Liquidity in Asset Markets with Search Frictions*

Ricardo Lagos
New York University

Guillaume Rocheteau
Federal Reserve Bank of Cleveland

June 2007

Abstract

We study how trading frictions in asset markets affect the distribution of asset holdings, asset prices, efficiency and standard measures of liquidity. To this end, we analyze the equilibrium and optimal allocations of a search-theoretic model of financial intermediation similar to Duffie, Gârleanu and Pedersen (2005). In contrast with the existing literature, the model we develop imposes no restrictions on asset holdings, so traders can accommodate frictions by varying their trading needs through changes in their asset postitions. We find that this is a critical aspect of investor behavior in illiquid markets. A reduction in trading frictions leads to an increase in the dispersion of asset holdings and trade volume. Transaction costs and intermediaries' incentives to make markets are non-monotonic in trade frictions. With entry of dealers these non-monotonicities give rise to an externality in liquidity provision that can lead to multiple equilibria. Tight spreads are correlated with large volume and short trading delays across equilibria. From a normative standpoint we show that the asset allocation across investors and the number of dealers are socially inefficient.

Keywords: bid-ask spread, execution delay, liquidity, search, trade volume
JEL Classification: D83, G1

*We are grateful to Gadi Barlevy, Darrell Duffie, Mariacristina De Nardi, Nicolae Gârleanu, Joe Haubrich, Neil Wallace and Pierre-Olivier Weill for comments. We also thank seminar participants at the Federal Reserve Bank of Cleveland, University of Basel, Indiana University, HEC Lausanne, Penn State University, University of Pennsylvania, Princeton University, Queens University, Rice University, Singapore Management University and UCLA. We thank Patrick Higgins for research assistance. Financial support from the C.V. Starr Center for Applied Economics at NYU is gratefully acknowledged. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

1 Introduction

Asset markets have traditionally been the realm of the Walrasian paradigm. Trade in these markets, e.g., the matching of buyers and sellers, is typically regarded as an instantaneous and costless process—and left unmodeled. The fact remains that the vast majority of real assets, such as houses and cars, and a large volume of financial assets, such as derivative securities, federal funds, unlisted stocks and most fixed-income securities, are traded in over-the-counter (OTC) markets. OTC markets operate in a completely decentralized manner: trade is bilateral, with prices and quantities negotiated by the parties involved in each trade.[1] In this paper we further the view that prying into the microstructure of decentralized asset markets by explicitly modelling the trading process is important to understand and assess the performance of these markets.

A recent literature pioneered by Duffie, Gârleanu and Pedersen (2005) (DGP hereafter) uses search theory to model the trading frictions characteristic of OTC markets.[2] The search-based approach is appealing because it can parsimoniously rationalize standard measures of liquidity—bid-ask spreads, execution delays and trade volume—and lends itself to study how the market structure and market conditions influence these measures. A virtue of DGP's search-based theory of liquidity is that it is analytically tractable, so that all these effects can be well understood.

The literature spurred by DGP, however, keeps the framework tractable by imposing a stark restriction on asset holdings, namely, that investors can only hold either 0 or 1 unit of the asset. In effect, the ability of market participants to respond to changes in market conditions is limited rather severely by this restriction. In this paper we develop a search-based model of liquidity in asset markets with no restrictions on investors' asset holdings. The model is close in structure and in spirit to DGP, but captures the heterogeneous responses of individual investors to changes in market conditions. From the broader perspective of search and matching

[1] For a description of OTC markets for corporate and municipal bonds, see Schultz (2001), Saunders, Srinivasan and Walter (2002) and Harris and Piwowar (2005). The functioning of the OTC market for federal funds is described in Hamilton (1996) and Ashcraft and Duffie (2007). Even in equity markets, where trading arrangements are often well developed, trading frictions exist and can be significant, see Boehmer (2005, Table 6), and Stoll (2006b).

[2] The search-theoretic literature on financial markets includes Duffie, Gârleanu and Pedersen (2005), Gârleanu (2006), Miao (2006), Rust and Hall (2003), Spulber (1996) and Weill (2007). There is also a large related literature, not search-based, which studies how exogenously specified transaction costs affect the functioning of asset markets. Recent examples include Lo, Mamaysky and Wang (2004) and Vayanos (1998). See Heaton and Lucas (1995) for a survey of this body of work.

theory, a striking feature of the model we develop is that it remains analytically tractable despite the large degree of heterogeneity among agents which is propagated endogenously by random matching with unrestricted asset holdings.[3] We provide a full characterization of the equilibrium—including the endogenous distribution of investors' asset positions—and are able to show how it depends on all the details of the market structure. The methodology we put forward allows us to analyze both steady-state and dynamic equilibria. Although we emphasize the application to OTC markets for financial securities, the structure and solution techniques we have developed should also prove useful for other applied issues in search theory and other macroeconomic models with idiosyncratic uncertainty.

Our methodological contribution provides new insights into how trading frictions affect outcomes in financial markets. We find that as a result of the restrictions on asset holdings, existing search-based theories of financial liquidity neglect a critical aspect of investor behavior in illiquid markets, namely the fact that market participants can mitigate trading frictions by adjusting their asset positions so as to reduce their trading needs.[4] The key theoretical observation is that an investor's asset demand in an OTC market depends on a weighted average of his current marginal utility from holding the asset at the time of the trade and his expected marginal valuation over future random trading times. A reduction in trading delays makes high valuation investors who subsequently draw low preference shocks less likely to remain locked into holding an undesirably large position for a long period of time. Thus, reductions in trading frictions induce investors with high valuations to put more weight on their current valuation and therefore to choose larger asset positions. The converse is true for investors with low valuations. This asset reallocation mechanism implies that reductions in trading frictions tend to increase the distribution of trade sizes (in the first-order stochastic sense). These endogenous responses of individual investors' asset demands have a significant impact on market efficiency and asset

[3] One can think of DGP as being to the search-based theory of financial markets what Kiyotaki and Wright (1993) is to search-based monetary theory. DGP restricted asset holdings for the same technical reasons why Kiyotaki and Wright restricted money holdings to $\{0, 1\}$, i.e., to keep the endogenous distribution of asset holdings manageable. The recent monetary literature, e.g., Lagos and Wright (2005), allows for unrestricted portfolios and keeps the analysis tractable by making assumptions that render the equilibrium distribution of money holdings degenerate. By way of comparison, the model we develop here allows for unrestricted portfolios and remains tractable even though we make no attempt to harness the heterogeneity that is generated by the model dynamics. In fact, we are even able to provide a closed-form characterization of the endogenous distribution of asset holdings not only in the steady state, but also along the dynamic equilibrium path.

[4] The importance of this mechanism in the context of another class of models—those with exogenous transaction costs—has been stressed by Constantinides (1986) (for the case of proportional transaction costs) and by Lo, Mamaysky and Wang (2004) (for the case of fixed trading costs).

prices as well as trade volume, bid-ask spreads and trading delays—precisely the dimensions of market liquidity which search-based theories of financial liquidity were designed to explain.

Trade volume is a manifestation of the extent to which the exchange mechanism is able to reallocate assets across investors. According to the theory, large trade volumes are characteristic of liquid markets, i.e., markets where investors are able to switch in and out of asset positions relatively fast. We find that trading frictions have a direct as well as two indirect general equilibrium effects on trade volume. If investors find trading opportunities more frequently, the number of investors who are able to trade rises, but the number of investors who are mismatched with their current portfolio falls. These two opposing effects of trading frictions on trade volume can be found in the existing literature. In addition, we find that a reduction in trading frictions shifts the equilibrium distribution of investors across desired and actual asset holdings in a way that tends to increase trade volume. This general equilibrium effect is implicitly shut off if one restricts asset holdings to lie in $\{0, 1\}$, which is why our model has different predictions for the behavior of trade volume in response to changes in the microstructure of the market. For example, in its basic formulation, DGP predicts that trade volume is independent of dealers' market power. In contrast, our model implies that trade volume will be lower in markets where dealers have more power to set the terms of trade.

From an investor's standpoint, bid-ask spreads represent the out-of-pocket transaction costs of trading in an illiquid market. In a search-based theory with bilateral bargaining, the bid-ask spread is increasing in the investor's value to immediate trade and decreasing in his value of searching for an alternative trade opportunity. In DGP, for example, where there is a unique trade size which is fixed and equal to unity, a decrease in trading frictions increases investors' value of searching for alternative trades, and the bid-ask spread narrows. We find that endogenizing investors' choices of asset holdings yields a richer set of empirical predictions for transaction costs. The fees or spreads that dealers charge still depend on the ease with which investors can find alternative trading partners, but in addition, they also depend on the extent of the mismatch between investors' endogenous asset positions and their valuations for the asset. Our model predicts a distribution of transaction costs, both across trade sizes—with spreads that increase with the size of the trade—as well as within a given trade-size category—across investors with different valuations. We also show that investors who are better informed about trading opportunities or have a higher bargaining power will trade larger quantities at a lower cost per dollar traded.

Trading delays, i.e., the fact that a counterpart for trade cannot be found instantaneously, is perhaps the fundamental distinguishing feature of the microstructure of an OTC market. The time it takes to execute a trade not only influences measures of liquidity such as trade volume and spreads, but it is often also used as a measure of market liquidity in itself. We endogenize trading delays by allowing free entry of dealers. This is a simple and natural financial market structure, yet one whose implications have not been explored so far.[5] In this context we again find that the model with unrestricted asset holdings bears new theoretical predictions, including a fundamental change of the equilibrium set. When interacted with the investors' unrestricted asset holding decisions, the dealers' incentives to make markets generate a liquidity externality which can give rise to multiple equilibria. Tight spreads are correlated with large volume and short trading delays across equilibria. Scarce liquidity can arise naturally as a self-fulfilling phenomenon in asset markets and a reduction in trading frictions can remove the multiplicity. Thus, perhaps counter to intuition, it is possible that a regulatory reform or a technological innovation that gives investors more direct access to the asset market (such as Electronic Communication Networks) leads to a relatively large increase in market liquidity and a higher volume of intermediated trades.

Finally, our model uncovers some novel insights regarding the welfare costs and inefficiencies associated with illiquid asset markets. With exogenous contact rates, the search equilibrium is efficient in the basic model of DGP. In particular, the dealers' market power has no effect on welfare: transaction costs are a pure transfer from investors to dealers. In contrast, the same ex-post bargaining protocol implies that asset holdings tend to be inefficient in our model because the dealers' market power distorts the investors' incentives to hold different asset positions. Investors with high valuations tend to invest too little, while those with low valuations tend to invest too much so the resulting equilibrium distribution of asset holdings is too concentrated. This inefficiency is eliminated if and only if dealers have no market power. With entry, dealers will not participate in market-making unless they have some market power. Thus, the inefficiencies on the intensive margin (investors' asset holdings) and the extensive margin (the number of dealers) cannot be corrected simultaneously: the size of the intermediation sector and investors' choices of asset holdings are generically inefficient.

The rest of the paper is organized as follows. We describe the basic environment in Section 2 and the equilibrium and its normative properties in Section 3. In Section 4 we analyze

[5] The practical relevance of this microstructure is described in Section 14.1 in Harris (2003).

the effects of trading frictions on the equilibrium distribution of asset holdings, trade volume and asset prices. Section 5 studies the effects of trading frictions on spreads. Section 6 deals with free entry. In Section 7 we use a parametrized version of the model to discuss some additional empirical predictions. In Section 8 we use the theory to analyze the effects of some recent technological and regulatory reforms in financial markets. Section 9 contrasts our main theoretical predictions to those of the related literature. Section 10 concludes.

2 Environment

Time is continuous, starts at $t = 0$ and goes on forever. There are two types of infinitely-lived agents: a unit measure of investors and a unit measure of dealers. There is one asset, one perishable consumption good called *fruit*, and another consumption good defined as *numéraire*. The asset is durable, perfectly divisible and in fixed supply, $A \in \mathbb{R}_+$. Each unit of the asset produces a unit flow of fruit. There is no market for fruit, so holding the asset is necessary to consume this good. The numéraire good is produced and consumed by all agents. The instantaneous utility function of an investor is $u_i(a) + c$, where $a \in \mathbb{R}_+$ represents the fruit consumption (which coincides with the investor's asset holdings), $c \in \mathbb{R}$ is the net consumption of the numéraire good ($c < 0$ if the investor produces more of these goods than he consumes), and $i \in \mathbb{X} = \{1, ..., I\}$ indexes a preference type.[6] The utility function $u_i(a)$ is continuously differentiable, strictly increasing and strictly concave.[7] Each investor receives a preference shock

[6] The fact that we assume a single type of asset is without loss of generality. As long as all assets are traded in the interdealer market we describe below, it would be easy to allow for any finite number of asset types. Formally, with m asset types an investor's asset holdings and utility function would be $\mathbf{a} \in \mathbb{R}_+^m$ and $u_i : \mathbb{R}_+^m \to \mathbb{R}$, respectively, but the analysis would essentially remain unchanged.

[7] Our specification associates a certain utility to the investor as a function of his asset holdings. This is a feature that we have borrowed from DGP. The utility the investor gets from holding a given asset position could be simply the value from enjoying the asset itself, as would be the case for real assets such as cars or houses. Alternatively, we can also think of the asset as being physical capital. Then, if each investor has linear utility over a single consumption good (as is the case in most search models), we can interpret $u_i(\cdot)$ as a production technology that allows the agent to use physical capital to produce the consumption good. The idiosyncratic component "i" can then be interpreted as a productivity shock that induces agents with low productivity to sell their capital to agents with high productivity in an OTC market. As yet another possibility, one could adopt the preferred interpretation of DGP, namely that $u_i(a)$ is in fact a reduced-form utility function that stands in for the various reasons why investors may want to hold different quantities of the asset, such as differences in liquidity needs, financing or financial-distress costs, correlation of asset returns with endowments (hedging needs), or relative tax disadvantages (as in Michaely and Vila (1996)). By now, several papers that build on the work of DGP have formalized the "hedging needs" interpretation. Examples include Duffie, Gârleanu and Pedersen (2006), Gârleanu (2006) and Vayanos and Weill (2007). (See also Lo, Mamaysky and Lang (2004).) These derivations typically start with investors who have CARA preferences, and then show that the risk-neutral approximation to this economy is essentially identical to the economy with linear reduced-form utility for the

with Poisson arrival rate δ. This process is independent across investors. Conditional on the preference shock, the probability the investor draws preference type i is π_i, with $\sum_{i=1}^{I} \pi_i = 1$.[8] These preference shocks capture the notion that investors will value the fruit from the asset differently over time thereby generating a need for investors to rebalance their asset positions. Dealers do not hold positions and their instantaneous utility is c, their consumption of the numéraire good.[9] All agents discount at rate $r > 0$.

There is a competitive market for the asset and dealers have continuous access to it. Investors do not have access to this competitive market but they contact dealers who can trade in this market on their behalf. Meetings with dealers occur at random according to a Poisson process with arrival rate α.[10] Once they have contacted each other, the dealer and the investor negotiate over the quantity of assets that the dealer will acquire for the investor and over the intermediation fee that the investor will pay the dealer for his services. After the transaction has been completed, the dealer and the investor part ways.

Asset holdings and preference types lie in the sets $\mathbb{R}_+$ and $\mathbb{X}$, respectively, and vary across investors and over time. We describe this heterogeneity with a probability space $(\mathbb{S}, \Sigma, H_t)$, where $\mathbb{S} = \mathbb{R}_+ \times \mathbb{X}$, Σ is a σ-algebra on the state space $\mathbb{S}$ and H_t is a probability measure on Σ which represents the distribution of investors across asset holdings and preference types at time t.

---

asset—a special case of our specification. For our purposes, the bottom line is that our preference structure is consistent with these formalizations, but moreover, that it is general enough to possibly accommodate other formalizations as well. Finally, notice that investors in DGP, and therefore the investors in our paper, are akin to the liquidity traders which are commonplace in the large body of the finance microstructure literature that uses asymmetric information instead of search frictions to rationalize bid-ask spreads, such as Glosten and Milgrom (1985) and Easley and O'Hara (1987).

[8] The assumption that the draw of the new preference shock is independent of the old preference shock allows us to solve for the value functions and the joint stationary distribution of portfolios and preference types in closed form. The assumption is otherwise inessential for our main results. See Gârleanu (2006) for an alternative formulation of these preference shocks.

[9] The restriction that dealers cannot hold assets is immaterial when analyzing steady-state equilibria. See Weill (2007) and Lagos, Rocheteau and Weill (2007) for dynamic equilibria where dealers can choose to hold positions.

[10] While our description of the trading process is stylized, it captures the salient features of the actual trading arrangements in OTC markets. We refer the interested reader to the discussion in Section 2.1 in Lagos and Rocheteau (2006).

3 Equilibrium

Let $V_i(a,t)$ denote the maximum expected discounted utility attainable by an investor who has preference type i and is holding portfolio a at time t. The value function $V_i(a,t)$ satisfies

$$V_i(a,t) = \mathbb{E}_i\Bigg[\int_t^{T_\alpha} e^{-r(s-t)}u_{k(s)}(a)ds + e^{-r(T_\alpha-t)}\{V_{k(T_\alpha)}[a_{k(T_\alpha)}(T_\alpha),T_\alpha] - p(T_\alpha)[a_{k(T_\alpha)}(T_\alpha)-a] - \phi_{k(T_\alpha)}(a,T_\alpha)\}\Bigg], \quad (1)$$

where T_α denotes the next time the investor contacts a dealer and $k(s) \in \mathbb{X}$ denotes the investor's preference type at time s. The expectations operator, $\mathbb{E}_i$, is taken with respect to the random variables T_α and $k(s)$ and is indexed by i to indicate that the expectation is conditional on $k(t) = i$. The first term on the right side of (1) contains the expected discounted utility flows enjoyed by the investor over the interval of time $[t, T_\alpha]$. The length of this interval, $T_\alpha - t$, is an exponentially distributed random variable with mean $1/\alpha$. The flow utility is indexed by the preference type of the investor, $k(s)$, which follows a compound Poisson process. The second term on the right side of (1) is the expected discounted utility of the investor from the time when he next contacts a dealer, T_α, onwards. At this time T_α the investor readjusts his asset holdings from a to $a_{k(T_\alpha)}(T_\alpha)$. In this event the dealer purchases $a_{k(T_\alpha)}(T_\alpha) - a$ in the market (or sells if this quantity is negative) at price $p(T_\alpha)$ on behalf of the investor. At this time the investor pays the dealer an intermediation fee $\phi_{k(T_\alpha)}(a, T_\alpha)$. Since the intermediation fee is determined in a bilateral meeting, it may depend on the investor's preference type and asset holdings.[11] Both the fee and the asset price are expressed in terms of the numéraire good.

Let $W(t)$ denote the maximum expected discounted utility attainable by a dealer. It satisfies

$$W(t) = \mathbb{E}\left\{e^{-r(T_\alpha-t)}\left[\int_{\mathbb{S}} \phi_i(a,T_\alpha)dH_{T_\alpha} + W(T_\alpha)\right]\right\},$$

where the expectations operator, $\mathbb{E}$, is taken with respect to T_α, which denotes the next time at which the dealer meets an investor. The random variable $T_\alpha - t$ is exponentially distributed with mean $1/\alpha$. Random matching implies that the investor whom the dealer meets is a random draw from the population of investors at time T_α. Thus, the dealer calculates the expected

[11] Our notation for the investor's new asset position, $a_{k(T_\alpha)}(T_\alpha)$, makes explicit that it may depend on time and on the investor's preference type at the time of the trade. In Lemma 1 we will show that the investor's new asset position is independent of the asset position he was holding at the time of the trade. To simplify the notation we anticipate this result and do not include the investor's asset holding at the time of the trade, a, as an argument of his new asset position.

intermediation fee using the measure of investors across preference types and asset holdings at time T_α, denoted H_{T_α}.

We turn to the determination of the terms of trade in bilateral meetings between dealers and investors. Consider a meeting at time t between a dealer and an investor of type i who is holding a. Let a' denote the investor's post-trade asset holdings and ϕ the intermediation fee. We take the pair (a', ϕ) to be the outcome corresponding to the Nash solution to a bargaining problem where the dealer has bargaining power $\eta \in [0, 1]$. The utility of the investor is $V_i(a', t) - p(t)(a' - a) - \phi$ if an agreement (a', ϕ) is reached, and $V_i(a, t)$ in case of disagreement. Therefore, the investor's gain from trade is $V_i(a', t) - V_i(a, t) - p(t)(a' - a) - \phi$. Analogously, the utility of the dealer is $W(t) + \phi$ if an agreement (a', ϕ) is reached and $W(t)$ in case of disagreement, so the dealer's gain from trade is the fee, ϕ.[12] The bargaining outcome is

$$[a_i(t), \phi_i(a, t)] = \arg\max_{(a', \phi)} [V_i(a', t) - V_i(a, t) - p(t)(a' - a) - \phi]^{1-\eta} \phi^\eta, \tag{2}$$

where the maximization is subject to the short-selling constraint $a' \geq 0$. The following lemma characterizes the bargaining solution taking the investor's value function as given.

Lemma 1 *The outcome of the bargaining problem (2) is*

$$a_i(t) = \arg\max_{a' \geq 0} \left[V_i(a', t) - p(t) a' \right], \tag{3}$$

$$\phi_i(a, t) = \eta \left\{ V_i[a_i(t), t] - V_i(a, t) - p(t)[a_i(t) - a] \right\}. \tag{4}$$

According to Lemma 1, the quantity of assets the investor buys, $a_i(t) - a$, maximizes the total surplus of the match (the sum of the dealer's and the investor's gains from trade). The intermediation fee is set to divide the total surplus according to each agent's bargaining power. From (3), it is immediate that the investor's new asset position, $a_i(t)$, is independent of a. Next, we use Lemma 1 to recast the investor's problem.

Substitute the terms of trade (3) and (4) into (1) to obtain

$$V_i(a, t) = \mathbb{E}_i \Bigg[\int_t^{T_\alpha} e^{-r(s-t)} u_{k(s)}(a) ds + e^{-r(T_\alpha - t)} \{ (1 - \eta) \max_{a' \geq 0} \left[V_{k(T_\alpha)}(a', T_\alpha) - p(T_\alpha)(a' - a) \right] + \eta V_{k(T_\alpha)}(a, T_\alpha) \} \Bigg]. \tag{5}$$

[12] It would be equivalent to set $\phi = (\hat{p} - p)(a' - a)$ and reformulate the bargaining problem as a choice of $(a' - a)$, the size of the order, and $\hat{p}$, the transaction price charged or paid by the dealer. So if $a' > a$ then the investor is a buyer and $\hat{p} > p$ can be interpreted as the *ask price* charged by the dealer. Conversely, if $a' < a$ then the investor is a seller and $\hat{p} < p$ is the *bid price* paid by the dealer.

It is apparent from (5) that the investor's payoff is the one he would get in an economy where he meets dealers according to a Poisson process with arrival rate α, but instead of bargaining, he readjusts his asset position and extracts the whole surplus with probability $1-\eta$, whereas with probability η he cannot readjust his asset position and enjoys no gain from trade. Thus, from the investor's standpoint, the stochastic trading process and the bargaining solution are payoff-equivalent to an alternative trading arrangement in which he has all the bargaining power in bilateral negotiations with dealers, but only gets to meet dealers according to a Poisson process with arrival rate $\kappa = \alpha(1-\eta)$. Let T_κ denote the next time the investor contacts a dealer in this economy. We can rewrite (5) as

$$V_i(a,t) = \bar{U}_i(a) + \mathbb{E}_i[e^{-r(T_\kappa - t)}\{p(T_\kappa)a + \max_{a' \geq 0}[V_{k(T_\kappa)}(a', T_\kappa) - p(T_\kappa)a']\}], \tag{6}$$

where

$$\bar{U}_i(a) = \mathbb{E}_i\left[\int_t^{T_\kappa} e^{-r(s-t)} u_{k(s)}(a) ds\right]. \tag{7}$$

The expectations operator, $\mathbb{E}_i$, is taken with respect to the random variables T_κ and $k(s)$, where $T_\kappa - t$ is exponentially distributed with mean $1/\kappa$.[13] From (6), the problem of an investor with preference shock i who gains access to the market at time t is given by

$$\max_{a' \geq 0}\left[\bar{U}_i(a') - \{p(t) - \mathbb{E}[e^{-r(T_\kappa - t)}p(T_\kappa)]\}a'\right]. \tag{8}$$

The investor chooses his asset holdings in order to maximize the expected present discounted value of his utility flow net of the expected present discounted value of the cost of holding the asset from time t until the next *effective time* T_κ when he can readjust his holdings. The following lemma offers a simpler, equivalent formulation of the investor's choice of asset holdings.

Lemma 2 *An investor with preference type i and asset holdings a who readjusts his asset position at time t solves*

$$\max_{a' \geq 0}\left[\bar{u}_i(a') - q(t)a'\right], \tag{9}$$

where

$$\bar{u}_i(a) = \frac{(r+\kappa)\, u_i(a) + \delta \sum_j \pi_j u_j(a)}{r + \kappa + \delta} \tag{10}$$

$$q(t) = (r+\kappa)\left[p(t) - \kappa \int_0^\infty e^{-(r+\kappa)s} p(t+s) ds\right]. \tag{11}$$

[13] As our notation makes explicit, $\bar{U}_i(a)$ is independent of t. This follows from the time-homogeneity of the Poisson meeting process and the Markovian process for $k(s)$. The right side of (7) only depends on t through the conditioning $k(t) = i$, which is captured by the "i" subscript.

Intuitively, $q(t) = (r+\kappa)\left\{p(t) - \mathbb{E}[e^{-r(T_\kappa - t)}p(T_\kappa)]\right\}$ is the opportunity cost plus the expected discounted capital loss, and $\bar{u}_i(a) = (r+\kappa)\bar{U}_i(a)$ the expected discounted utility (both expressed in flow terms) that the investor experiences by holding a from time t until his next opportunity to trade. Note that $\bar{u}_i(a)$ is a weighted average of the utilities in the various preference states. These weights depend on the transition rates α and δ, the discount rate r, the dealer's bargaining power, η, and the probability distribution $\{\pi_k\}_{k=1}^I$. It follows from Lemma 2 that an optimal choice of asset holdings $a_i(t)$ satisfies

$$\bar{u}_i'[a_i(t)] \leq q(t), \qquad \text{“}=\text{”} \quad \text{if } a_i(t) > 0. \tag{12}$$

Notice that we do not need to know the path for the price of the asset, $p(t)$, to solve for the investor's optimal asset holdings; $q(t)$ suffices. The following lemma establishes the relationship between $p(t)$ and $q(t)$.

Lemma 3 *(a) Condition (11) implies*

$$rp(t) - \dot{p}(t) = q(t) - \frac{\dot{q}(t)}{r+\kappa}. \tag{13}$$

(b) If $\lim_{t\to\infty} e^{-rt}p(t) = 0$*, then the price of the asset is*

$$p(t) = \int_t^\infty e^{-r(s-t)}\left[q(s) - \frac{\dot{q}(s)}{r+\kappa}\right]ds. \tag{14}$$

Part (a) of Lemma 3 provides additional insights into the investor's problem. Together with (12), (13) implies that if the investor holds the asset, his demand will satisfy $\bar{u}_i'[a_i(t)] + \dot{p}(t) - \frac{\dot{q}(t)}{r+\kappa} = rp(t)$. For example, if $\dot{q}(t) > 0$, then the investor will choose a smaller asset position than he would if he were not subject to trading delays (e.g., if he faced a very large κ for a given price trajectory). Part (b) shows how to recover the path of asset prices from the path of capital gains, $q(t)$. In Appendix B we show that $p(t)$ must satisfy $\lim_{t\to\infty} e^{-rt}p(t) = 0$ in any equilibrium, so we can appeal to this condition without loss of generality.

We can now simplify the expression for the intermediation fee that an agent in state i with asset holdings a pays the dealer who readjusts his asset position. From (4), $\phi_i(a,t) = \eta\{V_i[a_i(t),t] - V_i(a,t) - p(t)[a_i(t) - a]\}$, with $a_i(t)$ characterized by (12). If we use (6) to substitute for the value functions we arrive at

$$\phi_i(a,t) = \frac{\eta\{\bar{u}_i[a_i(t)] - \bar{u}_i(a) - q(t)[a_i(t) - a]\}}{r+\kappa}. \tag{15}$$

The intermediation fee depends on the dealer's bargaining power, η, the discount factor, r, and the transition rates α and δ. It also varies with the investor's asset holdings at the time the trade is executed, a, as well as with his desired asset holdings, a_i.

Next, consider the determination of investors' effective cost of holding the asset, $q(t)$. Since each investor faces the same probability of trade irrespective of his asset holdings, we appeal to the Law of Large Numbers to assert that over a small interval of time dt, the quantity of assets supplied in the interdealer market equals $\alpha dt A$.[14] Let $n_i(t)$ denote the measure of investors with preference type i at time t. The process for preference shocks implies $\dot{n}_i(t) = \delta\pi_i - \delta n_i(t)$ for all i. Hence,

$$n_i(t) = e^{-\delta t} n_i(0) + (1 - e^{-\delta t})\pi_i, \qquad \text{for } i \in \mathbb{X}. \tag{16}$$

The measure of type-i investors who trade through a dealer over a small interval of time dt is $\alpha n_i(t)dt$, so the demand for assets is $\alpha dt \sum_{i=1}^{I} n_i(t)a_i(t)$. The clearing condition for the asset market is

$$\sum_{i=1}^{I} n_i(t)a_i(t) = A. \tag{17}$$

If we use (12) to substitute $a_i(t)$ from (17), it becomes clear that this condition determines a unique $q(t)$.

Investors differ in their preference types and in their asset holdings. The heterogeneity in preference types is induced by the stochastic preference shocks and the heterogeneity in asset holdings is induced by the random trading process. Specifically, because prices vary over time, an investor's current asset holdings depend on the time that has elapsed since his last trade, as well as on his preference type at the time of his last trade. At every point in time there is a nondegenerate distribution of last contact times and of preference types at the last time of contact, and hence a nondegenerate distribution of asset holdings. Consider a set of asset holdings $\mathcal{A} \subseteq \mathbb{R}_+$ and a set of preference types $\mathcal{I} \subseteq \mathbb{X}$, then for all $(\mathcal{A}, \mathcal{I}) \in \Sigma$, $H_t(\mathcal{A}, \mathcal{I})$ defines the measure of investors whose asset holdings lie in $\mathcal{A}$ and whose preference types lie in $\mathcal{I}$. We characterize this probability measure in the following lemma, where we use $\mathbb{I}_{\{a \in \mathcal{A}\}}$ to denote an indicator function that equals 1 if $a \in \mathcal{A}$ and 0 otherwise.

Lemma 4 *The measure of investors across individual states at time t satisfies*

$$H_t(\mathcal{A}, \mathcal{I}) = \sum_{i \in \mathcal{I}} \sum_{j=1}^{I} \left[n_{ji}^0(\mathcal{A}, t) + \int_0^t \mathbb{I}_{\{a_j(t-\tau) \in \mathcal{A}\}} n_{ji}(\tau, t) d\tau \right] \tag{18}$$

[14] For a derivation of the Law of Large Numbers in random-matching environments, see Duffie and Sun (2007).

for all $(\mathcal{A},\mathcal{I}) \in \Sigma$,

$$n_{ji}(\tau,t) = \alpha e^{-\alpha\tau}\left(1-e^{-\delta\tau}\right)\pi_i n_j(t-\tau) \quad \text{if } i \neq j, \tag{19}$$

$$n_{ii}(\tau,t) = \alpha e^{-\alpha\tau}\left[\left(1-e^{-\delta\tau}\right)\pi_i + e^{-\delta\tau}\right] n_i(t-\tau) \tag{20}$$

and

$$n^0_{ji}(\mathcal{A},t) = e^{-\alpha t}\left[\left(1-e^{-\delta t}\right)\pi_i + e^{-\delta t}\mathbb{I}_{\{i=j\}}\right] H_0(\mathcal{A},\{j\}). \tag{21}$$

At time 0, the market starts off with investors distributed across preference types and asset holdings according to the initial probability measure H_0. At any subsequent time $t > 0$, there are two types of agents, those who have not contacted a dealer since time 0 and those who have. Among the former, the measure whose asset holdings and preference types lied in the set $(\mathcal{A},\{j\})$ at time 0 is $e^{-\alpha t}H_0(\mathcal{A},\{j\})$. At time t all these investors are holding the same asset position they were holding at time 0, but their preference types may have changed. The time-t measure of investors who started off with preference type j and assets in $\mathcal{A}$, whose preference type is i at the current time t, and who have never traded (so their asset holdings are still in $\mathcal{A}$) is $n^0_{ji}(\mathcal{A},t)$ as given in (21). Analogously, $n_{ji}(\tau,t)$ in (19) and (20) give the time-t density of investors who are holding asset position $a_j(t-\tau)$; i.e., those investors whose last trade was at time $t-\tau$ when their preference type was j, and who have preference type i at time t. We are now ready to define equilibrium.

Definition 1 *An equilibrium is a time-path* $\langle\{a_i(t)\}, q(t), p(t), \{\phi_i(a,t)\}, H_t\rangle$ *that satisfies (12), (14), (15), (17) and (18), given an initial condition* H_0.

Proposition 1 *There exists a unique equilibrium.*

The equilibrium can be found as follows. Equations (12) and (17) determine $\{a_i(t)\}$ and $q(t)$. Given $\{a_i(t)\}$ and $q(t)$, (14) and (15) imply $p(t)$ and $\phi_i(a,t)$, respectively. Finally, from $\{a_i(t)\}$ the distribution of investors' states is given by (18).

To illustrate how a reduction in trading frictions affects the equilibrium, consider the limiting case where trading delays vanish, i.e., $\alpha \to \infty$. From (10), $\bar{u}_i(a) \to u_i(a)$ and from (12) and (13), assuming an interior solution, we get $u_i'[a_i(t)] = q(t) = rp(t) - \dot{p}(t)$ for all i. Using (17) the effective cost of holding the asset converges to $q^*(t)$, which solves $\sum_{i=1}^{I} n_i(t)\, u_i'^{-1}[q^*(t)] = A$. From (15) we see that $\phi_i(a,t) \to 0$ for all a, i and t. With regards to the distribution of

investors, (21) implies that the measure of agents who have not contacted a dealer since time 0 vanishes; i.e., $n^0_{ji}(\mathcal{A}, t) \to 0$ for all i and j, all t and all $\mathcal{A} \subseteq \mathbb{R}_+$ as $\alpha \to \infty$. Also, as $\alpha \to \infty$, $H_t(\mathcal{A}, \mathcal{I}) \to \sum_{i \in \mathcal{I}} \mathbb{I}_{\{a_i(t) \in \mathcal{A}\}} n_i(t)$; i.e., every investor of every type i holds his desired portfolio $a_i(t)$ at all times.[15] Summarizing, as frictions vanish, investors choose $a_i(t)$ continuously by equating their current marginal utility from holding the asset to its effective cost—the flow opportunity cost minus the capital gain. The equilibrium fees, asset price and distribution of asset holdings are the ones that would prevail in a Walrasian economy.[16]

3.1 Efficiency

We now turn to the efficiency properties of the equilibrium. We study the problem of a social planner who maximizes the expected discounted sum of all agents' utilities. When choosing allocations, the planner is subject to the same frictions that investors and dealers face in the decentralized formulation studied above. Specifically, these frictions imply that over a small interval of time of length dt the planner can only reallocate assets among a measure αdt of investors chosen at random from the population. The planner chooses among allocations $\{a_i(t)\}_{i=1}^I$ that specify how to distribute the measure $\alpha dt A$ of assets among the measure αdt of investors whose asset positions he can reallocate at t.

Since the numéraire good enters linearly in the utility function of all agents, the consumption and production of these goods net out to 0 and can be ignored by the planner. Therefore, the planner only maximizes the investors' direct utilities from holding the asset. Given the initial measure H_0 of investors over asset holdings and preference types, the planner solves

$$\max_{\{a_i(t)\}_{i=1}^I} \left\{ K_0 + \int_0^\infty \sum_{i=1}^I e^{-rt} \alpha n_i(t) \hat{U}_i\left[a_i(t)\right] dt \right\}$$

$$\text{s.t.} \quad \sum_{i=1}^I \alpha n_i(t) a_i(t) \leq \alpha A, \tag{22}$$

$$\dot{n}_i(t) = \delta\left[\pi_i - n_i(t)\right] \tag{23}$$

[15] To see this, note that the time-t density of agents who have not contacted a dealer since time $t - \tau > 0$ is $n(\tau, t) = \sum_{i,j=1}^I n_{ji}(\tau, t)$. From (19) and (20), $\alpha \to \infty$ implies $n(\tau, t) \to 0$ for all $\tau > 0$, i.e., investors can find a dealer instantly when α is arbitrarily large, so the measure of investors who have not met a dealer between $t - \tau$ and t is zero for all $\tau > 0$. As for those investors who have met a dealer this "instant," we see from (19) and (20) that $n_{ji}(0, t) = 0$ for $i \neq j$ and $n_{ii}(0, t) = n_i(t)$.

[16] For related limiting results, but in stationary environments, see Duffie, Gârleanu and Pedersen (2005) and Miao (2006). In a different context, see Spulber (1996).

and $a_i(t) \geq 0$, for $i \in \mathbb{X}$, where

$$\hat{U}_i\left[a_i(t)\right] = \mathbb{E}_i\left[\int_t^{T_\alpha} e^{-r(s-t)} u_{k(s)}\left[a_i(t)\right] ds\right]$$

and $K_0 \equiv \int_{\mathbb{S}} \hat{U}_i(a) dH_0$. The expectations operator, $\mathbb{E}_i$, is taken with respect to the random variables T_α and $k(s)$, where $T_\alpha - t$ is exponentially distributed with mean $1/\alpha$. The constant K_0 captures the utility of all investors before they trade for the first time. The second term in the objective function states that over an interval of time of length dt, there is a measure $\alpha n_i(t) dt$ of investors of type i who can have their asset holdings rebalanced. An investor of type i is assigned a quantity of assets $a_i(t)$. The planner's choices must satisfy the resource constraint (23) and the law of motion for the measure of investors of each preference type.[17] The following proposition summarizes the efficiency properties of the equilibrium.

Proposition 2 *The equilibrium is efficient if and only if* $\eta = 0$.

The equilibrium with bargaining is efficient if and only if dealers have no bargaining power. This bargaining inefficiency is reminiscent of the traditional holdup problem emphasized in the investment literature. There is, however, a subtle difference. In our model the investor and the dealer bargain over both an intermediation fee and the quantity of the asset that the dealer trades on behalf of the investor. Hence, taking the behavior of the rest of the market as given, the investment decision is pair wise Pareto-efficient. The inefficiency arises because when conducting a trade, the investor anticipates that the intermediation fee he will have to pay to rebalance his asset holdings in his next encounter with a dealer will be increasing in the gains from that future trade. As a result, at the margin, investors are discouraged from taking positions that tend to lead to large asset reallocations in the future.

3.2 Steady state

Here we consider the limit of the equilibrium prices and allocations as $t \to \infty$.

[17] Since investors access the market according to independent stochastic processes with identical distributions, the measure of assets that can be reallocated among the α randomly drawn investors is $\alpha \int a dH_t = \alpha A$. Thus, the quantity of assets that can be reallocated among investors depends only on the mean of H_t, i.e., A, which is given. Consequently, the planner's decision of how to allocate assets at time t affects neither the measure of investors he will draw in the future nor the total measure of assets that these investors hold. In other words, H_t is not a state variable for the planner's problem.

Proposition 3 *For any H_0, the equilibrium allocations and prices described in Definition 1, $\langle \{a_i(t)\}, q(t), p(t), \{\phi_i(a,t)\}, H_t \rangle$, converge to the unique steady-state allocations and prices $\langle \{a_i\}, q, p, \{\phi_i(a)\}, H \rangle$, that satisfy*

$$\bar{u}_i'(a_i) \leq q \qquad \text{“}=\text{”} \quad \text{if } a_i > 0, \tag{24}$$

$$\sum_{i=1}^{I} \pi_i a_i = A, \tag{25}$$

$$p = \frac{q}{r}, \tag{26}$$

$$\phi_i(a) = \frac{\eta \left[\bar{u}_i(a_i) - \bar{u}_i(a) - q(a_i - a)\right]}{r + \kappa}, \tag{27}$$

$$H(\mathcal{A}, \mathcal{I}) = \sum_{i \in \mathcal{I}} \sum_{j=1}^{I} \left[\int_0^{\infty} \mathbb{I}_{\{a_j \in \mathcal{A}\}} n_{ji}(\tau, \infty)\, d\tau \right], \tag{28}$$

for all $(\mathcal{A}, \mathcal{I}) \in \Sigma$, where

$$n_{ji}(\tau, \infty) = \alpha e^{-\alpha\tau} \left[\left(1 - e^{-\delta\tau}\right) \pi_i + e^{-\delta\tau} \mathbb{I}_{\{i=j\}} \right] \pi_j. \tag{29}$$

Our notational convention is to omit the “t” argument in an endogenous variable whenever we refer to its steady-state value. The expressions (24), (25) and (26) are the steady-state counterparts of (12), (17) and (13), respectively.

In general, the individual state of an investor is a pair $(a, j) \in \mathbb{R}_+ \times \mathbb{X}$, where a is his current portfolio and j his current preference type. But according to (24), investors are only distributed among I levels of asset holdings in the steady state. The reason is that any level of asset holdings a such that $a \neq a_i$ for $i \in \mathbb{X}$ is transient, since whenever an investor adjusts his portfolio he chooses $a \in \{a_i\}_{i=1}^{I}$. Thus, the set of ergodic states is $\{a_i\}_{i=1}^{I} \times \mathbb{X}$. In other words, the steady-state measure $H(\mathcal{A}, \mathcal{I})$ is characterized by I^2 mass points. When analyzing the steady state we simplify the exposition and denote an individual investor's state $(a_i, j) \in \{a_i\}_{i=1}^{I} \times \mathbb{X}$ by $(i, j) \in \mathbb{X}^2$. Hence, for state (i, j), i represents the portfolio the investor currently has (i.e., the one corresponding to the preference shock he had at the time he last rebalanced his asset holdings), and j represents his current preference shock. We use n_{ij} to denote the steady-state measure of investors in state ij, i.e., $n_{ij} = H(\{a_i\}, \{j\}) = \int_0^{\infty} n_{ij}(\tau, \infty)\, d\tau$. From (28) and (29),

$$n_{ij} = \frac{\delta \pi_i \pi_j}{\alpha + \delta}, \quad \text{for } j \neq i, \tag{30}$$

$$n_{ii} = \frac{\delta \pi_i^2 + \alpha \pi_i}{\alpha + \delta}. \tag{31}$$

Notice that $\partial n_{ij}/\partial\alpha < 0$ if $j \neq i$ and $\partial n_{ii}/\partial\alpha > 0$, i.e., the measure of investors who are matched to their desired asset positions increases with the rate at which investors get to rebalance their asset holdings.

4 Asset positions, prices and trade volume

In this section we study the effects of search frictions on individual asset holdings and derive their implications for asset prices and trade volume. We focus on the steady state and specialize the analysis to utility functions of the form $u_i(a) = \varepsilon_i u(a)$. For this class of preferences, $\bar{u}_i(a) = \bar{\varepsilon}_i u(a)$, where $\bar{\varepsilon}_i = \frac{(r+\kappa)\varepsilon_i + \delta\bar{\varepsilon}}{r+\kappa+\delta}$ and $\bar{\varepsilon} = \sum_{j=1}^{I} \pi_j \varepsilon_j$ and (24) becomes[18]

$$\bar{\varepsilon}_i u'(a_i) = rp. \tag{32}$$

Let $a_i = g_i(\kappa; p)$ denote the choice of asset holdings characterized by (32) and differentiate it to get

$$\frac{\partial g_i(\kappa;p)}{\partial\kappa} = \frac{\delta(\bar{\varepsilon} - \varepsilon_i)}{(r+\kappa+\delta)^2} \frac{[u'(a_i)]^2}{rpu''(a_i)}. \tag{33}$$

The asset price, p, is kept fixed in this calculation, so we are isolating the partial equilibrium effect of κ on individual demand. Note that $\partial g_i(\kappa;p)/\partial\kappa$ has the same sign as $\varepsilon_i - \bar{\varepsilon}$. That is, investors with a preference shock above average increase their demand when κ increases. An agent with $\varepsilon_i > \bar{\varepsilon}$ anticipates that his preferences are likely to revert toward $\bar{\varepsilon}$ in the future, and that when this happens, he may be unable to rebalance his asset position for some time. Consequently, from (32), his choice of a_i is lower than $u'^{-1}(rp/\varepsilon_i)$, his choice of asset holdings in a world with no trading delays. A larger α means that it will be easier for the investor to find a dealer in the future; a lower η implies that it will be less costly to readjust his asset holdings in the future. In both cases the investor assigns more weight to his current marginal utility from holding the asset relative to its expected value. Conversely, investors with a preference shock below average reduce their demand when κ increases. These endogenous responses of individual investors' asset demands have important implications for the way illiquid markets operate. In Proposition 2 we have shown that this mechanism leads to allocative inefficiency if dealers have any degree of bargaining power. Next, we show how these reallocation effects shape the implications of search frictions for asset prices and trade volume.

[18] For notational simplicity, we focus on interior solutions unless otherwise specified.

Standard frictionless models emphasize two sets of factors that affect the determination of equilibrium asset prices, i.e., intrinsic properties of the asset and the characteristics of investors who buy it. Search theory identifies a third element: the manner in which the asset is traded, i.e., the details of the micro structure of the asset market, such as the rate at which investors meet dealers and the bargaining power of dealers. For example, suppose that the same asset is traded in two segmented OTC markets that differ only in the degree of market power of the dealers that participate in each market. All else equal, would we expect the asset price to be higher or lower in the market where dealers have more market power? In order to answer these types of questions we offer following proposition, which characterizes the effects of search frictions on asset prices.

Proposition 4 *If* $-\left[u'(a)\right]^2/u''(a)$ *is strictly increasing in* a*, then* $dp/d\kappa > 0$*. If* $-\left[u'(a)\right]^2/u''(a)$ *is strictly decreasing in* a*, then* $dp/d\kappa < 0$*. If* $-\left[u'(a)\right]^2/u''(a)$ *is independent of* a*, then* $dp/d\kappa = 0$*.*

For a given p, the demands of investors with relatively low valuations ($\varepsilon_i < \bar{\varepsilon}$) fall, while those of investors with high valuations ($\varepsilon_i > \bar{\varepsilon}$) rise. Whether an increase in κ raises the asset price depends on the curvature of the individual demand for the asset as a function of $\bar{\varepsilon}_i$, i.e., $\partial a_i/\partial\bar{\varepsilon}_i = -\left[u'(a_i)\right]^2/\left[u''(a_i)rp\right]$, and hence on the curvature of the utility function. If $u(a) = \log a$ then a_i is linear in $\bar{\varepsilon}_i$, and as one aggregates the individual changes in demands induced by an increase in κ, the increases in a_i (for investors with values of ε_i larger than $\bar{\varepsilon}$) and the decreases in a_i (for investors with values of ε_i lower than $\bar{\varepsilon}$) cancel each other out. As a result, κ has no effect on the aggregate demand for assets nor on the equilibrium price. This finding is consistent with the idea that trading frictions need not be reflected in asset prices.[19] If u is not too concave, a_i is a convex function of $\bar{\varepsilon}_i$. For this case, Jensen's inequality implies that the increases in a_i for relatively large values of ε_i outweigh the decreases in a_i for relatively low values of ε_i and the aggregate demand for the asset increases in response to an increase in κ. In turn, this implies that the equilibrium price of the asset increases with κ. Conversely, the asset price is decreasing in κ if u is sufficiently concave. By specializing preferences further we obtain the following corollary to Proposition 4.

Corollary 1 *Let* $u_i(a) = \varepsilon_i a^{1-\sigma}/(1-\sigma)$ *with* $\sigma > 0$*. If* $\sigma > 1$ (< 1) *then* $dp/d\kappa < 0$ (> 0) *and if* $u(a) = \log a$ *then* $dp/d\kappa = 0$*.*

[19] For a related result, see Constantinides (1986), Gârleanu (2006) and Heaton and Lucas (1996).

It is clear from (33) that regardless of the ultimate effect of search frictions on the asset price, an increase in κ makes high-valuation investors take on larger positions and low-valuation investors take smaller positions. This seems to suggest that the distribution of asset holdings will spread out if frictions are reduced. But this intuition based on (33) is only partial, because (33) keeps the equilibrium asset price constant. We characterize the full equilibrium effect of search frictions on the distribution of asset holdings in the following proposition.

Proposition 5 *Let $u_i(a) = \varepsilon_i a^{1-\sigma}/(1-\sigma)$ with $\sigma > 0$. An increase in κ causes the equilibrium distribution of asset holdings to become riskier, in the second-order stochastic sense.*

Proposition 5 confirms that the equilibrium distribution of asset holdings across investors becomes more disperse when trading frictions are reduced. This result is important to understand the relationship between search frictions and trade volume.

Trade volume is a manifestation of the extent to which the market mechanism is able to reallocate assets across investors. Large trade volumes are characteristic of liquid markets, i.e., markets where investors are able to switch in and out of asset positions relatively fast. Let $\mathcal{V}$ denote the volume of trade, defined as

$$\mathcal{V} = \frac{\alpha}{2} \sum_{i,j=1}^{I} n_{ij} \left| a_j - a_i \right|. \tag{34}$$

An increase in κ has three distinct effects on trade volume. First, the measure of investors in any individual state $(i, j) \in \mathbb{X}^2$ who gain access to the market and are therefore *able to trade* increases, which tends to increase trade volume. Second, the proportion $1 - \sum_{i=1}^{I} n_{ii}$ of agents who are mismatched to their asset position—and hence the fraction of agents who *wish to trade*—decreases, which tends to reduce trade volume. Finally, the distribution of asset holdings spreads out, which tends to increase the quantity of assets traded in many individual trades. With (30) and (34), it is possible to show that the first two effects combined lead to an increase in $\mathcal{V}$. While it is difficult to sign the third effect in general due to the general equilibrium effects of the price on asset holdings, we provide analytical results for two special cases.

First, consider the model with $I = 2$—the case analyzed by DGP. In this case it is possible to show that an increase in κ unambiguously leads to an increase in overall trade volume. This is formalized in the following result, which is a corollary of Proposition 5.

Corollary 2 *Let $u_i(a) = \varepsilon_i a^{1-\sigma}/(1-\sigma)$ with $\sigma > 0$, and $I = 2$. Trade volume, $\mathcal{V}$, increases with κ.*

The second special case allows for richer heterogeneity in types, but adopts a specification of preferences for which the equilibrium asset price is independent of trading frictions. In this case, we can get a sharp characterization of the effects of market structure on the distribution of trade sizes and on trade volume.

Proposition 6 *Let $u_i(a) = \varepsilon_i \ln a$. For any pair (κ, κ') such that $\kappa' > \kappa$, the distribution of trade sizes associated with κ' dominates the one associated with κ in the first-order stochastic sense.*

The proof of Proposition 6 consists of showing that with logarithmic preferences, a reduction in trading frictions increases the size of every trade, $|a_i - a_j|$. As a result, the volume of trade unambiguously increases with κ.

5 Transaction costs

The transaction costs borne by investors in OTC markets include the intermediation fees they are charged by the dealers who intermediate their trades.[20] In this section we study how changes in trading frictions affect intermediation fees. Our interest in these relationships is twofold. First, intermediation fees and the implied bid-ask spreads are among of the most common measures of market liquidity, since they quantify the out-of-pocket costs borne by the investors who trade in illiquid markets.[21] Second, these fees are an important source of revenue to the dealers who operate in these markets, and hence a key determinant of their incentives to make markets and provide liquidity, a theme that we will explore in detail in Section 6.

Intermediation fees depend on the rate at which investors can contact alternative dealers, on their bargaining power in bilateral negotiations and on the size of the trade (see (27)). The following lemma shows that, keeping the characteristics of an investor and a dealer constant, transaction costs—both total and per unit of asset traded—increase with the size of the trade. As it turns out, this link between intermediation fees and trade size shapes the general equilibrium effects of changes in trading frictions on transaction costs.

[20] Transaction costs in OTC markets also include trading delays, which are the focus of Section 6.

[21] See footnote 12 and Section 7 for the theoretical link between intermediation fees and bid-ask spreads.

Lemma 5 *Consider an investor who holds asset position $a \geq 0$ and wishes to trade $a_i - a > 0$. (i) $\partial\phi_i(a)/\partial a$ has the same sign as $a - a_i$ and (ii) $\frac{\partial}{\partial a}\left[\frac{\phi_i(a)}{a_i - a}\right] < 0$.*

In the general equilibrium, trading frequencies and bargaining power affect transaction costs through three channels. Consider, for example, the fee paid by an investor who currently has preference type i, and whose preference type was j at the time of his last trade. A larger α tends to reduce the fees that dealers can extract for any given trade size (it increases the denominator of (27)). Intuitively, a larger α implies better search options for the investor—the *competition effect* of reduced trading frictions. An increase in α also changes the investor's expected utility from holding his current asset position, a_j, relative to the expected utility from holding his desired asset position, a_i (it changes $\bar{u}_i$ in (27)). This effect may decrease or increase the fee he pays depending on the specific values of a_j and a_i. Finally, α affects the equilibrium levels of the actual and desired asset positions a_j and a_i themselves. A larger α induces investors to conduct larger asset reallocations every time they trade (see, e.g., Corollary 2 or Proposition 6). By Lemma 5, this translates into larger fees for dealers, on average—the *reallocation effect* of reduced trading frictions. These three effects can give rise to nonmonotonicities in the dealers' incentives to make markets in response to changes in the degree of trading frictions. We prove this result analytically for the case of "patient" traders, both for intermediation fees for individual trades (Proposition 7) and for average intermediation fees (Corollary 3). Notice that along a stationary equilibrium the only transactions that investors carry out involve trading $a_i - a_j$, for $(i, j) \in \mathbb{X}^2$. We use this observation to simplify the exposition and use ϕ_{ji} to denote $\phi_i(a_j)$.

Proposition 7 *For each $(i, j) \in \mathbb{X}^2$, there exists $\bar{r} > 0$, such that for all $r < \bar{r}$ and $\eta \in (0, 1)$, ϕ_{ji} is non-monotonic in κ and it is largest for some $\kappa \in (0, \infty)$.*

To interpret the finding in Proposition 7 consider the fees, ϕ_{ji}, borne by investors who currently have preference type i, and whose preference type was j at the time of their last trade. Notice that such an investor holds asset position a_i and engages in a trade that leaves him with asset position a_j, and that these asset positions are themselves functions of the degree of trading frictions, κ. In very illiquid markets ($\kappa \to 0$) investors hedge against future preference shocks by choosing asset holdings that reflect their average utility from holding the asset rather than their current utility at the time they trade. Consequently trade sizes are small, which makes the intermediation fees that these agents pay, ϕ_{ji}, also small. In very liquid

markets ($\kappa \to \infty$) investors trade large quantities but the intermediation fees that they pay are still small, because of favorable search options. For intermediate values of κ, trade sizes are considerable and dealers effectively have a significant degree of market power which results in relatively large intermediation fees.

Proposition 7 has implications for measures of market-wide transaction costs. Consider, for instance, the average fee charged by dealers across the various types of trades, denoted Φ. This average fee is also the expected revenue of an individual dealer conditional on meeting an investor; it equals $\sum_{i,j=1}^{I} n_{ji}\phi_{ji}$, or using (15),

$$\Phi = \eta \sum_{i,j=1}^{I} n_{ji} \frac{\bar{u}_i(a_i) - \bar{u}_i(a_j)}{r+\kappa}. \tag{35}$$

The average fee, Φ, depends on the mismatch between investors' desired and actual asset positions, as measured by $\bar{u}_i(a_i) - \bar{u}_i(a_j)$, as well as on the frequency with which they gain access to the asset market. The following corollary of Proposition 7 is useful to understand the dealers' incentives to make markets, which will be the focus of the following section.

Corollary 3 *There exists $\hat{r} > 0$, such that for all $r < \hat{r}$ and $\eta \in (0,1)$, the dealers' expected revenue, Φ, is non-monotonic in κ and it is largest for some $\kappa \in (0,\infty)$.*

According to Corollary 3, dealers are better off when they trade in markets which are neither too liquid nor too illiquid, i.e., when κ is neither to large not too small. If κ is very large, dealers would find it profitable to shift the trading activity to more illiquid markets, i.e., markets with larger η or smaller α. Conversely if κ is very small, perhaps surprisingly, dealers would benefit from reductions in η and increases in α.

6 Endogenous execution delays

In the previous sections we have shown how investors' endogenous choices of asset positions determine their effective demand for liquidity. For instance, if frictions are severe (κ small), desired and actual asset positions, $\{a_i\}$, tend to be very close to each other, so investors' trading needs, and hence their demand for liquidity services, are small. In this section we allow for free entry of dealers in order to endogenize the supply of liquidity services and the length of the trading delays. We formalize the notion—common in the finance microstructure literature—

that a dealer's expected profit depends on the competition for order flow that he faces from other dealers.[22]

Suppose that the Poisson rate at which an investor contacts a dealer, α, is a continuously differentiable function of the measure of dealers in the market, υ, with $\alpha(\upsilon)$ a strictly increasing and $\alpha(\upsilon)/\upsilon$ a strictly decreasing function of υ. We specify that $\alpha(0) = 0$, $\lim_{\upsilon\to\infty} \alpha(\upsilon) = \infty$ and $\lim_{\upsilon\to\infty} \alpha(\upsilon)/\upsilon = 0$. Since all matches are bilateral and random, the Poisson rate at which a dealer serves an investor is $\alpha(\upsilon)/\upsilon$. For larger υ, each investor contacts dealers faster, but the order flow decreases for each individual dealer.

There is a large measure of dealers who can choose to participate in the market. Dealers who choose to operate incur a flow cost $\gamma > 0$ that represents the ongoing costs of running the dealership, e.g., exchange membership dues, the cost of searching for investors, advertising their services and so on. Free-entry implies $\frac{\alpha(\upsilon)}{\upsilon}\Phi = \gamma$, i.e., that the expected instantaneous profit of a dealer equals his flow operation cost.[23] With (35), the free-entry condition becomes

$$\frac{\alpha(\upsilon)}{\upsilon}\eta \sum_{i,j=1}^{I} n_{ji} \frac{\bar{u}_i(a_i) - \bar{u}_i(a_j)}{r + \alpha(\upsilon)(1-\eta)} = \gamma. \tag{36}$$

A steady-state equilibrium with free entry is a list $\langle \{a_i\}, q, p, \{\phi_i(a)\}, \{n_{ji}\}, \upsilon \rangle$ that satisfies (24)–(27), (30), (31) and (36), with $\alpha = \alpha(\upsilon)$.

Proposition 8 *Assume $\eta > 0$. There exists a steady-state equilibrium with free entry of dealers, and it has $\upsilon > 0$.*

Proposition 8 establishes the existence of a steady-state equilibrium with free entry provided that dealers have some bargaining power. (Otherwise, intermediation fees would equal 0 in every trade and dealers would be unable to cover their operation costs.) Figure 1 provides a typical representation of a dealer's expected profit net of operation costs, $\frac{\alpha(\upsilon)}{\upsilon}\Phi - \gamma$, as a function of

[22] See, for example, Harris (2003, p. 298):

> "In competitive dealer markets, dealer spreads ultimately depend on the costs that dealers incur in running their business. The free entry and exit of dealers ensures that spreads will adjust so that dealers just earn normal profits. When spreads are too high, their competition for order flow will cause spreads to fall, and as spreads fall, so do expected profits."

[23] Our free entry of dealers is akin to the free entry of firms in Pissarides (2000). Rubinstein and Wolinsky (1987) also assume free entry of dealers (or *middlemen*). See Wahal (1997) or Weston (2000) for an empirical study of the determinants of entry and exit of market-makers in NASDAQ and their impact on spreads and the level of trading activity, e.g., trade volume and the number of trades.

the measure of dealers in the market. The average fee, Φ, is positive and bounded for all values of υ, while the dealers' contact rate goes to infinity as υ approaches 0 and to zero as υ goes to infinity. Therefore, a dealer's expected profit is strictly positive for small υ and approaches $-\gamma$ for large υ, so continuity ensures it must equal zero somewhere in between.

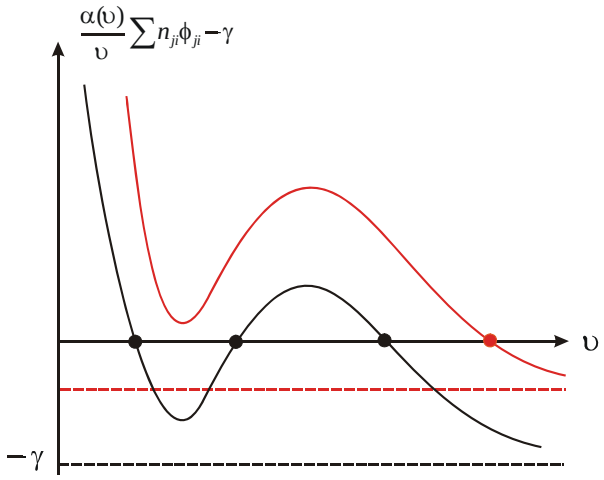

Figure 1: Multiple steady-state equilibria

The steady-state equilibrium with free entry need not be unique. While the measure of dealers, υ, is strictly increasing in Φ, according to Corollary 3 the dealers' expected revenue, Φ, can itself be a non-monotonic function of $\alpha(\upsilon)$. Faster trade means more competition among dealers, which tends to reduce intermediation fees. But an increase in $\alpha(\upsilon)$ also induces investors to take on more extreme asset positions which means that on average, dealers will intermediate larger asset reallocations and earn larger fees. The model will exhibit multiple steady states if the second effect is strong enough. (But for a given value υ, the rest of the equilibrium, $\langle\{a_i\}, q, p, \{\phi_i(a)\}, \{n_{ji}\}\rangle$, is uniquely determined as in the previous sections.) Generically there is an odd number of steady-state equilibria.[24]

The type of strategic complementarity that leads to multiple equilibria in this model is of a different kind than the one often found in other models of search equilibrium. In particular it is not due to increasing returns to scale in the meeting technology, as in Diamond (1982) or

[24] In our numerical examples we typically find either one or three equilibria. (See Lagos and Rocheteau (2007)). Note that the lowest and highest steady states in Figure 1 are "stable" in the following heuristic sense: if one perturbates slightly the measure of dealers from its steady-state value, free entry tends to bring the measure of dealers back towards its steady-state value.

Vayanos and Weill (2007). In fact, our assumption $\partial\left[\alpha\left(\upsilon\right)/\upsilon\right] < 0$ implies that dealers *reduce* the rate at which other dealers contact investors. What is key in our model is the liquidity externality generated by the way in which the dealers' incentives to make markets interact with the investors' unrestricted asset holding decisions. The equilibrium would be unique without this general equilibrium effect that operates through the endogenous shifts in investors' asset holdings in response to changes in the degree of trading frictions.

In the case of multiple equilibria, the market can be stuck in a low-liquidity equilibrium where few dealers enter and investors engage in relatively small transactions. The low-liquidity equilibrium exhibits large spreads, small trade volume and long trade-execution delays. Thus, tight spreads are correlated with large volume and short trading delays across equilibria. Steady-state welfare across equilibria increases with the measure of dealers. The high and low equilibria share the following comparative statics: a decrease in the participation cost of dealers raises the measure of dealers in the market. If the decrease in the participation cost is large enough, the multiplicity of equilibria can be removed. (The expected profits curve in Figure 1 shifts upward.) For the case of patient traders, the following proposition shows that the model necessarily exhibits multiple steady-state equilibria if $\alpha(\upsilon)/\upsilon$ is not too elastic and the dealer's cost of operation is in some intermediate range.

Proposition 9 *Assume* $\eta \in (0,1)$ *and* $\alpha\left(\upsilon\right) = \upsilon^{\theta}$*, with* $\theta \in (0,1)$*. There exist* $\tilde{r} > 0$*,* $\tilde{\theta} \in (0,1)$*,* $\overline{\gamma} > 0$ *and* $\underline{\gamma} \in (0, \overline{\gamma})$ *such that for all* $(r,\theta) \in (0,\tilde{r}) \times (\tilde{\theta},1)$*, there are multiple steady-state equilibria if* $\gamma \in (\underline{\gamma}, \overline{\gamma})$*.*

The presence of multiple equilibria due to the strategic interactions between investors and intermediaries may not be a mere theoretical curiosity. Biais and Green (2005), for example, document that the liquidity of the bond market on the NYSE dried up in the 1920's for municipal bonds and in the 1940's for corporate bonds, and attribute the ensuing shift in the structure of the market for bonds to "externalities in liquidity provision and strategic behavior by financial intermediaries."[25]

[25] Pagano (1989) provides a well-known model of multiple equilibria in a financial market. Both the model and the economic mechanisms that can give rise to multiplicity in his setup are quite different from the ones we are presenting here.

6.1 Efficiency

We investigate the efficiency properties of equilibrium with free entry of dealers. The dynamic planner's problem is complex since the measure of dealers at any point in time will typically depend on the whole distribution $H_t(a, i)$, not just its mean. To keep the analysis manageable, here we consider the case where the discount rate is close to 0, i.e., we characterize the allocation chosen by a social planner who maximizes steady-state welfare. In this case, the planner solves

$$\begin{aligned} &\max_{\{a_i\},\upsilon} \sum_{i,j} n_{ij} u_j(a_i) - \upsilon\gamma \\ &\quad \text{s.t.} \sum_{i,j} n_{ij} a_i = A, \end{aligned} \tag{37}$$

where the steady-state distribution $\{n_{ij}\}$ satisfies (30) and (31). The planner maximizes the population-weighted sum of investors' utilities from holding the asset, net of the participation costs of the dealers and taking into account that the stationary distribution $\{n_{ij}\}$ depends on the measure of dealers, υ.

Proposition 10 *Assume $r \approx 0$. An equilibrium with free-entry is efficient if and only if $\eta = 0$ and $\upsilon\alpha'(\upsilon)/\alpha(\upsilon) = \eta$.*

As before, investors' asset holdings are efficient if and only if dealers have no bargaining power. Entry introduces an additional inefficiency: when a dealer enters the market, he imposes a negative externality on other dealers' order flow. As it is well-known since Hosios (1990), these externalities are internalized if and only if the elasticity of the contact technology $\alpha(\upsilon)$ coincides with dealers' bargaining power. Since there is no free-entry equilibrium with $\upsilon > 0$ when $\eta = 0$, an equilibrium with entry is always inefficient.

7 More on spreads

The bid-ask spread is a key dimension of financial liquidity: it constitutes the investors' main out-of-pocket cost of trading and it determines the dealers' incentive to make markets. In this section we parametrize the steady state of the model of Section 2 and conduct numerical simulations that complement our previous analysis of transaction costs.

Consider an investor with asset holdings a_i who trades quantity $a_j - a_i \neq 0$. The *effective transaction price* he pays (or receives if $a_j - a_i$ is negative) is

$$\hat{p}_{ij} = p + \frac{\phi_{ij}}{a_j - a_i}$$

per unit of the asset. This effective transaction price can be interpreted as a *bid price* if the investor sells ($a_j - a_i < 0$) and as an *ask price* if he buys.[26] The price in the interdealer market is a natural benchmark against which to assess the cost of a trade since it is the price that an investor would pay if he had direct access to the market. The transaction cost to the investor per dollar traded is then $(\hat{p}_{ij} - p)/p = \phi_{ij}/p(a_j - a_i)$, which is sometimes referred to as the *liquidity premium*.

In our model there is not a single bid-ask pair: bid and ask prices vary with an investor's asset holdings and preference type. The average *effective spread*, $\mathcal{S}$, is a measure of marketwide trading costs often used in empirical work. It averages the bid and ask prices (expressed as a proportion of the asset price) across all trades, weighting each type of trade by its share in trade volume:

$$\mathcal{S} = \frac{1}{p}\sum_{i,j}\frac{n_{ij}\left|a_i - a_j\right|}{\sum_{k,\ell} n_{k\ell}\left|a_k - a_\ell\right|}\frac{\phi_{ij}}{\left|a_i - a_j\right|} = \frac{\alpha}{2}\frac{\Phi}{p\mathcal{V}}. \tag{38}$$

We study the effects of changes in α on $\mathcal{S}$ by means of a numerical example. We normalize the stock of assets by setting $A = 1$, let a unit of time correspond to a day and take the rate of time preference to be 10 percent per year, i.e., $r = 0.1/360$. We set $\delta = 1/7$ so that investors receive one preference shock every week on average, and take the average execution delay to be one day, i.e., $\alpha = 1$.[27] We assume that dealers and investors have equal bargaining power, i.e., $\eta = 0.5$. We let $u_i(a) = \varepsilon_i \ln a$.[28] The support for the values of ε_i is $\{\varepsilon_i\}_{i=1}^{I} = \{\frac{i-1}{I-1}\}_{i=1}^{I}$ with $I = 50$. The preference shock ε_i is drawn with probability

$$\pi_i = \frac{\lambda^{i-1}/(i-1)!}{\sum_{j=1}^{I}\lambda^{j-1}/(j-1)!}, \tag{39}$$

with $\lambda = 25$, which approximates a Normal distribution.

[26] This is in line with the equivalent formulation of the bargaining problem discussed in footnote 12.

[27] Trading delays in corporate bond markets range from a minute to a day, according to Saunders, Srinivasan and Walter (2002, p. 97).

[28] We know from Corollary 1 that in this case the asset price is independent of α. This specification is convenient because it will allow us to interpret the results corresponding to different values of α as corresponding to different markets with various degrees of trading frictions, or to different groups of investors with various individual contact rates with dealers.

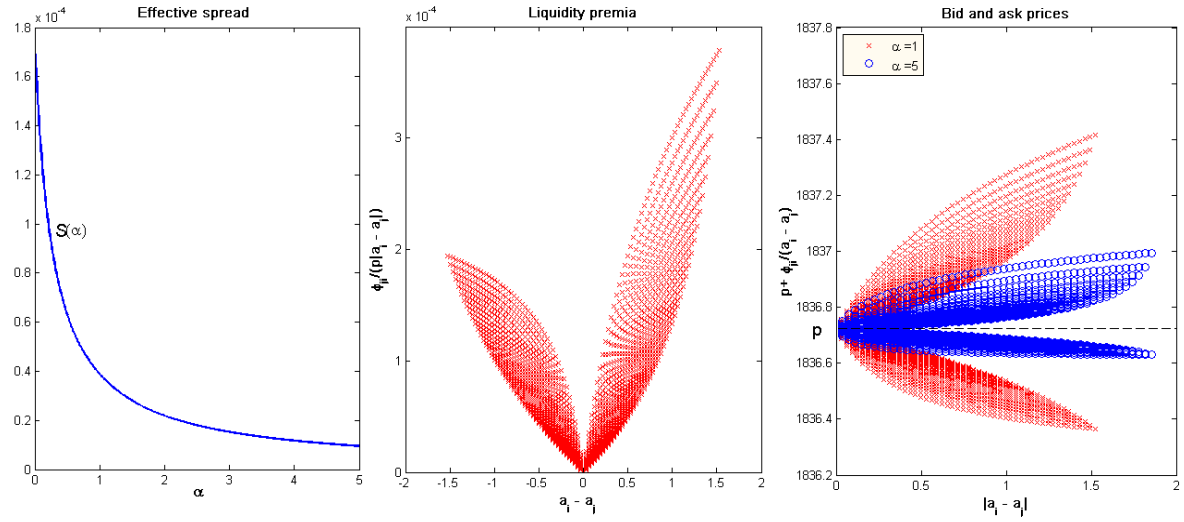

Figure 2: Spreads

The left panel of Figure 2 plots $\mathcal{S}$ for values of α ranging from 0 to 5. The effective spread decreases as trading delays are reduced. If we think of bond markets as being more "opaque" (i.e., they have lower α and higher η) than equity markets, this is consistent with several studies that have found that bond trades are substantially more expensive than equity trades.[29] The theoretical prediction is also in accordance with the observation that the adoption of electronic trading in equity markets has led to smaller spreads.[30]

Our model also has a rich set of predictions for individual transaction costs. The middle panel of Figure 2 displays the liquidity premium corresponding to each transaction, i.e., the pairs $\{(a_j - a_i, \phi_{ij}/p\,|a_i - a_j|)\}_{i,j=1,...,I}$. The right panel of Figure 2 plots the bid and ask prices for all trades as a function of the size of the trade, i.e., the pairs $\{(|a_j - a_i|\,, \hat{p}_{ij})\}_{i,j=1}^{I}$, for two different values of α (the crosses correspond to $\alpha = 1$ and the circles to $\alpha = 5$).

As illustrated in the middle panel of Figure 2, the liquidity premia, $\phi_{ji}/p\,|a_i - a_j|$, increase with the distance between a_j and a_i. Similarly, the right panel shows that the model generates an increasing relationship between trade size and the bid and ask prices for given α. These results are consistent with the empirical evidence on equity markets (e.g., Boehmer (2005, Table

[29] See Harris and Piwowar (2006) and Edwards, Harris and Piwowar (2005) for evidence on municipal and corporate bond markets, respectively.

[30] See Stoll (2006) and Section 8 below.

7, Panel B)) and the foreign exchange markets (Burnside et al. (2006, Table 12)), where larger trades pay larger transaction costs. These findings can be interpreted as instances of *price concession*—the commonly recognized fact that traders in illiquid markets often move prices against themselves in order to fill large orders.

Some empirical studies on municipal and corporate bond markets document that larger trades tend to be executed at a discount (Harris and Piwowar (2006) and Edwards, Harris and Piwowar (2004)). This pattern is often attributed to the fact that larger trades tend to be conducted by more "sophisticated" traders, i.e., traders who are better informed about trading opportunities or have stronger bargaining positions.[31] One can interpret the right panel of Figure 2 as a plot of the collection of individual spreads in an economy with heterogenous investors, some of which can contact dealers faster than others. Interestingly, sophisticated investors (here those who contact dealers with a larger Poisson intensity) trade larger quantities but pay lower spreads than less sophisticated investors. So our model can rationalize the effects on trade sizes and spreads that have been attributed to heterogeneity in the investors' degree of sophistication.[32]

Finally, note that the middle and right panels show that the model generates a distribution of transaction costs, not only across trade size categories, but also among trades of equal size, which is in accordance with the evidence from the market for municipal bonds (e.g., Green, Hollifield and Schurhoff (2006a)). This heterogeneity arises in the model because—with the two key features of OTC markets, trading delays and bargaining—two trades of equal size can pay different per-unit fees since the associated gains from trade may be different for the two trades.[33]

8 Market reforms

Over the last few years financial markets have been in the midst of a technological revolution. The advent of Electronic Communication Networks (ECNs)—private electronic screen-based

[31] According to Green, Hollifield and Schurhoff (2006a, p.1), "... some buyers appear to know which bonds are on 'sale' at a given point in time, and others do not." Green, Hollifield and Schurhoff (2006b) estimate dealers' bargaining power and find that it is higher for small to medium sized trades.

[32] In the theories of the bid-ask spread based on informational asymmetries, informed traders tend to trade larger quantities but they also face larger spreads. See, e.g., Easley and O'Hara (1987).

[33] Notice that in our parametrized example, $|a_i - a_j| = |a_k - a_\ell|$ whenever $|i - j| = |k - \ell|$, so there are many trades of identical size which have different associated gains from trade. Other parametrizations may have the property that a different $|a_i - a_j|$ corresponds to each (i, j) pair with $i \neq j$. But in such cases, analogously to what we find here, there will typically be trades of similar size but very different gains from trade.

trading systems built around computer algorithms that match buy and sell orders through an open limit-order book—is allowing investors to find trading opportunities more rapidly, and sometimes even directly, without the intervention of traditional intermediaries. The widespread use of these new technologies has the potential to accelerate trade execution and drastically reduce the transaction costs borne by investors.[34]

While these technological innovations were underway, several regulatory order-handling reforms were introduced in financial markets to foster competition among securities dealers and reduce their market power.[35] These reforms have in many cases effectively granted investors direct access to interdealer markets, opening up the possibility that they may trade directly with other investors, circumventing dealers.[36] In this section we analyze the implications of these technological and regulatory transformations and relate our theoretical findings to the existing empirical literature. We first extend our model by allowing investors periodic direct access to the interdealer market. We then discuss the effects of changes in the degree of market power of dealers, which captures the effects of the various efforts (e.g., decimalization, pressures to reduce spreads) to reduce their ability to charge large spreads.

8.1 ECNs

Suppose that in addition to periodically meeting dealers, investors get periodic direct access to the competitive interdealer market. This formulation captures the increased competition for order flow from ECNs faced by dealers as a result of the order-handling regulatory changes. Specifically, suppose that investors can access the competitive market either through a dealer,

[34] See Stoll (2006) and Allen et al. (2001) for an account of the emergence of ECNs in U. S. equity and fixed-income markets, respectively.

[35] See Barclay et al. (1999) for a detailed description of the New Order Handling Rules that were introduced by the Securities Exchange Commission. These regulatory reforms followed the Christie and Schultz (1994) finding that NASDAQ dealers avoided odd-eighth quotes in 70 of the largest 100 NASDAQ stocks in 1991, which led them to argue that dealers tacitly colluded to keep bid-ask spreads wide.

[36] Before the implementation of the order-handling reforms, all NASDAQ order flow had to be routed to some NASDAQ dealer who could trade through or ahead of the investors' orders, so investors were unable to bypass dealers who quoted wide spreads. In addition, NASDAQ dealers had exclusive access to an ECN (SuperMontage) that effectively acted as an interdealer market outside the reach of regular investors, allowing dealers to quote one set of prices for retail customers on NASDAQ while offering more favorable prices to other marketmakers on the ECN. Under the new order-handling rules, if a dealer places an order on an ECN, the price and quantity are incorporated in the ECN quote displayed on NASDAQ. Also, the investors' orders can now be routed to the interdealer ECN, and can compete directly with the NASDAQ dealers' quotes. These days many other interdealer markets are open to investors, examples include the trading platforms BrokerTec and E-Speed.

whom they contact with Poisson rate α, or directly, with Poisson rate β.[37] The Bellman equation for $V_i(a)$ can be shown to satisfy (6) with $\kappa = \alpha(1-\eta)+\beta$. The steady-state distribution $(n_{ij})_{i,j=1}^{I}$ is given by (30)-(31) where α is replaced by $\alpha+\beta$. Equilibrium exists and is unique by Proposition 1.

From Proposition 7 and Corollary 3, we know that an increase in β can have non-monotonic effects on transaction costs and dealers' revenue. In particular, if the market is initially very illiquid (α small or η large), then a regulatory or technological change that grants investors direct access to the interdealer market will increase the dispersion of investors' asset holdings and increase the distribution of trade sizes. In turn, this can increase the dealers' incentives to make markets and hence the endogenous level of intermediation—despite the fact that they are subjected to more intense competition for order flow. We formalize this as follows.

Proposition 11 *If $\eta = 1$, there exists a unique equilibrium with entry, and it has $\upsilon > 0$. There is $\bar{r} > 0$ such that for all $r < \bar{r}$, υ is nonmonotonic in β. Moreover, for all $r < \bar{r}$, υ is largest for some $\beta \in (0, \infty)$.*

The proof follows immediately from Proposition 8 and Corollary 3, so we omit it.[38] When β is small, investors of all types choose asset holdings very close to A so dealers have no incentive to participate in market-making. As β increases, the resulting increase in the dispersion of the distribution of asset holdings leads to larger trades on average, which stimulates the entry of dealers. So in contrast to what casual intuition might suggest, the level of intermediation need not be decreasing in the degree of competition for order flow faced by dealers. This finding is interesting because in the midst of the recent regulatory and technological changes, concerns have been raised that increased competition from alternative trading networks could

[37] In this formulation, whether a trade is conducted through a dealer or directly in the interdealer market is determined exogenously. Notwithstanding, notice that the effects of competition between these modes of trading manifest themselves in the equilibrium prices and allocations. For example, if β increases, this subjects dealers to more competition and affects the distribution of trade sizes and bid-ask spreads. Miao (2006) studies a model where investors can choose to trade in a centralized market intermediated by market-makers or in a decentralzied market where traders search for counterparties, and therefore he makes the competition among these markets explicit. Another difference is that in our model investors continuously cycle between being buyers and sellers, while in Miao (2006) agents trade once, exit the market and get replaced by new agents.

[38] To highlight the main point, the proposition specializes to the case of $\eta = 1$ because equilibrium is unique in this case. To see this notice that $\eta = 1$ implies that $\{\bar{u}_i(\cdot)\}$ and $\{a_i\}$ are independent of α. In this case the average fee only depends on $\alpha(\upsilon)$ through the distribution of investors. As the number of dealers increases, a larger measure of investors hold their desired portfolios, which reduces dealers' opportunities to intermediate trades. Thus, Φ is strictly decreasing in α (and υ) and the left side of (36) is strictly decreasing in υ, which implies uniqueness of the steady-state equilibrium with entry.

reduce dealers' incentives to make markets, adversely affecting the liquidity of the market. Weston (2000, 2002) finds that the increase in competition resulting from the growth of trading through ECNs in NASDAQ has resulted in larger trade volumes, tighter bid-ask spreads per dollar traded and net *entry* of market-makers, which can be rationalized by Proposition 11.[39]

Interestingly, in cases in which multiple equilibria exist (see Proposition 9), one can show that a reduction in trading frictions can remove the multiplicity. Thus, perhaps counter to intuition, it is possible that a mild regulatory reform or a technological innovation that gives investors more direct access to the asset market leads to a relatively large increase in market liquidity and even result in a higher volume of *intermediated* trades.

8.2 Dealers' market power

As we mentioned above, many of the regulatory reforms implemented in the 1990's were intended to reduce the dealers' ability to charge large spreads.[40] Within the context of our model, η captures the effects of regulations and other details of the market structure that determine a dealer's ability to extract a rent in their trades with investors. As we noted in the previous section, $\kappa = \alpha(1-\eta)+\beta$, so the effects of a decrease in η are essentially the same as the effects of an increase in β, namely, a larger trade volume and a smaller effective spread.[41]

So far we have focused on the steady-state effects of changes in policy and market structure on market liquidity. We now use the model with a fixed measure of dealers to illustrate how market reforms, in particular changes in the degree of market power of dealers, affect the dynamics of the different dimensions of market liquidity. We focus on the behavior of trade

[39] Barclay et al. (1999) also find that spreads fell significantly in NASDAQ in response to the order handling reforms of the 1990's without adversely affecting quality of execution. Similarly, Stoll (2006) documents that the widespread use of electronic trading in stock markets has led to tighter bid-ask spreads per dollar traded, larger trade volume and larger total revenue for securities firms. See Allen, Hawkins and Sato (2001) and Weston (2002) for references to related work.

[40] *Decimalization* is an example of such a policy. Up to the 1990's, US stocks had been priced in units of 1/8 of a dollar. Partly in response to the impact of the findings of Christie and Schultz (1994), Congress passed the Common Cents Pricing Act of 1997, which required the minumum tick size to be a penny. The tick size sets a floor on how narrow the spreads can become. So, in principle, this decimalization would foster competition among dealers. During that period there were also more direct pressures on dealers to reduce spreads. For more on this, see the accounts in Christie, Harris and Schultz (1994) and Hasbrouck (2004).

[41] In Lagos and Rocheteau (2007) we use a special case of the model with free entry to show that a reduction in η can lead to a decrease in effective spreads, an increase in trade volume and at the same time stimulate dealer entry and reduce execution delays. There we also show that provided that the initial market power of dealers is large enough, a reduction in η also increases welfare, which may lend some theoretical support for the regulatory reforms of the 1990's.

volume and the effective spread, which in a dynamic context equal

$$\mathcal{V}(t) = \frac{\alpha}{2} \int_{\mathbb{S}} |a_i(t) - a| \, dH_t \qquad \text{and} \qquad \mathcal{S}(t) = \frac{\int_{\mathbb{S}} \phi_i(a,t) \, dH_t}{p(t) \int_{\mathbb{S}} |a_i(t) - a| \, dH_t}.$$

Consider an economy where dealers have market power η^0, which is initially at the corresponding steady state. Let a_i^0 denote the asset position chosen by an agent with preference type i when he reallocates his asset holdings in this steady state. The initial measure of investors with asset holdings a_i^0 and preference type j is n_{ij} given by (30) and (31). We study an unanticipated regulatory change at time $t = 0$ that permanently reduces the dealers' market power from $\eta^0 = 0.75$ to 0.5 (the rest of the parameters are as in Section 7). Figure 3 shows the trajectories for trade volume and the effective spread.[42]

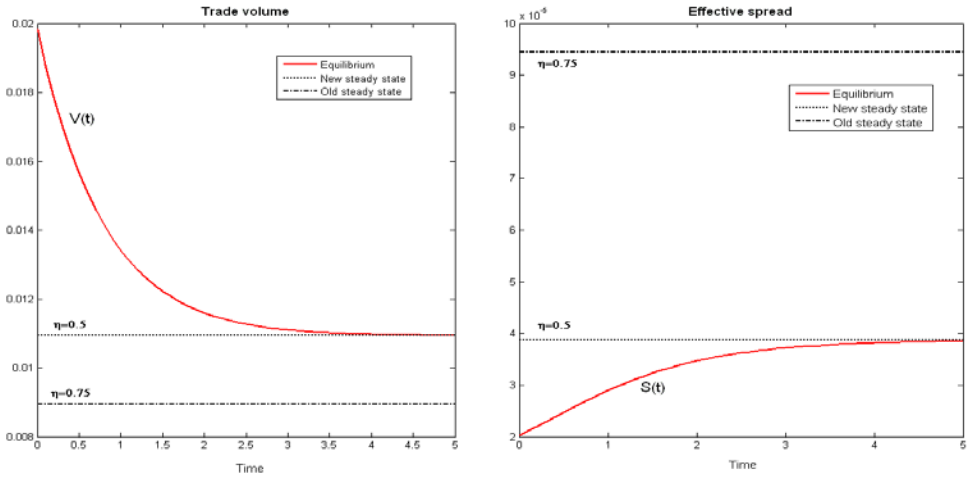

Figure 3: Effects of a permanent and unanticipated change in η on market liquidity

Trade volume is higher in the new steady state, as illustrated in the left panel of Figure 3. The reason is that the distribution of trade sizes associated with η dominates in a first-order-stochastic-dominance sense the one associated with $\eta^0 > \eta$ (Proposition 6). Trade volume overshoots its new steady-state level on impact. This is due to the evolution of the distribution of trade sizes over time. At time 0 the support of the distribution of asset positions is $\{a_i^0\}_{i=1}^I$,

[42] Since the distribution of preference types remains constant, $n_i(t) = \pi_i$ for all t, the effective cost of holding the asset typically jumps instantly to its new steady-state level. In the case with $\sigma = 1$, the asset price, p, is independent of η, so it remains constant. Since $p(t)$ is constant for all $t > 0$, $a_i(t)$ and $\phi_i(a,t)$ are also independent of t.

which coincides with the set of desired portfolios before the policy change, but the set of investors' desired portfolios becomes $\{a_i\}_{i=1}^I$ in response to the change in dealers' bargaining power. Hence, all investors adjust their asset holdings as they get access to the market, even those who were holding their desired asset holdings before the market reform. Trade volume approaches its new steady-state level as the measure of investors who have not contacted a dealer since the market reform goes to 0. The right panel of Figure 3 shows the trajectory for the effective spread. Along the transition, the effective spread undershoots its new, lower steady-state level. This behavior is driven by the time-paths for trade volume and the distribution of trade sizes.[43] The adjustment process to the new steady state occurs fast in this example: the equilibrium has essentially converged to the steady state after 5 days.

9 Related literature

Traders who operate in markets with OTC-style frictions will seek to mitigate these trading frictions by adjusting their asset positions so as to reduce their trading needs. Our previous analysis has shown that this a critical aspect of investor behavior in illiquid markets. To illustrate this point, in this section we derive the main predictions of a version DGP's model and contrast them with those of a special case of our formulation. This comparison will underscore the fact that the type of "liquidity hedging" that we have identified—and that only becomes possible with unrestricted asset holdings—generates new insights on how trading frictions shape the various dimensions of market liquidity, alters the empirical predictions of the theory, and leads to a different assessment of their normative implications.

We will contrast the empirical predictions of DGP's model with those of a special case of our model with $\mathbb{X} = \{1, 2\}$ and $u_i(a) = \varepsilon_i \frac{a^{1-\sigma}}{1-\sigma}$ for $i \in \mathbb{X}$ and $\sigma > 0$. We focus on the version of DGP's model with no inter-investor meetings (e.g., the version that DGP use in their Theorem 4 and part (i) of Theorem 6). DGP restrict $a \in \{0, 1\}$ and let u_{ij} denote the flow utility of an investor with asset position $i \in \{0, 1\}$ and preference type $j \in \{0, 1\}$.[44] DGP assume

[43] To get a sense for the dynamics of the distribution of trade sizes, first note that all investors are mismatched with their asset positions in the aftermath of the market reform. The trades conducted along the transition path by investors who in the absence of the policy change would have been holding their desired asset position have sizes $\left|a_i - a_i^0\right|$, which tend to be smaller than the trade sizes $|a_i - a_j|$, i.e., the only ones that will be carried out in the new steady state. Since spreads increase with trade size (Lemma 5), the dynamics of the distribution of trade sizes tends to make the effective spread low initially.

[44] DGP state their restriction on asset holdings as $a \in [0, 1]$ but only study equilibria in which agents hold either 0 or 1 unit of the asset, which is effectively equivalent to imposing the restriction $a \in \{0, 1\}$.

$u_{00} = u_{01} = 0$, so for comparison purposes, we do the same hereafter. To simplify the notation, in both models we let π denote the steady-state fraction of investors with high valuation.[45]

Trade volume. Trade volume is $\mathcal{V} = \alpha \frac{\delta\pi(1-\pi)}{\alpha+\delta} \frac{(\bar{\varepsilon}_2)^{1/\sigma} - (\bar{\varepsilon}_1)^{1/\sigma}}{\pi(\bar{\varepsilon}_2)^{1/\sigma} + (1-\pi)(\bar{\varepsilon}_1)^{1/\sigma}} A$ in our model and $\mathcal{V}_{DGP} = \alpha \frac{\delta\pi(1-\pi)}{\alpha+\delta} \min\{\frac{A}{\pi}, \frac{1-A}{1-\pi}\}$ in DGP. The latter is independent of the dealers' bargaining power, η, and of all preference parameters and holding payoffs (e.g., r, k). In contrast, these parameters are critical determinants of trade volume in our theory, as they influence the investors' choices of asset holdings (the second factor in $\mathcal{V}$). As discussed in Section 8, our model predicts—in accordance with available evidence—that markets in which dealers have less market power will tend exhibit larger trade volume.[46]

Price. Since asset holdings are indivisible in DGP, equilibrium in the interdealer market requires investors who are on the long side of the market to be indifferent between trading and not trading. It is easy to show that in steady state investors who want to sell are on the short side if and only if $A < \pi$. The equilibrium price in the interdealer market is

$$p = \begin{cases} \frac{1}{r} \frac{(r+\kappa)u_{11} + \delta\bar{u}}{r+\kappa+\delta} & \text{if } A < \pi \\ \frac{1}{r} \frac{(r+\kappa)u_{10} + \delta\bar{u}}{r+\kappa+\delta} & \text{if } \pi < A, \end{cases} \tag{40}$$

where $\bar{u} \equiv \pi_1 u_{11} + \pi_0 u_{10}$.[47]

The asset holding restrictions in DGP are also the reason why the asset price in their theory is independent of the stock of assets, A, for any $A < \pi$ and for any $A > \pi$, with a discontinuity at $A = \pi$. In contrast, the asset price in our model is smooth and decreasing in A. For example, in the special case of our model that we are considering in this section, $p = \left(\sum_i \pi_i \bar{\varepsilon}_i^{1/\sigma}\right)^{\sigma} / rA^{\sigma}$.[48] The behavior of the asset price in response to changes in the trading frictions in DGP depends critically on the level of A. From (40), p is increasing in α (decreasing in η) if $A < \pi$ but decreasing in α (increasing in η) if $A > \pi$. In contrast, with unrestricted asset holdings these extensive-margin considerations are irrelevant to assess the impact of trading frictions on the asset price (recall Proposition 4).

[45] "High valuation" corresponds to the index "2" in our formulation and "1" in DGP.

[46] Apart from these qualitative differences, the theory with unrestricted portfolios also has different quantitative implications for the relationship between trade volume and trading frictions. For example, DGP's model has a sharp empirical implication: the elasticity of trade volume with respect to trading frictions equals $\frac{\delta}{\alpha+\delta} \in (0,1)$. In contrast, in the model with unrestricted asset holdings the corresponding elasticity is larger by an amount that equals the elasticity of $(a_2 - a_1)$ with respect to α—which is positive, capturing the notion that each investor wishes to conduct a larger trade when frictions are reduced.

[47] If $A = \pi$, $p \in \left[\frac{(r+\kappa)u_{10} + \delta\bar{u}}{r(r+\kappa+\delta)}, \frac{(r+\kappa)u_{11} + \delta\bar{u}}{r(r+\kappa+\delta)}\right]$ and the equilibrium price in the interdealer market is indeterminate.

[48] Notice that we obtain DGP's formulation with $A < \pi$ as a special case of ours when $\sigma \to 0$.

A paper that is closely related to ours is an independent contribution by Gârleanu (2006), which studies the asset pricing implications of infrequent (Poisson) trading opportunities. Some of our findings are similar: like us, he finds that investors take more extreme positions when trading delays are short. Also, Gârleanu stresses that the asset price is not affected by the trading frictions—which is true in our model for a particular specification of the utility function (see Corollary 1). In terms of differences, trades in Gârleanu (2006) are not intermediated by dealers ($\alpha = 0$ in our formulation) so he could not consider the implications of execution delays for transaction costs and dealers' incentives to provide liquidity, which are at the center of our analysis.

Transaction costs. DGP's transaction costs can be expressed in terms of the intermediation fees ϕ_{01} and ϕ_{10} that dealers charge investors who want to buy and sell, respectively. The equilibrium spread is $s = \frac{\eta(u_{11}-u_{10})}{r+\kappa+\delta}$.[49] Conditional on having contacted an investor, the expected intermediation fee that accrues to a dealer in DGP is $\Phi_{DGP} = \frac{\delta\pi(1-\pi)}{\alpha+\delta}\min\{\frac{A}{\pi}, \frac{1-A}{1-\pi}\}s$. This key determinant of dealers' incentives to make markets is decreasing in the investors' contact rate with dealers, α, and increasing in the dealers' bargaining power, η. In contrast, as we have shown analytically in Proposition 3, in our model with no restrictions on asset holdings it is natural for the average fee to be nonmonotonic in α and η. Our theory suggests that these nonmonotonicities can be important. From an applied standpoint, they help explain how OTC markets have reacted to recent changes in their market structure (Section 8). From a theoretical standpoint, they can generate self-fulfilling liquidity shortages in markets with free entry of dealers (Section 6).[50]

Another key difference with DGP is the fact that since the equilibrium in the model with unrestricted portfolios implies a nondegenerate distribution of trade sizes, our theory has predictions for the relationship between transaction costs and transaction sizes. As we showed

[49] Since asset holdings in DGP are restricted to lie in $\{0, 1\}$, every trade is of size 1 and hence $\phi_{01} + \phi_{10} = s$. In addition, the indivisibility assumption implies that dealers either charge a fee on asset sales or on asset purchases, but not both. Specifically, if $A < \pi$ then $\phi_{01} = 0$ and investors only pay a fee $\phi_{10} = s$ when they sell. Conversely, if $\pi < A$, $\phi_{10} = 0$ and investors only pay a fee $\phi_{01} = s$ when they buy.

[50] The spread, s, is decreasing in α and increasing in η in this version of DGP with no inter-investor meetings. One can also verify that the average effective spread weighted by the sizes of each trade and expressed as a proportion of the price is also decreasing in α and increasing in η. The behavior of this measure of the marketwide spread, i.e., (38), is much more complicated in our model, where the investors' expected holding payoffs, their individual asset demands, the asset price, and the whole distribution of asset holdings change in response to a change in α. Our numerical work, some of which we have reported in Section 7, is in accordance with the predictions of DGP.

in Lemma 5 and illustrated in Section 7, transaction costs are increasing in the size of the transaction. Thus, if $a_i - a_j > a_i - a_k > 0$, then the effective price at which the investor buys is $\hat{p}_{ji} > \hat{p}_{ki}$, i.e., he effectively pays higher prices when he conducts larger purchases. Conversely, $\hat{p}_{ji} < \hat{p}_{ki}$ if $a_i - a_j < a_i - a_k < 0$, i.e., he effectively receives lower prices when he conducts larger sales. In other words, the theory with unrestricted asset holdings naturally generates instances of *price concession* which are commonplace in OTC markets.[51]

Execution delays. DGP endogenized trading delays by allowing a single monopolist dealer to choose search intensity once-and-for-all at the beginning of time. Free entry of competing dealers or market-makers is a feature of most OTC markets, however, the implications of this microstructure have not yet been explored in the literature. We find that allowing for free entry of dealers is a natural way to endogenize execution delays and the amount of liquidity supplied by dealers, and that it provides an important channel through which changes in market conditions affect transaction costs and trade volume (Section 8). In addition, the interaction between free entry and unrestricted asset holdings leads to a natural kind of strategic complementarity that can help rationalize self-fulfilling liquidity shortages in markets with OTC-style frictions (Section 6).

Welfare. The equilibrium allocation is always constrained efficient in the baseline model of DGP—regardless of the value of η—which stands in contrast to the finding we report in Proposition 2. The reason is that in our model investors choose asset holdings, while this intensive margin is absent in DGP. For the same reason, the inefficiency result we find in Proposition 10 also has no counterpart in DGP.

10 Conclusion

We developed a search-theoretic model of an asset market and have used it to analyze the relationship between the fundamental trading frictions characteristic of OTC markets (trading delays, dealers' market power) and standard measures of financial liquidity, such as the size of bid-ask spreads, trade volume and execution delays. We have shown that the theory can be used to analyze the positive and normative implications of recent regulatory and technological innovations in trading.

[51] See Section 4.3 in Harris (2003).

From a methodological standpoint, our work shows that by imposing severe restrictions on asset holdings, existing search-based theories of financial liquidity neglect a critical aspect of investor behavior in illiquid markets, namely the fact that market participants can mitigate trading frictions by adjusting their asset positions so as to reduce their trading needs. We have found that this mechanism, which effectively amounts to incorporating a demand for liquidity at the investor level absent in previous work, has important implications for market efficiency and the way in which trading frictions shape asset prices as well as trade volume, bid-ask spreads and trading delays—precisely the dimensions of market liquidity which search-based theories of financial liquidity were designed to explain.

The model we have developed allows for fairly general forms of investor heterogeneity and it has relatively few parameters that map naturally into observables. One could easily imagine calibrating or estimating the model using data on trade execution in OTC markets. We think that much could be learned from such exercises. For example, one could quantify the welfare gains associated with a given reduction in trading frictions, and the impact that the introduction of electronic trading networks will have on bid-ask spreads, average execution times, trade volume and other standard measures of liquidity. Various extensions are worth considering. First, there are many issues, such as the dynamic provision of liquidity by dealers who can hold asset positions, that would require a more detailed study of the model dynamics. Second, as an alternative to bilateral bargaining, one could explore alternative trading mechanisms that combine price-posting and directed search which would correspond to the more transparent market structures in the OTC spectrum.

References

[1] Allen, H., J. Hawkins, and S. Sato (2001): "Electronic Trading and its Implications for Financial Systems," *BIS papers*, 7, 30–52.

[2] Ashcraft, A., and D. Duffie (2007): "Systemic Illiquidity in the Federal Funds Market," *American Economic Review* (Papers and Proceedings), 97, 221–225.

[3] Barclay, M. J., W. G. Christie, J. H. Harris, E. Kandel, and P. H. Schultz (1999): "Effects of Market Reform on the Trading Costs and Depths of Nasdaq Stocks," *The Journal of Finance*, 65, 1–34.

[4] Biais, B., and R. Green (2005): "The Microstructure of the Bond Market in the 20th Century," unpublished manuscript.

[5] Boehmer, E. (2005): "Dimensions of Execution Quality: Recent Evidence for U.S. Equity Markets," *Journal of Financial Economics*, 78, 463–704.

[6] Burnside, C., M. Eichenbaum, I. Kleshchelski, and S. Rebelo (2006): "The Returns to Currency Speculation," unpublished manuscript.

[7] Christie, W. G., J. H. Harris, and P. H. Schultz (1994): "Why did NASDAQ Market Makers Stop Avoiding Odd-eigth Quotes?," *Journal of Finance*, 49, 1841–1860.

[8] Christie, W. G., J. H. Harris, and P. H. Schultz (1994): "Why do NASDAQ Market Makers Avoid Odd-eigth Quotes?," *Journal of Finance*, 49, 1813–1840.

[9] Costantinides, G. (1986): "Capital Market Equilibrium with Transaction Costs," *Journal of Political Economy*, 94, 842–862.

[10] Diamond, P. A. (1984): "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, 90, 881–894.

[11] Duffie, D., N. Gârleanu, and L. H. Pedersen (2005): "Over-the-counter Markets," *Econometrica*, 73, 1815–1847.

[12] Duffie, D., N. Gârleanu, and L. H. Pedersen (2006): "Valuation in Over-the-counter Markets," *Review of Financial Studies* (forthcoming).

[13] Duffie, D., and Y. Sun (2007): "Existence of Independent Random Matching," *Annals of Applied Probability*, 17, 386–419.

[14] Easley, D., and M. O'Hara (1987): "Price, Trade Size, and Information in Securities Markets," *Journal of Financial economics,* 19, 69–90.

[15] Edwards, A., L. Harris, and M. Piwowar (2004): "Corporate Bond Market Transparency and Transaction Costs," unpublished manuscript.

[16] Gârleanu, N. (2006): "Portfolio Choice and Pricing in Illiquid Markets," unpublished manuscript.

[17] Glosten, L., and P. Milgrom (1985): "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, 13, 71–100.

[18] Green, R., B. Hollifield, and N. Schurhoff (2006a): "Dealer Intermediation and Price Behavior in the Aftermarket for New Bond Issues," unpublished manuscript.

[19] Green, R., B. Hollifield, and N. Schurhoff (2006b): "Financial Intermediation and Costs of Trading in an Opaque Market," *Review of Financial Studies* (forthcoming).

[20] Hamilton, J. (1996): "The Daily market for Federal Funds," *Journal of Political Economy*, 104, 26–56.

[21] Harris, L. (2003): *Trading & Exchanges: Market Microstructure for Practitioners*. Oxford University Press.

[22] Harris L., and M. Piwowar (2006): "Secondary Trading Costs in the Municipal Bond Market," *Journal of Finance*, 61, 1361–1397.

[23] Hasbrouck, J. (2004): *Empirical Microstructure: Economic and Statistical Perspectives on the Dynamics of Trade in Securities Markets*, unpublished teaching notes, Department of Finance, Stern School of Business, New York University.

[24] Heaton, J., and D. J. Lucas (1995): "The Importance of Investor Heterogeneity and Financial Market Imperfections for the Behavior of Asset Prices," *Carnegie-Rochester Conference Series on Public Policy*, 42, 1–32.

[25] Hosios, A. J. (1990): "On the Efficiency of Matching and Related Models of Search and Unemployment," *Review of Economic Studies,* 57, 279–98.

[26] Kiyotaki, N., and R. Wright (1993): "A Search-Theoretic Approach to Monetary Economics," *American Economic Review*, 83, 63–77.

[27] Lagos, R., and G. Rocheteau (2006): "Search in Asset Markets," Federal Reserve Bank of Cleveland Working Paper 06-07.

[28] Lagos, R., and G. Rocheteau (2007): "Search in Asset Markets: Market Structure, Liquidity, and Welfare," *American Economic Review* (Papers and Proceedings), 97, 198-202.

[29] Lagos, R., G. Rocheteau, and P. O. Weill (2007): "Crashes and Recoveries in Illiquid Markets," unpublished manuscript.

[30] Lagos, R., and R. Wright (2005): "A Unified Framework for Monetary Theory and Policy Analysis," *Journal of Political Economy*, 113, 463–484.

[31] Lo, A. W., H. Mamaysky, and J. Wang (2004): "Asset Prices and Trading Volume under Fixed Transactions Costs," *Journal of Political Economy*, 112, 1054–1090.

[32] Miao, J. (2006): "A Search Model of Centralized and Decentralized Trade," *Review of Economic Dynamics*, 9, 68–92.

[33] Michaely, R., and J. L. Vila (1996): "Trading Volume with Private Valuation: Evidence from the Ex-dividend Day," *Review of Financial Studies*, 9, 471–509.

[34] Pagano, M. (1989): "Trading Volume and Asset Liquidity," *Quarterly Journal of Economics*, 104, 255–274.

[35] Pissarides, C. A. (2000): *Equilibrium Unemployment Theory.* MIT Press, 2nd edition, Cambridge.

[36] Rubinstein, A., and A. Wolinsky (1987): "Middlemen," *Quarterly Journal of Economics*, 102, 581–594.

[37] Rust, J., and J. Hall (2003): "Middlemen versus Market Makers: A Theory of Competitive Exchange," *Journal of Political Economy,* 111, 353–403.

[38] Saunders, A., A. Srinivasan, and W. Ingo (2002): "Price Formation in the OTC Corporate Bond Markets: A Field Study of the Interdealer Market," *Journal of Economics and Business*, 54, 95–113.

[39] Schultz, P. (2001): "Corporate Bond Trading Costs: A Peek Behind the Curtain," *Journal of Finance*, 41, 677–698.

[40] Spulber, D. (1996): "Market Making by Price Setting Firms," *Review of Economic Studies*, 63, 559–580.

[41] Stoll, H. (2006): "Electronic Trading in Stock Markets," *Journal of Economic Perspectives*, 20, 153–174.

[42] Vayanos, D. (1998): "Transaction Costs and Asset Prices: A Dynamic Equilibrium Model," *Review of Financial Studies*, 11, 1–58.

[43] Vayanos, D., and P. O. Weill (2007): "A Search-Based Theory of the On-the-Run Phenomenon," *Journal of Finance* (forthcoming).

[44] Wahal, S. (1997): "Entry, Exit, Market Makers, and the Bid-ask Spread," *Review of Financial Studies*, 10, 871–901.

[45] Weill, P. O. (2007): "Leaning Against the Wind," *Review of Economic Studies* (forthcoming).

[46] Weston, J. (2000): "Competition on the NASDAQ and the Impact of Recent Market Reforms," *Journal of Finance,* 55, 2565–2598.

[47] Weston, J. (2002): "Electronic Communication Networks and Liquidity on the NASDAQ," *Journal of Financial Services Research*, 22, 125–139.

A Proofs

Proof of Lemma 1. The Nash solution requires the outcome to be Pareto efficient. Since agents' payoffs are linear in ϕ, $a_i(t)$ must maximize the surplus from the match, namely $V_i\left[a_i(t),t\right]-V_i(a,t)-p\left[a_i(t)-a\right]$. This gives (3). Differentiating the Nash product in (2) with respect to ϕ and equating to zero gives (4). ■

Proof of Lemma 2. We proceed in two steps: (i) derive a simpler expression for $\bar{U}_i(a)$ and (ii) compute $\mathbb{E}\left[e^{-rT_\kappa}p(t+T_\kappa)\right]$.

(i). Changing variables, (7) can be written as

$$\bar{U}_i(a)=\mathbb{E}_i\left[\int_0^{T_\kappa-t}e^{-rs}u_{k(t+s)}(a)ds\right]. \tag{41}$$

Let T_δ denote the next time the investor draws a new preference shock and $T_{\delta\kappa}=\min\left(T_\delta,T_\kappa\right)$. Since preference shocks and effective contacts with dealers follow independent Poisson processes, $T_\delta-t$, $T_\kappa-t$ and $T_{\delta\kappa}-t$ are exponentially distributed random variables with means $1/\delta$, $1/\kappa$, and $1/\left(\kappa+\delta\right)$, respectively. We write (41) recursively,

$$\bar{U}_i(a)=\mathbb{E}_i\left[\int_0^{T_{\delta\kappa}-t}e^{-rs}u_{k(t)}(a)ds+\mathbb{I}_{\{T_\delta<T_\kappa\}}e^{-r(T_\delta-t)}\bar{U}_{k(T_\delta)}(a)\right],$$

where the expectation is over the random variables $T_\delta-t$, $T_\kappa-t$ and $T_{\delta\kappa}-t$, conditional on $k\left(t\right)=i$. Notice that

$$\begin{aligned}\mathbb{E}_i\left[\int_0^{T_{\delta\kappa}-t}e^{-rs}u_{k(t)}(a)ds\right]&=\int_0^\infty\left[\int_0^{\tau_{\delta\kappa}}e^{-rs}u_i(a)ds\right](\kappa+\delta)\,e^{-(\kappa+\delta)\tau_{\delta\kappa}}d\tau_{\delta\kappa}\\&=\frac{u_i(a)}{r+\kappa+\delta}.\end{aligned} \tag{42}$$

Since $T_\delta-t$ and $T_\kappa-t$ are independent random variables, and $k\left(T_\delta\right)=j$ with probability π_j for all $T_\delta-t\geq0$, we have

$$\begin{aligned}\mathbb{E}_i\left[\mathbb{I}_{\{T_\delta<T_\kappa\}}e^{-r(T_\delta-t)}\bar{U}_{k(T_\delta)}(a)\right]&=\int_0^\infty\int_0^\infty\sum_{j=1}^I\mathbb{I}_{\{\tau_\delta<\tau_\kappa\}}e^{-r\tau_\delta}\bar{U}_j(a)\pi_j\delta e^{-\delta\tau_\delta}\kappa e^{-\kappa\tau_\kappa}d\tau_\delta d\tau_\kappa\\&=\left[\int_0^\infty\int_0^{\tau_\kappa}e^{-r\tau_\delta}\delta e^{-\delta\tau_\delta}\kappa e^{-\kappa\tau_\kappa}d\tau_\delta d\tau_\kappa\right]\sum_{j=1}^I\pi_j\bar{U}_j(a)\\&=\frac{\delta}{r+\kappa+\delta}\sum_{j=1}^I\pi_j\bar{U}_j(a).\end{aligned} \tag{43}$$

Combine (42) and (43) to get

$$\bar{U}_i(a) = \frac{u_i(a)}{r+\kappa+\delta} + \frac{\delta}{r+\kappa+\delta}\sum_{j=1}^{I}\pi_j\bar{U}_j(a). \tag{44}$$

Multiply (44) through by π_i, add over i, solve for $\sum_j \pi_j\bar{U}_j(a)$ and substitute this is expression back into (44) to obtain

$$\bar{U}_i(a) = \frac{\bar{u}_i(a)}{r+\kappa}, \tag{45}$$

where $\bar{u}_i(a)$ is as in (10).

(ii). The expected discounted price of the asset at the next time when the investor gets an opportunity to trade is

$$\mathbb{E}[e^{-r(T_\kappa - t)}p(T_\kappa)] = \kappa\int_0^\infty e^{-(r+\kappa)s}p(t+s)\,ds. \tag{46}$$

Finally, substitute (45) and (46) into (8) and multiply through by $(r+\kappa)$ to obtain the formulation of the investor's problem in the statement of the lemma. ■

Proof of Lemma 3. (a) To obtain (13)**,** rewrite (11) as

$$q(t) = (r+\kappa)\,p(t) - \kappa e^{(r+\kappa)t}\int_t^\infty (r+\kappa)\,e^{-(r+\kappa)s}p(s)ds \tag{47}$$

and differentiate with respect to t.

(b) To arrive at (14), integrate (13) forward using the condition $\lim_{t\to\infty} e^{-rt}p(t) = 0$. ■

Proof of Lemma 4. We proceed in three steps: (i) derive $n_{ji}(\tau,t)$, (ii) derive $n^0_{ji}(\mathcal{A},t)$ and (iii) obtain $H_t(\mathcal{A},\mathcal{I})$ for an arbitrary $(\mathcal{A},\mathcal{I}) \in \Sigma$.

Step (i). Since investors meet dealers according to a Poisson process with arrival rate α, the length of the time period between any time t and the next time the investor meets a dealer is an exponentially distributed random variable with mean $1/\alpha$. Thus, the density measure of investors who last readjusted their asset holdings at time $t-\tau > 0$ is $\alpha e^{-\alpha\tau}$. The compound Poisson process for preference shocks implies that the probability that an investor who last contacted a dealer at time $t-\tau$ has a history of preference types involving $k(t-\tau) = j$ and $k(t) = i$ is $\left(1-e^{-\delta\tau}\right)\pi_i + \mathbb{I}_{\{i=j\}}e^{-\delta\tau}$. Since the measure of investors with preference type j at time $t-\tau$ is $n_j(t-\tau)$, and the Poisson process for meeting dealers and the compound Poisson process for preference shocks are independent, the density measure of investors who last traded

at time $t-\tau$ and who have a history of preferences involving $k(t-\tau)=j$ and $k(t)=i$, is $n_{ji}(\tau,t)=\alpha e^{-\alpha\tau}\left[\left(1-e^{-\delta\tau}\right)\pi_i+\mathbb{I}_{\{i=j\}}e^{-\delta\tau}\right]n_j(t-\tau)$, as given by (19) and (20).

Step (*ii*). Let T_α denote the first time an investor contacts a dealer. Since T_α is an exponentially distributed random variable, $\Pr(T_\alpha>t)=e^{-\alpha t}$. Thus, $e^{-\alpha t}$ is the measure of investors who have not contacted a dealer up to time t. Since the Poisson meeting process is independent of investors' individual states, the time-t measure of investors whose asset holdings and preference types lied in the set $(\mathcal{A},\{j\})$ at time 0 and who have not yet met a dealer at time t is $e^{-\alpha t}H_0(\mathcal{A},\{j\})$. The measure of investors who were of preference type j at time 0 and are of type i at time t is $\left(1-e^{-\delta t}\right)\pi_i+e^{-\delta t}\mathbb{I}_{\{j=i\}}$. Thus, the time-$t$ measure of investors who at time 0 had preference type j and assets in $\mathcal{A}$, whose preference type is i at the current time t, and who have never traded (so their asset holdings are still in $\mathcal{A}$) is $n^0_{ji}(\mathcal{A},t)=e^{-\alpha t}\left[\left(1-e^{-\delta t}\right)\pi_i+e^{-\delta t}\mathbb{I}_{\{j=i\}}\right]H_0(\mathcal{A},\{j\})$, as given in (21).

(*iii*). $H_t(\mathcal{A},\mathcal{I})$ is the measure of investors who have an individual state $(a,i)\in(\mathcal{A},\mathcal{I})$ at time t. The first term in $H_t(\mathcal{A},\mathcal{I})$ is $\sum_{i\in\mathcal{I}}\sum_{j=1}^{I}n^0_{ji}(\mathcal{A},t)$, namely those investors who never contacted dealers but who were holding asset positions in the set $\mathcal{A}$ at time 0 and whose preference types at t lie in $\mathcal{I}$. The time-t measure of investors of type i who chose an asset position in the set $\mathcal{A}$ the last time they traded, given that their preference type at that time was j, is $\int_0^t\mathbb{I}_{\{a_j(t-\tau)\in\mathcal{A}\}}n_{ji}(\tau,t)d\tau$. Thus, the second term in $H_t(\mathcal{A},\mathcal{I})$, namely the measure of investors who the last time they traded chose asset positions that belong to the set $\mathcal{A}$ and whose preference types at time t lie in $\mathcal{I}$, is $\sum_{i\in\mathcal{I}}\sum_{j=1}^{I}\int_0^t\mathbb{I}_{\{a_j(t-\tau)\in\mathcal{A}\}}n_{ji}(\tau,t)d\tau$. ■

Proof of Proposition 1. For all $t\geq 0$, the distribution $\{n_i(t)\}_{i=1}^I$ is unique and given by (16). Given that u_i is strictly concave and continuously differentiable, (12) implies that any interior choice $a_i(t)$ is a strictly decreasing, continuous function of $q(t)$ for every i. Therefore, the market-clearing condition (17) determines a unique $q(t)$ for each $t\geq 0$. Given $q(t)$, there is a unique $\{a_i(t)\}_{i=1}^I$ that solves (12). Given $q(t)$, (15) gives the fee $\phi_i(a,t)$ for every i and a. Finally, given $\{a_i(t)\}_{i=1}^I$ the distribution H_t is given by (18). ■

Proof of Proposition 2. Calculations similar to those contained in part (*i*) of the proof of Lemma 2 imply $\hat{U}_i(a)=\bar{u}_i(a)/(r+\alpha)$. Substitute this expression into the planner's objective function to get

$$\max_{\{a_i(t)\}}\int_0^\infty\frac{\alpha}{r+\alpha}\left\{\sum_{i=1}^{I}\left[\frac{r+\alpha}{r+\alpha+\delta}u_i[a_i(t)]+\frac{\delta}{r+\alpha+\delta}\sum_{j=1}^{I}\pi_j u_j\left[a_i(t)\right]\right]n_i(t)\right\}e^{-rt}dt$$

subject to $\sum_{j=1}^{I} n_j(t)a_j(t) \leq A$ and $a_i(t) \geq 0$ for all i. Let

$$\mathcal{L}(t) = \sum_{i=1}^{I} \left[\frac{r+\alpha}{r+\alpha+\delta} u_i[a_i(t)] + \frac{\delta}{r+\alpha+\delta} \sum_{k=1}^{I} \pi_k u_k[a_i(t)] \right] n_i(t) + \lambda(t) \left[A - \sum_{i=1}^{I} n_i(t)a_i(t) \right],$$

where $\lambda(t)$ is the Lagrange multiplier associated with the feasibility constraint. The planner's problem then reduces to finding, for each t, the sequence $\{a_i(t)\}_{i=1}^{I}$ that solves $\max_{\{a_i(t)\}} \mathcal{L}(t)$. Since $\mathcal{L}(t)$ is strictly jointly concave in $\{a_i(t)\}_{i=1}^{I}$, the first-order necessary and sufficient conditions for this problem are

$$\frac{(r+\alpha)\, u_i'\left[a_i(t)\right] + \delta \sum_{k=1}^{I} \pi_k u_k'\left[a_i(t)\right]}{r+\alpha+\delta} \leq \lambda(t), \qquad \text{“}=\text{”} \quad \text{if } a_i(t) > 0, \tag{48}$$

for $i = 1, ..., I$. The resource constraint (23) at equality is

$$\sum_{i=1}^{I} n_i(t) a_i^* \left[\lambda(t)\right] = A \tag{49}$$

where $a_i^*(\lambda)$ is the a_i that satisfies (48). Comparing (49) with (17), (48) with (12), and setting $q(t) = \lambda(t)$, it becomes clear that (12) coincides with (48) if and only if $\eta = 0$. ■

Proof of Proposition 3. From (16), $\lim_{t\to\infty} n_i(t) = \pi_i$ for each i. Thus, condition (17) becomes $\sum_{i=1}^{I} \pi_i a_i(t) = A$, where according to (12), $a_i(t) = \max\left\{\bar{u}_i'^{-1}\left[q(t)\right], 0\right\}$. With this, the market clearing condition can be written as $\sum_{i=1}^{I} \pi_i \max\left\{\bar{u}_i'^{-1}\left[q(t)\right], 0\right\} = A$. Given that u_i is continuously differentiable for each i, this condition defines a unique q which is time-invariant. Given this q, (12) implies a unique set of time-invariant optimal asset holdings $\{a_i\}_{i=1}^{I}$. Thus, $\{a_i\}_{i=1}^{I}$ and q satisfy (24) and (25). Given the fact that $q(t) = q$ for all t, (13) implies (26). Given q and $\{a_i\}_{i=1}^{I}$, (15) implies (27), which determines the time invariant fees $\{\phi_i(a)\}_{i=1}^{I}$. To derive the time-invariant limit of the measure of investors across individual states, note that $\lim_{t\to\infty} n_{ji}^0(\mathcal{A}, t) = 0$ for all i, $j \in \mathbb{X}$ and all $\mathcal{A} \subseteq \mathbb{R}_+$. Also, $\lim_{t\to\infty} n_{ji}(\tau, t) = \alpha e^{-\alpha\tau}\left[\left(1 - e^{-\delta\tau}\right)\pi_i + e^{-\delta\tau}\mathbb{I}_{\{i=j\}}\right]\pi_j \equiv n_{ji}(\tau, \infty)$ and $\lim_{t\to\infty} a_j(t-\tau) = a_j$, so $\lim_{t\to\infty} H_t(\mathcal{A}, \mathcal{I}) = H(\mathcal{A}, \mathcal{I})$ for all $(\mathcal{A}, \mathcal{I}) \in \Sigma$. ■

Proof of Proposition 4. Differentiate (25) to obtain

$$\frac{dp}{d\kappa} = \frac{\sum_{i=1}^{I} \pi_i \partial a_i / \partial \kappa}{-\sum_{i=1}^{I} \pi_i \partial a_i / \partial p}.$$

From (32), we know that the denominator of this expression is strictly positive, so we focus on the sign of the numerator. Differentiate (32) to obtain $\partial a_i/\partial\kappa$, multiply by π_i, and add over all i to arrive at

$$\sum_{i=1}^{I}\pi_i\frac{\partial a_i}{\partial\kappa}=\frac{\delta}{(r+\kappa+\delta)^2 rp}\sum_{i=1}^{I}\pi_i\frac{[u'(a_i)]^2}{-u''(a_i)}\left(\varepsilon_i-\bar{\varepsilon}\right).$$

Suppose $-[u'(a)]^2/u''(a)$ is strictly increasing in a. Let $\bar{a}$ denote the a that solves (32) for $\bar{\varepsilon}_i=\bar{\varepsilon}$. Then, note that $-[u'(a_i)]^2(\varepsilon_i-\bar{\varepsilon})/u''(a_i)\geq-[u'(\bar{a})]^2(\varepsilon_i-\bar{\varepsilon})/u''(\bar{a})$ for each i, with strict inequality for all i such that $\varepsilon_i\neq\bar{\varepsilon}$. Thus, $\sum_{i=1}^{I}\pi_i\frac{\partial a_i}{\partial\kappa}>0$ and consequently, $\frac{dp}{d\kappa}>0$. Similar reasoning implies $\frac{dp}{d\kappa}<0$ if $-[u'(a)]^2/u''(a)$ is strictly decreasing and $\frac{dp}{d\kappa}=0$ if $-[u'(a)]^2/u''(a)$ is constant in a. ■

Proof of Proposition 5. With $u_i(a)=\varepsilon_i a^{1-\sigma}/(1-\sigma)$, the model can be solved in closed form:

$$a_i=\frac{\bar{\varepsilon}_i^{1/\sigma}}{\sum_{j=1}^{I}\pi_j\bar{\varepsilon}_j^{1/\sigma}}A \tag{50}$$

$$q=\frac{\left(\sum_{j=1}^{I}\pi_j\bar{\varepsilon}_j^{1/\sigma}\right)^{\sigma}}{A^{\sigma}}. \tag{51}$$

From (50), the individual demand for the asset by an agent whose current preference shock is ε_i in an economy where the direct effective access rate to the asset market is κ is

$$a_i(\kappa)=\frac{A}{\sum_{j=1}^{I}\pi_j\left[\frac{(r+\kappa)\varepsilon_j+\delta\bar{\varepsilon}}{(r+\kappa)\varepsilon_i+\delta\bar{\varepsilon}}\right]^{1/\sigma}}. \tag{52}$$

Consider $\kappa'>\kappa$. One can verify that there exists a unique $\tilde{\varepsilon}\in(\varepsilon_1,\varepsilon_I)$ such that $a_i(\kappa')>a_i(\kappa)$ for all $\varepsilon_i>\tilde{\varepsilon}$, $a_i(\kappa')<a_i(\kappa)$ for all $\varepsilon_i<\tilde{\varepsilon}$ and $a_i(\kappa')=a_i(\kappa)\equiv\tilde{a}$ if $\varepsilon_i=\tilde{\varepsilon}$. With (30) and (31), the cumulative distribution of assets across investors for the economy indexed by κ, is $\mathbb{G}_\kappa(a)=\sum_{j=1}^{I}\mathbb{I}_{\{a_j(\kappa)\leq a\}}\pi_j$. This, and the fact that $(\kappa'-\kappa)[a_i(\kappa')-a_i(\kappa)]>0$ iff $\varepsilon_i>\tilde{\varepsilon}$ implies that $\mathbb{G}_{\kappa'}(a)\geq\mathbb{G}_\kappa(a)$ for all $a<\tilde{a}$ and $\mathbb{G}_{\kappa'}(a)\leq\mathbb{G}_\kappa(a)$ for all $a>\tilde{a}$. Thus, given that both distributions have the same mean and that $a_I(\kappa')>a_I(\kappa)$, the fact that the cumulative density functions cross only once implies that $\mathbb{G}_\kappa$ dominates $\mathbb{G}_{\kappa'}$ in the second-order stochastic sense. ■

Proof of Corollary 2. For $I=2$, we have $\mathbb{X}=\{1,2\}$ and

$$\mathcal{V}=\frac{\alpha\delta\pi_1\pi_2}{\alpha+\delta}\left[a_2(\kappa)-a_1(\kappa)\right],$$

where $a_i(\kappa)$ is given by (52). Since $\varepsilon_1 < \varepsilon_2$, we have $a_1(\kappa) < a_2(\kappa)$, and differentiating (52) with respect to κ implies $\frac{da_1(\kappa)}{d\kappa} < 0 < \frac{da_2(\kappa)}{d\kappa}$. To find $\frac{d\mathcal{V}}{d\kappa}$, we consider two cases. (i) An increase in κ caused by a decrease in η (keeping α constant). For this case, $\frac{d\mathcal{V}}{d\kappa} = \frac{da_2(\kappa)}{d\kappa} - \frac{da_1(\kappa)}{d\kappa} > 0$. (ii) An increase in κ caused by an increase in α, which implies $\frac{d\mathcal{V}}{d\kappa} = \left(\frac{\delta}{\alpha+\delta}\right)^2 \pi_1\pi_2 \left[a_2(\kappa) - a_1(\kappa)\right] + \frac{\alpha\delta\pi_1\pi_2}{\alpha+\delta}\left[\frac{da_2(\kappa)}{d\kappa} - \frac{da_1(\kappa)}{d\kappa}\right] > 0$. ■

Proof of Proposition 6. Since $u_i(a) = \varepsilon_i \ln a$, we have $a_i > 0$ for all i, and $a_i \neq a_j$ unless $i = j$. From (30), the proportion of trades that involve buying a_i and selling a_j or vice versa (for $i \neq j$) is $(n_{ij} + n_{ji})/(1 - \sum_{i=1}^I n_{ii}) = 2\pi_i\pi_j/(1 - \sum_{i=1}^I \pi_i^2)$, which is independent of κ. From Corollary 1, $dp/d\kappa = 0$, so differentiating (32),

$$\frac{d\left[g_i(\kappa;p) - g_j(\kappa;p)\right]}{d\kappa} = \frac{\delta(\varepsilon_i - \varepsilon_j)}{rp(r+\kappa+\delta)^2}.$$

Thus, $|a_i - a_j| = |g_i(\kappa;p) - g_j(\kappa;p)|$ increases with κ for all $i \neq j$. The measure of trades of size less than $z \geq 0$ is

$$\sum_{i=1}^I \sum_{j\neq i} \frac{\pi_i\pi_j}{1 - \sum_{i=1}^I \pi_i^2} \mathbb{I}_{\{|a_i - a_j| \leq z\}},$$

which is decreasing in κ. ■

Proof of Lemma 5. Differentiate (27) to obtain

$$\frac{\partial \phi_i(a)}{\partial a} = -\frac{\eta}{r+\kappa}\left[\bar{u}_i'(a) - q\right].$$

Suppose that the nonnegativity constraint on a_i is slack. Then, since $\bar{u}_i$ is strictly concave and $\bar{u}_i'(a_i) - q = 0$, we know that $\bar{u}_i'(a) - q < 0$ if and only if $a - a_i > 0$, and $\frac{\partial\phi_i(a)}{\partial a}$ has the same sign as $a - a_i$. If $a_i = 0$, then $a > a_i$ and $\bar{u}_i'(a) - q < \bar{u}_i'(a_i) - q \leq 0$, so $\frac{\partial\phi_i(a)}{\partial a} > 0$ which is the same sign as $a - a_i = a > 0$. This establishes part (i). To show part (ii), divide (27) by $(a_i - a)$ and differentiate the resulting expression to get

$$\frac{\partial}{\partial a}\left[\frac{\phi_i(a)}{a_i - a}\right] = \frac{\eta}{r + \alpha(1-\eta)}\left[\frac{\bar{u}_i(a_i) - \bar{u}_i(a) - \bar{u}_i'(a)(a_i - a)}{(a_i - a)^2}\right],$$

which is negative for all $a \neq a_i$, since $\bar{u}_i$ is strictly concave. ■

Proof of Proposition 7. Let $q(\kappa, r)$, $a_i(\kappa, r)$ and $\phi_{ji}(\kappa, r)$ denote, respectively, the equilibrium q, a_i and ϕ_{ji} that solve the system (24), (25) and (27) for all $i \in \mathbb{X}$. We proceed

in three steps: (i) show that $\phi_{ji}(\kappa, r) > 0$ for all $\kappa \in (0, \infty)$ and all $r \in [0, \infty)$ provided $a_i(\kappa, r) \neq a_j(\kappa, r)$ and $\eta > 0$; (ii) establish that $\lim_{\kappa \to \infty} \phi_{ji}(\kappa, r) = 0$ for any $r \geq 0$ and all $(i, j) \in \mathbb{X}^2$; (iii) show that for each $\kappa \in (0, \infty)$ there is $\bar{r} > 0$ such that $\phi_{ji}(0, r) < \phi_{ji}(\kappa, r)$ for all $r \in (0, \bar{r})$. The nonmonotonicity of $\phi_{ji}(\kappa, r)$ with respect to κ for all $r \in [0, \bar{r})$ will then follow from steps (i) through (iii).

(i). Since $\phi_{ij} = \frac{\eta}{r+\kappa}\{\max_{a'}[\bar{u}_i(a') - qa'] - [\bar{u}_i(a_j) - qa_j]\}$, we have $\phi_{ij}(\kappa, r) > 0$ for all $\kappa \in (0, \infty)$ and all $r \in [0, \infty)$, provided $\eta > 0$ and $a_j \neq \arg\max_{a' \geq 0}[\bar{u}_i(a') - qa']$ (i.e., provided the investor trades).

(ii). $\lim_{\kappa \to \infty} q(\kappa, r) = \bar{q}$ and $\lim_{\kappa \to \infty} a_i(\kappa, r) = \arg\max_{a' \geq 0}[u_i(a') - \bar{q}a'] \equiv h_i^{\infty}(\bar{q})$, where $\bar{q}$ solves $\sum_{i=1}^{I} \pi_i h_i^{\infty}(\bar{q}) = A$, which in turn implies $\bar{q} \in (0, \infty)$ and $h_i^{\infty}(\bar{q}) < \infty$. Therefore $\lim_{\kappa \to \infty} \phi_{ji}(\kappa, r) = 0$ for any $r \geq 0$ and all $(i, j) \in \mathbb{X}^2$.

(iii). Let $\kappa \to 0$ to obtain $q(0, r) = \tilde{q}$ and $a_i(0, r) = \arg\max_{a' \geq 0}[\tilde{u}_i(a') - \tilde{q}a'] \equiv h_i^0(\tilde{q})$, where $\tilde{u}_i(a) = \frac{r}{r+\delta} u_i(a) + \frac{\delta}{r+\delta} \sum_{k=1}^{I} u_i(a)$ and $\tilde{q}$ solves $\sum_{i=1}^{I} \pi_i h_i^0(\tilde{q}) = A$. Observe that $\lim_{r \to 0} a_i(0, r) = \tilde{u}_i'^{-1}(\tilde{q}) = A$, for each $i \in \mathbb{X}$. With this, apply L'Hôpital's rule to find $\lim_{r \to 0} \phi_{ji}(0, r) = 0$.

Our assumptions on primitives imply that $q(\kappa, r)$ and $a_i(\kappa, r)$ are continuous functions, so $\phi_{ji}(\kappa, r)$ is continuous. Hence, for each (i, j) with $i \neq j$ and an each $\kappa \in (0, \infty)$, there is some $\bar{r} > 0$ such that for all $r \in [0, \bar{r})$, we have $\lim_{\kappa \to \infty} \phi_{ji}(\kappa, r) = 0 < \phi_{ji}(\kappa, r)$ (by (i) and (ii)) and $\phi_{ji}(0, r) < \phi_{ji}(\kappa, r)$ (by (i) and (iii)), which establishes the nonmonotonicity of ϕ_{ij} with respect to κ. ■

Proof of Corollary 3. Write $\Phi(\alpha, \eta, r) = \sum_{i,j=1}^{I} n_{ji}(\alpha) \phi_{ji}[\alpha(1-\eta), r]$, where $n_{ji}(\alpha)$ is given by (30)–(31). Fix an arbitrary $(\alpha, \eta) \in (0, \infty) \times (0, 1)$. From Proposition 7,

$$\min_{\{(i,j) \in \mathbb{X}^2 : i \neq j\}} \phi_{ji}[\alpha(1-\eta), r] > 0$$

for all $r \in [0, \infty)$, and there is $r_0 > 0$ such that for all $r \in [0, r_0)$, $\max_{(i,j) \in \mathbb{X}^2} \phi_{ji}(0, r) < \min_{\{(i,j) \in \mathbb{X}^2 : i \neq j\}} \phi_{ji}[\alpha(1-\eta), r]$. Then, for any $r \in [0, r_0)$, we have $\lim_{\alpha' \to \infty} \Phi(\alpha', \eta, r) = 0 < \Phi(\alpha, \eta, r)$ (by (ii)) and $\Phi(0, \eta, r) < \Phi(\alpha, \eta, r)$ (by (iii)), which establishes the nonmonotonicity of Φ with respect to α, and therefore with respect to $\kappa = \alpha(1-\eta)$. ■

Proof of Proposition 8. Using (10), we can write (35) as

$$\Phi = \frac{\eta\delta}{(\alpha+\delta)[r + \alpha(1-\eta) + \delta]} \sum_{i,j=1}^{I} \pi_i \pi_j [u_i(a_i) - u_i(a_j)].$$

From (24) we know that a_i is a continuous function of υ and q, i.e., $a_i = a_i(\upsilon, q)$. From (25), there is a unique q that clears the asset market and it is a continuous function of υ, i.e., $q = q(\upsilon)$. Thus, $a_i = a_i[\upsilon, q(\upsilon)]$ is a continuous function of υ. Define the map $\Gamma(\upsilon)$ as

$$\Gamma(\upsilon) \equiv \frac{[\alpha(\upsilon)/\upsilon]\,\eta\delta\sum_{i,j=1}^{I}\pi_i\pi_j\left\{u_i[a_i(\upsilon)] - u_i[a_j(\upsilon)]\right\}}{[\alpha(\upsilon)+\delta]\,[r+\alpha(\upsilon)(1-\eta)+\delta]}. \tag{53}$$

This is the left-hand side of the free-entry condition (36). First, we establish that $\lim_{\upsilon\to 0}\Gamma(\upsilon) = \infty$. Recall that $a_i = \arg\max_{a'\geq 0}[\bar{u}_i(a') - qa']$, therefore,

$$\frac{(r+\kappa)\,u_i(a_i) + \delta\sum_{k=1}^{I}\pi_k u_k(a_i)}{r+\kappa+\delta} - qa_i \geq \frac{(r+\kappa)\,u_i(a_j) + \delta\sum_{k=1}^{I}\pi_k u_k(a_j)}{r+\kappa+\delta} - qa_j \tag{54}$$

holds for every i and j. Since (24) implies $a_i = a_j$ if and only if $a_i = a_j = 0$, (54) holds with strict inequality for any i such that $a_i > 0$. Multiplying this inequality through by $\pi_i\pi_j$ and summing over all i and j implies $\sum_{i,j=1}^{I}\pi_i\pi_j\left\{u_i[a_i(\upsilon)] - u_i[a_j(\upsilon)]\right\} > 0$. The inequality is strict since for every υ we have $a_i > 0$ at least for $i = I$. Then, $\lim_{\upsilon\to 0}\Gamma(\upsilon) = \infty$ follows from $\eta > 0$ and the fact that

$$\lim_{\upsilon\to 0}\frac{\alpha(\upsilon)/\upsilon}{[\alpha(\upsilon)+\delta]\,[r+\alpha(\upsilon)(1-\eta)+\delta]} = \infty.$$

Next, note that the fact that $\sum_{i,j=1}^{I}\pi_i\pi_j\left\{u_i[a_i(\upsilon)] - u_i[a_j(\upsilon)]\right\}$ is bounded (because $a_i(\upsilon)$ must be bounded for (25) to hold), together with

$$\lim_{\upsilon\to\infty}\frac{\alpha(\upsilon)/\upsilon}{[\alpha(\upsilon)+\delta]\,[r+\alpha(\upsilon)(1-\eta)+\delta]} = 0$$

implies that $\lim_{\upsilon\to\infty}\Gamma(\upsilon) = 0$. Finally, since Γ is continuous, there exists some $\upsilon \in \mathbb{R}_+$ such that $\Gamma(\upsilon) = \gamma$. ■

Proof of Proposition 9. In an equilibrium with entry, the measure of dealers satisfies

$$\Phi[\alpha(\upsilon), \eta, r] = \gamma\upsilon^{1-\theta}. \tag{55}$$

From Proposition **3,** there is $\tilde{r} > 0$ such that $\underline{\gamma} \equiv \Phi(0,\eta,r) < \sup_{\upsilon}\Phi[\alpha(\upsilon),\eta,r] \equiv \overline{\gamma}$ for all $r \in [0,\tilde{r})$, and $\lim_{\upsilon\to\infty}\Phi[\alpha(\upsilon),\eta,r] = 0 < \underline{\gamma}$. Note that as $\theta \to 1$, $\gamma\upsilon^{1-\theta}$ converges uniformly to γ on any closed interval $[\upsilon_0,\upsilon_1] \subseteq (0,\infty)$. Thus, for any $\gamma \in (\underline{\gamma},\overline{\gamma})$, there is a $\tilde{\theta}$ such that for for all $\theta \in (\tilde{\theta},1)$, there are multiple (at least three) values of $\upsilon > 0$ that satisfy (55). ■

Proof of Proposition 10. The Lagrangian associated with this problem is

$$\mathcal{L} = \frac{\alpha}{\alpha+\delta}\sum_{i=1}^{I}\pi_i u_i(a_i) + \frac{\delta}{\alpha+\delta}\sum_{i,j=1}^{I}\pi_i\pi_j u_j(a_i) - \upsilon\gamma + \lambda\left(A - \sum_{i=1}^{I}\pi_i a_i\right),$$

where $\lambda \in \mathbb{R}_+$ is the Lagrange multiplier associated with the resource constraint $\sum_{i,j=1}^{I} n_{ij} a_i = A$. The first-order condition with respect to a_i is

$$\frac{\alpha}{\alpha+\delta} u_i'(a_i) + \frac{\delta}{\alpha+\delta} \sum_{k=1}^{I} \pi_k u_k'(a_i) \leq \lambda, \qquad \text{“} = \text{”} \quad \text{if } a_i > 0. \tag{56}$$

As $r \to 0$, the left-hand side of (24) approaches $\frac{\alpha(1-\eta)}{\alpha(1-\eta)+\delta} u_i'(a_i) + \frac{\delta}{\alpha(1-\eta)+\delta} \sum_{k=1}^{I} \pi_k u_k'(a_i)$, which coincides with the left-hand side of (56) if and only if $\eta = 0$. The first-order condition for the measure of dealers is

$$\frac{\alpha'(\upsilon)}{\alpha(\upsilon)+\delta} \sum_{i,j=1}^{I} \frac{\delta \pi_i \pi_j \left[u_i(a_i) - u_j(a_i)\right]}{\alpha(\upsilon)+\delta} = \gamma. \tag{57}$$

From (53) we know that, as $r \to 0$, the equilibrium condition for entry of dealers approaches

$$\frac{\left[\alpha(\upsilon)/\upsilon\right]\eta}{\alpha(\upsilon)(1-\eta)+\delta} \sum_{i,j=1}^{I} \frac{\delta \pi_i \pi_j \left[u_i(a_i) - u_i(a_j)\right]}{\alpha(\upsilon)+\delta} = \gamma,$$

which converges to (57) as $\eta \to 0$ if and only if $\alpha'(\upsilon)\upsilon/\alpha(\upsilon) = \eta$. ■

Proof of Proposition 11. Define

$$\Gamma(\upsilon) \equiv \frac{\left[\alpha(\upsilon)/\upsilon\right]\delta \sum_{i,j=1}^{I} \pi_i \pi_j \left\{u_i[a_i(\beta,r)] - u_i[a_j(\beta,r)]\right\}}{\left[\alpha(\upsilon)+\beta+\delta\right](r+\beta+\delta)},$$

where $a_i(\beta,r) = \arg\max_{a' \geq 0} \left[\bar{u}_i(a') - qa'\right] \equiv h_i(q)$ for $\eta = 1$ and q solves $\sum_{i=1}^{I} \pi_i h_i(q) = A$. Note that $\eta = 1$ implies $a_i(\beta,r)$ is independent of υ. Let $\upsilon(\beta,r)$ denote the equilibrium measure of participating dealers; it solves $\Gamma(\upsilon) - \gamma = 0$. With arguments similar to those used in Proposition 8, it can be shown that such a $\upsilon(\beta,r)$ exists and $\upsilon(\beta,r) > 0$ for $r > 0$. Moreover, $\upsilon(\beta,r)$ is unique since $\Gamma'(\upsilon) < 0$. Note that $\lim_{\beta\to\infty} \upsilon(\beta,r) = 0$, and since $\lim_{r\to 0} a_i(0,r) = A$ for all i, $\lim_{r\to 0} \upsilon(0,r) = 0$. Since Γ is continuous in r, for a given β there is $\bar{r} > 0$ such that $\upsilon(0,r) < \upsilon(\beta,r)$ and $\lim_{\beta'\to\infty} \upsilon(\beta',r) < \upsilon(\beta,r)$ for all $r \in (0,\bar{r})$. This establishes the nonmonotonicity of υ with respect to β. ■

B Transversality condition

In this appendix we show that an equilibrium, the asset price $p(t)$ necessarily satisfies the condition $\lim_{t\to\infty} e^{-rt}p(t) = 0$, which we used in part (b) of Lemma 2. The proof we offer here is an adapted version of a similar proof that appears in Lagos, Rocheteau and Weill (2007).

Consider an investor who effectively contacts the market with Poisson intensity κ. Let $\{T_n\}_{n=1}^{\infty}$ denote the sequence of contact times and N_t the number of contacts over the time interval $[0,t)$. We adopt the convention that $T_0 = 0$ (but T_0 is not a contact time). An asset plan, $\mathbf{a}$, for the investor specifies his asset holdings as a function time, s, and his history of preference shocks and contact times, $\langle k(s), \{T_n\}_{n=1}^{\infty}\rangle$ for $s \geq 0$. Let $\mathbf{a} = a(s)$ denote an asset plan. An asset plan is feasible if $a(s) = a(T_n)$ for all $s \in [T_n, T_{n+1})$ and $a(0) = a_0 > 0$, which is given. Let $V_i^t(\mathbf{a}, 0)$ be the expected discounted utility over the time interval $[0,t)$ of an investor with preference type i at time 0 who follows an asset plan $\mathbf{a}$. It satisfies

$$V_i^t(\mathbf{a},0) = \mathbb{E}_i\left\{\sum_{n=0}^{\infty}\left[\int_{T_n}^{T_{n+1}} e^{-rs}u_{k(s)}\left[a(T_n)\right]\mathbb{I}_{\{T_{n+1}\leq t\}}ds + \int_{T_n}^{t} e^{-rs}u_{k(s)}\left[a(T_n)\right]\mathbb{I}_{\{T_n\leq t<T_{n+1}\}}ds\right]\right\}$$
$$- \mathbb{E}_i\left\{\sum_{n=1}^{\infty} e^{-rT_n}p(T_n)\left[a(T_n) - a(T_{n-1})\right]\mathbb{I}_{\{T_n\leq t\}}\right\},$$

where the expectations operator, $\mathbb{E}_i$, is taken with respect to the random variables $\langle k(s), \{T_n\}_{n=1}^{\infty}\rangle$ for $s \geq 0$ and is indexed by i to indicate that the expectation is conditional on $k(0) = i$. Collect terms to arrive at

$$V_i^t(\mathbf{a},0) = \mathbb{E}_i\left\{\mathbb{I}_{\{0\leq t<T_1\}}\int_0^t e^{-rs}u_{k(s)}(a_0)\,ds + \mathbb{I}_{\{T_1\leq t\}}\left[\int_0^{T_1} e^{-rs}u_{k(s)}(a_0)\,ds + e^{-rT_1}p(T_1)a_0\right]\right\}$$
$$+ \mathbb{E}_i\left\{\sum_{n=1}^{\infty}\mathbb{I}_{\{T_{n+1}\leq t\}}\int_{T_n}^{T_{n+1}} e^{-rs}u_{k(s)}\left[a(T_n)\right]ds\right\}$$
$$- \mathbb{E}\left\{\sum_{n=1}^{\infty}\mathbb{I}_{\{T_{n+1}\leq t\}}e^{-rT_n}\left[p(T_n) - e^{-r(T_{n+1}-T_n)}p(T_{n+1})\right]a(T_n)\right\}$$
$$+ \mathbb{E}_i\left\{\int_{T_{N_t}}^{t} e^{-rs}u_{k(s)}\left[a(T_{N_t})\right]ds\right\} - \mathbb{E}\left\{e^{-rT_{N_t}}p(T_{N_t})a(T_{N_t})\right\}, \tag{58}$$

where the expectations operator, $\mathbb{E}$, is taken with respect to $\{T_n\}_{n=1}^{\infty}$. It is shown in Lagos, Rocheteau and Weill (2007, Lemma 2) that $V_i^t(\mathbf{a},0)$ converges to a finite limit $V_i^{\infty}(\mathbf{a},0)$ as

$t \to \infty$. After taking this limit we find

$$\begin{aligned} V_i^{\infty}(\mathbf{a}, 0) = \mathbb{E}_i \left\{ \int_0^{T_1} e^{-rs} u_{k(s)} \left[a(0) \right] ds + e^{-rT_1} p(T_1) a(0) \right\} \\ + \mathbb{E}_i \left\{ \sum_{n=1}^{\infty} e^{-rT_n} \left\{ \bar{U}_{k(T_n)} \left[a(T_n) \right] - q(T_n) a(T_n) \right\} \right\} \\ - \lim_{t \to \infty} \mathbb{E} \left\{ e^{-rT_{N_t}} p(T_{N_t}) a(T_{N_t}) \right\}. \end{aligned} \tag{59}$$

To arrive at (59), note that $T_n = \sum_{k=1}^{n} (T_k - T_{k-1})$ is the sum of n independent exponentially-distributed random variables, so $\lim_{t\to\infty} \mathbb{I}_{\{T_{n-1} \leq t < T_n\}} = 0$ and $\lim_{t\to\infty} \mathbb{I}_{\{T_n \leq t\}} = 1$ almost surely for all finite $n \geq 1$. The former implies that the first term on the right-side of (58) converges to 0 as $t \to \infty$. The latter implies that the second term of (58) converges to the first term of (59) and that the second and third terms of (58) converge to the second term of (59). To see that the first term on the last line of the right side of (58) goes to 0 as $t \to \infty$, write it as

$$\mathbb{E}_i \left\{ e^{-rT_{N_t}} \int_0^{t - T_{N_t}} e^{-rs} u_{k(s+T_{N_t})} \left[a(T_{N_t}) \right] ds \right\}. \tag{60}$$

Any asset plan that is consistent with equilibrium must be bounded, hence the integrand of (60) is bounded above. This integrand is also bounded below, since either u is bounded below or else it satisfies the Inada condition which ensures that any optimal plan has $a(s) > 0$ for all s. The fact that $t - T_{N_t} < \infty$ almost surely (because $t - T_{N_t}$ is exponentially distributed) implies that the integral in (60) is bounded. Finally, note that $\Pr\left(T_{N_t} < \tau\right) = e^{-\kappa(t-\tau)}$ for any $\tau < t$, so $T_{N_t} \to \infty$ almost surely as $t \to \infty$, which means that (60) goes to 0 as $t \to \infty$.

Now consider an optimal asset plan, $\mathbf{a}$, and scale it down by $1 - \varepsilon$. Define $\Delta_\varepsilon \equiv V_i^{\infty}(\mathbf{a}, 0) - V_i^{\infty}\left[(1-\varepsilon)\mathbf{a}, 0\right]$; then,

$$\begin{aligned} \Delta_\varepsilon = \mathbb{E}_i \left\{ \sum_{n=1}^{\infty} e^{-rT_n} \left\{ \bar{U}_{k(T_n)} \left[a(T_n) \right] - \bar{U}_{k(T_n)} \left[(1-\varepsilon)\, a(T_n) \right] - \varepsilon q(T_n) a(T_n) \right\} \right\} \\ - \lim_{t \to \infty} \mathbb{E} \left\{ \varepsilon e^{-rT_{N_t}} p(T_{N_t}) a(T_{N_t}) \right\}. \end{aligned}$$

Divide the previous expression by ε to get

$$\begin{aligned} \frac{\Delta_\varepsilon}{\varepsilon} = \mathbb{E}_i \left\{ \sum_{n=1}^{\infty} e^{-rT_n} \frac{\bar{U}_{k(T_n)} \left[a(T_n) \right] - \bar{U}_{k(T_n)} \left[a(T_n)(1-\varepsilon) \right] - \varepsilon q(T_n) a(T_n)}{\varepsilon} \right\} \\ - \lim_{t \to \infty} \mathbb{E} \left\{ e^{-rT_{N_t}} p(T_{N_t}) a(T_{N_t}) \right\}. \end{aligned}$$

Since the asset plan $\mathbf{a}$ is optimal, we can take the limit as $\varepsilon \to 0$, apply L'Hôpital's Rule and use the first-order condition for the investor's problem (e.g., (8)) to find that it must satisfy

$$\lim_{\varepsilon \to 0} \frac{\Delta_\varepsilon}{\varepsilon} = -\lim_{t \to \infty} \mathbb{E}\left\{e^{-rT_{N_t}} p(T_{N_t}) a(T_{N_t})\right\} \geq 0.$$

Since $e^{-rt} p(t) a(t) \geq 0$ for all t, the previous condition can be rewritten as

$$\lim_{t \to \infty} \mathbb{E}\left\{e^{-rT_{N_t}} p(T_{N_t}) a(T_{N_t})\right\} = 0 \tag{61}$$

for each investor. We can use the market-clearing condition to write

$$\int_\Omega a^\omega(t) d\omega = A, \quad \forall t,$$

where $a^\omega(t)$ is investor ω's asset demand at time t and Ω denotes the set of investors. Hence,

$$A \lim_{t \to \infty} \mathbb{E}\left\{e^{-rT_{N_t}} p(T_{N_t})\right\} = \lim_{t \to \infty} \mathbb{E}\left\{\int_\Omega e^{-rT_{N_t}} p(T_{N_t}) a^\omega(T_{N_t}) d\omega\right\} = 0,$$

since (61) holds for each ω. Then $T_{N_t} \to \infty$ almost surely as $t \to \infty$, so $\lim_{t \to \infty} e^{-rt} p(t) = 0$. ■