly power.

When large traders affect prices, information about their future trades is valuable to others. This is so even when large traders themselves have no information about asset payoffs. Cao, Evans, and Lyons (2006) label information about future large trades "inventory information." Brunnermeier and Pedersen (2005) assume that a large trader needs to sell because of financial distress, and show that other traders exploit this information by selling at the same time as him. Such "predatory" behavior benefits these traders because they cause the distressed trader to sell at low prices, at which they can buy. Pritsker (2005) studies predatory behavior in a multi-asset setting. Attari, Mello, and Ruckes (2005), Fardeau (2011) and Venter (2011) model the financial constraints of distressed traders and examine how predatory behavior by others can bring them closer to distress by moving asset prices against them. Carlin, Lobo, and Viswanathan (2007) derive predatory behavior as a breakdown of collaboration in a repeated game.

Most papers mentioned so far emphasize non-price-taking behavior by liquidity demanders; liquidity suppliers, such as market makers, are assumed to behave competitively. Biais (1993) studies how oligopolistic market makers bid for an order, depending on whether or not they know the inventories of their competitors. He relates the quality of market makers' information to whether the market is centralized or fragmented. Earlier papers on oligopolistic market makers include Ho and Stoll (1980) and Copeland and Galai (1983).

[8] Bertsimas and Lo (1998), Almgren and Chriss (1999), Almgren (2003) and Huberman and Stanzl (2005) study the optimal policy of large traders in partial-equilibrium settings, under exogenous price dynamics. Obizhaeva and Wang (2006) derive the price dynamics faced by large traders from a model of the limit-order book, which describes how new limit orders arrive after existing ones are consumed.

Glosten (1989) shows that when information is asymmetric, a market with a monopoly market maker can dominate one with competitive market makers. This is because the market can break down with competitive market makers, but breakdown can be avoided with a monopoly market maker. Glosten (1989) models perfect competition between market makers in terms of a zero-profit condition. Glosten (1994) derives this condition as the equilibrium of a game in which market makers post price-quantity schedules. Bernhardt and Hughson (1997) study this game in the case of two oligopolistic market makers. Biais, Martimort, and Rochet (2000) study the game for a general number of market makers and provide a full characterization of the equilibrium. For a finite number of market makers the equilibrium has a Cournot flavor, and it converges to the competitive case characterized by Glosten (1989) as the number goes to infinity. Back and Baruch (2011) provide an alternative characterization of the same game.

Models with non-price-taking behavior study the interaction between small numbers of agents. In this sense, they are related to models of sequential order arrival, in which traders arrive in the market one at a time and remain there for a short period. The latter models assume implicitly participation costs since agents are not present in the market until when they arrive. Early models in that spirit include Garman (1976), Amihud and Mendelson (1980) and Ho and Stoll (1981), in which market makers receive an exogenous flow of orders.

Glosten and Milgrom (1985) propose a highly tractable model of sequential order arrival with asymmetric information. Some of the agents receive a private signal about the asset payoff, while others do not and trade for liquidity reasons. Upon arriving in the market, agents can execute a buy or sell transaction of a fixed size with market makers. As in Kyle (1989), market makers are risk neutral and competitive. Therefore, they compete a la Bertrand and absorb orders at a price equal to their conditional expectation of the asset payoff. In equilibrium, the bid price that market makers quote to buy from other agents is lower than the ask price that they quote to sell to them. This is because when market makers buy, they suspect that other agents might have sold to them because of negative information. Glosten and Milgrom (1985) thus link the bid-ask spread to asymmetric information, building on earlier work by Bagehot (1971) and Copeland and Galai (1983).

The models of Glosten and Milgrom (1985) and Kyle (1985) give rise to different measures of illiquidity. Illiquidity in Glosten and Milgrom (1985) is measured by the bid-ask spread since all transactions are assumed to be of a fixed size. By contrast, in Kyle (1985) transactions can be of any size since probability distributions are normal and trading strategies linear. Illiquidity is measured by the sensitivity of price to quantity, which corresponds to λ in our model. While the bid-ask spread and λ are derived within different models, the share the basic property of

being increasing in the degree of asymmetric information. Easley and O'Hara (1987) consider a hybrid model in which agents arrive in the market one at a time and can execute transactions of variable size with market makers. The prices that market makers post depend on quantity, in a spirit similar to Kyle (1985). Easley and O'Hara (1992) allow the time when private information arrives to be stochastic. They show that the bid-ask spread increases following a surge in trading activity because market makers infer that private information has arrived.

A recent literature studies sequential order arrival in limit-order markets, where there are no designated market makers and liquidity is supplied by the arriving agents. Agents can execute a buy or sell transaction of a fixed size. Impatient agents execute this transaction immediately upon arrival through a market order, and hence demand liquidity. Patient agents submit instead a limit order, i.e., a price-elastic demand function, which is executed against future market orders. Hence, they supply liquidity to future agents. Papers in that literature include Parlour (1998), Foucault (1999), Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005) and Rosu (2009).[9] These papers determine the bid-ask spread that results from the submitted limit orders, the choice of agents between market and limit orders, the expected time for limit orders to execute, etc. A positive bid-ask spread arises even in the absence of asymmetric information, and is decreasing in the degree of competition between limit-order suppliers. This parallels our result that λ is larger when liquidity suppliers behave monopolistically than when they are competitive.

2.6 Funding Constraints

Agents' portfolios often involve leverage, i.e., borrow cash to establish a long position in a risky asset, or borrow a risky asset to sell it short. In the perfect-market benchmark, agents can borrow freely provided that they have enough resources to repay the loan. But as the Corporate Finance literature emphasizes, various frictions can limit agents' ability to borrow and fund their positions. These frictions can also influence the supply of liquidity in the market.

Since in our model consumption is allowed to be negative and unbounded from below, agents can repay a loan of any size by reducing consumption. Negative consumption can be interpreted as a costly activity that agents undertake in Period 2 to repay a loan. We derive a funding constraint by assuming that agents cannot commit to reduce their consumption below a level $-A \leq 0$. This nests the case of full commitment assumed in the rest of this paper

[9]These papers assume that agents have market power. Biais, Hombert, and Weill (2011) assume instead that agents are competitive and observe their valuation for an asset only infrequently. They show that the optimal orders that agents submit at the observation times can be price-contingent and concern future execution.

($A = \infty$), and the case where agents can walk away from a loan rather than engaging in negative consumption ($A = 0$). Because our focus is on how the funding constraint influences the supply of liquidity, we impose it on liquidity suppliers only, i.e., assume that the lack of commitment concerns only them.

For simplicity, we assume that loans must be fully collateralized in the sense that agents must be able to commit enough resources to cover any losses in full. To ensure that full collateralization is possible, we replace normal distributions by distributions with bounded support. We denote the support of the asset payoff D by $[\bar{D} - b_D, \bar{D} + b_D]$ and that of the liquidity shock z by $[-b_z, b_z]$. We assume that D and z are distributed symmetrically around their respective means, D is positive (i.e., $\bar{D} - b_D \geq 0$), and agents receive a positive endowment B of the riskless asset in Period 0.

We first describe the equilibrium, then examine how the funding constraint affects the illiquidity measures and the illiquidity discount, and finally survey the literature on funding constraints. To derive the funding constraint, we note that losses from investing in the risky asset can be covered by wealth W_1 or negative consumption. Since liquidity suppliers must be able to cover losses in full, and cannot commit to consume less than $-A$, losses cannot exceed $W_1 + A$, i.e.,

$$\theta_1^s(S_1 - D) \leq W_1 + A \quad \text{for all } D.$$

This yields the constraint

$$m|\theta_1^s| \leq W_1 + A, \tag{24}$$

where

$$m \equiv S_1 - \min_D D \quad \text{if} \quad \theta_1^s > 0, \tag{25a}$$

$$m \equiv \max_D D - S_1 \quad \text{if} \quad \theta_1^s < 0. \tag{25b}$$

The constraint (24) requires that a position of θ_1^s shares is backed by capital $m|\theta_1^s|$. This limits the size of the position as a function of the capital $W_1 + A$ available to suppliers in Period 1. Suppliers' capital is the sum of the capital W_1 that they physically own in Period 1, and the capital A that they can access through their commitment to consume $-A$ in Period 2. The parameter m is the required capital per share, and can be interpreted as a margin or haircut. The margin is equal to the maximum possible loss per share. For example, the margin (25a) for a long position does not exceed the asset price S_1, and is strictly smaller if the asset payoff D has a positive lower bound (i.e., $\min_D D = \bar{D} - b_D > 0$).

Intuitively, the constraint (24) can bind when there is a large discrepancy between the price S_1 and the expected payoff $\bar{D}$, since this is when liquidity suppliers want to hold large positions. If (24) binds, then the demand function θ_1^s of a liquidity supplier in Period 1 is determined by (24) together with the requirement that θ_1^s has the same sign as $\bar{D} - S_1$. If (24) does not bind, then the demand function is

$$\theta_1^s = \left(f'\right)^{-1}(\bar{D} - S_1), \tag{26}$$

where

$$f(\theta) \equiv \frac{\log \mathrm{E} \exp\left[-\alpha\theta(D - \bar{D})\right]}{\alpha}. \tag{27}$$

Eq. (26) generalizes (1a) to the case where asset payoffs are not normal. The function $f(\theta)$ can be interpreted as a cost of bearing risk, and is positive, convex, and equal to $\frac{1}{2}\alpha\theta^2$ under normality (which is ruled out by the bounded-support requirement). Note that since $f(\theta)$ is convex, $f'(\theta)$ is increasing, and so the demand θ_1^s is a decreasing function of the price S_1. The demand function of a liquidity demander for the risky asset in Period 1 is

$$\theta_1^d = \left(f'\right)^{-1}(\bar{D} - S_1) - z. \tag{28}$$

Combining (24)-(28) with the market clearing equation (2), we can characterize the equilibrium in terms of two regions: the *abundant-capital* region and the *scarce-capital* region. The abundant-capital region is defined by the condition that the endowment B of the riskless asset that agents receive in Period 0, plus the quantity A that liquidity suppliers can access through their commitment to consume $-A$ in Period 2, exceeds a threshold K^*. If $B + A > K^*$, then liquidity suppliers are well-capitalized and their funding constraint never binds. If instead $B + A < K^*$, then the constraint binds for large positive and possibly large negative values of the liquidity shock z. For example, when z is large and positive, the price S_1 is low and liquidity suppliers are constrained because they want to hold large long positions. Note that in both regions, the constraint does not bind for $z = 0$. Indeed, the unconstrained outcome for $z = 0$ is that liquidity suppliers maintain their endowments $\bar{\theta}$ of the risky asset and B of the riskless asset. Since this yields positive consumption, the constraint is met.

An increase in the liquidity shock z lowers the price S_1 and raises the liquidity suppliers' position θ_1^s. These results are the same as in the perfect-market benchmark of Section 2.1, but the intuition is more complicated when the funding constraint binds. Suppose that capital is scarce (i.e., $B + A < K^*$), and z is large and positive, in which case suppliers hold long

positions and are constrained. The intuition why they can buy more, despite the constraint, when z increases is as follows. Since the price S_1 decreases, suppliers realize a capital loss on the $\bar{\theta}$ shares of the risky asset that they carry from Period 0. This reduces their wealth in Period 1 and tightens the constraint. At the same time, a decrease in S_1 triggers an equal decrease in the margin (25a) for long positions because the maximum possible loss on these positions decreases. This effect loosens the constraint. It also dominates the first effect since it is equivalent to a capital gain on the θ_1^s shares that suppliers hold in Period 1, and suppliers are net buyers for $z > 0$ (i.e., $\theta_1^s > \bar{\theta}$). Hence, suppliers can buy more in response to an increase in z. The price in Period 0 can be computed in closed form when the risk-aversion coefficient α is small. The form of the price is different in the abundant and in the scarce capital region.

We next examine how the funding constraint impacts the illiquidity measures and the illiquidity discount. We compute these variables in the abundant-capital region, and compare with the scarce-capital region.

Result 11 *Suppose that α is small or z is drawn from a two-point distribution. Price impact λ is higher when capital is scarce than when it is abundant.*

Result 12 *Price reversal γ is higher when capital is scarce than when it is abundant.*

The intuition is as follows. When the liquidity shock z is close to zero, the constraint does not bind in both the abundant- and scarce-capital regions, and therefore price and volume are identical in the two regions. For larger values of z, the constraint binds when capital is scarce, impairing suppliers' ability to accommodate an increase in z. As a result, an increase in z has a larger effect on price and a smaller effect on volume. Since the effect on price is larger, so is the price reversal γ. Since the effect on price per unit of volume is also larger, so is the price impact λ. Note that λ measures an average price impact, i.e., the average slope of the relationship between return and signed volume. This relationship exhibits an important non-linearity when capital is scarce: the slope increases for large values of z, which is when the constraint binds.

The illiquidity discount, and hence the asset's expected return, is higher when capital is scarce. This is because the funding constraint binds asymmetrically: it is more likely to bind when liquidity demanders sell ($z > 0$) than when they buy ($z < 0$). Indeed, the constraint binds when the suppliers' position is large in absolute value—and a large position is more likely when suppliers buy in Period 1 because this adds to the long position $\bar{\theta}$ that they carry from Period 0. Since price movements in Period 1 are exacerbated when the constraint binds, and the constraint is more likely to bind when demanders sell, the average price in Period 1 is lower

when capital is scarce. This yields a lower price in Period 0.

Result 13 *Suppose that α is small. The asset's expected return $E(R)$ is lower when capital is scarce than when it is abundant.*

The literature on funding constraints in financial markets can be viewed as part of a broader literature on the limits of arbitrage. Indeed, both literatures emphasize the idea that some traders rely on external capital, which is costlier than internal capital, and this affects liquidity and asset prices. External capital can take the form of collateralized debt, as in our model, or other forms such as equity. We first survey work that derives funding constraints from collateralized debt, and then survey more briefly the broader theoretical literature on the limits of arbitrage. An extensive survey of the latter literature is Gromb and Vayanos (2010a).

The effects of funding constraints have been studied in macroeconomic settings, starting with Bernanke and Gertler (1989) and Kiyotaki and Moore (1997). In these papers, adverse shocks to economic activity depress the collateral values of productive assets, and this reduces lending and amplifies the drop in activity. Similar amplification effects arise in financial-market settings, as we point out below.

A number of papers link the tightness of funding constraints to the volatility of the collateral. Hart and Moore (1994, 1995) show that uncertainty about assets' liquidation values impairs agents' ability to borrow. Shleifer and Vishny (1992) endogenize liquidation values and the ability to borrow in market equilibrium. Geanakoplos (1997, 2003) defines collateral equilibrium, in which agents borrow to buy financial assets and post the assets as collateral. The amount of collateral is determined endogenously in equilibrium, and is increasing in asset volatility. Moreover, if volatility increases following adverse shocks, funding constraints tighten, and this causes agents to sell assets, amplifying the shocks. The link between volatility and ability to borrow is also present in our model. Indeed, an increase in the parameter b_D, which measures the dispersion of the asset payoff distribution, raises the margins in (25a) and (25b), holding the price S_1 constant. The funding constraint (24) in our model is derived along the lines of Geanakoplos (2003), who also provides conditions under which full collateralization is an equilibrium outcome.

Gromb and Vayanos (2002) link market liquidity to the capital of financial intermediaries and their funding constraints—a link that we also derive in Results 11 and 12. Investors are subject to liquidity shocks and can realize gains from trade across segmented markets by trading with intermediaries. Intermediaries exploit price discrepancies, and in doing so supply liquidity to investors: they buy low in a market where investors are eager to sell, and sell

high in a market where investors are eager to buy, thus supplying liquidity to both sets of investors. Intermediaries fund their position in each market using collateralized debt, and face a funding constraint along the lines of (24). Shocks to asset prices that trigger capital losses by intermediaries, tighten the intermediaries' funding constraints and force them to reduce their positions. This lowers market liquidity and amplifies the shocks.

Amplification effects do not arise in our model because liquidity suppliers increase their position θ_1^s in Period 1 following an increase in the liquidity shock z. Amplification effects require instead that suppliers decrease their position, hence becoming demanders of liquidity. Recall that suppliers in our model are able to increase their position following an increase in z because while their wealth decreases, there is a stronger countervailing effect caused by a decrease in the margin. Amplification effects arise when the margin instead increases, as in Geanakoplos (1997, 2003), or stays constant. They can arise even when the margin decreases but there are multiple periods, as in Gromb and Vayanos (2002). Kondor (2009) shows that amplification effects can arise even in the absence of shocks. Indeed, if a price discrepancy between two assets were to remain constant or decrease over time, intermediaries would exploit it and reduce it to a level from which it could increase. Gromb and Vayanos (2010b) determine conditions under which arbitrageurs stabilize or destabilize prices in a simple static setting.

Liu and Longstaff (2004) study how funding-constrained intermediaries exploit price discrepancies under an exogenous price process. They show that a funding constraint, along the lines of (24), prevents drastically intermediaries from exploiting opportunities that appear to be perfect arbitrages. Other papers on optimal portfolio policy under funding constraints are Grossman and Vila (1992), Jurek and Yang (2007) and Milbradt (2011).

A number of papers study the effects of funding constraints in the presence of multiple risky assets. Brunnermeier and Pedersen (2009) show in a static setting that funding constraints generate not only amplification, but also contagion, whereby shocks to one asset are transmitted to otherwise unrelated assets through changes in intermediaries' positions. Moreover, a tightening of funding constraints has the largest impact on the prices of more volatile assets because these require more collateral. Pavlova and Rigobon (2008) derive contagion in a dynamic international-economy setting with portfolio constraints, of which funding constraints are a special case. Gromb and Vayanos (2011a, 2011b) derive the joint dynamics of intermediary capital, asset volatility, correlations and liquidity. They show that amplification and contagion are stronger when intermediary capital is neither too high nor too low. Related results are shown in Danielsson, Shin, and Zigrand (2011), who derive funding constraints from value-at-risk requirements of banks.

Amplification and contagion can also be derived in models without explicit funding constraints but where risk aversion depends on wealth. This is done in Kyle and Xiong (2001) and Xiong (2001), who endow intermediaries with logarithmic utility, under which the coefficient of absolute risk aversion decreases in wealth. Following adverse shocks, intermediaries reduce their positions because they become more risk averse and not because they hit funding constraints. The analysis has similarities to that with funding constraints, e.g., amplification and contagion are stronger when intermediary capital is neither too high nor too low. An important difference is in the welfare and policy implications: funding constraints can create inefficiencies and the scope for welfare-improving policies, while wealth effects preserve the Pareto optimality of equilibrium.

Early work on the limits of arbitrage does not consider funding constraints explicitly, but argues that such constraints can shorten traders' horizons, and this can affect asset prices. De Long, Shleifer, Summers, and Waldmann (1990) show that short horizons can cause deviations from the law of one price. They assume an infinite-horizon economy, two assets with identical payoffs, and stochastic shocks to the demand for one of the assets. They show that when traders have short horizons there exist two equilibria: one in which the assets trade at the same price and one in which they trade at different prices. The intuition for the latter equilibrium is that agents do not trade aggressively against price discrepancies between the two assets for fear that they might widen in the short run. As a consequence, demand shocks can cause price discrepancies and render traders' belief self-fulfilling.

Tuckman and Vila (1992, 1993) show that short horizons can arise endogenously because of holding costs. Moreover, holding costs can render traders unwilling to exploit price discrepancies between assets with similar payoffs for fear that they might widen in the short run. Dow and Gorton (1994) assume short horizons and show that holding costs can generate large mispricings. Casamatta and Pouget (2011) endogenize short horizons based on moral hazard between fund managers and investors, and show that they cause prices to be less informative.

Shleifer and Vishny (1997) model the reliance of traders on external capital and its implications for traders' horizons and asset pricing. They assume that traders can buy a underpriced asset but run the risk that the mispricing might worsen in the short run. Traders can raise external funds to buy the asset, but the suppliers of the funds can request them back if the trade performs poorly in the short run. This assumed performance-flow relationship can generate amplification effects: following demand shocks that cause the mispricing to worsen in the short run, traders are deprived of funds and must sell the asset, causing the mispricing to worsen further.

Shleifer and Vishny (1997) derive the funding constraint from equity finance: traders can be interpreted as managers of an open-end fund raising equity from fund investors. Yet, the amplification effects that they find are similar to those in the papers that derive funding constraints from collateralized debt. Recent work on the limits of arbitrage seeks to derive funding constraints from optimal contracts, instead of assuming an exogenous contract form. Examples are Acharya and Viswanathan (2011), He and Krishnamurthy (2011), Hombert and Thesmar (2011), and Biais, Heider, and Hoerova (2012). Endogenizing the constraints would help identify whether the common results, such as amplification, are driven by a single underlying friction, or whether the constraints are fundamentally different. Recent work on the limits of arbitrage also seeks to develop tractable dynamic multi-asset models that can address empirical puzzles. The survey by Gromb and Vayanos (2010a) provides more details and references.

Funding constraints can interact with other market imperfections. Yuan (2005) and Albagli (2011) consider the interaction with asymmetric information, and impose funding constraints on informed agents. Yuan (2005) shows that when prices drop, informed agents become constrained and hence prices become less informative. The resulting increase in uncertainty exacerbates the price drop, causing volatility to be asymmetric and higher on the downside. Albagli (2011) derives multiple equilibria, through a mechanism that is reminiscent of De Long, Shleifer, Summers, and Waldmann (1990) but does not require an infinite horizon. When future demand shocks are expected to have a large effect on prices, funding-constrained agents do not trade aggressively on their information. This makes prices less informative, hence reducing the willingness of future agents to absorb demand shocks.

Diamond and Verrecchia (1987) and Bai, Chang, and Wang (2006) study the interaction between asymmetric information and short-sale constraints, which are related to funding constraints. Diamond and Verrecchia (1987) show that short-sale constraints prevent investors with negative private signals from trading. But even though only investors with positive signals are trading, the market adjusts for this, and short sales do not cause overpricing. Bai, Chang, and Wang (2006) show that short-sale constraints can instead cause underpricing because they generate uncertainty about the extent of negative private information.

2.7 Search

In the perfect-market benchmark, the market is organized as a centralized exchange. Many markets, however, have a more decentralized form of organization. For example, in over-the-counter markets, investors negotiate prices bilaterally with dealers. Locating suitable counter-

parties in these markets can take time and involve search.

To model decentralized markets, we assume that agents do not meet in a centralized exchange in Period 1, but instead must search for counterparties. When a liquidity demander meets a supplier, they bargain bilaterally over the terms of trade, i.e., the number of shares traded and the share price. We assume that bargaining leads to an efficient outcome, and denote by $\phi \in [0,1]$ the fraction of transaction surplus appropriated by suppliers. We denote by N the measure of bilateral meetings between demanders and suppliers. This parameter characterizes the efficiency of the search process, and is bounded by $\min\{\pi, 1-\pi\}$ since there cannot be more meetings than demanders or suppliers. Assuming that all meetings are equally likely, the probability of a demander meeting a supplier is $\pi^d \equiv N/\pi$, and of a supplier meeting a demander is $\pi^s \equiv N/(1-\pi)$.

We first describe the equilibrium, then examine how the search friction affects the search illiquidity measures and the expected return, and finally survey the literature on search frictions. Prices in Period 1 are determined through pairwise bargaining between liquidity demanders and suppliers. Agents' outside option is not to trade and retain their positions from Period 0, which in equilibrium are equal to $\bar{\theta}$. The consumption in Period 2 of a liquidity supplier who does not trade in Period 1 is $C_2^{sn} = W_0 + \bar{\theta}(D - S_0)$. This generates a certainty equivalent

$$CEQ^{sn} = W_0 + \bar{\theta}(\bar{D} - S_0) - \tfrac{1}{2}\alpha\sigma^2\bar{\theta}^2, \tag{29}$$

where the first two terms are the expected consumption, and the third a cost of bearing risk that is quadratic in position size. If the supplier buys x shares at price S_1, the certainty equivalent becomes

$$CEQ^{s} = W_0 + \bar{\theta}(\bar{D} - S_0) + x(\bar{D} - S_1) - \tfrac{1}{2}\alpha\sigma^2(\bar{\theta} + x)^2 \tag{30}$$

because the position becomes $\bar{\theta} + x$. Likewise, the certainty equivalent of a liquidity demander who does not trade in Period 1 is

$$CEQ^{dn} = W_0 + \bar{\theta}(\bar{D} - S_0) - \tfrac{1}{2}\alpha\sigma^2(\bar{\theta} + z)^2, \tag{31}$$

and if the demander sells x shares at price S_1, the certainty equivalent becomes

$$CEQ^{d} = W_0 + \bar{\theta}(\bar{D} - S_0) - x(\bar{D} - S_1) - \tfrac{1}{2}\alpha\sigma^2(\bar{\theta} + z - x)^2. \tag{32}$$

Under efficient bargaining, x maximizes the sum of certainty equivalents $CEQ^s + CEQ^d$. The maximization yields $x = z/2$, i.e., the liquidity shock is shared equally between the two agents.

The price S_1 is such that the supplier receives a fraction ϕ of the transaction surplus, i.e.,

$$CEQ^s - CEQ^{sn} = \phi \left(CEQ^s + CEQ^d - CEQ^{sn} - CEQ^{dn} \right). \tag{33}$$

Substituting (29)-(32) into (33), we find that the price in Period 1 is

$$S_1 = \bar{D} - \alpha\sigma^2 \left[\bar{\theta} + \tfrac{1}{4} z(1+2\phi) \right]. \tag{34}$$

Eq. (34) implies that the impact of the liquidity shock z on the price in Period 1 increases in the liquidity suppliers' bargaining power ϕ. When, for example, $z > 0$, liquidity demanders need to sell, and greater bargaining power by suppliers results in a lower price. Comparing (34) to its centralized-market counterpart (3) reveals an important difference: price impact in the search market depends on the distribution of bargaining power within a meeting, characterized by the parameter ϕ, while price impact in the centralized market depends on aggregate demand-supply conditions, characterized by the measures $(\pi, 1-\pi)$ of demanders and suppliers. The price in the centralized market in Period 0 is given by the same equation (4) as in the perfect-market benchmark, but for different constants (M, Δ_1).

We next examine how the search friction impacts the illiquidity measures and the expected return. We perform two related but distinct exercises: compare the search market with the centralized market of Section 2.1, and vary the measure N of meetings between liquidity demanders and suppliers.

When N decreases, the search process becomes less efficient and trading volume decreases. At the same time, the price in each meeting remains the same because it depends only on the distribution of bargaining power within the meeting. Since λ measures the price impact of volume, it increases. One would conjecture that λ in the search market is higher than in the centralized market because only a fraction of suppliers are involved in bilateral meetings and provide liquidity ($N \leq 1 - \pi$). Result 14 confirms this conjecture when bargaining power is symmetric ($\phi = 1/2$). The conjecture is also true when suppliers have more bargaining power than demanders ($\phi > 1/2$) because the liquidity shock has then larger price impact. Moreover, the result extends to all values of ϕ when less than half of suppliers are involved in meetings ($N \leq (1-\pi)/2$).

Result 14 *Price impact λ is*

$$\lambda = \frac{\alpha\sigma^2(1+2\phi)}{2N}, \tag{35}$$

and increases when the measure N of meetings decreases. It is higher than in the centralized market if $\phi + 1/2 \geq N/(1-\pi)$.

Because the price in the search market is independent of N, so is the price reversal γ. Moreover, γ in the search market is higher than in the centralized market if ϕ is large relative to π.

Result 15 *Price reversal γ is*

$$\gamma = \frac{\alpha^2 \sigma^4 \sigma_z^2 (1+2\phi)^2}{16}, \tag{36}$$

and is independent of the measure N of meetings. It is higher than in the centralized market if $\phi + 1/2 \geq 2\pi$.

When the measure N of meetings decreases, agents are less likely to trade in Period 1, and a natural conjecture is that the illiquidity discount increases. Result 16 confirms this conjecture under the sufficient condition $\phi \leq 1/2$. Intuitively, if $\phi \approx 1$, a decrease in the measure of meetings does not affect liquidity demanders because they extract no surplus from a meeting. Since, however, liquidity suppliers become worse off, the risk-neutral probability of being a demander decreases, and the price can increase.[10]

Result 16 *A decrease in the measure N of meetings lowers the price in Period 0 if $\phi \leq 1/2$.*

Early work modeling search frictions in asset markets and their implications for equilibrium prices includes Burdett and O'Hara (1987), Pagano (1989b) and Keim and Madhavan (1996). These papers focus on the market for large blocks of shares (known as the "upstairs" market in the New York Stock Exchange).

Duffie, Garleanu and Pedersen (2002, 2005, 2008) model price formation in asset markets building on the search framework of Diamond (1982), Mortensen (1982) and Pissarides (1985), in which a continuum of agents negotiate prices in bilateral meetings over an infinite horizon and continuous time. Duffie, Garleanu, and Pedersen (2002) focus on the repo market, where traders can borrow or lend assets. In a centralized market with no frictions, lenders of positive-supply assets would compete their rent down to zero. Indeed, equilibrium requires that some

[10]The illiquidity discount in the search market is higher than in the centralized market if ϕ is large relative to π. This property is the same as for λ and γ, but the calculations are more complicated.

agents hold the assets, and hence would be willing to lend them as long as they earn any non-zero rent. With search frictions, however, lenders can earn a rent because they can extract some of the borrowers' surplus when bargaining in bilateral meetings. The rent is an additional payoff from holding the assets and raises their price in the spot market.

Duffie, Garleanu, and Pedersen (2008) focus on the spot market and assume that the valuation of agents for a risky asset switches over time between high and low. Agents with high valuation who do not own the asset seek to buy it. Conversely, agents with low valuation who own the asset seek to sell it. The equilibrium prices that emerge in the bilateral meetings depend not only on the measures of buyers and sellers, as in a centralized market, but also on their relative bargaining power. Our model yields an extreme version of this result: the price in Period 1 depends only on the bargaining power parameter ϕ and not on the measures $(\pi, 1-\pi)$ of liquidity demanders and suppliers. An implication of this result is that an increase in search frictions can raise or lower the asset price, with the price decreasing when there are more buyers than sellers. Indeed, with larger frictions, the price responds less to the aggregate demand/supply conditions, and hence decreases when these conditions are favorable to the sellers. Finally, following a positive shock to the measure of sellers, which moves the market away from steady state, prices drop and recover gradually with the drop being larger when frictions increase.

Duffie, Garleanu, and Pedersen (2005) introduce market makers who intermediate trade. Market makers differ from other agents, who we term investors, because they can be contacted more easily. If investors are better able to contact each other, then market makers face more competition and post lower bid-ask spreads. Moreover, if investors are heterogeneous in their ability to contact market makers, then market makers post lower spreads for investors with higher such ability. Weill (2007) studies the dynamics of an intermediated search market away from steady state. He shows that following a positive shock to the measure of sellers, market makers build up inventories, which they gradually unload to buyers. Market makers acquire the asset despite having lower valuation for it than other agents because they are more efficient in passing it to the buyers.

Vayanos and Wang (2007) and Weill (2008) extend the analysis to multiple assets, and show that search frictions can generate price discrepancies between assets with identical payoffs. Buyers choose one of two assets to search for, and then can only meet sellers of that asset. In equilibrium, they can locate one asset more easily, and are hence willing to pay a higher price for it. The asset that is easier to locate has a higher number of sellers either because it attracts endogenously high-turnover agents in Vayanos and Wang (2007), or because it is in larger supply

in Weill (2008). Note that one-asset models, such as Duffie, Garleanu, and Pedersen (2008), yield the opposite prediction that assets in larger supply trade at lower prices.

Vayanos and Weill (2008) show that deviations from the law of one price can arise even under simultaneous search, i.e., buyers can meet sellers of all assets. Key to this result is the presence of short sellers, who borrow an asset in the repo market, then sell it in the spot market, and then buy it back again to unwind the short sale. In equilibrium, short sellers endogenously concentrate in one asset, making it more liquid. That asset trades at a higher price because its superior liquidity is priced by the longs, i.e., the buyers who seek to establish long positions. Moreover, the higher concentration of short-sellers in one asset makes it profitable for longs to lend the asset in the repo market, and further raises its price as in Duffie, Garleanu, and Pedersen (2002).

A number of papers relax the assumption that agents can hold zero or one unit of an asset. Garleanu (2009) and Lagos and Rocheteau (2009) show that an increase in search frictions makes agents less willing to change their positions in response to short-run shocks to their valuation for the asset. This is because they are aware that it will take them time to change their positions back should an offsetting shock hit. Since agents become less responsive to shocks in either direction, search frictions have an ambiguous effect on the price, consistent with Duffie, Garleanu, and Pedersen (2008). Lagos, Rocheteau, and Weill (2011) study the effects of shocks that move the market away from steady state, and show that the speed of recovery is non-monotonic in search frictions. Afonso and Lagos (2011) study price formation in the interbank market, and determine how the Federal Funds Rate depends on the search frictions and on Federal Reserve policy actions.

Search models emphasize the idea that matching buyers and sellers takes time. In their work on participation costs, Grossman and Miller (1988) model a related idea: a liquidity shock experienced by some agents is absorbed first by a small set of market participants and then by a larger set of agents who gradually arrive in the market. The market participants who first absorb the shock act as intermediaries, building up inventories and then unwinding them. Search models provide a natural setting to study the process through which assets are reallocated across agents via the temporal variation in intermediaries' inventories. This is done, for example, in Weill (2007), where intermediaries are modeled as a special class of agents who can be contacted more easily than others. It is also done in Afonso and Lagos (2011), where agents engage endogenously in intermediation when they meet others with large liquidity shocks: they absorb more than their final share of a shock knowing that they can unload it to others in future bilateral meetings. Duffie and Strulovici (2011) model the process through which new

agents slowly become informed about liquidity shocks in one market and bring their capital into that market. Mitchell, Pedersen, and Pulvino (2007) and Duffie (2010) emphasize the idea that capital moves slowly across markets in response to profitable investment opportunities.

All papers mentioned so far assume that agents have symmetric information about the asset payoff. If some agents receive private signals, then these can be revealed gradually through the bilateral meetings, as agents learn the information of those they meet and of those that their meeting partners have met in the past. Papers studying the transmission of private information in decentralized markets include Wolinksy (1990), Blouin and Serrano (2001), Duffie and Manso (2007), Duffie, Malamud, and Manso (2009), Golosov, Lorenzoni, and Tsyvinski (2011) and Zhu (2011).

Finally, some papers study portfolio choice under the assumption that agents can trade only after a lag, which could reflect unmodeled search frictions or market breakdowns. For example, Longstaff (2001) restricts trading strategies to be of bounded variation, while Ang, Papanikolaou, and Westerfield (2011) assume that investors can trade only at exogenous random times. Both papers take prices as given and compute the utility loss from infrequent trading. This exercise is in the spirit of the one performed in Constantinides (1986) in the case of transaction costs, but the utility loss is larger in the case of infrequent trading. Longstaff (2009) shows in an equilibrium model that infrequent trading has large effects on asset prices.

3 Empirical Evidence

In this section we survey the empirical literature on market liquidity. In Section 3.1 we present various measures of illiquidity, and discuss their inter-relationships from a theoretical viewpoint. We also use the theory to examine how well these measures reflect the underlying market imperfections. In Section 3.2 we survey the empirical evidence on how illiquidity measures vary across assets and over time, how they relate to asset characteristics such as supply and volatility, and how they relate to each other. In Section 3.3 we examine the link between illiquidity measures and expected asset returns.

3.1 Empirical Measures of Illiquidity

Empirical papers employ a wide variety of illiquidity measures. Some measures, such as lambda and price reversal, are motivated by theory, while others, such as bid-ask spread, market depth, turnover and trade size, are more intuitive or heuristic. Within our unified model, we compute

lambda and price reversal, and study their relationship with a variety of market imperfections. Price impact λ is defined as the regression coefficient of returns on signed volume, and is based on the idea that trades in illiquid markets should have large price impact. Price reversal γ is defined as minus the autocovariance of returns, and is based on the idea that trades in illiquid markets should generate transitory deviations between price and fundamental value. The measures λ and γ capture two fundamental and distinct aspects of illiquidity, so we start by surveying the empirical work on them. We then review a number of additional measures.

An early paper that emphasizes the link between systematic price reversals and market imperfections is Niederhoffer and Osborne (1966). Using stock ticker data, they find that prices in consecutive transactions revert on average, and argue that these reversals are linked to the mechanics of the market making process. Roll (1984) shows that the bouncing of prices between the bid and the ask causes systematic reversals, and derives analytically an increasing relationship between γ and the bid-ask spread. Based on this relationship, he argues that γ can be used to estimate the bid-ask spread when data on the latter are not available. He computes this estimator using daily and weekly returns. One difficulty with the estimation is that γ can occasionally be negative because of, e.g., sampling noise, while a positive bid-ask spread implies a positive γ. Harris (1990) derives statistical properties of the estimator in Roll (1984), and shows that its small-sample bias can be large. Hasbrouck (2009) derives a Bayesian Gibbs estimator, which is in the spirit of Roll (1984) but has better statistical properties. Bao, Pan, and Wang (2011) estimate γ for corporate bonds.

Campbell, Grossman, and Wang (1993) show that the autocovariance of daily stock returns is more negative when trading volume is large. They explain this finding theoretically in a model similar to Grossman and Miller (1988), in which risk-averse market makers absorb liquidity shocks. Price movements without volume are caused by fundamentals, while movements with volume are caused by buying or selling pressure. Since the latter movements are transitory, price reversals are larger conditional on volume.

The analysis in Campbell, Grossman, and Wang (1993) suggests that illiquidity should be measured by γ conditional on trading volume. This measure is more precise than the unconditional γ because it focuses more sharply on the effects of price pressure. Llorente, Michaely, Saar, and Wang (2002) show that conditional γ correlates negatively with measures of asymmetric information, consistent with the theoretical model of Wang (1994). As in Campbell, Grossman, and Wang (1993), conditional γ is derived as minus the coefficient of a regression of a stock's daily return on the product of volume times return on the previous day, controlling for previous-day return. Pastor and Stambaugh (2003) define a similar measure, but use returns in

excess of a market index instead of raw returns, and volume signed by return instead of volume times return. They aggregate their measure across stocks to construct a proxy for aggregate illiquidity, which they show is a priced risk factor.

Refined measures of γ can be derived using data not only on trading volume but also on market maker inventories. Indeed, since market makers absorb liquidity shocks, changes to their inventories away from the long-run average should signal transitory price movements and hence predict future returns. Even more importantly, data on inventories can help determine the horizon of liquidity effects and hence better estimate those effects. Indeed, computing γ with daily returns assumes implicitly that the horizon is one day. The horizon, however, could be longer if, for example, a liquidity shock is generated by a trader breaking a large trade into many smaller ones over several days (a phenomenon shown theoretically in models of imperfect competition, surveyed in Section 2.5, and empirically in, e.g., Chan and Lakonishok (1993)). Data on inventories could help identify such effects. Ho and Macris (1980) and Madhavan and Smidt (1993) show that changes in the inventories of market makers in the stock market correlate positively with contemporaneous movements in stocks' transaction prices. Hendershott and Menkveld (2011) use inventories to predict future stock returns.

Kyle (1985) defines price impact λ and links it to the degree of asymmetric information. While price impact in Kyle (1985) is permanent because market makers are risk neutral and competitive, in general it consists of a transitory and a permanent component. The transitory component measures profits earned by market makers. These can be a compensation for risk, information asymmetry, transaction costs or participation costs, or rents from monopoly power. Glosten and Harris (1988) attempt to separate the permanent and transitory component using stock transaction data. They observe only trade size, and identify whether a trade is buyer- or seller-initiated through the Lee and Ready (1991) algorithm, i.e., trades taking place above the midpoint of the bid-ask spread are classified as buyer-initiated while those taking place below are seller-initiated. Because they observe only transactions and not the bid-ask spread, they use maximum likelihood to infer buyer- and seller-initiated trades and to estimate the parameters of their model. They further allow each of the transitory and permanent components to be affine in transaction size. The linear part of the sum of the two affine functions is λ, and the fixed (constant) part is related to the bid-ask spread. Sadka (2006) estimates a similar model using a larger cross-section of stocks and data that include not only transactions but also quotes. This allows for a more precise estimation of the model parameters, e.g., the variable and fixed parts of the permanent and transitory components.

Measures related to λ can be constructed by comparing volume and absolute returns.

An early such measure is Aminvest (e.g., Dubofsky and Groth (1984)) which divides volume aggregated over twenty trading days by the sum of absolute values of returns during those days. More recently, Amihud (2002) divides absolute daily return by daily volume and averages over a number of days. Since high price impact means high absolute return per unit of volume, it is associated with a low value of the Aminvest measure and a high value of the Amihud measure. An advantage of the two measures over the estimates of λ in Glosten and Harris (1988) and Sadka (2006) is that they can be constructed using only daily data on volume and returns, rather than intraday data on returns, transactions, and quotes.

Because λ is the coefficient of a linear regression, it measures the average slope of the price impact function, which can be non-linear. Loeb (1983) and Keim and Madhavan (1996) find that the function is concave, i.e., price impact per unit trade is smaller for large orders. Earlier studies of the price effect of large orders include Kraus and Stoll (1972) and Holthausen, Leftwich, and Mayers (1990).

How do γ-based measures compare to λ-based ones? Each of γ and λ captures a fundamental and distinct aspect of illiquidity, so the two measures are not substitutes but can be taken together to give a more complete picture. At the same time, each measure has some advantages over the other, from the viewpoint of empirical estimation or theory.

Estimating γ is simpler than λ for several reasons. First, γ requires only data on returns, while λ requires additional data on transactions and quotes. Second, λ can fail to reflect a causal effect of volume of prices. For example, if prices or other public news cause volume, then λ can be positive even if volume has no causal effect on prices.[11] Third, even if causality goes only from volume to prices, λ requires identifying the signed trades of liquidity demanders. Such identification is typically done using the Lee and Ready (1991) algorithm (e.g., Glosten and Harris (1988) and Sadka (2006)), but the algorithm can misclassify trades. For example, Asquith, Oman, and Safaya (2010) find that short sales are often classified mistakenly as buyer-initiated because of a requirement that they are executed at an uptick or above the existing bid. This can underestimate the short-sales' price impact.

Identifying the signed trades of liquidity demanders becomes easier when data on the identity of traders are available. Such data often exist for designated market makers, present in many markets. Under the assumption that market makers mostly supply rather than demand liquidity, signed trades of liquidity suppliers can be identified. Examples of papers that estimate

[11] The causality problem does not arise in our unified model. Indeed, volume is generated by shocks observable only to liquidity demanders, such as the liquidity shock z and the signal s. Since these shocks can affect prices only through the liquidity demanders' trades, λ measures correctly the price impact of these trades.

price impact using data on market maker trades or inventories are Hendershott and Menkveld (2011) for stocks and Edwards, Harris, and Piwowar (2007) for corporate bonds. As is the case of γ, data on inventories can help determine the horizon of liquidity effects and hence better estimate those effects.

The difficulties in estimating λ should be set against an important theoretical advantage of that measure. This is that λ captures not only the transitory component of price impact, as does γ, but also the permanent component, driven by the information that trades convey. The latter component is an important aspect of illiquidity and should not be ignored. The model of Kyle (1985) provides a simple illustration of this point. Since market makers are assumed risk neutral and competitive, γ is equal to zero. Yet, this does not mean that liquidity is perfect: a trader entering into the market would suffer from price impact. The imperfect liquidity is reflected correctly in λ, which is positive.

Our unified model delivers a similar message concerning the theoretical advantage of λ. While participation costs, transaction costs, and funding constraints increase both λ and γ (Results 1, 2, 3, 11 and 12), we show that λ reflects better than γ the remaining imperfections. Indeed, both asymmetric information and imperfect competition increase λ (Results 5 and 8) but can decrease γ (Results 6 and 9). And while search frictions can decrease both λ and γ, the conditions for λ to decrease are more stringent than for γ (Results 14 and 15).

An additional measure of illiquidity is the probability of informed trading (PIN), introduced by Easley, Kiefer, and O'Hara (1997). As with λ and γ, PIN is motivated by theory: it is derived from a structural estimation of the Easley and O'Hara (1992) model of sequential order arrival. PIN is high when there is a large imbalance between buyer- and seller-initiated trades, as classified by the Lee and Ready (1991) algorithm. This follows from two assumptions on liquidity traders: they are equally likely to trade in either direction, and their trades are independent over time. These assumptions imply that long sequences of buying or selling are more likely to come from informed traders. The assumption that liquidity trades are independent over time is strong: for example, a liquidity trader can break a large trade into a long sequence of small trades in the same direction. Hence, it is important to investigate the performance of PIN outside the Easley and O'Hara (1992) model and for a variety of market imperfections—a task that we undertake for λ and γ in this survey. Some results in this direction are in Buss and Dumas (2011), who show that PIN is positive in a symmetric-information model with transaction costs.

An intuitive and widely used measure of illiquidity is the bid-ask spread. Several versions of this measure exist. One is the quoted spread, defined as the difference between the quoted

ask and bid prices. A drawback of quoted spread is that many trades are executed inside the spread, i.e., at more favorable prices. A measure remedying this drawback is the effective spread, defined as the difference between transaction price and mid-point of the quoted spread. This difference is taken in absolute value, and is multiplied by two so that it is expressed in the same terms as the quoted spread. A third measure is the realized spread, defined as the reversal between consecutive (or near-consecutive) transactions: the price of the current transaction minus of the future transaction if the current transaction is above the mid-point, and the opposite if the current transaction is below. This difference is multiplied by two, as with the effective spread. The realized spread measures profits earned by liquidity suppliers. These can be a compensation for risk, transaction costs or participation costs, or rents from monopoly power. If instead liquidity suppliers are risk neutral, competitive, and incur no costs, then the realized spread is zero. The effective spread can still be non-zero, however, because of asymmetric information, as in Glosten and Milgrom (1985). A detailed discussion of quoted, effective, and realized spread is in Huang and Stoll (1996), who compare the liquidity of the New York Stock Exchange (NYSE) and the NASDAQ based on these measures. Biais, Glosten, and Spatt (2005) develop a unified model showing how the bid-ask spread is affected by asymmetric information, market maker risk aversion, order-processing costs, and imperfect competition. The bid-ask spread in their model is closely related to λ. An earlier paper linking the bid-ask spread to many of the same variables is Stoll (1978b).

While the bid-ask spread is an intuitive measure of illiquidity, it has some limitations. First, its estimation requires detailed data on transactions and quotes. Second, because the spread is valid only for transactions up to a certain size, it provides no information on the prices at which larger transactions might take place. By the same token, it provides no information on how the market might respond to a long sequence of transactions in the same direction. Such a sequence could be generated, for example, by a trader breaking a large trade into many smaller ones, and could span several days. The market's response to large buying or selling pressure is an important aspect of illiquidity. The measures λ and γ capture that aspect better than the bid-ask spread, especially if returns are computed over a horizon matching that of the liquidity effects. The bid-ask spread is related to these measures but imperfectly as we show in Section 3.2. This could be because the relative effects of the various market imperfections on the measures can differ. For example, in line with the theoretical analysis of Section 2, the bid-ask spread could be more affected by order-processing costs, while λ and γ could be influenced more heavily by the risk aversion of liquidity suppliers.

A measure of illiquidity which is related to the bid-ask spread is market depth, defined as the quantity of limit orders at the bid and the ask. Higher market depth is a sign of higher

liquidity since the bid and ask prices are valid for larger transaction sizes. The limitations of market depth are similar to those of the bid-ask spread: it is data intensive, and it gives only a partial picture of the market's response to large buying or selling pressure.

In addition to the measures of illiquidity described so far, empirical papers have also used a number of more heuristic measures. Some of these measures relate to trading activity. For example, Bhushan (1994) measures illiquidity by the inverse of trading volume. Lesmond, Ogden, and Trzcinka (1999) propose the LOT measure, which is based on the number of non-trading days. Mahanti, Nashikkar, Subrahmanyam, Chacko, and Mallik (2008) propose "latent liquidity," defined as the turnover of investors holding an asset weighted by their asset holdings. The idea behind this measure is that if investors with large holdings trade frequently, then they can absorb large shocks on short notice. Other measures relating to trading activity include turnover, trade size, trade frequency, and number of non-trading days. Fleming (2003) uses trading volume and trade size, alongside bid-ask spread and price impact, to measure the illiquidity of Treasury bonds. Chen, Lesmond, and Wei (2007) use LOT, number of non-trading days, and bid-ask spread to measure the illiquidity of corporate bonds.

Measuring illiquidity by the inverse of trading activity can be motivated by theory. Indeed, many of the models surveyed in Section 2 yield the intuitive result that market imperfections reduce trading volume.[12] Moreover, models with transaction costs, surveyed in Section 2.3, show that the costs give rise to a no-trading region and hence reduce trading frequency. But while a negative effect of market imperfections on trading activity is intuitive, using the inverse of trading activity to measure illiquidity has some drawbacks. First, trading activity does not provide a direct estimate of the costs of trading. A direct estimate requires translating trading activity into a measure of costs, as is done in the case of LOT. Second, trading activity can be influenced by variables other than the market imperfections, such as the supply of an asset, the number of investors holding it, and the size of their trading needs. These variables can imply the opposite relationship between illiquidity, as measured by λ and γ, and trading activity. For example, the parameter σ_z^2, which characterizes the trading needs of investors in our unified model, increases both trading volume and γ. Hence, it implies a positive relationship between illiquidity, as measured by γ, and trading activity. We return to this issue in Section 3.2, where we discuss the relationship between measures of illiquidity and asset characteristics.

Other heuristic measures of illiquidity include asset characteristics such as the supply of

[12]The result does not arise in all the models. In Kyle (1985), for example, asymmetric information increases volume because it triggers trades by the insider but has no effect on the activity of noise traders, which is exogenous. Endogenizing noise trading, however, can restore the negative effect of asymmetric information on volume. An extreme version of this result is the possibility of market breakdown shown in Section 2.5.

an asset, the volatility of its payoffs, and its time from issuance. For example, Karpoff and Walkling (1988) measure the illiquidity of stocks under the assumption that it is positively related to price volatility and negatively related to the price level, market capitalization, and number of shares. Bao, Pan, and Wang (2011) use issuance size, time from issuance, time to maturity, rating, and γ to gauge the illiquidity of corporate bonds. These measures are imperfect proxies of illiquidity, and do not provide direct estimates of the costs of trading. Their use, however, can be motivated by theory in some cases. For example, as we point out in Section 3.2, volatility typically increases both λ and γ.

An aspect of liquidity that most of the measures presented so far do not capture is market resiliency, the speed at which liquidity recovers from shocks. Consider, for example, the dynamics of liquidity in a limit-order market. When a market order hits the existing limit orders, it causes the price to change. Moreover, as existing limit orders are consumed, new orders arrive gradually and liquidity recovers. Most measures of illiquidity do not capture this gradual recovery; price reversal γ does but only to a limited extent because it measures the reversal of price rather than of liquidity. Obizhaeva and Wang (2006) show that the intraday dynamics of liquidity in limit-order markets have an important effect on the optimal execution of large orders. Biais, Hillion, and Spatt (1995) study these dynamics empirically and show that market resiliency is finite. Other empirical studies of intraday resiliency include Coppejans, Domowitz, and Madhavan (2004), Degryse, De Jong, Van Ravenswaaij, and Wuyts (2005) and Dong, Kempf, and Yadav (2007). The survey by Parlour and Seppi (2008) contains more references to empirical work on limit-order markets.

The measures of illiquidity presented so far concern an individual asset. Illiquidity, however, can vary over time in a correlated manner across assets and markets, as we show in Section 3.2. Hence, it is useful to also measure it at a more aggregate level. A number of papers, surveyed in Section 3.3, measure the aggregate illiquidity of an asset class by averaging the measures presented so far over individual assets within the class. Yet, it is possible to construct more direct measures. For example, Longstaff (2004) uses the yield spread between RefCorp and US Treasury bonds to measure the "flight to liquidity" premium. Since both types of bonds are guaranteed by the US Federal government, the spread should arise mainly from their relative liquidity.

Fontaine and Garcia (2011) and Hu, Pan, and Wang (2011) measure the illiquidity of the US Treasury market by the deviations of observed Treasury yields from a fitted term structure. The logic behind their measures derives from funding constraints and limited arbitrage, which we study theoretically in Section 2.6. When arbitrageurs are well capitalized, they can supply

ample liquidity, and so can eliminate deviations between Treasury yields and their fundamental values which are assumed to lie on the fitted term structure. When instead capital is scarce, liquidity is imperfect and substantial deviations can appear.

Fontaine and Garcia (2011) assume that the residuals from the fit depend on a latent liquidity factor, which they estimate under parametric assumptions on how the residuals depend on that factor and on a bond's maturity and time from issuance. They show that consistent with the funding-constraint interpretation, the illiquidity of the Treasury market is negatively related to the asset size of the shadow banking sector. Hu, Pan, and Wang (2011) use the sum of squared residuals as a measure of illiquidity, and not just for the Treasury market but also for the overall financial market. They show that their measure rises sharply during crises even outside the Treasury market. Thus, the illiquidity of Treasuries seems to reflect broader market-wide conditions. Consistent with this evidence, Hu, Pan, and Wang (2011) find that their measure is a priced risk factor for a broad set of assets.

3.2 Properties of Illiquidity Measures

In this section we survey empirical evidence on the cross-section and time-series variation of illiquidity measures. To set the stage for what follows, we report in Table 1 means, medians, and standard deviations for some of the measures presented in Section 3.1. These are taken from Goyenko, Holden, and Trzcinka (2009), and concern 400 randomly selected US stocks over 1993-2005. Each measure is computed for each stock and month, and the summary statistics concern stock-month observations.

The effective and realized spread are expressed as (approximate) percentages of the mid-point because prices are in logs. The median effective spread is 1.6%, implying an average difference between transaction price and mid-point of 0.8%. The mean effective spread is 2.9%, significantly higher than the median because of a large tail of stocks with low liquidity. The median realized spread is about one third of the median effective spread, meaning that one third of the effective spread reflects a transitory component associated to price reversal. Note that the reversal is measured over a very short horizon of five minutes, and can fail to capture reversal over longer horizons.

The price impact λ is computed as in Hasbrouck (2009): it is the coefficient of a regression of the return over a five-minute interval on the signed squared-root dollar volume during that interval. Signed squared-root dollar volume is the sum over transactions of the squared root of the dollar transaction size times an indicator which is one if the transaction is below the

Table 1: Illiquidity measures for US stocks.

Effective spread is the difference between transaction price and mid-point of the quoted spread, in absolute value and multiplied by two. Realized spread is the price of a current transaction minus that of a transaction five minutes later if the current transaction is above the mid-point, and the opposite if the current transaction is below. This difference is multiplied by two. Price impact λ is the coefficient of a regression of the return over a five-minute interval on the signed squared-root dollar volume during that interval. Signed squared-root dollar volume is the sum over transactions of the squared root of the dollar transaction size times an indicator which is one if the transaction is below the mid-point and minus one if it is above. Realized price impact is the current mid-point minus the mid-point five minutes later if the current transaction is above the current mid-point, and the opposite if the current transaction is below. The Amihud measure is an average of daily ratios of absolute return over volume. The Pastor-Stambaugh measure is minus the coefficient of a regression of a stock's daily return in excess of the market index on the previous day's signed volume, controlling for the previous day's return. Volume on a given day is signed using the excess return on that day. Prices are in logs. The sample consists of 400 randomly selected stocks traded in NYSE/NASDAQ/AMEX over 1993-2005. All averages and regression coefficients are computed for each stock and month, and the summary statistics concern stock-month observations. The data in the table are taken from Goyenko, Holden, and Trzcinka (2009), but we define Pastor-Stambaugh with the opposite sign.

Measure	Mean	Median	Standard deviation
Effective Spread	0.029	0.016	0.040
Realized Spread	0.015	0.005	0.032
Price impact λ	130.425×10^{-6}	15.793×10^{-6}	2446.202×10^{-6}
Realized price impact	0.016	0.010	0.038
Amihud	6.314×10^{-6}	0.104×10^{-6}	91.957×10^{-6}
Pastor-Stambaugh	0.179	0.000	10.129

mid-point and minus one if it is above. The median price impact is 15.793. To map this into an actual price change, suppose that there is only one transaction during the five-minute interval and it is for \$10000. The price change then is $15.793 \times \sqrt{10000} \times 10^{-6} = 0.0016$, i.e., 16 basis points. The "realized price impact" measures the typical impact of volume over the five-minute interval, and is defined in terms of the movement of the mid-point. Its median is 1%, i.e., 100 basis points. The median of the Amihud measure is 0.104×10^{-6}, meaning that if daily volume is \$10000, then the price change is $0.104 \times 10000 \times 10^{-6} = 0.001$, i.e., 10 basis points. This is comparable to the value derived using λ.

Table 1 shows that all measures of illiquidity have high standard deviation compared to their mean and median. Some of this variation reflects a secular downward trend, i.e., increase in liquidity over time. For example, Comerton-Forde, Hendershott, Jones, Moulton, and Seasholes (2010) report that the value-weighted effective spread in the NYSE in 2005 was about one tenth of its 1994 value. Yet, time-series variation around the trend as well as cross-sectional variation are also important, and we examine them next.

An important source of cross-sectional variation in illiquidity is the size of a stock, as

measured, for example, by market capitalization or number of shares. Large stocks are typically more liquid that small stocks. For example, Loeb (1983) and Stoll and Whaley (1983) find that market capitalization correlates negatively with the bid-ask spread. Demsetz (1968), Benston and Hagerman (1974) and Hamilton (1976) find a negative correlation between the number of shares and the bid-ask spread. Roll (1984) finds that market capitalization correlates negatively with the γ-based estimator of the bid-ask spread, and Sadka (2006) finds that it also correlates negatively with the price impact λ. The effect of size is economically significant: for example, Hendershott and Moulton (2007) find that the effective spread for the lowest and second-lowest market-capitalization quintile of US stocks is about six and three times, respectively, that for the highest quintile.

Many theoretical models imply a negative relationship between illiquidity and size, provided that the former is measured by λ and the latter by the aggregate trading needs of investors holding an asset. For example, in models with participation costs, surveyed in Section 2.2, higher trading needs by liquidity demanders induce more participation by liquidity suppliers, and this reduces λ. In models with asymmetric information, surveyed in Section 2.4, higher trading needs reduce the informational content per unit trade, also reducing λ. These effects also hold in our unified model, in which the parameter σ_z^2 measures the trading needs of liquidity demanders. The effects of σ_z^2 on γ, however, are generally in the opposite direction than for λ: an increase in σ_z^2 increases γ except under participation costs, where there is no effect. Intuitively, higher trading needs cause larger price effects, and while the effect per unit trade (corresponding to λ) can be smaller, the total effect (corresponding to γ) is generally larger.

A negative relationship between λ and trading needs translates to a negative one between λ and market capitalization or number of shares if the latter two variables are positively related to trading needs. Such a positive relationship arises naturally in most of the search models surveyed in Section 2.7, which assume that the probability that an asset holder needs to sell is independent of the total number of asset holders.

In addition to size, an important source of cross-sectional variation in illiquidity is the volatility of asset payoffs. For example, Stoll (1978a) finds that more volatile stocks have higher bid-ask spreads. Stoll (1978a) also finds that more than 80% of the cross-sectional variation in stocks' bid-ask spreads can be explained by volatility combined with measures of trading activity, number of market makers, and price level. Chen, Lesmond, and Wei (2007) find that illiquidity as measured by bid-ask spread, LOT, and number of non-trading days is higher for corporate bonds with lower rating or higher maturity, characteristics which are associated with higher price volatility. Edwards, Harris, and Piwowar (2007) and Bao, Pan, and Wang (2011)

find similar results measuring illiquidity by a price-impact based estimate and γ, respectively.

A positive relationship between illiquidity and volatility is implied by many theoretical models, as well as by our unified model. Intuitively, liquidity suppliers trading high-volatility assets are exposed to more risk and possibly to more asymmetric information. Therefore, they require a larger price movement to absorb liquidity shocks, which means that these shocks have larger price impact and cause larger transitory deviations between price and fundamental value.

Significant variation in illiquidity arises not only within asset classes, e.g., stocks or corporate bonds, but also across classes. Spiegel (2008) summarizes related evidence, drawing on Goyenko, Subrahmanyam, and Ukhov (2011) for Treasury bonds, Chen, Lesmond, and Wei (2007) for corporate bonds, and Hendershott and Moulton (2007) for stocks. Bid-ask spreads for Treasury bonds are generally smaller than for stocks, which are generally smaller than for corporate bonds. As Spiegel (2008) points out, the theory can readily explain some but not all of these comparisons. In particular, the lower bid-ask spreads of Treasury bonds relative to stocks and corporate bonds can be due to their less volatile payoffs. At the same time, the higher bid-ask spreads for corporate bonds relative to stocks are puzzling, given that the former have less volatile payoffs. Possible explanations are that the corporate-bond market is less transparent (e.g., Edwards, Harris, and Piwowar (2007)) and there is less competition between market makers. A negative effect of market-maker competition on bid-ask spreads has first been shown in Tinic and West (1972).

Papers studying the time-series variation of illiquidity document that there is some commonality across assets, i.e., the illiquidity of different assets moves up and down together. Chordia, Roll, and Subrahmanyam (2000) show that the effective bid-ask spread of the average NYSE stock increases by 0.778% for each 1% increase in the average effective bid-ask spread of the other stocks. At the same time, the R-squared of this regression is only 1.4%, meaning that the importance of the common factor relative to idiosyncratic ones in driving time-series variation is small. The explanatory power of the common factor increases for stocks with large market capitalization. Hasbrouck and Seppi (2001) also find a common factor with small explanatory power for the stocks in the Dow Jones index, using both effective spread and λ. Huberman and Halka (2001) provide additional evidence for commonality, and show that it persists even after controlling for market-wide returns and volatility.

The commonality in illiquidity can be linked to a number of aggregate variables. Huberman and Halka (2001) show that illiquidity increases following negative market returns and increased market volatility. Chordia, Roll, and Subrahmanyam (2001), Chordia, Sarkar, and

Subrahmanyam (2005) and Hameed, Kang, and Viswanathan (2010) also find strong increases in illiquidity following market drops. This evidence is consistent with the increase in illiquidity during market crises, such as the stock market crash of 1987, the debt market crisis of 1998, and the financial crisis of 2007-9. Using γ as a measure of illiquidity, Bao, Pan, and Wang (2011) find that the illiquidity of corporate bonds is not influenced by volatility in the bond market, but increases following negative returns and increased volatility in the stock market. This suggests that factors influencing illiquidity in the stock market matter also for bonds, and possibly vice-versa, a result also found in Chordia, Sarkar, and Subrahmanyam (2005).

While commonality in illiquidity can be linked to some aggregate variables, the variation cannot be fully explained by these variables, as Huberman and Halka (2001) show. Comerton-Forde, Hendershott, Jones, Moulton, and Seasholes (2010) provide evidence helping to account for the unexplained variation. They show that bid-ask spreads in the NYSE increase following periods when market makers have realized losses, even after controlling for market returns and volatility. Thus, the funding constraints on market makers could account for a significant fraction of the commonality in illiquidity and its sharp increase following crises. This is consistent with the theoretical models of funding constraints, surveyed in Section 2.6 and in Gromb and Vayanos (2010b).[13]

The common variation of illiquidity explored in the previous papers concerns how a given measure computed for different assets comoves over time. We next examine the extent to which different measures of illiquidity comove, both across assets and over time. Table 2 reports correlations between the measures included in Table 1. These are taken from Goyenko, Holden, and Trzcinka (2009) and concern 400 randomly selected US stocks over 1993-2005. Each measure is computed for each stock and month. Cross-sectional correlations are computed for each month and are averaged across months. Time-series correlations concern an equally weighted portfolio of the stocks.

In terms of the taxonomy of the previous section, the measures in Table 2 can be divided into three groups: (a) λ-based, consisting of Amihud, price impact λ, and realized price impact, (b) γ-based, consisting of Pastor-Stambaugh, and (c) spread-based, consisting of effective and realized spread. An additional taxonomy concerns the horizon at which these measures are computed: (i) measures computed using high-frequency intraday data, consisting of effective spread, realized spread, price impact λ, and realized price impact, which correspond to the rows of Table 2, and (ii) measures computed using low-frequency daily data, consisting of Amihud

[13]Funding constraints are not the only imperfection that can generate time-varying aggregate illiquidity. For example, Eisfeldt (2004) shows that aggregate illiquidity can decrease during downturns because asymmetric information becomes more severe.

Table 2: Correlation across illiquidity measures for US stocks.

The definition of the measures is in Table 1. The sample consists of 400 randomly selected stocks traded in NYSE/NASDAQ/AMEX over 1993-2005. All measures are computed for each stock and month. Cross-sectional correlations are computed for each month and are averaged across months. Time-series correlations concern an equally weighted portfolio of the stocks. The data in the table are taken from Goyenko, Holden, and Trzcinka (2009), but we define Pastor-Stambaugh with the opposite sign.

	Amihud		Pastor-Stambaugh	
	Cross Section	Time Series	Cross Section	Time Series
Effective Spread	0.571	0.608	0.118	0.366
Realized Spread	0.305	0.511	0.031	0.351
Price impact λ	0.317	0.400	0.064	0.192
Realized price impact	0.516	0.511	−0.035	0.230

and Pastor-Stambaugh, which correspond to the columns.

All measures of illiquidity in Table 2 covary positively. Yet, the correlations are relatively low: they range from 30-60% for Amihud and 0-40% for Pastor-Stambaugh. Moreover, the correlations appear to be driven mainly by the second taxonomy because they are roughly the same regardless of whether the high-frequency measures are λ- or spread-based. One interpretation of these results is that the low-frequency measures are imperfect proxies of illiquidity, which is measured more precisely by the high-frequency measures. Under that interpretation, low-frequency measures should be used only when high-frequency data are not available, which is often the case for the long samples needed for asset-pricing analysis (Section 3.3). Moreover, Amihud seems to capture illiquidity better than Pastor-Stambaugh, given its higher correlation with the high-frequency measures.

An alternative interpretation of the results in Table 2 is that different frequencies capture different phenomena. For example, and as pointed out in Section 3.1, the bid-ask spread and the price impact evaluated over a horizon of five minutes might fail to capture how the market would respond to a long sequence of transactions in the same direction. Thus, high-frequency measures might not be capturing the most relevant aspect of illiquidity. Moreover, the lower correlations for Pastor-Stambaugh relative to Amihud are not necessarily evidence that the former measure is less suitable than the latter in capturing illiquidity. Instead, it could be capturing different aspects. Differences in behavior across λ- and γ-based measures are, indeed, to be expected given the differences in the theoretical properties of λ and γ, discussed in Sections 3.1 and 3.2.

Sadka (2006) computes correlations involving measures based on trading activity. He finds

that trading activity is inversely correlated with price impact λ, consistent with the use of its inverse as a proxy for illiquidity. The correlations are low, however. For example, the correlation between trading volume and λ is -0.18, and that between turnover and λ is -0.11. The correlations become -0.06 and -0.05 when λ is replaced by Amihud. The low correlations are consistent with the theoretical discussion in Section 3.1. Yet, while the theory can provide suggestions as to why correlations between various measures of illiquidity can be low, more work is needed to explain the exact properties of these correlations.

3.3 Illiquidity and Asset Returns

Many of the papers linking illiquidity to asset prices and expected returns focus on the level of illiquidity rather than its time variation. Amihud and Mendelson (1986) sort stocks into portfolios according to quoted bid-ask spreads and market betas, and regress portfolio returns on these characteristics. They find that returns are increasing and concave in transaction costs, a result which they also derive theoretically. The effects are significant, as can be seen by comparing returns across the seven portfolios they consider when sorting along the bid-ask spread dimension. The average annual return of the stocks in the middle spread portfolio exceeds that in the lowest spread portfolio by 2.88%. The average bid-ask spread difference between the stocks in the two portfolios is 0.66%, implying an expected return premium of 4.36 per unit of spread. The corresponding quantities in the comparison between the highest and the middle spread portfolios are 5.22% for the return, 2.06% for the bid-ask spread, and 2.53 for the expected return premium per unit of spread. The return-per-spread ratio is higher for the less liquid stocks, consistent with a concave return-spread relationship.

Amihud and Mendelson (1986) show additionally that illiquidity can help explain the small-firm effect, namely that small stocks earn higher expected returns than large stocks (Banz (1981) and Reinganum (1981)). Including size in their regression, they show that the higher returns of small stocks can be largely explained by these stocks' higher bid-ask spreads. The idea that transaction costs can help explain the size effect is also explored in Stoll and Whaley (1983). Eleswarapu and Reinganum (1993) question the findings of Amihud and Mendelson (1986) by showing that the effect of bid-ask spread on expected returns holds only in January. Eleswarapu (1997) finds, however, a strong effect for both January and non-January months on a sample of NASDAQ stocks.

Brennan and Subrahmanyam (1996) measure the illiquidity of stocks by price impact λ instead of bid-ask spread. They employ the model of Glosten and Harris (1988) to separate

price impact into a variable part, equal to λ, and a fixed part. Unlike the previous papers, they control for risk using the Fama-French three-factor model instead of the CAPM. They show that returns are positively related to both the variable and the fixed part of price impact. Curiously, however, the bid-ask spread is negatively related to returns. They attribute this result to a possible correlation between the spread and a risk factor not captured by the Fama-French model.

Additional papers linking the level of illiquidity to expected stock returns include Brennan, Chordia, and Subrahmanyam (1998), who use the inverse of trading volume as a measure of illiquidity, Datar, Naik, and Radcliffe (1998), who use inverse turnover, and Chalmers and Kadlec (1998), who use amortized spread. Amortized spread measures the aggregate cost of the bid-ask spread to investors, and is roughly equal to spread times turnover. These papers find a positive relationship between illiquidity and expected returns.

Spiegel and Wang (2005) provide a more skeptical assessment of the effects of illiquidity. They regress stock returns on idiosyncratic risk and various illiquidity measures, and show that idiosyncratic risk renders many of these measures insignificant. Idiosyncratic risk, however, could also be a measure of illiquidity, especially when measured using short-horizon returns. Indeed, transitory deviations between price and fundamental value, caused by liquidity shocks, are larger for illiquid stocks. Moreover, these deviations are more likely to manifest themselves on the idiosyncratic component of returns, which liquidity suppliers cannot hedge away, than on the systematic component. Hence, illiquid stocks are likely to be characterized by higher idiosyncratic volatility.

A different criticism of illiquidity effects comes from Ben-Rephael, Kadan, and Wohl (2010). They find that while these effects were strong in the distant past, they have become weak more recently. For example, the coefficient of a cross-sectional regression of expected returns of NYSE stocks on Amihud's illiquidity measure has declined by a factor of 16 from the period 1964-1974 to 1997-2008, and its value in the latter period is statistically insignificant. This suggests a strong decline of illiquidity effects: not only transaction costs have been declining, as pointed out in the previous section, but liquidity premia per unit of transaction costs seem to have been declining as well.

Measuring the effects of illiquidity on stock prices and expected returns requires separating them from the effects of risk. The papers mentioned so far control for risk using the CAPM or a multi-factor model, but such adjustments are likely to be imperfect. Sharper tests can be derived by identifying stocks with identical cash flows that differ only in liquidity. One such example is restricted versus publicly-traded stocks. Restricted stocks cannot be traded publicly

for a given period after issuance, which was two years until 1997, because of regulatory reasons. The firms issuing them, however, issue publicly-traded stocks as well, which are identical to restricted stocks in all aspects other than the trading restriction. Empirical studies on the pre-1997 period (e.g., Silber (1991)) find that restricted stocks were priced on average 35% below their publicly-traded counterparts.[14]

Assets with similar cash flows can be found not only in the stock market but also among Treasury bonds. A number of papers examine the yield differential between on- and off-the-run Treasury securities with similar time to maturity. On-the-run securities are newly issued and actively traded, while off-the-run ones have been issued in the more distant past and trade less actively. Amihud and Mendelson (1991) find that off-the-run Treasury notes trade at lower prices and hence higher yields than on-the-run Treasury bills with same time to maturity. The effect is significant: the average yield is 6.52% for notes and 6.09% for bills, implying a yield differential of 43 basis points (bps). Further supporting evidence is in Kamara (1994). Warga (1992) compares the returns of on- and off-the-run portfolios that have matched duration and are rebalanced annually. The average annual return of on-the-run portfolios is 55bps below that of their off-the-run counterparts. Additional evidence on the on-the-run phenomenon is in Krishnamurthy (2002), Goldreich, Hanke, and Nath (2005), and Strebulaev (2007). Boudoukh and Whitelaw (1993) document a similar effect in Japan: a highly liquid "benchmark" government bond trades at a yield of 60bps below other bonds with similar characteristics.

While on-the-run effects are of significant magnitude, they arise not only because of liquidity but also because of repo specialness: on-the-run bonds are more expensive than off-the-run bonds partly because they constitute better collateral for borrowing funds in the repo market. Duffie (1996) shows that specialness can arise because of the higher liquidity of on-the-run bonds, and derives the price premium due to specialness. Vayanos and Weill (2008) derive the higher liquidity and specialness of on-the-run bonds endogenously, and decompose the on-the-run premium into a liquidity and a specialness part. Banerjee and Graveline (2011) provide a model-free decomposition of the premium.

A market that is particularly suited for measuring the effects of illiquidity on prices is the corporate-bond market. Indeed, corporate bonds are less liquid than Treasury bonds and stocks (Section 3.2); they typically carry no specialness premia; and they are less risky than stocks so

[14]Large discounts have also been found for assets other than stocks. For example, Brenner, Eldor, and Hauser (2001) report that currency options that are issued by central banks and cannot be traded prior to maturity were priced on average 21% below exchange-traded options. Comment (2012) argues, however, that restricted-stock discounts are driven primarily by factors other than illiquidity. For example, restricted stocks are placed privately, but private placements typically occur at prices lower than public placements even for freely-traded stocks.

the confounding effects of risk could be smaller. Chen, Lesmond, and Wei (2007) study how the yield spreads of corporate bonds relative to Treasuries are affected by illiquidity, measured by bid-ask spread, LOT, and number of non-trading days. They find that more illiquid corporate bonds have higher yield spreads. Bao, Pan, and Wang (2011) find a robust and even stronger positive relationship between yield spreads and illiquidity, when the latter is measured by γ.

The empirical papers mentioned so far examine whether a variety of illiquidity measures are positively related to expected asset returns. Our unified model suggests, however, that the relationship is more complex: it depends on the underlying cause of illiquidity, on the measure of illiquidity being used, and on the sources of cross-sectional variation. Accounting for these complexities can both shed light on existing empirical findings and help with the design of new tests.

To illustrate some of the complexities, suppose that illiquidity is caused by asymmetric information. If it is measured by λ, then its empirical relationship with expected returns will be positive since asymmetric information raises both λ and expected returns (Results 5 and 7). If, however, it is measured by γ, then the relationship can turn negative since asymmetric information can reduce γ (Result 6). Furthermore, if the imperfection is imperfect competition, then a negative relationship can arise even if illiquidity is measured by λ. This is because imperfect competition raises λ but can lower expected returns (Results 8 and 10). Finally, if cross-sectional variation is driven by the variance σ_z^2 of liquidity shocks rather than by the imperfections, then the relationship between λ and expected returns will be negative. Indeed, under both asymmetric information and imperfect competition, larger σ_z^2 lowers λ and raises expected returns (Vayanos and Wang (2011)). In summary, a positive relationship between illiquidity and expected returns is more likely to arise when illiquidity is measured by λ rather than γ, and when cross-sectional variation in trading needs is appropriately controlled for.

Evidence consistent with our theory comes from recent studies of the corporate-bond market. Dick-Nielsen, Feldhutter, and Lando (2012) examine how yield spreads are linked to λ, as approximated by the Amihud measure, and to γ. They find that the positive relationship between spreads and λ is more robust than that between spreads and γ, both across different rating categories and across the pre- and post-2008-crisis sample periods. Moreover, for the post-crisis period, the relationship between spreads and γ becomes insignificant except for AAA-rated bonds. For speculative-grade bonds the relationship becomes even negative. Given that speculative-grade bonds are more likely to be subject to asymmetric information, this finding appears consistent with our theory. Rayanankorn and Wang (2012) provide additional evidence along these lines.

Recent work linking illiquidity to expected returns focuses on the time variation of illiquidity and its possible role as a priced risk factor. Amihud (2002) examines how time variation in illiquidity affects stock returns. He computes the illiquidity of the aggregate stock market by averaging the Amihud measure over stocks. He shows that in a year when aggregate illiquidity increases, returns are low, consistent with a negative effect of illiquidity on prices. Moreover, returns are predicted to be high over the next year, consistent with a positive effect of illiquidity on expected returns.

The finding that movements in aggregate illiquidity affect stock returns opens up the possibility that illiquidity might be a priced risk factor. Indeed, investors might prefer to avoid stocks that go down when illiquidity goes up because they become exposed to a systematic risk. As a consequence, such stocks with "high liquidity risk" might be earning high expected returns, controlling for other risk characteristics. Pastor and Stambaugh (2003) test for this effect. They compute aggregate illiquidity by averaging the Pastor-Stambaugh measure over stocks, and show that stocks with high liquidity risk earn abnormally high expected returns. The effect is significant: stocks in the highest liquidity-risk decile outperform stocks in the lowest decile by 7.5% annually, controlling for the three Fama-French factors and a momentum factor.

Acharya and Pedersen (2005) allow illiquidity to affect expected returns both through its level and as a risk factor. They derive these effects in a theoretical model, which shows that liquidity risk affects expected returns through three covariances: between a stock's return and aggregate illiquidity, as in Pastor and Stambaugh (2003), between a stock's illiquidity and the return on the aggregate stock market, and between a stock's illiquidity and aggregate illiquidity. Using the Amihud measure for illiquidity, they find that all three covariances matter in a way consistent with their model.

Pastor and Stambaugh (2003) find that liquidity risk can explain about 50% of the momentum anomaly (Jegadeesh and Titman (1993)). Sadka (2006) explores further the link between liquidity risk and asset-pricing anomalies, using price impact as a measure of illiquidity. He decomposes price impact into a variable part λ and a fixed part, and shows that only λ is a priced risk factor. This risk factor can explain more than 50% of momentum and of the post-earnings-announcement drift anomaly (Bernard and Thomas (1989)). Korajczyk and Sadka (2008) find that only the common component across illiquidity measures is a priced risk factor. Hasbrouck (2009) finds weak evidence that illiquidity is a priced risk factor when measured using his Bayesian Gibbs estimator. Watanabe and Watanabe (2008) find that the pricing of liquidity risk is time varying. Lin, Wang, and Wu (2011) find that liquidity risk is priced in

the corporate bond market, while Bongaerts, De Jong, and Driessen (2012) find that only the level of illiquidity is priced. Sadka (2010) and Franzoni, Nowak, and Phalippou (2012) find that liquidity risk is priced in the cross section of hedge funds and private equity, respectively.

Most of the papers studying illiquidity as a priced risk factor focus on a single asset class. This leaves open the question whether liquidity risk is truly systematic and common to many asset classes. Hu, Pan, and Wang (2011), which measure aggregate illiquidity based on pricing errors of US Treasury bonds, provide evidence in favor of a systematic risk factor. They show that the covariance with aggregate illiquidity helps explain the returns of two sets of assets that are sensitive to market-wide conditions: hedge funds and currency carry trades.

A challenge for the empirical literature on liquidity risk is to establish that the effects are driven by risk and not by the level of illiquidity. Indeed, an asset's illiquidity level is typically positively correlated with the asset's loading on the illiquidity risk factor: this correlation is particularly high for stocks as shown in Acharya and Pedersen (2005), and lower for corporate bonds as shown in Bongaerts, De Jong, and Driessen (2012). Some papers on liquidity risk side-step the multi-collinearity problem by omitting the level of illiquidity.

While empirical papers provide suggestive evidence that illiquidity is a priced risk factor, more theoretical work is needed to clarify the nature of that factor and interpret the evidence. Does the illiquidity factor reflect variation in illiquidity only, or is such variation caused by more fundamental factors which might be affecting asset prices through additional channels? And what fundamental characteristics of an asset determine its liquidity risk, i.e., its sensitivity to the illiquidity factor? For example, if the illiquidity factor is a crisis factor, it could be affecting asset prices also through the risk premium because risk aversion increases during crises. Moreover, an asset's sensitivity to that factor could be determined by sensitivities to both risk aversion and illiquidity. Work on funding constraints, surveyed in Section 2.6, has begun to address these issues. Further work in that area could provide more comprehensive answers and put empirical research on liquidity risk on a firmer theoretical foundation.

4 Conclusion

Illiquidity can be viewed as a consequence of various forms of market imperfections. A large theoretical literature shows that even simple imperfections can break the clean properties of the perfect-market model and lead to rich but complex behavior. Moreover, this behavior can be sensitive to the particular form of imperfection and the specification of the model. The situation is reminiscent of the saying "Happy families are all alike; every unhappy family is unhappy in

its own way." The lack of a unified framework and robust predictions makes it difficult not only to advance our theoretical understanding of illiquidity, but also to provide guidance for empirical work, e.g., how to measure illiquidity, what theoretical predictions to test, and how to interpret the empirical findings.

In this survey we hope to demonstrate that a framework can be constructed to unify the existing theoretical work. Our framework allows us to examine in a consistent manner how various forms of market imperfections affect illiquidity and expected asset returns. It also shows how well different empirical measures of illiquidity capture the underlying imperfections. Furthermore, it provides new insights in interpreting existing empirical findings and guiding further analysis. Needless to say, the framework has a number of limitations, some of which are pointed out in the Introduction. But this only suggests that more research is needed; and the limitations of the framework may well point us to new and fruitful directions.

References

Acharya, Viral, and Lasse Pedersen, 2005, Asset pricing with liquidity risk, *Journal of Financial Economics* 77, 375–410.

Acharya, Viral, and S Viswanathan, 2011, Leverage, moral hazard, and liquidity, *Journal of Finance* 66, 99–138.

Admati, Anat, 1985, A noisy rational expectations equilibrium for multi-asset securities markets, *Econometrica* 53, 629–658.

——— , and Paul Pfleiderer, 1988, A theory of intraday patterns: Volume and price variability, *Review of Financial Studies* 1, 3–40.

——— , 1991, Sunshine trading and financial market equilibrium, *Review of Financial Studies* 4, 443–481.

Afonso, Gara, and Ricardo Lagos, 2011, Trade dynamics in the market for Federal Funds, working paper New York University.

Aiyagari, Rao, and Mark Gertler, 1991, Asset returns with transaction costs and uninsurable individual risks: A stage III exercise, *Journal of Monetary Economics* 27, 309–331.

Akerlof, George A., 1970, The market for lemons : Quality uncertainty and the market mechanism, *The Quaterly Journal of Economics* 84, 359–369.

Albagli, Elias, 2011, Amplification of uncertainty in illiquid markets, working paper University of Southern California.

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

Almgren, Robert, 2003, Optimal execution with nonlinear impact functions and trading-enhanced risk, *Applied Mathematical Finance* 10, 1–18.

——— , and Neil Chriss, 1999, Value under liquidation, *Risk* 12, 61–63.

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

——— , and Haim Mendelson, 1980, Dealership market: Market-making with inventory, *Journal of Financial Economics* 8, 31–53.

——— , 1986, Asset pricing and the bid-ask spread, *Journal of Financial Economics* 17, 223–249.

——— , 1991, Liquidity, maturity, and the yield on US treasury securities, *Journal of Finance* 46, 479–486.

Amihud, Yacov, Haim Mendelson, and Lasse Pedersen, 2005, Liquidity and asset pricing, *Foundations and Trends in Finance* 1, 269–364.

Ang, Andrew, Dimitris Papanikolaou, and Mark Westerfield, 2011, Portfolio choice with illiquid assets, working paper Columbia University.

Asquith, Paul, Rebecca Oman, and Chris Safaya, 2010, Short sales and trade classification algorithms, *Journal of Financial Markets* 13, 157–173.

Attari, Mukarram, Antonio Mello, and Martin Ruckes, 2005, Arbitraging arbitrageurs, *Journal of Finance* 60, 2471–2511.

Back, Kerry, 1992, Insider trading in continuous time, *Review of Financial Studies* 5, 387–409.

——— , and Shmuel Baruch, 2004, Information in securities markets: Kyle meets Glosten and Milgrom, *Econometrica* 72, 433–465.

——— , 2011, Strategic liquidity provision in limit order markets, working paper Rice University.

Back, Kerry, Charles Cao, and Gregory Willard, 2000, Imperfect competition among informed traders, *Journal of Finance* 55, 2117–2155.

Back, Kerry, and Hal Pedersen, 1998, Long-lived information and intraday patterns, *Journal of Financial Markets* 1, 385–402.

Bagehot, Walter, 1971, The only game in town, *Financial Analysts Journal* 22, 12–14.

Bai, Yang, Eric C. Chang, and Jiang Wang, 2006, Asset prices and short-sale constraints, working paper.

Balduzzi, Pierluigi, and Anthony Lynch, 1999, Transaction costs and predictability: Some utility cost calculations, *Journal of Financial Economics* 52, 47–78.

Banerjee, Snehal, and Jeremy Graveline, 2011, The cost of short-selling liquid securities, *Journal of Finance* forthcoming.

Banz, Rolf, 1981, The relationship between return and market value of common stocks, *Journal of Financial Economics* 9, 3–18.

Bao, Jack, Jun Pan, and Jiang Wang, 2011, The illiquidity of corporate bonds, *Journal of Finance* 66, 911946.

Baruch, Shmuel, 2002, Insider trading and risk aversion, *Journal of Financial Markets* 5, 451–464.

Basak, Suleyman, and Domenico Cuoco, 1998, An equilibrium model with restricted stock market participation, *Review of Financial Studies* 11(2), 309–341.

Beber, Alessandro, Joost Driessen, and Patrick Tuijp, 2012, Pricing liquidity risk with heterogeneous investment horizons, working paper Cass Business School.

Ben-Rephael, Azi, Ohad Kadan, and Avi Wohl, 2010, The diminishing liquidity premium, working paper Tel-Aviv University.

Benston, George, and Robert Hagerman, 1974, Determinants of bid-ask spreads in the over-the-counter market, *Journal of Financial Economics* 1, 353–364.

Bernanke, Ben, and Mark Gertler, 1989, Agency costs, net worth, and business fluctuations, *The American Economic Review* 1, 14–31.

Bernard, Victor, and Jacob Thomas, 1989, Post-earnings-announcement drift: Delayed price response or risk premium?, *Journal of Accounting Research* 27, 1–48.

Bernhardt, Dan, and Eric Hughson, 1997, Splitting orders, *Review of Financial Studies* 10, 69–102.

Bertsimas, Dimitris, and Andrew Lo, 1998, Optimal control of execution costs, *Journal of Financial Markets* pp. 1–50.

Bhattacharya, U, and M Spiegel, 1991, Insiders, outsiders, and market breakdowns, *Review of Financial Studies* 4, 255–282.

Bhushan, Ravi, 1994, An informational efficiency perspective on the post-earnings drift, *Journal of Accounting and Economics* 18, 46–65.

Biais, Bruno, 1993, Price formation and equilibrium liquidity in fragmented and centralized markets, *Journal of Finance* 48, 157–185.

——— , Lawrence Glosten, and Chester Spatt, 2005, Market microstructure: A survey of microfoundations, empirical results and policy implications, *Journal of Financial Markets* 8, 217–264.

Biais, Bruno, Florian Heider, and Marie Hoerova, 2012, Risk-sharing or risk-taking? counterparty risk, incentives and margins, working paper University of Toulouse.

Biais, Bruno, Pierre Hillion, and Chester Spatt, 1995, An empirical analysis of the limit order book and the order flow in the paris bourse, *Journal of Finance* 50, 1655–1689.

Biais, Bruno, Johan Hombert, and Pierre-Olivier Weill, 2011, Trading and liquidity with limited cognition, .

Biais, Bruno, David Martimort, and Jean-Charles Rochet, 2000, Competing mechanisms in a common value environment, *Econometrica* 68, 799–837.

Blouin, Max, and Roberto Serrano, 2001, A decentralized market with common values uncertainty: Non-steady states, *Review of Economic Studies* 68, 323–346.

Bongaerts, Dion, Frank De Jong, and Joost Driessen, 2012, An asset pricing approach to liquidity effects in corporate bond markets, working paper Tilburg University.

Boudoukh, Jacob, and Robert Whitelaw, 1993, Liquidity as a choice variable: A lesson from the Japanese government bond market, *Review of Financial Studies* 6, 265–292.

Brennan, Michael, Tarun Chordia, and Avanidhar Subrahmanyam, 1998, Alternative factor specifications, security characteristics, and the cross-section of expected returns, *Journal of Financial Economics* 49, 345373.

Brennan, Michael, and Avanidhar Subrahmanyam, 1996, Market microstructure and asset pricing: On the compensation for illiquidity in stock returns, *Journal of Financial Economics* 41, 441–464.

Brenner, Menachem, Rafi Eldor, and Shmuel Hauser, 2001, The price of options illiquidity, *Journal of Finance* 56, 789–805.

Brown, David, and Robert Jennings, 1990, On technical analysis, *Review of Financial Studies* 2, 527–552.

Brunnermeier, Markus, and Lasse Pedersen, 2005, Predatory trading, *Journal of Finance* 60, 1825–1863.

——— , 2009, Market liquidity and funding liquidity, *Review of Financial Studies* 22, 2201–2238.

Buffa, Andrea, 2011, Insider trade disclosure, market efficiency, and liquidity, working paper London Business School.

Burdett, Kenneth, and Maureen O'Hara, 1987, Building blocks: An introduction to block trading, *Journal of Banking and Finance* 11, 193–212.

Buss, Adrian, and Bernard Dumas, 2011, The equilibrium dynamics of liquidity and illiquid asset prices, working paper Goethe University Frankfurt.

Buss, Adrian, Raman Uppal, and Grigory Vilkov, 2011, Asset prices in general equilibrium with transactions costs and recursive utility, working paper Goethe University Frankfurt.

Caldentey, Rene, and Ennio Stacchetti, 2010, Insider trading with a random deadline, *Econometrica* 78, 245283.

Campbell, John, Sanford Grossman, and Jiang Wang, 1993, Trading volume and serial correlation in stock returns, *Quarterly Journal of Economics* 108, 905–939.

Cao, H., M. Evans, and R. Lyons, 2006, Inventory information, *Journal of Business* 79, 325–364.

Carlin, Bruce, Miguel Lobo, and S. Viswanathan, 2007, Episodic liquidity crises: Cooperative and predatory trading, *Journal of Finance* 62, 2235–2274.

Casamatta, Catherine, and Sebastien Pouget, 2011, Fund managers' contracts and financial markets' short-termism, working paper University of Toulouse.

Cespa, Giovani, and Thierry Foucault, 2011, Learning from prices, liquidity spillovers and endogenous market segmentation, working paper HEC Paris.

Chalmers, John, and Gregory Kadlec, 1998, An empirical investigation of the amortized spread, *Journal of Financial Economics* 48, 159–188.

Chan, Louis, and Josef Lakonishok, 1993, Institutional trades and intraday stock price behavior, *Journal of Financial Economics* 33, 173199.

Chau, Minh, and Dimitri Vayanos, 2008, Strong form efficiency with monopolistic insiders, *Review of Financial Studies* 21, 2275–2306.

Chen, Long, David Lesmond, and Jason Wei, 2007, Corporate yield spreads and bond liquidity, *Journal of Finance* 62, 119–149.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2000, Commonality in liquidity, *Journal of Financial Economics* 56, 3–28.

——— , 2001, Market liquidity and trading activity, *Journal of Finance* 56, 501–530.

Chordia, Tarun, Asani Sarkar, and Avanidhar Subrahmanyam, 2005, An empirical analysis of stock and bond market liquidity, *Review of Financial Studies* 18, 85129.

Chowhdry, B., and V. Nanda, 1991, Multimarket trading and market liquidity, *Review of Financial Studies* 4, 483–511.

Cohen, Kalman J., Steven F. Maier, Robert A. Schwartz, and David K. Whitcomb, 1981, Transaction costs, order placement strategy, and the existence of the bid-ask spread, *Journal of Political Economy* 89, 287–305.

Comerton-Forde, Carole, Terrence Hendershott, Charles Jones, Pamela Moulton, and Mark Seasholes, 2010, Time variation in liquidity: The role of market-maker inventories and revenues, *Journal of Finance* 65, 295–331.

Comment, Robert, 2012, Revisiting the illiquidity discount for private companies: A new (and "skeptical") restricted-stock study, *Journal of Applied Corporate Finance* 24, 80–91.

Constantinides, G. M., 1986, Capital market equilibrium with transaction costs, *Journal of Political Economy* 94, 842–862.

Copeland, Thomas, and Dan Galai, 1983, Information effects on the bid-ask spread, *Journal of Finance* 38, 1457–1469.

Coppejans, Mark, Ian Domowitz, and Ananth Madhavan, 2004, Resiliency in an automated auction, working paper Barclays Global Investors.

Danielsson, Jon, Hyun Song Shin, and Jean-Pierre Zigrand, 2011, Balance sheet capacity and endogenous risk, working paper London School of Economics.

Datar, Vinay, Narayan Naik, and Robert Radcliffe, 1998, Liquidity and stock returns: An alternative test, *Journal of Financial Markets* 1, 203–219.

Davis, Mark H. A., and Andrew Norman, 1990, Portfolio selection with transaction costs, *Mathematics of Operations Research* 15, 676–713.

De Long, Bradford, Andrei Shleifer, Lawrence Summers, and Robert Waldmann, 1990, Noise trader risk in financial markets, *Journal of Political Economy* 98, 703738.

Degryse, Hans, Frank De Jong, Maarten Van Ravenswaaij, and Gunther Wuyts, 2005, Aggressive orders and the resiliency of a limit order market, *Review of Finance* 9, 201–242.

DeMarzo, Peter, and Branko Urosevic, 2006, Ownership dynamics and asset pricing with a large shareholder, *Journal of Political Economy* 114, 774–815.

Demsetz, Harold, 1968, The cost of transacting, *The Quarterly Journal of Economics* 82, 33–53.

Diamond, Douglas, and Robert Verrecchia, 1981, Information aggregation in a noisy rational expectations economy, *Journal of Financial Economics* 9, 221–235.

——— , 1987, Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics* 18, 277–311.

Diamond, Peter A., 1982, Aggregate demand management in search equilibrium, *Journal of Political Economy* 90, 881–894.

Dick-Nielsen, Jens, Peter Feldhutter, and David Lando, 2012, Corporate bond liquidity before and after the onset of the subprime crisis, *Journal of Financial Economics* 103, 471–492.

Dong, Jiwei, Alexander Kempf, and Pradeep Yadav, 2007, Resiliency, the neglected dimension of market liquidity: Empirical evidence from the new york stock exchange, working paper University of Oklahoma.

Dow, James, and Gary Gorton, 1994, Arbitrage chains, *Journal of Finance* 49, 819–49.

Dubofsky, David, and John Groth, 1984, Exchange listing and stock liquidity, *Journal of Financial Research* 7(4), 291–302.

Duffie, Darrell, 1996, Special repo rates, *Journal of Finance* 51, 493–526.

——— , 2010, Presidential address: Asset price dynamics with slow-moving capital, *Journal of Finance* 65, 12371267.

——— , Nicolae Garleanu, and Lasse Pedersen, 2002, Securities lending, shorting, and pricing, *Journal of Financial Economics* 66, 307–339.

——— , 2005, Over-the-counter markets, *Econometrica* 73, 1815–1847.

——— , 2008, Valuation in over-the-counter markets, *Review of Financial Studies* 20, 1865–1900.

Duffie, Darrell, Semyon Malamud, and Gustavo Manso, 2009, Information percolation with equilibrium search dynamics, *Econometrica* 77, 1513–1574.

Duffie, Darrell, and Gustavo Manso, 2007, Information percolation in large markets, *American Economic Review Papers and Proceedings* 97, 203–209.

Duffie, Darrell, and Bruno Strulovici, 2011, Capital mobility and asset pricing, working paper Stanford University.

Dumas, Bernard, and Elisa Luciano, 1991, An exact solution to a dynamic portfolio choice problem under transactions costs, *Journal of Finance* 46, 577–595.

Easley, David, Nicholas Kiefer, and Maureen O'Hara, 1997, One day in the life of a very common stock, *Review of Financial Studies* 10, 805–835.

Easley, David, and Maureen O'Hara, 1987, Price, trade size, and information in securities markets, *Journal of Financial Economics* 19, 69–90.

——— , 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 576–605.

——— , 2004, Information and the cost of capital, *Journal of Finance* 59, 1553–1583.

Edwards, Amy, Lawrence Harris, and Michael Piwowar, 2007, Corporate bond market transaction costs and transparency, *Journal of Finance* 62, 1421–1451.

Eisfeldt, Andrea, 2004, Endogenous liquidity in asset markets, *Journal of Finance* 59, 1–30.

Eleswarapu, Venkat, 1997, Cost of transacting and expected returns in the NASDAQ markets, *Journal of Finance* 52, 2113–2127.

——— , and Marc Reinganum, 1993, The seasonal behavior of liquidity premium in asset pricing, *Journal of Financial Economics* 34, 373–386.

Ellul, Andrew, and Marco Pagano, 2006, IPO underpricing and after-market liquidity, *Review of Financial Studies* 19, 381–421.

Fardeau, Vincent, 2011, Strategic liquidity provision and predatory trading, working paper London School of Economics.

Fishman, Michael, and Kathleen Hagerty, 1992, Insider trading and the efficiency of stock prices, *RAND Journal of Economics* 23, 106–122.

Fleming, Michael, 2003, Measuring treasury market liquidity, *Federal Reserve Bank of New York Economic Policy Review* September, 83–108.

Fleming, Wendell, Sanford Grossman, Jean-Luc Vila, and Thalia Zariphopoulou, 1990, Optimal portfolio rebalancing with transactions costs, working paper Brown University.

Fontaine, Jean-Sebastien, and Rene Garcia, 2011, Bond liquidity premia, *Review of Financial Studies* forthcoming.

Foster, Douglas, and S. Viswanathan, 1996, Strategic trading when agents forecast the forecasts of others, *Journal of Finance* 51(4), 1437–1478.

Foucault, Thierry, 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* 2, 99–134.

——— , Ohad Kadan, and Eugene Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.

Franzoni, Francesco, Eric Nowak, and Ludovic Phalippou, 2012, Private equity performance and liquidity risk, *Journal of Finance* forthcoming.

Garleanu, Nicolae, 2009, Portfolio choice and pricing in illiquid markets, *Journal of Economic Theory* 144, 532–564.

——— , and Lasse Pedersen, 2004, Adverse selection and the required return, *Review of Financial Studies* 17, 643–665.

Garman, Mark, 1976, Market microstructure, *Journal of Financial Economics* 3, 257–275.

Geanakoplos, John, 1997, Promises, promises, in B. Arthur, S. Durlauf, and D. Lane, ed.: *The Economy as an Evolving Complex System II* . pp. 285–320 (Addison-Wesley: Reading, MA).

——— , 2003, Liquidity, default and crashes: Endogenous contracts in general equilibrium, in M. Dewatripont, L. Hansen, and S. Turnovsky, ed.: *Advances in Economics and Econometrics: Theory and Applications II, Econometric Society Monographs: Eighth World Congress* . pp. 170–205 (Cambridge University Press: Cambridge, UK).

Glosten, Lawrence, 1989, Insider trading, liquidity and the role of the monopolist specialist, *Journal of Business* 62, 211–235.

——— , 1994, Is the electronic open limit order book inevitable?, *Journal of Finance* 49, 1127–1161.

——— , and Lawrence Harris, 1988, Estimating the components of the bid-ask spread, *Journal of Financial Economics* 21, 123–142.

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

Goettler, Ronald, Christine Parlour, and Uday Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149–2192.

Goldreich, David, Bernd Hanke, and Purnendu Nath, 2005, The price of future liquidity: Time-varying liquidity in the US treasury market, *Review of Finance* 9, 1–32.

Goldsmith, David, 1976, Transaction costs and the theory of portfolio selection, *Journal of Finance* 31, 1127–1139.

Golosov, Mikhail, Guido Lorenzoni, and Aleh Tsyvinski, 2011, Decentralized trading with private information, working paper Massachusetts Institute of Technology.

Goyenko, Ruslan, Craig Holden, and Charles Trzcinka, 2009, Do liquidity measures measure liquidity?, *Journal of Financial Economics* 92, 153–181.

Goyenko, Ruslan, Avanidhar Subrahmanyam, and Andrey Ukhov, 2011, The term structure of bond market liquidity and its implications for expected bond returns, *Journal of Financial and Quantitative Analysis* 46, 111–139.

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

——— , 2010a, Limits of arbitrage, *Annual Review of Financial Economics* 2, 251–275.

——— , 2010b, A model of financial market liquidity based on intermediary capital, *Journal of the European Economic Association, Papers and Proceedings* pp. 456–466.

——— , 2011a, The dynamics of financially constrained arbitrage, working paper INSEAD.

——— , 2011b, Financially constrained arbitrage and cross-market contagion, working paper INSEAD.

Grossman, Sanford, 1976, On the efficiency of competitive stock markets when traders have diverse information, *Journal of Finance* 31, 573–585.

——— , and Merton Miller, 1988, Liquidity and market structure, *Journal of Finance* 43, 617–637.

Grossman, Sanford, and Joseph Stiglitz, 1980, On the impossibility of informationnaly efficient markets, *American Economic Review* 70, 393–408.

Grossman, Sanford, and Jean-Luc Vila, 1992, Optimal investment strategies with leverage constraints, *Journal of Financial and Quantitative Analysis* 27, 151–168.

Grundy, Bruce, and Maureen McNichols, 1989, Trade and the revelation of information through prices and direct disclosure, *Review of Financial Studies* 2, 495–526.

Guo, Ming, and Hui Ou-Yang, 2010, A continuous-time model of risk-averse strategic trading with dynamic information, working paper Chueng-Kong Graduate School of Business.

Hameed, Allaudeen, Wenjin Kang, and S. Viswanathan, 2010, Stock market declines and liquidity, *Journal of Finance* 65, 257294.

Hamilton, James, 1976, Competition, scale economies, and transaction cost in the stock market, *Journal of Financial and Quantitative Analysis* 11, 779–802.

Harris, Lawrence, 1990, Statistical properties of the roll serial covariance bid/ask spread estimator, *Journal of Finance* 45(2), 579–590.

Hart, Oliver, and John Moore, 1994, A theory of debt based on the inalienability of human capital, *Quarterly Journal of Economics* 101, 841–879.

——— , 1995, An analysis of the role of hard claims in constraining management, *American Economic Review* 85, 567–585.

Hasbrouck, Joel, 2007, *Empirical Market Microstructure* (Oxford University Press: Oxford).

——— , 2009, Trading costs and returns for US equities: Estimating effective costs from daily data, *Journal of Finance* 64, 1445–1477.

——— , and Duane Seppi, 2001, Common factors in prices, order flows and liquidity, *Journal of Financial Economics* 59, 383–411.

He, H., and J. Wang, 1995, Differential information and dynamic behavior of stock trading volume, *Review of Financial Studies* 8, 919972.

He, Z., and A. Krishnamurthy, 2011, A model of capital and crises, *Review of Economic Studies* forthcoming.

Heaton, John, and Deborah J. Lucas, 1996, Evaluating the effects of incomplete markets on risk sharing and asset pricing, *Journal of Political Economy* 104, 443–487.

Hellwig, Martin, 1980, On the aggregation of information in competitive markets, *Journal of Economic Theory* 22, 477–498.

Hendershott, Terrence, and Albert Menkveld, 2011, Price pressures, working paper University of California, Berkeley.

Hendershott, Terrence, and Pamela Moulton, 2007, The shrinking New York Stock Exchange floor and the hybrid market, working paper University of California, Berkeley.

Ho, Thomas, and Richard Macris, 1980, Dealer bid-ask quotes and transaction prices: An empirical study of some amex options, *Journal of Finance* 39, 23–45.

Ho, Thomas, and Hans Stoll, 1980, On dealer markets under competition, *Journal of Finance* 35, 259–267.

——— , 1981, Optimal dealer pricing under trading transactions and return uncertainty, *Journal of Financial Economics* 9, 47–73.

——— , 1983, The dynamics of dealer markets under competition, *Journal of Finance* 38, 1053–1074.

Holden, Craig, and Avanidhar Subrahmanyam, 1992, Long-lived private information and imperfect competition, *Journal of Finance* 47, 247–270.

——— , 1994, Risk aversion, imperfect competition, and long-lived information, *Economic Letters* 44, 181–190.

Holthausen, Robert, Richard Leftwich, and David Mayers, 1990, Large-block transactions, the speed of response, and temporary and permanent stock-price effects, *Journal of Financial Economics* 26, 71–95.

Hombert, Johan, and David Thesmar, 2011, Overcoming limits of arbitrage: Theory and evidence, working paper HEC.

Hu, Xing, Jun Pan, and Jiang Wang, 2011, Noise as information for illiquidity, working paper Massachusetts Institute of Technology.

Huang, Jennifer, and Jiang Wang, 2009, Liquidity and market crashes, *Review of Financial Studies* 22, 2607–1643.

——— , 2010, Market liquidity, asset prices, and welfare, *Journal of Financial Economics* 95, 107–127.

Huang, Ming, 2003, Liquidity shocks and equilibrium liquidity premia, *Journal of Economic Theory* 109, 104–129.

Huang, Roger, and Hans Stoll, 1996, Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE, *Journal of Financial Economics* 41, 313–357.

Huberman, Gur, and Dominika Halka, 2001, Systematic liquidity, *Journal of Financial Research* 24, 161–178.

Huberman, Gur, and Werner Stanzl, 2005, Optimal liquidity trading, *Review of Finance* 9, 165–200.

Huddart, Steven, John Hughes, and Carolyn Levine, 2001, Public disclosure and dissimulation of insider trades, *Econometrica* 69, 665–681.

Jackson, Matthew, 1991, Equilibrium, price formation, and the value of private information, *Review of Financial Studies* 4, 1–16.

Jang, Bong-Gyu, Hyeng Keun Koo, Hong Liu, and Mark Loewenstein, 2007, Liquidity premia and transaction costs, *Journal of Finance* 62, 2329–2366.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–91.

Jurek, Jakub, and Halla Yang, 2007, Dynamic portfolio selection in arbitrage, working paper Princeton University.

Kamara, Avraham, 1994, Liquidity, taxes, and short-term treasury yields, *Journal of Financial and Quantitative Analysis* 29, 403–417.

Karpoff, Jonathan, and Ralph Walkling, 1988, Short-term trading around ex-dividend days, *Journal of Financial Economics* 21, 291–298.

Keim, Donald, and Ananth Madhavan, 1996, The upstairs market for large-block transactions: Analysis and measurement of price effects, *Review of Financial Studies* 9, 1–36.

Kiyotaki, Nobuhiro, and John Moore, 1997, Credit cycles, *Journal of Political Economy* 105, 211–248.

Klemperer, Paul, and Margaret Meyer, 1989, Supply function equilibria in oligopoly under uncertainty, *Econometrica* 57, 1243–1277.

Kondor, Peter, 2009, Risk in dynamic arbitrage: Price effects of convergence trading, *Journal of Finance* 64, 638–658.

Korajczyk, Robert, and Ronnie Sadka, 2008, Pricing the commonality across alternative measures of liquidity, *Journal of Financial Economics* 87, 45–72.

Kraus, Alan, and Hans Stoll, 1972, Price impacts of block trading on the new york stock exchange, *Journal of Finance* 27, 569–588.

Krishnamurthy, Arvind, 2002, The bond/old-bond spread, *Journal of Financial Economics* 66, 463–506.

Kyle, Albert, 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1336.

——— , 1989, Informed speculation with imperfect competition, *Review of Economic Studies* 56, 317–356.

——— , and Wei Xiong, 2001, Contagion as a wealth effect, *Journal of Finance* 56, 1401–1440.

Laffont, Jean-Jacques, and Eric Maskin, 1990, The efficient market hypothesis and insider trading on the stock market, *Journal of Political Economy* 98, 70–93.

Lagos, Ricardo, and Guillaume Rocheteau, 2009, Liquidity in asset markets with search frictions, *Econometrica* 77, 403–426.

——— , and Pierre-Olivier Weill, 2011, Crises and liquidity in over the counter markets, *Journal of Economic Theory* forthcoming.

Lee, Charles, and Mark Ready, 1991, Inferring trade direction from intraday data, *Journal of Finance* 46, 733–746.

Leland, Hayne, 1992, Insider trading: Should it be prohibited, *Journal of Political Economy* 100, 859–887.

Lesmond, David, Joseph Ogden, and Charles Trzcinka, 1999, A new estimate of transaction costs, *Review of Financial Studies* 12, 1113–1141.

Lin, Hai, Junbo Wang, and Chunchi Wu, 2011, Liquidity risk and expected corporate bond returns, *Journal of Financial Economics* pp. 628–650.

Liu, Hong, 2004, Optimal consumption and investment with transaction costs and multiple risky assets, *Journal of Finance* 59, 289–338.

——— , and Mark Loewenstein, 2002, Optimal portfolio selection with transaction costs and finite horizons, *Review of Financial Studies* 15, 805–835.

Liu, Jun, and Francis Longstaff, 2004, Losing money on arbitrage: Optimal dynamic portfolio choice in markets with arbitrage opportunities, *Review of Financial Studies* 17, 611–641.

Llorente, Guillermo, Roni Michaely, Gideon Saar, and Jiang Wang, 2002, Dynamic volume-return relation of individual stocks, *Review of Financial Studies* 15, 1005–1047.

Lo, Andrew, Harry Mamaysky, and Jiang Wang, 2004, Asset prices and trading volume under fixed transactions costs, *Journal of Political Economy* 112, 1054–1090.

Loeb, Thomas, 1983, Trading cost: The critical link between investment information and results, *Financial Analysts Journal* 39, 39–44.

Longstaff, Francis, 2001, Optimal portfolio choice and the valuation of illiquid securities, *Review of Financial Studies* 14(2), 407–431.

——— , 2004, The flight-to-liquidity premium in US treasury bond prices, *Journal of Business* 77, 511–526.

——— , 2009, Portfolio claustrophobia: Asset pricing in markets with illiquid assets, *American Economic Review* 99, 1119–1144.

Lynch, Anthony, and P. Balduzzi, 2000, Predictability and transaction costs: The impact on rebalancing rules and behavior, *Journal of Finance* 55, 2285–2310.

Lynch, Anthony, and Sinan Tan, 2011, Explaining the magnitude of liquidity premia: The role of return predictability, wealth shocks and state-dependent transaction costs, *Journal of Finance* 66, 1329–1368.

Madhavan, Ananth, 2000, Market microstructure: A survey, *Journal of Financial Markets* 3, 205–258.

——— , and Seymour Smidt, 1993, An analysis of changes in specialist inventories and quotations, *Journal of Finance* 48, 1595–1627.

Mahanti, Sriketan, Amrut Nashikkar, Marti Subrahmanyam, George Chacko, and Gaurav Mallik, 2008, Latent liquidity: A new measure of liquidity, with an application to corporate bonds, *Journal of Financial Economics* 88, 272–298.

Mankiw, N.G., and S. Zeldes, 1991, The consumption of stockholders and nonstockholders, *Journal of Financial Economics* 29, 97–112.

Mayshar, Joram, 1979, Transaction costs in a model of capital market equilibrium, *Journal of Political Economy* 87, 673–700.

Mehra, Rajnish, and Edward C. Prescott, 1985, The equity premium: A puzzle, *Journal of Monetary Economics* 15(2), 145–161.

Merton, Robert, 1971, Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* 3, 373–413.

——— , 1987, Presidential address: A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483–510.

Milbradt, Konstantin, 2011, Level 3 assets: Booking profits, concealing losses, *Review of Financial Studies* forthcoming.

Mildenstein, Eckart, and Harold Schleef, 1983, The optimal pricing policy of a monopolistic marketmaker in equity market, *Journal of Finance* 38, 218–231.

Mitchell, Mark, Lasse Pedersen, and Todd Pulvino, 2007, Slow moving capital, *American Economic Review, Papers and Proceedings* 97, 215–220.

Mortensen, Dale, 1982, Property rights and efficiency in mating, racing, and related games, *American Economic Review* 72, 968–979.

Naik, Narayan, Anthony Neuberger, and S. Viswanathan, 1999, Trade disclosure regulation in markets with negotiated trade, *Review of Financial Studies* 12, 873–900.

Niederhoffer, Victor, and M. Osborne, 1966, Market making and reversal on the stock exchange, *Journal of the American Statistical Association* 61, 897–916.

Obizhaeva, Anya, and Jiang Wang, 2006, Optimal trading strategy and supply/demand dynamics, working paper University of Maryland.

O'Hara, Maureen, 1995, *Market Microstructure Theory* (Blackwell Publishers: Cambridge).

——— , 2003, Liquidity and price discovery, *Journal of Finance* 58, 1335–1354.

Pagano, Marco, 1989a, Endogenous market thinness and stock price volatility, *Review of Economic Studies* 56, 269–287.

——— , 1989b, Trading volume and asset liquidity, *Quarterly Journal of Economics* 104, 255–274.

——— , and Ailsa Roell, 1996, Transparency and liquidity: A comparison of auction and dealer markets with informed trading, *Journal of Finance* 51, 579–611.

Parlour, Christine, 1998, Price dynamics in limit order markets, *Review of Financial Studies* 1, 789–816.

——— , and Duane Seppi, 2008, Limit order markets: A survey, in Arnoot Boot, and Anjan Thakor, ed.: *Handbook of Financial Intermediation and Banking* (North Holland).

Pastor, Lubos, and Robert Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111(3), 642–685.

Pavlova, Anna, and Roberto Rigobon, 2008, The role of portfolio constraints in the international propagation of shocks, *Review of Economic Studies* 75, 1215–1256.

Pissarides, Christopher, 1985, Short-run equilibrium dynamics of unemployment, vacancies, and real wages, *American Economic Review* 75, 676–690.

Pritsker, Matthew, 2005, Large investors: Implications for equilibrium asset, returns, shock absorption, and liquidity, working paper Board of Governors of the Federal Reserve.

Qiu, Weiyang, and Jiang Wang, 2010, Asset pricing under heterogeneous information, working paper Massachusetts Institute of Technology.

Rayanankorn, Surapap, and Jiang Wang, 2012, Different aspects of corporate bond illiquidity and bond yields, working paper Massachusetts Institute of Technology.

Reinganum, Marc, 1981, Misspecification of capital asset pricing: Empirical anomalies based on earnings yields and market values, *Journal of Financial Economics* 9, 19–46.

Repullo, Rafael, 1999, Some remarks on Leland's model of insider trading, *Economica* 66, 359–374.

Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* 39, 1127–1139.

Rostek, Marzena, and Marek Weretka, 2011, Dynamic thin markets, working paper University of Wisconsin.

Rosu, Ioanid, 2009, A dynamic model of the limit-order book, *Review of Financial Studies* 22, 4601–4641.

Sadka, Ronnie, 2006, Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk, *Journal of Financial Economics* 80, 309–349.

——— , 2010, Liquidity risk and the cross-section of hedge fund returns, *Journal of Financial Economics* 98, 54–71.

Shleifer, A., and R. Vishny, 1992, Liquidation values and debt capacity: A market equilibrium approach, *Journal of Finance* 47, 1343–1366.

——— , 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.

Silber, William, 1991, Discounts on restricted stock: The impact of illiquidity on stock prices, *Financial Analysts Journal* pp. 60–64.

Spiegel, M, 2008, Patterns in cross market liquidity, *Financial Research Letters* 5, 2–10.

Spiegel, Matthew, and Xiaotong Wang, 2005, Cross-sectional variation in stock returns: Liquidity and idiosyncratic risk, working paper Yale University.

Stoll, Hans, 1978a, The pricing of security dealer services: An empirical study of NASDAQ stocks, *Journal of Finance* 33, 1153–1172.

——— , 1978b, The supply of dealer services in securities markets, *Journal of Finance* 33, 1133–1151.

——— , and Robert Whaley, 1983, Transaction costs and the small firm effect, *Journal of Financial Economics* 12, 57–79.

Strebulaev, Ilya, 2007, Liquidity and asset pricing: Evidence from the US treasury securities market, working paper Stanford University.

Suominen, Matti, and Kalle Rinne, 2011, A structural model of short-term reversals, working paper Aalto University.

Tinic, Seha, and Richard West, 1972, Competition and the pricing of dealer services in the over-the-counter market, *Journal of Financial and Quantitative Analysis* 7, 1707–1727.

Tuckman, Bruce, and Jean-Luc Vila, 1992, Arbitrage with holding costs: A utility-based approach, *Journal of Finance* 47, 1283–1302.

——— , 1993, Holding costs and equilibrium arbitrage, working paper 1153 Anderson Graduate School of Management, UCLA.

Vayanos, Dimitri, 1998, Transaction costs and asset prices: A dynamic equilibrium model, *Review of Financial Studies* 11, 1–58.

——— , 1999, Strategic trading and welfare in a dynamic market, *Review of Economic Studies* 66, 219–254.

——— , 2001, Strategic trading in a dynamic noisy market, *Journal of Finance* 56, 131–171.

——— , 2004, Flight to quality, flight to liquidity, and the pricing of risk, working paper London School of Economics.

——— , and Jean-Luc Vila, 1999, Equilibrium interest rate and liquidity premium with transaction costs, *Economic Theory* 13, 509–539.

Vayanos, Dimitri, and Jiang Wang, 2010, Liquidity and asset prices: A unified framework, working paper London School of Economics.

——— , 2011, Liquidity and expected returns under asymmetric information and imperfect competition, *Review of Financial Studies* forthcoming.

Vayanos, Dimitri, and Tan Wang, 2007, Search and endogenous concentration of liquidity in asset markets, *Journal of Economic Theory* 136, 66–104.

Vayanos, Dimitri, and Pierre-Olivier Weill, 2008, A search-based theory of the on-the-run phenomenon, *Journal of Finance* 63, 1361–1398.

Venter, Gyuri, 2011, Financially constrained strategic arbitrage, working paper Copenhagen Business School.

Vives, Xavier, 1995, The speed of information revelation in a financial market mechanism, *Journal of Economic Theory* 67, 178–204.

Wang, Jiang, 1993, A model of intertemporal asset prices under asymmetric information, *Review of Economics Studies* 60, 249–282.

——— , 1994, A model of competitive stock trading volume, *Journal of Political Economy* 102, 127–168.

Warga, Arthur, 1992, Bond returns, liquidity, and missing data, *Journal of Financial and Quantitative Analysis* 27, 605–617.

Watanabe, Akiko, and Masahiro Watanabe, 2008, Time-varying liquidity risk and the cross section of stock returns, *Review of Financial Studies* 21, 2449–2486.

Weill, Pierre-Olivier, 2007, Leaning against the wind, *Review of Economic Studies* 74, 1329–1354.

——— , 2008, Liquidity premia in dynamic bargaining markets, *Journal of Economic Theory* 140, 66–96.

Wilson, Robert, 1979, Auctions of shares, *Quarterly Journal of Economics* 93, 675–689.

Wolinksy, Asher, 1990, Information revelation in a market with pairwise meetings, *Econometrica* 58, 1–23.

Xiong, Wei, 2001, Convergence trading with wealth effects: An amplification mechanism in financial markets, *Journal of Financial Economics* 62, 247–292.

Yuan, Kathy, 2005, Asymmetric price movements and borrowing constraints: A REE model of crisis, contagion, and confusion, *Journal of Finance* 60, 379–411.

Zhu, Xiaotong, 2011, Finding a good price in opaque over-the-counter markets, *Review of Financial Studies* forthcoming.